

TOWARDS AN UNDERSTANDING OF THE IMPACT OF UNCERTAINTY  
AND EMOTIONAL CONTENT ON USERS' DECISION-MAKING ABOUT  
MISINFORMATION

by

Alireza Karduni

A dissertation submitted to the faculty of  
The University of North Carolina at Charlotte  
in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in  
Computing and Information Systems

Charlotte

2020

Approved by:

---

Dr. Wenwen Dou

---

Dr. Samira Shaikh

---

Dr. Doug Markant

---

Dr. Isaac Cho

---

Dr. Sara Levens

---

Professor Eric Sauda



## ABSTRACT

ALIREZA KARDUNI. Towards an Understanding of the Impact of Uncertainty and Emotional Content on Users’ Decision-making About Misinformation. (Under the direction of DR. WENWEN DOU)

Misinformation on social media is a phenomenon with considerable impacts on our societies. To effectively mitigate its effects, we need to employ a cross-disciplinary effort to holistically address the critical elements involved in the process, including the source, the content, and the consumers of misinformation. However, the vast majority of computational approaches to addressing this problem focus on detecting and flagging misinformation content. Other important aspects, such as behaviors and intents of sources and the consumers’ decision-making processes, are often neglected. In this thesis, we address this gap through a series of studies around how users make decisions about content and sources of misinformation facilitated by Visual Analytics.

First, We introduce a Visual Analytic system, Verifi, that combines temporal, language, and network analysis features and enables users to assess the veracity of multiple news sources. Using Verifi, we conducted a controlled experiment that highlighted how uncertainty and conflicting cues in information impacts users’ perceptions of source credibility. Next, we extended Verifi to a more comprehensive multi-modal system enabling users to study sources through social network analysis, text, and images. Through a qualitative domain expert study conducted on Verifi, we learned valuable lessons about the importance of users’ trust in sources and how emotional content in images might impact users’ judgments and perceptions.

Inspired by the lessons learned from our work on Verifi, we present two studies on

the effects of emotions in images on users' perception of content bias and source credibility. In the first study, we investigated the impact of happy and angry portrayals in facial imagery on users' decisions. In the second study, we explore the interaction between users' prior attitude towards multiple personalities and how those personalities are portrayed visually on their decisions. Our results show that the systematic usage of angry facial emotions in images increases users' perceived content bias and decreases the perceived source credibility. These results highlight how implicit visual propositions by news sources impact our judgements and pave the way for visual analytic systems that are sensitive to users' individual and group interactions with such visual information.



## ACKNOWLEDGMENTS

After eight years of being away from my parents, siblings, and friends in Iran, I wrote this document. I am thankful to many who helped me throughout the ups and downs of these years.

I am thankful to my family:

To my mother, Vahideh Gandomikal, and my father, Habibollah Kardooni, who both supported me unconditionally. You both are pillars for everything I am. I cannot describe how grateful I am to them and how excited I am to be physically close to them soon. Any achievements and progress I have ever made are thanks to them.

My sister, Parivash, whom I lost just about when I first came to the US. My memory of her unconditional love and care for us siblings helped me push through.

To my siblings, Neda, Hoda, and Roozbeh, who looked after me every day and every moment, They are great scholars and enormous inspirations for me. I am beyond thankful to have sisters I can always talk to and a brother who is there for all of us.

To Bisharafa for playing online games with me every day during COVID: My nephews, Reza and Ali, My cousins Hossein and Hassan. And Sina, who is working his way through the family.

I am thankful to my friends in the US, who are like family to me. Alireza Bahrami-rad, who is always there for me like a true brother. And who yelled at me countless times so I would finish my dissertation in time. Dr. Hamed Pakatchi, who explained complicated concepts to me and helped me formulate my thoughts like a teacher. Dr.

Amirhassan Kermanshah, who, in pajamas, we wrote my first ever academic article together. Mitra Mostafavi, for helping me with designing my studies and bringing Kambucha every once in a while. Dr. Ashkan Mahdavi, Dr. Mehdi Sharifzadeh, Dr. Poorya Naderi, Dr. Saman Mostafavi, and Dr. Armin Sarabi, who not only were my virtual poker buddies during COVID but at several points through my dissertation research, have helped me through their friendship and expertise. I am thankful to my friends, Hana Maleki, for her unconditional friendship. Elina Moghani, for making sure I do things right. Sogol Moshtaghi, for her rare but life-long sharing of our laughter and depressions. Niloofar Salavati, hard to put into words, but especially for “Verdi ke barre ha mikhanand”. Hac Tran and Sara Ellis, for being my TRUE AMERICAN friends. Noushin, Ashkan, Ali Fatemi, and Atefeh my Charlotte friends who made it bearable to live there.

And, Sheida, my dearest friend, whom I lost the moment I came to Charlotte and whose memory and love motivated me to work hard throughout my life at UNC-Charlotte.

...

I’m incredibly grateful to my Adviser, Dr. Wenwen Dou, for her constant support and mentorship. I couldn’t ask for a better adviser during my Ph.D studies.

This dissertation is, in essence, a collaborative work. It would not be possible without my dear co-authors and collaborators: Ryan Wesslen, Dr. Isaac Cho, Dr. Samira Shaikh, Sashank Santhanam, Dr. Doug Markant, and Dr. Wenwen Dou.

I am thankful to Professor Eric Sauda for being a fantastic mentor and collaborator and ensuring I understand that research should always be in conjunction with

enjoyment.

I am thankful to my doctoral committee members, Dr. Isaac Cho, Dr. Sara Levens, Dr. Samira Shaikh, Dr. Wenwen Dou, Dr. Doug Markant, and Professor Eric Sauda, for their guidance and comments throughout this process.

And I am thankful to anyone whose name I might have forgotten to mention as I wrote this messy acknowledgment.

## TABLE OF CONTENTS

LIST OF FIGURES	xii
LIST OF TABLES	xv
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: BACKGROUND	6
2.1. Introduction	6
2.2. A framework for studying misinformation	6
2.2.1. Defining Misinformation	7
2.2.2. Elements involved in the production of misinformation	9
2.2.3. Elements of misinformation and their relationships	14
2.3. Misinformation and computation: current approaches	19
2.3.1. Content: Automatic Fact-Checking and Misinformation detection	20
2.3.2. Measuring and categorizing trustworthiness and veracity of sources	24
2.3.3. Human-Misinformation interaction: Interactive tools, Visual Analytic Systems, Cognitive bias mitigation efforts	28
2.4. discussion and conclusion	32
CHAPTER 3: Studying Uncertainty and Decision-Making About Misinformation using Visual Analytics	35
3.1. Introduction	35
3.2. Related Work	38

3.3. Verifi: A Visual Analytic System for Investigating Misinformation	41
3.3.1. Dataset	42
3.3.2. The Verifi User Interface	45
3.4. Experiment Design	47
3.4.1. Research Questions	47
3.4.2. Experiment Tasks	48
3.4.3. Experiment Procedure and Participants	52
3.5. Data Analysis Methods	53
3.6. Analyses Results	54
3.6.1. RQ2: Measuring the Impact of Uncertainty	55
3.7. Discussion and Future Work	59
3.8. Conclusion	61
CHAPTER 4: Visual Analytics for Multimodal Exploration of Misinformation Source Behavior	62
4.1. Introduction	62
4.2. Production and Identification of Misinformation	64
4.2.1. Why and how we believe misinformation	67
4.2.2. Battling misinformation	69
4.3. The Design of Verifi2	71
4.3.1. Task Characterization	72
4.4. Verifi2	73
4.4.1. Data Analysis and Computational Methods	76

	x
4.4.2. Visual Interface	81
4.5. Usage Scenario: exploring news related to an organization	85
4.6. Expert Interviews	90
4.6.1. A spectrum of trustworthiness rather than binary classes	91
4.6.2. Verifi2 as a potential tool for education on misinformation	94
4.6.3. Positive and negative usage of facial expressions in images	98
4.7. Discussion	99
4.8. Conclusion	101
CHAPTER 5: IMPACT OF EMOTIONAL FACIAL EXPRESSIONS IN IMAGES ON USERS' JUDGEMENTS OF CONTENT BIAS AND SOURCE CREDIBILITY	102
5.1. Introduction	102
5.2. Study motivation, design, and implementation	105
5.2.1. Dependent variables and elicitation method	106
5.2.2. Dataset	108
5.2.3. Study interface and procedures	109
5.2.4. Study 1 & 2 models:	110
5.2.5. Topic modelling and qualitative analysis of topics:	110
5.3. Study 1	111
5.3.1. Hypotheses:	113
5.3.2. Dependent & independent variables:	114
5.3.3. Model specification:	114

	xi
5.3.4. participants	115
5.3.5. Results	115
5.3.6. Analysis of users' comments:	118
5.3.7. Discussion:	122
5.4. Study 2	125
5.4.1. Experiment design:	126
5.4.2. Hypotheses:	127
5.4.3. Dependent & independent variables:	128
5.4.4. Study 2 model specification:	129
5.4.5. Participants:	129
5.4.6. Results	129
5.4.7. Qualitative analysis of users' comments:	131
5.4.8. Study 2 discussion:	135
5.5. Discussion	138
CHAPTER 6: CONCLUSION AND FUTURE DIRECTIONS	142
REFERENCES	145

## LIST OF FIGURES

FIGURE 1: A reconstructed diagram describing the categorization of false information by Wardell and Derakhshan[201].	10
FIGURE 2: Elements of misinformation and their relations.	19
FIGURE 3: The Verifi interface: Account View (A), Social Network View (B), Tweet Panel (C), Map View (D), and Entity Word Cloud (E). The interface can be accessed at <a href="http://Verifi.Herokuapp.com">Verifi.Herokuapp.com</a> .	40
FIGURE 4: Top 20 most predictive language features of Fake and Real news outlets as measured by each feature's average effect on Accuracy. 't' prefix indicates the feature is normalized by the account's tweet count and 'n' indicates normalization by the account's word count (summed across all tweets). Features with borders are included in Verifi.	45
FIGURE 5: Available cues for selected accounts (column) and users' response regarding the importance of these cues (row, Q1-Q6). Left: Shows each of the eight selected accounts as well as the cues available for each of them. Right: Shows average of importance for each cue per account based on participants' responses. Values in gray circles below each account name show average accuracy for predicting that account correctly. The left figure is purely based on the (conflicting) information presented in the cues and is independent from user responses. The right figure based on the user responses on the importance of each cue coincides with the information in the left table.	49
FIGURE 6: A sample of users' comments about their decisions. Highlighted text shows users' mention of either a qualitative or quantitative reason. Green denotes reasons/cues pointing to the account being real while red pointing to being fake.	57
FIGURE 7: The Verifi2 interface is designed to support the investigation of misinformation sources on social media. The interface consists of 5 views: A) Account view, B) Social Network view, C) Entity Cloud view, D) Word/Image Comparison view and E) Tweet view.	68



- FIGURE 8: Verifi2 pipeline. Tweets, meta information, and images from 285 accounts are collected from Twitter streaming API. Language features, named entities, and word embeddings are extracted from the tweets. A social network is built based on the mention/retweet relationships. The accounts are clustered together using community detection of a bipartite account/named entity network. Image similarities are calculated based on feature vectors. All of these analysis are stored in a database and used for the visual interface. 74
- FIGURE 9: Word Comparison view. The query word is “North Korea”, the top related keywords from real news accounts are shown on the left while the keywords from suspicious accounts are displayed on the right. 84
- FIGURE 10: Top: Illustrating comparison between top semantically related words to the entity “GOP” (The Republican Party). Bottom: Community of news accounts in the social network that most mention the term GOP. 86
- FIGURE 11: comparison of images/tweet pairs between real and suspicious news shows how these groups use images differently. 87
- FIGURE 12: The Image Comparison view highlights how suspicious accounts frame images to convey messages. 87
- FIGURE 13: The Line + Range elicitation method. 107
- FIGURE 14: Study 1 conditions and process 112
- FIGURE 15: Study 1 fixed effects odds ratios for bias choice (left) and bias uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: \*\*\* 99.9%, \*\* 99%, \* 95%. For image\_emotion, the reference category is happy. For image\_shown, the reference category is no, left is the reference condition for source\_orientation, and mainstream is the reference condition for source\_type. 117
- FIGURE 16: Study 1 fixed effects coefficients for credibility choice (left) and credibility uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: \*\*\* 99.9%, \*\* 99%, \* 95%. For image\_emotion, the reference category is mixed. For image\_shown, the reference category is no, left is the reference condition for source\_orientation, and mainstream is the reference condition for source\_type. 118

FIGURE 17: Mean and bootstrapped 95% confidence interval of users' responses for each account.	123
FIGURE 18: Study 2 conditions and process	126
FIGURE 19: Study 2 fixed effects Odds Ratios for bias choice (left) and bias uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: *** 99.9%, ** 99%, * 95%. For image_emotion, the reference category is no image.	130
FIGURE 20: Study 2 fixed effects Odds Ratios for credibility choice (left) and credibility uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: *** 99.9%, ** 99%, * 95%. For image_emotion, the reference category is no image.	132
FIGURE 21: Mean and 95% bootstrapped confidence interval of choices for each condition/politician.	136

## LIST OF TABLES

TABLE 1: Distribution of types of news outlets	42
TABLE 2: 34 candidate language features from five sources.	44
TABLE 3: Eight accounts with masked account names. Background colors indicate real (green) and fake (red).	48
TABLE 4: User accuracy and Fake prediction across conditions.	55
TABLE 5: Log odds ratios for each independent variable in two logistic regressions. The Accuracy column is 1 = Correct, 0 = Incorrect Decision. The Fake column is the user’s prediction: 1 = Fake, 0 = Real. The @accounts variables use @XYZ as the reference level and the Group variables use the Control Group as the reference level.	56
TABLE 6: 8 Twitter accounts that users evaluate in study 1. Each user goes through content sorted based on emotions in facial expressions collected from tweets’ images.	113
TABLE 7: Study 1 topic model of users’ comments. Shows a table of 20 extracted topics sorted based on number of unique users with comments assigned to each topic. [41].	119
TABLE 8: Politicians selected for study 2	128
TABLE 9: Study 2 topic model of users’ comments. Top shows a table of 20 extracted topics sorted based on number of unique users who had comments associated with each of the topics. [41].	133

## CHAPTER 1: INTRODUCTION

Throughout history, misinformation has been used intentionally to manipulate people's opinions and beliefs [201]. Technological advancements have been a crucial element in the development of misinformation. Its earliest traces can be traced back to societies with the earliest writing systems through which rulers would falsify written record to glorify themselves and demean enemies [117, 180]. The effects of information manipulation for political or economic gain only increased by the invention of new technologies such as print, press, and later in the 20th century by the explosive utilization of mass media such as television and the radio [180, 76, 23]. The birth of the Internet, Information Communication Technologies (ICTs), and social media drastically increased the rate and means of information production, curation, and sharing.

Consequently, these massive amounts of misinformation affect many people on a global scale [108, 78, 131, 20]. The prevalence of false information in societies can make democracies ungovernable [29]. As the advancements in computation and communication technologies have played a significant role in the growth of misinformation and its threats to our societies, we now face a dire need to develop new technological efforts and tools for effectively battling this growing problem.

The issue of misinformation today is more complex and multi-faceted than ever. Individuals and organizations can now easily create and disseminate information

on social media. Numerous agents with malicious or non-malicious intents create a phenomenon often labeled as “fake news” [201, 108]. Furthermore, misinformation sources can use automated/semi-automated bots on social media platforms to rapidly spread misinformation [56, 166]. Consumers are likely to believe and share without rigorous fact-checking. Cognitive scientists have highlighted various factors such as prior exposure to news[139], selective exposure, and confirmation bias, causing audiences to believe misinformation [108]. News outlets take advantage of these psychological factors and introduce slants, falsities, or political biases into their news [23, 18].

Additionally, social media platforms tend to introduce algorithms to curate information that might create filter bubbles or echo chambers through which audiences are less likely to be exposed to news they disagree with [136, 26]. There are many terms used by scholars, journalists, and politicians that refer to the accuracy and intentions of information and media, including propaganda, disinformation, misinformation, and fake news. However, there are no agreed-upon definitions for these types of information manipulations [201, 179, 82]. These complexities in sources, types, means of production, and definitions of misinformation along with different psychological and social factors highlight the need for a comprehensive and multi-disciplinary approach towards preventing and intervening misinformation.

Even though combating misinformation is at the forefront of many political, journalistic, and scholarly discussions, our attempts on combating misinformation are still limited. Most of the existing computational approaches fall under Automatic Fact-Checking (AFC) and misinformation detection, including automatic or semi-

automatic identification, verification, and correction of the misinformation. To date, the effectiveness of these methods without human supervision remains very limited [67]. Furthermore, correcting information does not necessarily result in a change in belief [133, 60, 122]; while repeating misinformation even for fact-checking might prove to be counterproductive [175]. Lazer et al. suggest the necessity of a comprehensive strategy on education, empowering individuals, along with a collaborative approach between industry and academia on battling misinformation [108]. Furthermore, they emphasize the importance of shifting focus from the veracity of content (e.g., if a piece of news is factual or not) to evaluating sources' trustworthiness.

In this dissertation, we aim to develop a deeper understanding of how users' judgments and decisions are shaped by the content they observe. This requires an effort to assemble and develop state-of-the-art computational and visualization methods that highlight sources' biased and malicious behavior and understand how these methods can affect users' decision-making processes about misinformation. Thus, we situate this dissertation at the intersection of visual analytics and human decision-making.

To do this, we first need to develop a robust framework and knowledge to address this problem. In chapter 2, with the goals of finding gaps in the computer science research on misinformation, I first provide an overview of different aspects of misinformation discussed in other disciplines, which allows us to develop a simple but useful framework for studying misinformation. The framework identifies four main elements in the ecosystem of misinformation: Source, Content, Consumers, and Truth. Furthermore, by focusing on the relationships between these elements, we identify gaps in the computational efforts towards combating misinformation.

In chapter 3, we describe our work on developing a first visual analytic system focusing on multimodal exploration of misinformation. Verifi allows users to study misinformation by analyzing language features and social behaviors of sources of news. Using Verifi, we conducted an in-lab user study to understand the effects of uncertainty and users' biases on their decision-making processes. The user study results help us identify how the uncertainty of information negatively affects users' ability to identify trustworthy news sources. Moreover, we show how qualitative and analytical analysis helps users achieve better accuracy in determining sources' veracity.

In chapter 4, we transform Verifi into a Visual Analytics System with more robust multimodal features. Verifi2 is among the first systems that enable users to study and explore news sources' behavior using social networks, linguistic features, word, and image similarity. We evaluate Verifi2 by interviewing experts on misinformation and psychology to understand how such systems can improve users' understanding and when and how users decide to trust misinformation sources. Our interviews reiterate the importance of uncertainties in this field. Furthermore, they highlight the importance of creating a system that remains simple and easy to use while allowing users to engage with the uncertainties inherent within misinformation content. Moreover, our domain experts highlighted the primacy of understanding how visual information impact users' decision making around misinformation sources.

In chapter 5, inspired by the lessons learned from our domain expert study, we conduct two consecutive controlled experiments on how positive (happy) and negative (angry) images might impact users' judgements about content and sources of misin-

formation. Throughout both studies, we used a novel elicitation method based on our prior work on belief updating of correlations [94]. The method is called Line + Range and allows us to elicit beliefs and uncertainty of users' decisions as continuous values. Our controlled experiments provide evidence of the impact of images with angry emotions on users' perceptions of bias and credibility of sources. Moreover, qualitative analysis of users' self-described rationale for their decisions highlights a complex set of heuristics primarily focusing on text and possibly reinforced by systematically emotionalized images.

The interdisciplinary work in this thesis contributes to Visual Analytics, Human-Computer Interaction, and Decision Making. It offers both technological advancements through new visual analytics systems for combating misinformation and novel knowledge about how humans make decisions under uncertainty and based on multiple modes of information such as text and images. This thesis concludes by describing multiple future paths for research at the intersection of these disciplines on mitigating the effects of misinformation on users.



## CHAPTER 2: BACKGROUND

### 2.1 Introduction

In this chapter, I will provide a survey of literature on Misinformation both from a computational perspective and from other related disciplines concerned with misinformation. The structure of this survey is as follows: in section 2.2.1 I offer an overview of different terms and definitions related to misinformation. In section 2.2.2, I describe my adopted framework for studying misinformation which calls for broadening our focus from fact-checking the content of misinformation to include interactions between content, sources, and consumers of misinformation. In section 2.2.3, I overview related literature from psychology, cognitive and social sciences categorized by each element. In section 2.3, I survey existing misinformation-related literature in computer science related to content (section 2.3.1), sources (section 2.3.2), and consumers (section 2.3.3). Finally, in section 2.4, I discuss the lacking areas of the current literature on computation and misinformation and propose an agenda for future interdisciplinary research.

### 2.2 A framework for studying misinformation

In this section, I propose a framework for studying and categorizing misinformation. I do so by first describing the different definitions needed to describe the misinformation phenomena better. These definitions help us understand how we can

categorize misinformation. Then, I describe the different elements involved in the production and consumption of misinformation including producers, the content of misinformation, and the audience. By combining different definitions and recognizing the elements involved in misinformation, we will be able to critically analyze the existing literature involving computation and misinformation.

### 2.2.1 Defining Misinformation

After United State’s presidential election at 2016, the term “fake news” has become a topic of interest in many journalistic and political circles around the globe [?, 145, 201]. Fake news has been defined as “deliberately constructed lies, in the form of news articles, meant to mislead the public” [173]. Many scholars from different disciplines have elected to use this term as they discuss the problem of misinformation [108, 20, 198, 102]. A survey of 34 scholarly articles using the term fake news showed that it is used to describe a wide range of concepts including news satire, news parody, news fabrication, photo manipulation, advertising in the guise of news reports, and propaganda. The authors conclude that the common feature between all mentions of “fake news” is that they all “appropriate the look and feel of real news; from how websites look; to how articles are written; to how photos include attributions.” [179].

Other scholars suggest refraining from using the term fake news; as it generally fails in correctly and accurately describing the complexities of misinformation [201, 82, 200, 173, 213]. Starbird focuses on a dichotomy of “alternative” and “mainstream” media outlets. She utilizes conspiracy theories and alternative facts as the term to describe misinformation and “fake news” [172]. However, scholars such as

Castells and Chomsky have shown various ways mainstream media utilize misinformation to influence peoples' opinions [76, 23]. Caroline Jack from Data and Society calls for more accurate terminology for discussing "problematic information" because each word might infer assumptions about the producer, the type of message, and the persons receiving the information. The report emphasizes the importance of paying attention to these factors as they greatly affect the strategies needed for combating them. She differentiates between misinformation and disinformation by defining misinformation as "information whose inaccuracy is unintentional" while disinformation describes "purposefully falsifying information". She also highlights the difficulties in differentiating between publicity and propaganda. While both are geared towards influencing audiences, propaganda is often referred to as attempts to deliberately manipulate or deceive people. Moreover, Jackson describes a third category which differs from disinformation or propaganda. While these concepts mostly aim to gain support for different beliefs or idea, some events, described by the term "gaslighting" use falsified information to create uncertainty and tension in various societies. [82].

A report on "Information Disorder" by Wardell and Derakhshan produced for the Council of Europe, categorizes problematic information into three groups by asking two questions: Whether a piece of information is harmful or not and whether it is false. These categories are Misinformation, Disinformation, and Malinformation [201]. Misinformation is information that is false but is not produced with an intent to harm. It is created when journalists or individuals share misinterpretation about an event or a rumor without realizing the information is not accurate. For example, after the Bombing at the Ariana Grande concert in Manchester, images of several individuals

were tweeted by multiple agencies as missing. Several of those individuals, in reality, had nothing to do with the bombings [203]. Disinformation, on the other hand, is false information that is created with the intention to harm groups or individuals. A well-known example of Disinformation was a conspiracy known as Pizzagate created by Alex Jones of Infowars that Hillary Clinton had sexually abused children in satanic rituals [157]. Malinformation is information that is not false but is released, often illegally, with the intent to harm. Two examples of such malinformation are the release of personal emails before two elections in the United States from Hillary Clinton, which is believed to have affected the results of the presidential election, and Emanuel Macron, which was not effective in its intent [16]. Figure ?? shows the three categories as defined by Wardell and Derakhshan.

As noted by many scholars whose work involves misinformation, fake news is not an adequate terminology for describing the complexity of the problem. For consistency with these suggestions, I define misinformation as *information that is false, with or without the intent to harm or manipulate consumers*. My definition corresponds to the combined definition of Misinformation and Disinformation offered by Wardell et al. [201]. In the next section, I describe the elements involved in the production of misinformation that serve as a framework for studying related computational methods addressing this phenomenon.

### 2.2.2 Elements involved in the production of misinformation

The defining factor in the study of misinformation is its relationship to “facts” and “truth”. Most technology companies and news outlets approach misinformation

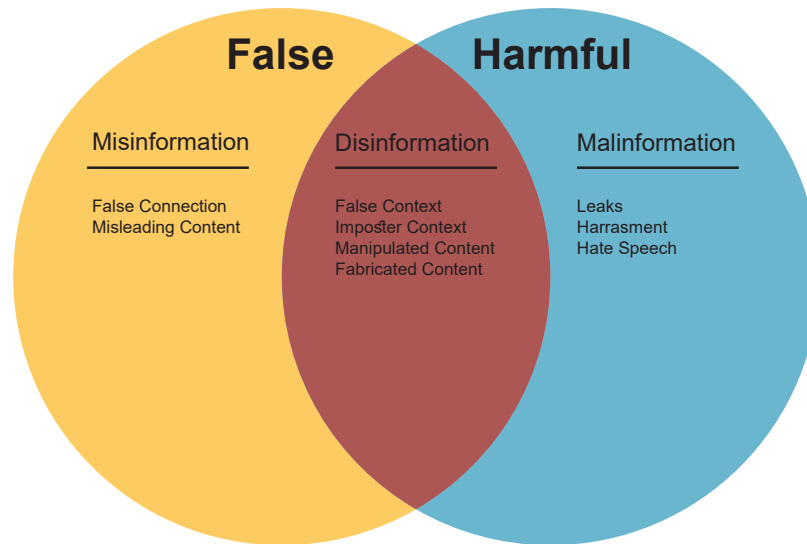


Figure 1: A reconstructed diagram describing the categorization of false information by Wardell and Derakhshan[201].

by providing rigorous analysis and “fact-checking” of news. However, Lazer et al. suggest that communicating the results of fact-checking by itself is not enough and in some cases can prove to be counterproductive [108]. One of the biggest challenges in addressing the problem of misinformation is the multiplicity of elements involved in the process of production and consumption of misinformation. Scholars from multiple disciplines have highlighted these elements.

In an analysis of the social production of misinformation in the United States during the Iraq War in 2004, Arsenault and Castells describe a complex model of misinformation that includes media organizations, political actors, general psychological climate, and the mental frames of the audience. One of the interesting factors they discuss is misinformation producers’ usage of specific language using framing, biases, slants, and metaphors. The authors discuss in detail, how these elements synthesize and result

in social misperceptions [23]. The significance of their study is how they describe a model of misinformation that involves the producers of misinformation, their agenda, their usage of emotions and language, and the audiences emotional and mental state. In a study of "fake news" in the US presidential election of 2016, Bakir and McStay argue that we should see the current social media driven landscape of misinformation in light of the systematic, political and commercial efforts in liberal democracies to influence opinions and beliefs of populations through propaganda. They count five main elements as crucial in the current landscape of misinformation: "The financial decline of legacy news; the news cycle's increasing immediacy; the rapid circulation of misinformation and disinformation via user-generated content and propagandists; the increasingly emotionalized nature of online discourse; and the growing number of people financially capitalizing on algorithms used by social media platforms and internet search engines" [25]. They also recognize the multiplicity of elements and the importance of observing the economic benefits and intentions of sources of misinformation, as well as the emotional state of the audience.

Vargo et al. study the agenda-setting power of "fake news" and consider partisan media, "fake news" media, and fact-checkers as three separate entities that influence people's opinions consequently have power in setting domestic and international policy agenda [186]. Lazer et al. emphasize the importance of focusing not only on the message but also paying attention to the source of misinformation as well as taking an educational approach to address the consumers of misinformation [108, 122]. Wardell and Derakhshan, describe three essential elements including Agent ( or creator ) of misinformation, the message, and the interpreter. For each of these elements, the

authors describe multiple vital factors to consider including the intent and type of agent, the emotional content and type of the message, and the mental state of the interpreter[201].

In the report by Wardel and Derakhshan curated for the council for Europe, the authors define three main elements in the process of “information disorder”: 1) The agent or the actors who initially distribute the message, 2) the message which is the content of misinformation and encompasses text, images, and other types of media, 3) and the interpreter who is the person consuming the message. Inspired by these three elements while taking into account other scholarly definitions [108, 179, 23, 122], I adopt three main elements of misinformation for conducting critical analysis on means of computationally battling misinformation: source, content, and consumers of misinformation. The reason I chose this terminology is that it more accurately represents different elements of misinformation. In contrast to Wardel and Derakhshan who describe the agent as actors who initially “create and produce and distribute the message” [201]; I use the term source to more broadly describe any source including bots and redistributors of misinformation that have not initially created a message. I also use content to more broadly refer to the textual, visual and multimedia content of misinformation. Finally, I use consumers as a broader definition to include groups or individuals who might be the target of misinformation.

- **Source:** The outlet through which misinformation is being consumed. these outlets can include:

- professional News accounts or agencies

- Journalists, bloggers, social media personalities, or news aggregates
  - Bots that disseminate information whether detected or undetected.
  - Other individuals who share misinformation
- **Content:** the content of misinformation that is being distributed. The content can include:
    - textual messages on social media
    - news online from news agencies and blog
    - visual information including images and videos
  - **Consumer:** Individuals or groups who are exposed to or effected by misinformation who make various decisions in regards to misinformation:
    - they make many decisions when exposed to these content: to trust the content, to trust the source, and to share and propagate the content.

Using these elements as a framework for analyzing misinformation allows us to systematically study misinformation from creation or dissemination by sources to consumption by the audience. Through inquiries about the relationship between these elements, we can raise important questions about misinformation. By looking at the relationship between sources and contents of misinformation, we can start to question the means, and methods sources use to create misinformation. Also, by looking at the relationship between consumers and the content, we can explore the psychological and social factors of why individuals believe in misinformation. In the next sections, I explore literature related to each of these elements.



### 2.2.3 Elements of misinformation and their relationships

#### 2.2.3.1 content-truth relationship: facts, fact-checking, and truth

The relationship between misinformation and facts is likely the most prominent aspect of fighting misinformation. Fact-checking is the primary means of dealing with this relationship and is the main approach by social media platforms, technology companies, media outlets, and third-party organizations that specialize in this task. Social Media platforms take part in the process by taking various measures such as employing third party fact-checkers and developing new technologies to automatically detect fake news [14, 1, 8]. Many third-party organizations exist that specialize in flagging misinformation content and sources and communicating the results with consumers [4, 6, 9, 5]. These organizations provide a variety of information such as different scales of rating and labeling news pieces. They cover political statements, claims, TV ads, and articles. These organizations tag news pieces as suspicious, completely false, misleading, or out of context. There are also numerous fact-checking attempts that use crowd-sourcing as their primary means of battling misinformation[7, 3]. These systems are still early in development, and their efficacy is yet to be examined. Even though fact-checking misinformation is a critical task, its effectiveness in battling misinformation has been questioned [108, 107, 57].

#### 2.2.3.2 source-consumer relationship: sources' intent

When studying Misinformation from the perspective of sources, the first question that comes to mind is the intent of source or why the source is delivering misinformation to consumers. Lazer et al. argue that understanding the intent of sources

is a primary factor in understanding misinformation and should be given prevalence over the factuality of a single news story [108]. Different motivations and intents have been identified about sources of misinformation. These include financial, economic, and advertisement [25, 23, 34, 76, 177]; ideological or partisan[20, 176]; political and power [23, 176, 179], and satire and parody[46, 28, 201]. Based on these intents, sources take different strategies of content propagation and manipulation to influence Consumers.

### 2.2.3.3 Source-content relationship: means of propagation, verbal and visual frames, strategies

The means of which sources propagate misinformation can be an indicator of its intentions. Besides prominent misinformation sources with strong editorial staff, many take extensive use of social bots. Social bots are emerging phenomena that act as sources of misinformation [32]. They impersonate real sources of misinformation and are often automatic or semi-automatic. They have the power to rapidly propagate misinformation and pollute the information space [56]. Social bots are active in the early stages in the life cycle of misinformation and also interact with high profile accounts on social media [166]. Social bots tend to behave similarly to real individuals and sources on social media by liking, sharing, and commenting on other sources of news [108]. By amplifying the spread of misinformation, social bots take part in the creation of echo-chambers and activating consumers' cognitive biases [107].

Sources do not treat information uniformly, and they take different manipulation strategies and framing techniques to produce believable stories. News media that

produce misinformation often share information in a biased manner. They focus on partisan topics and issues, as well as inflammatory topics, emotional content, specific moral foundations, or specific geographies and political figures. They often focus on topics from online partisan news media and also have the power to set the issue agenda for those accounts[185]. Moreover, these sources tend to disseminate news about topics that are important to specific populations. These focuses include biases towards extreme ideologies, political parties and figures, and inflammatory issues[171, 23, 20]. These topics are chosen based on the psychological climate of audiences often focusing on fear-mongering, amplifying anger and outrage in consumers.[23, 25]. Sources of Misinformation, often focus on specific moral visions and foundations of different groups, to take advantage of their mental framing.[53, 23, 106]. Moreover, they tend to cover news related to only a subset of geographic places, as well as political figures [18, 20].

Misinformation sources take extensive use of visual information and images to mislead consumers. They use images to convey different biases not necessarily detectable in textual content[61]. Images contain implicit visual propositioning that can communicate various ideological or stereotypical messages that might receive greater resistance when are put in words. Thus viewers and consumers are more likely to be unaware of the implicit biases and frames in visual content. [17, 123]. This powerful tool has been used in various contexts including the negative portrayal of different political candidates from different parties[64], out-of-context usage and altering of images[115], or by merging multiple images to change the implicit message [36]. Images of people, either prominent or not, has significant effects on consumers' opinions.

It has been shown that altering images can produce significant differences in people’s assessments of a political figure [43].

#### 2.2.3.4 Consumer-content relationship: social factors and cognitive processes

Sources produce misinformation with different intents and using different strategies. However, the reason they are often successful in communicating their stories is only partially due to their approach. Humans are affected by multiple social, cognitive, and psychological processes that make them prone to believe and further share misinformation.

Various social phenomena have been identified to influence consumers’ opinions and decisions toward misinformation. Allcott and colleagues showed that propagation and belief of misinformation were influenced by ideological polarization of consumers [20]. Moreover, Pennycook and Rand’s research on individuals receptivity towards fake news showed that conservative, right-leaning individuals were more likely to believe misinformation [142]. This polarization can be a direct result of “echo chambers” influenced by the “filter bubble” phenomenon where algorithms by technology companies shape the information consumers are exposed to [171]. Even though some studies show that there is no reliable evidence for the existence of filter bubbles [214] we can find many examples of echo chambers on various social media platforms [174, 58, 27]. Moreover, individuals are more likely to be affected by misinformation when in a social setting[62]. Moreover, a study on students’ tendencies to share misinformation showed that many social reasons such as “sharing eye-catching messages” or “interacting with friends” ranked high as causes of sharing misinformation [38].

Overall, it has been noted that we are more likely to believe a piece of information if our social circles also accept them [169, 107, 49].

Many causes of why consumers believe misinformation, has been attributed to psychological and cognitive factors. These factors include different cognitive biases as well as different degrees of analytical thinking. Kahneman and Tversky describe a form of cognitive bias popularized as Availability Bias in which a person evaluates a probability based on how easily relevant information come to mind [183, 171]. This form of bias highlights how access and exposure to specific information can have an impact on opinions and decisions. On a study on causes of misinformation, Pennycook and colleagues showed that prior exposure indeed significantly affects consumers perception of the accuracy of misinformation [140]. Another aspect of why consumers believe misinformation is selective exposure which can be described as consumers' tendency to believe information that aligns well with their views and beliefs and also avoid information that is against their pre-conceived notions [171, 107, 97]. One of the main causes of this selective exposure is known to be confirmation bias, or our tendency to privilege evidence that confirms our existing hypothesis over all possible hypotheses [128, 107]. Consumers' emotional proximity to different topics and events also increases their susceptibility to believing misinformation [79]. In contrast to consumer's cognitive and emotional state, the ability to perform analytically thinking and problem-solving has been shown to increase our resiliency to misinformation [141, 142].

Focusing on elements involved in the production and consumption of misinformation and their relationships provides us with a systematic framework derived from

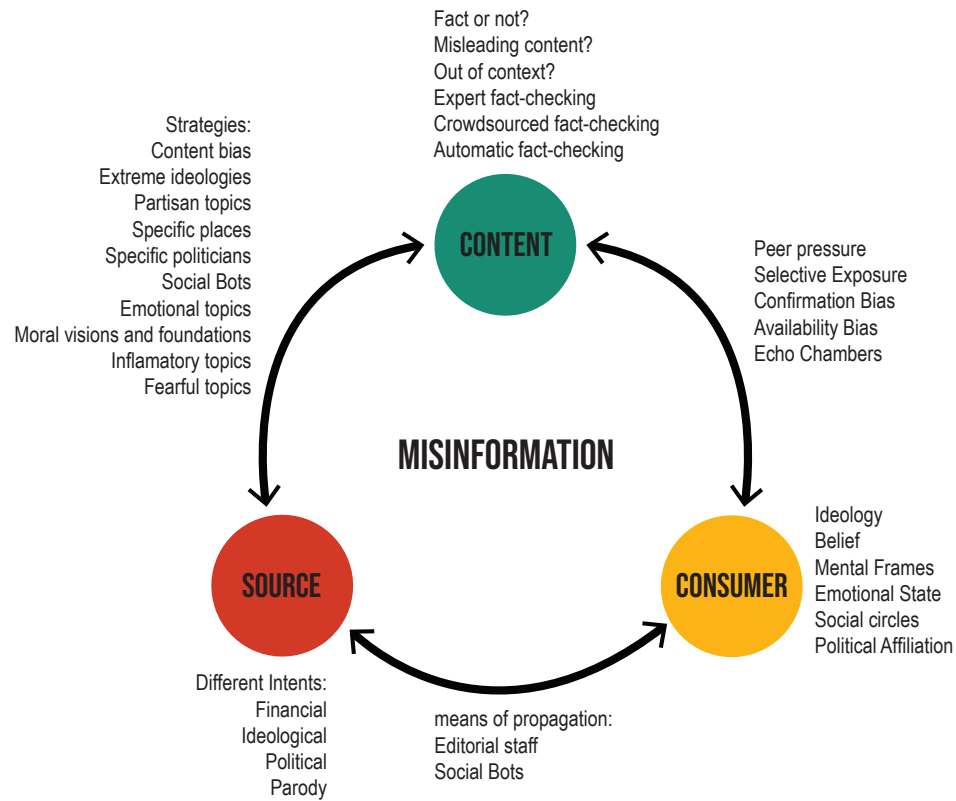


Figure 2: Elements of misinformation and their relations.

multiple disciplines to study how we can use computation to battle misinformation (see figure ??). In the next section, I provide a summary of the state-of-the-art literature in computation related fields that focus on the problem of fake news and misinformation. section 2.2.2 will serve as a framework for categorizing existing works on misinformation and highlighting any existing gaps in our current approaches.

### 2.3 Misinformation and computation: current approaches

In this section, I provide an overview of the current scholarly studies on misinformation within the fields related to Computer Science and Information Technology. Each section follows the categorization based on the framework and elements discussed in

the previous section. It is important to note that many of the mentioned methods focus on multiple elements of misinformation. However, the categorization is based on whether the primary focus of the scholarly work was on either of the elements. In the coming section, I first discuss efforts on Automatic Fact-checking as well as fake news detection. Next, I provide research that focuses on sources of misinformation and their relationship to misinformation. Finally, I focus on research that focuses on the human aspects of misinformation and computation.

### 2.3.1 Content: Automatic Fact-Checking and Misinformation detection

One of the main topics in the mainstream media in regards to misinformation is fact-checking the veracity and credibility of news articles. The report for the Reuters Institute for the study of journalism defines Automatic Fact-Checking (AFC) as using technologies to "deciding the truth of public claims and separating legitimate views from misinformation." This task, also dubbed as deception detection is defined as the "prediction of the chances of a particular news article (e.g., news report, editorial, expose) being intentionally deceptive ( fake, fabricated, staged news, or a hoax)" [159]. In the report, three main elements are defined for Automated Fact-Checking: 1) Identification which involves monitoring news media and sources, identifying factual statements, and prioritizing claims to check; 2) verification which involves checking with other existing fact-checks, checking against authoritative sources, and unstructured credibility scoring; and 3) correction which involves flagging repeated falsehoods, providing contextual data, and publishing new fact-checks [66]. Several systems exist that try to address one or more from these categories.

A unique example of a system that provides claim detection, verification, and correction is ClaimBuster which combines a combination of Natural Language Processing and Supervised Learning to classify and score whether sentences are "check-worthy" or not [72, 74]. They classify claims by presidential candidates into Non-Factual Sentences, Unimportant Factual Sentences, and Check-worthy Factual Sentences. The system produces a score that reflects the extent to which a sentence belongs to the check-worthy group. They extract features from text and use Random Forest to find the most discriminating ones. They use an SVM model combined with words, Part of Speech tags and Entity extractions and achieve a precision of 72% and a recall of 67% [72]. Later iterations of the System, assesses check-worthy claims with databases of third-party fact-checks based on similarity. Moreover, it provides supporting and debunking evidence from knowledge bases and web [74]. Other attempts in fact-checking generally use external sources to check specific claims or news pieces[168]. They use either open web sources using statistical scoring [114] or knowledge graphs that can act as ground-truth for misinformation claims [48].

Other approaches to misinformation detection deviate from fact-checking and focus on categorizing news based on features from their content. Numerous research projects have used writing styles and language features visible in the text. A group of scholarly work uses a "bag of words" representation to analyze and distinguish misinformation[158]. These methods include lexicon based dictionaries of moral foundations and subjectivity [191, 138, 192] and location-based words [135]. [168]. Other works have used more complex syntactic analysis such as the concept of deep syntax [144] to detect deception with high accuracy in multiple datasets [55]. Another



approach to using linguistic features of news content is by building classifiers based on various language features. Oraby et al. compare an automated bootstrapping method of extracting discriminating features from annotated “fact” vs. “feeling” arguments with a Naive Bayes classifier and show that both methods perform well in discriminating these features [134].

Potthast et al. have used a modified version of unmasking, a method to determine whether to articles are from a single author or not [149]. The method iteratively removes most distinguishing features from two articles and observes the rates of which cross-validation accuracy drops [104]. The authors show that various stylistic features can distinguish real news content from hyper-partisan media, satire, and fake news. Bourgonje, Schneider, and Rehm approach the problem of misinformation by focusing on Headlines [34]. They use a logistic regression classifier for detecting the stance of headlines concerning the body of their articles. Various word level features including punctuation marks and part of speech tags were used to train an SVM classifier for predicting satire and real news articles with 82% accuracy [158].

Scholars have also been able to detect misinformation based on features extracted from their visual information. A few studies so far have attempted to detect misinformation based on image information. Gupta et al. used a dataset of “fake” images distributed during hurricane sandy [70]. The authors used only non-visual features such as URLs, propagation patterns, user features, as well as tweet features. Using these features, they were able to achieve high accuracy in predicting fake images from real images, although the authors discuss that this high accuracy might be due to the similarity of many images. Another notable study uses visual attributes from tweet

images to classify real and fake news images.

Another notable example is research by Jin et al., focuses on features extracted from images [84]. In their study, the authors use images in the context of news events. They define news events as mentions of certain keywords in a specific time span. They then extract a set of features from images including visual clarity (distribution difference between two image sets), Visual Coherence (how coherent images in specific news event are), Visual Similarity (pairwise similarity distribution histogram in an event), visual diversity (visual difference in the image set of a target news event), and Visual Clustering (number of clusters in a news event). They use four classification models: SVM, Logistic Regression, KStar, and Random forest. The authors were able to achieve an accuracy of 83.6 percent using the Random Forest algorithm which hints towards the promise of using image features from misinformation. However, in non-computer science literature, it has been shown that many other more complex contextual features might be good signals for deception and fake news including facial features and emotions [123].

Another approach to analyzing the content of misinformation is through patterns of diffusion and propagation. These approaches are often conducted through temporal analysis of news sharing or through building different types of networks [168]. Liang Wu and Huan Liu developed a message characterization method that focuses purely on content propagation [209]. They propose an LSTM-RNN model that purely uses social proximity and community structures as features to characterize fake news. In another study, unsupervised topic models were used to construct a network of stance-based network and mining conflicting viewpoints; while using an iterative method,

the authors would score characterize news with high levels of conflicting topics[83]. Also without looking at the verbal content of news, Ma et al. argue that focusing on temporal patterns and time series of news spread can be an essential feature in detecting rumors [111]. Their method, called Dynamic-Series-Time Structure explores the variation of various social context features over time and can be used to detect suspicious news from credible news.

These examples, are among the methods of using news content or relationships between content to categorize specific pieces of news as misinformation or not. These methods include supervised models, network models, network and temporal models, and they use both text and image features to achieve the goal of categorizing news content. As mentioned in the previous sections, the source and consumers of misinformation are also essential elements in the process of misinformation. In the next section, we study computational approaches in categorizing and understanding sources of misinformation.

### 2.3.2 Measuring and categorizing trustworthiness and veracity of sources

As emphasized by Lazer et al. focusing on sources of misinformation can be more useful in combating misinformation because often repeating misinformation even in the context of fact-checking might prove to be damaging [108, 139]. Even though the majority of computational approaches specifically focus on content, various scholars have put their primary focus on understanding and categorizing sources of misinformation.

One of the most significant red flags in regards to sources of misinformation is

whether the sources use automated methods to propagate news. Social bots amplify the reach of misinformation and are built to exploit consumers' cognitive and social biases [107]. These bots are built with different intents. They are often benign and have been proved to be helpful in certain situations [56]. However, more often than not, bots distribute information without verification and can result in circulating false accusations [69]. In some cases, bots have been shown to focus on specific topics supporting specific political parties or candidates [152]. Moreover, bots can behave similarly to humans by liking and sharing news as well as communicating with humans [81]. These behaviors and features have been used to develop many different computational methods to detect and battle social bots.

Ferrera et al. propose a taxonomy of social bot detection systems that include graph-based, crowdsourced, and feature-based bot detection methods [56]. Graph-based methods often take advantage of the fact that malicious bot accounts are highly connected to other malicious accounts and have used community detection algorithms to detect clusters of social bots. However, in networks with well-formed clusters, these community detection methods perform poorly [187]. Alvisi et al. take note of these shortcomings and offer a community detection that takes a local approach rather than a global community detection algorithm that is more resilient to real-world social bot clusters [21]. Crowd-sourced methods assume a human's ability to differentiate social bots and legitimate accounts [56]. Wang et al. conduct a study using both experts and Mechanical Turk and find that while Mechanical Turk workers vary in their efficiency, experts achieve "near-optimal" accuracy in detecting bots [197]. This is the reason why many large companies hire groups of experts to take charge of detecting

social bots. However, using crowd-sourced information is not always cost-effective. The Feature-based category focuses on feature engineering from bots' behaviors and content and uses Machine Learning to predict Social Bots. "Bot or Not?" is a well-known example that utilizes such methods [45]. The system uses a combination of linguistic and sentiment features with a Random Forest model to score the likelihood of a Twitter account being a bot or not.

Other scholars have developed computational methods to go beyond the concept of social bots. Less concerned with misinformation, in an attempt to help journalism experts detect and assess sources of misinformation at the time of breaking news, Diakopoulos et al. create two classifiers. The first classifier categorizes sources into organizations, journalists/blogger and ordinary people [47]. A second classifier uses a dictionary-based approach to identify eyewitnesses and achieves low false-positive rates (89% precision) but a high false negative rate (32% recall). The authors provide other cues to help users with the task of source classification including named entities, URL categorization, and spatial information [47]. Castillo et al. combine features from the content of tweets with source level features such as registration age. Statuses count, the number of followers and friends, their verified status, and the existence of description and URL in their user profile to classify news accounts into credible or not [37]. Other source based features such as the location of users and also the client used to produce news are also shown to be useful in predicting the credibility of sources [210].

Volkova et al. created a neural network classifier that utilizes various linguistic features including Biased Language, Subjectivity, and Moral Foundations, along

with signals from the news sources’ social activities on Twitter to predict Suspicious Vs. Real news as well as multiple misinformation categories such as clickbaits, Hoaxes, and propaganda [191]. They developed a Convolutional Neural Network model that achieves high levels of accuracy at the binary classification of misinformation sources (real vs. fake). In their model, including mention/retweet interactions between sources greatly improved the performance of the results. The authors also found differences in language cues between different types of accounts. For example, “verified” news sources contain significantly fewer markers for bias language, as well as harm, loyalty, and authority moral cues. The authors also find that users retweeting misinformation sources, send high volumes of tweets of shorter periods of time. Inspired by the findings of Volkova et al. [191], Karduni et al. developed a model using a random forest classifier to separate misinformation from real news and found features such as fear, anger, and negativity to be highly correlated with misinformation account, while features such as fairness and loyalty were highly correlated with real news accounts [97].

Characterizing sources of misinformation has been the primary goal of various computational methods. Network-based, crowd-sourced based and feature based methods have been used to detect malicious social bots. Others focused on building classifiers using source related features that can categorize other sources of misinformation based on their behavior and usage of content. Developing robust methods to detect and characterize the source and content of misinformation is extremely valuable. However, the question of whether these methods can be useful to reduce the harm of misinformation remains unanswered. Some user-facing tools have been created that aim

to assist general users or experts in detecting and rebuking misinformation. These systems were evaluated based on different criteria. In the next section, I move to the third element in the triad of misinformation. By focusing on Consumers, I offer an overview of the existing user-facing systems and their evaluation attempts.

### 2.3.3 Human-Misinformation interaction: Interactive tools, Visual Analytic Systems, Cognitive bias mitigation efforts

The vast majority of computational efforts on battling misinformation has been in detecting misinformation from content and source perspective. However, arguably, the ultimate goal of misinformation mitigation and detection efforts would be to help reduce the effect of the information on users. It has been suggested that providing automated or authoritative fact-checking results along with misinformation might prove counterproductive [108, 107]. There have been numerous studies in the fields of psychology, cognitive science, and social sciences that highlight the complexities of the relationship between humans and misinformation [139, 141, 142, 20, 23]. However, the efforts in the fields of computation have mostly neglected the human aspects dealing with information. In this section, I offer a highlight of some of the notable efforts in human-computer-interaction (HCI). Moreover, I also introduce notable interactive interfaces built to combat misinformation.

Some studies have examined how users ability to detect misinformation while using computational tools. Flintham et al. conducted a survey and set of interviews to understand consumer behavior and attitude, as well as their strategies to detect misinformation on social media. Within their study, they ask users to find ‘fake news’

on Facebook while thinking aloud. Using a thematic analysis of qualitative data, they found that users had different approaches towards sources of misinformation, some giving complete primacy to the authenticity of the source, while others decided to disregard the source's reputation. The authors also found that authors interest in a specific "kind" of news played an important role in their interest in separating facts from fake news. Finally, they found that their participants were reluctant to trust a tool that would allow them to determine the veracity of misinformation on social media [59]. Pourghomi et al. study the interaction techniques utilizes by Facebook to help users fact-check news posts and compare those methods with another proposed method called "right click authenticate". The method presents facts related and editorial pieces formatted similar to Wikipedia and suggests that it might be more useful to engage in authenticating news themselves rather than relying on third-party fact-checkers [150]. In another notable study, Kasra et al. conduct a focus group study on users to understand how well they can detect and rebuke doctored and faked images. Their findings suggest that users do not perform well at identifying fake online images. They also found that the users' main strategy to rebuke images was to refer to non-visual attributes such as the source and the accompanying description. Moreover, they found that users failed to identify cues in images when specifically asked for [98].

To build a visual system to help journalists asses sources of news, Diakopoulos et al. develop an interface specifically built for journalism experts. The interface offers a variety of information regarding sources including the number of friends and followers, the location they mostly tweet from, Whether the account is an eyewitness



to specific news, and named entities extracted from the content of the tweets. They conducted a series of expert interviews and found intriguing results concerning their needs. Even though the system was not built explicitly for sources of misinformation, the interviewees expressed their interest in features to detect misinformation [47]. Narwal et al. develop an automated assistant called UnbiasedCrowd that aims to help users understand biases in visual information on Twitter. The system collects images, clusters them using by extracting fisher vectors and K-Means clustering. The system then allows users to highlight biases and share to others using automated bots or manually. The authors conducted a study with experts and the general public. One of the interesting findings was the need for providing context to clustered images. Moreover, their general public study found two groups, one who actively propagated the information about biased images and a group that took a defensive stance which highlights users different strategies towards content verification [129].

Several systems and interfaces have been developed with varying amounts of interactivity. Gupta et al. develop TweetCred that provides real-time credibility assessment of content on Twitter. The system uses an SVM classifier on produces a visual score between 1 and 7 for the credibility of each news content on Twitter. The system does not offer other interactive tools for users to deal with misinformation[68]. Emergent is a dataset of rumors and a real-time rumor tracking website. The system mostly serves as a fact-checking source that offers fact-checking by simply tagging claims as True, False, or unverified and offering users extra information such as the originating source, number of shares, and topic tags [2].

Twittertrails is an interactive web-based system that affords users to view fact-

checking information on specific claims. The system allows for searching based on keywords, category, and levels of spread and skepticism. Linearly, each claim is tagged with two radial encoding, one for the spread and one for skepticism. Selecting each rumor opens a new page that provides different interactive visualizations and discussions on propagation, temporal nature, level of visibility, topics related to the rumor, and images used in tweets related to the rumor. The system does not go beyond fact-checking to test usability and relationship to how consumers would use the system[15]. RumorLens is one of the first systems that combine computational tools such as keyword-based classifiers for rumors with interactive visualizations to help humans in the task of rumor detection. The system includes a network diagram for highlighting the propagation of rumors, a snakey diagram paired with a timeline of the lifetime of the rumor. The system allows users to explore the progression of a rumor detected from twitter using a combination of different visualizations [155]. Hoaxy is a search engine and a dashboard that combines various scores on misinformation accounts, as well as interactive visualizations of the timeline of propagation and a network visualization to allow users to explore misinformation on social media. Hoaxy offers information on Accounts that share misinformation as well as information about the content of misinformation from these accounts [165]. Finally, RumorFlow is a Visual Analytics tool for understanding and analyzing how rumors are disseminated and discussed by users of social media. It uses Reddit as a source for data. It uses semantic similarity, sentiment analysis, and Wikipedia Entity Linking as methods of extracting extra information from the content. The visual analytics system includes a theme river visualization to highlight the development of rumors,

a word cloud, and topic cloud, as well as a snakey diagram showing relationships between topics related to rumors [44].

Most of these systems and studies share a common goal: they provide preliminary studies on how consumers approach misinformation and offer computational tools in the form of interfaces that offer a combination of automated scores, basic visualizations, and contextual information. However, the majority of the efforts do not include efforts to understand the usefulness of the system. Moreover, none of these efforts go in-depth into the issues that cause consumers to be affected by misinformation. Overall, there seems to be a real gap in computer science literature that utilize computational efforts to help the complex decision-making process of consumers. Future efforts, require careful studies on how computational methods and tools can be used in a real-world context to help consumers make informed decisions about misinformation online.

In the next section, I offer some discussions on the overall landscape of misinformation research in computer science and information technology and conclude with remarks on some potential interdisciplinary research paths for combating misinformation.

## 2.4 discussion and conclusion

Studying computational methods of battling misinformation through a systematic framework with three main elements of Source, Content, and Consumer allows us to categorize these approaches into three categories of work. The first category, which comprises majority of computational effort, is misinformation detection and verifica-

tion. These works generally focus purely on determining the veracity of news content. The work in this category generally focus on a combination of tasks: To Automatically fact-check claims and provide scores, context, alternative claims, etc; and to detect and classify verbal and visual misinformation of various kinds ( deception, clickbaits, rumors, propaganda, etc ). Automatic Fact-checking which has proved to be a very difficult task [67] can be done though crowdsourcing, checking with external sources, or a combination of classifiers and knowledge-bases [67]. On the other hand, detection approaches have used a variety of methods to distinguish misinformation. These methods include using language and styles and features as a signal, using visual features, and through temporal and propagation patterns of misinformation.

Another group of studies focus on determining and verifying sources of misinformation. As one of the biggest modern challenges of misinformation is utilization of social bots, in this category of work, a lot of effort has been put on developing methods to automatically classify sources as bot or human accounts. Bot detection is approached using graph-based methods, crowd-sourced methods, and by creating classifiers that use sources' behavior and aggregate content as features. A group of scholars have also developed methods to score and classify the trustworthiness of news sources. These works utilize aggregate signals from the content of different sources including language and writing styles, features extracted from social network behavior of sources, as well as other metadata such as location and registration age.

The third category which arguably can be the most important one, is consumer-facing systems and studies. These works aim to provide tools to users to understand, detect, and mitigate the effects of misinformation. A number of interactive systems

have been developed that mostly allow users to fact-check misinformation as well as to visualize the behavior of sources. Empirical studies on the usefulness and efficacy of these systems and methods are extremely sparse. Moreover, the existing studies do not deal with findings from other disciplines that are known to be extremely important in the process of battling misinformation. These findings include confirmation biases, prior exposure, social and peer pressure, echo-chambers and filter bubbles.

Even though the potential of using computational methods to battle misinformation has been discussed [107, 108], there still hasn't been studies that provide insights on how computational tools and visual systems can be used to moderate the effects of multiple cognitive biases and social pressures. Even though there has been great advancements in computational detection of misinformation content and sources, in order to truly battle misinformation through computational tools, one of the biggest next steps should be to develop methods and tools that are designed to deal with the complex psychological and social process of how humans consume misinformation and why they are effected by it. This, in fact, requires a collaborative interdisciplinary attempt bringing together experts from psychology, cognitive sciences, education, social and political sciences, and computer science.

## CHAPTER 3: STUDYING UNCERTAINTY AND DECISION-MAKING ABOUT MISINFORMATION USING VISUAL ANALYTICS

This chapter is a slight modification of our paper with my co-authors Ryan Wesslen, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou titled “Can You Verifi This? Studying Uncertainty and Decision-Making about Misinformation Using Visual Analytics” published in the proceedings of the Twelfth International AAAI Conference on Web and Social Media [97].

### 3.1 Introduction

The spread of misinformation on social media is a phenomena with global consequences, one that, according to the World Economic Forum, poses significant risks to democratic societies [78]. The online media ecosystem is now a place where false or misleading content resides on an equal footing with verified and trustworthy information [105]. In response, social media platforms are becoming “content referees,” faced with the difficult task of identifying misinformation internally or even seeking users’ evaluations on news credibility.<sup>1</sup> On the one hand, the news we consume is either wittingly or unwittingly self-curated, even self-reinforced [182]. On the other hand, due to the explosive abundance of media sources and the resulting information overload, we often need to rely on heuristics and social cues to make decisions about the credibility of information [122, 166]. One such decision-making heuristic is con-

---

<sup>1</sup><https://www.wsj.com/articles/facebook-to-rank-news-sources-by-quality-to-battle-misinformation-1516394184>

firmation bias, which has been implicated in the selective exposure to and spread of misinformation [19]. This cognitive bias can manifest itself on social media as individuals tend to select claims and consume news that reflect their preconceived beliefs about the world, while ignoring dissenting information [122].

While propaganda and misinformation campaigns are not a new phenomenon [170], the ubiquity and virality of the internet has lent urgency to the need for understanding how individuals make decisions about the news they consume and how technology can aid in combating this problem [168]. Visual analytic systems that present coordinated multiple views and rich heterogeneous data have been demonstrably useful in supporting human decision-making in a variety of tasks such as textual event detection, geographic decision support, malware analysis, and financial analytics [193, 199]. **Our goal is to understand how visual analytics systems can be used to support decision-making around misinformation and how uncertainty and confirmation bias affect decision-making within a visual analytics environment.**

In this work, we seek to answer the following overarching research questions: *What are the important factors that contribute to the investigation of misinformation? How to facilitate decision-making around misinformation by presenting the factors in a visual analytics system? What is the role of confirmation bias and uncertainty in such decision-making processes?*

To this aim, we first leveraged prior work on categorizing misinformation on social media (specifically Twitter) [191] and identified the dimensions that can distinguish misinformation from legitimate news. We then developed a visual analytic system, Verifi, to incorporate these dimensions into interactive visual representations. Next,

we conducted a controlled experiment in which participants were asked to investigate news media accounts using Verifi. Through quantitative and qualitative analysis of the experiment results, we studied the factors in decision-making around misinformation. More specifically, we investigated how **uncertainty, conflicting signals manifested in the presented data dimensions**, affect users' ability to identify misinformation in different experiment conditions. Our work is thus uniquely situated at the intersection of the psychology of decision-making, cognitive biases, and the impact of socio-technical systems, namely visual analytic systems, that aid in such decision-making.

Our work makes the following important contributions:

- *A new visual analytic system:* We designed and developed Verifi<sup>2</sup>, a new visual analytic system that incorporates dimensions critical to characterizing and distinguishing misinformation from legitimate news. Verifi enables individuals to make informed decisions about the veracity of news accounts.
- *Experiment design to study decision-making on misinformation:* We conducted an experiment using Verifi to study how people assess the veracity of the news media accounts on Twitter and what role confirmation bias plays in this process. To our knowledge, our work is the first experimental study on the determinants of decision-making in the presence of misinformation in visual analytics.

As part of our controlled experiment, we provided cues to the participants so that they would interact with data for the various news accounts along various dimensions

---

<sup>2</sup><http://verifi.herokuapp.com>; open source data and code provided at <https://github.com/wesslen/verifi-icwsm-2018>



(e.g., tweet content, social network). Our results revealed that conflicting information along such cues (e.g., connectivity in social network) significantly impacts the users' performance in identifying misinformation.

### 3.2 Related Work

We discuss two distinct lines of past work that are relevant to our research. First, we explore cognitive biases, and specifically the study of confirmation bias in the context of visual analytics. Second, we introduce prior work on characterizing and visualizing misinformation in online content.

#### Confirmation bias:

Humans exhibit a tendency to treat evidence in a biased manner during their decision-making process in order to protect their beliefs or pre-conceived hypothesis [86], even in situations where they have no personal interest or material stake [132]. Research has shown that this tendency, known as confirmation bias, can cause inferential error with regards to human reasoning [54]. Confirmation bias is the tendency to privilege information that confirms one's hypotheses over information that disconfirms the hypotheses. Classic laboratory experiments to study confirmation bias typically present participants with a hypothesis and evidence that either confirms or disconfirms their hypothesis, and may include cues that cause uncertainty in interpretation of that given evidence. Our research is firmly grounded in these experimental studies of confirmation biases. We adapt classic psychology experimental design, where pieces of evidence or *cues* are provided to subjects used to confirm or disconfirm a given hypothesis [202, 132].

### Visualization and Cognitive Biases:

Given the pervasive effects of confirmation bias and cognitive biases in general on human decision-making, scholars studying visual analytic systems have initiated research on this important problem. [194] categorized four perspectives to build a framework of all cognitive biases in visual analytics. [39] presented a user study and identified an approach to measure anchoring bias in visual analytics by priming users to visual and numerical anchors. They demonstrated that cognitive biases, specifically anchoring bias, affect decision-making in visual analytic systems, consistent with prior research in psychology. However, no research to date has examined the effects of confirmation bias and uncertainty in the context of distinguishing information from misinformation using visual analytic systems - we seek to fill this important gap. Next, we discuss what we mean by misinformation in the context of our work.

### Characterizing Misinformation:

Misinformation can be described as information that has the camouflage of traditional news media but lacks the associated rigorous editorial processes [122]. Prior research in journalism and communication has demonstrated that news outlets may slant their news coverage based on different topics [52]. In addition, [20] show that the frequency of sharing and distribution of fake news can heavily favor different individuals. In our work, we use the term fake news to encompass misinformation including ideologically slanted news, disinformation, propaganda, hoaxes, rumors, conspiracy theories, clickbait and fabricated content, and even satire. We chose to use “fake news” as an easily accessible term that can be presented to the users as a label

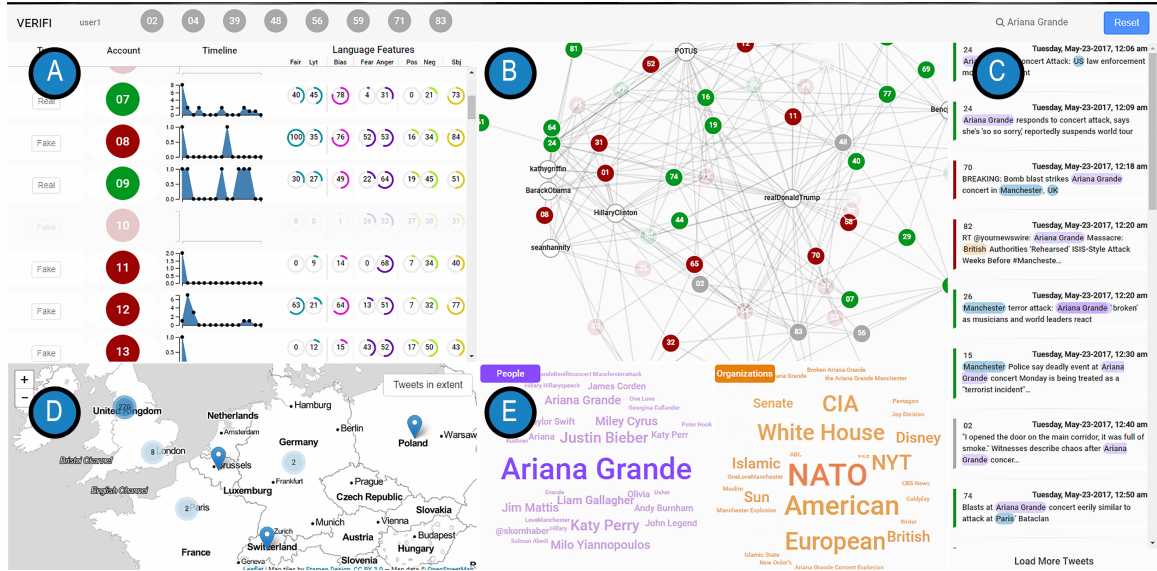


Figure 3: The Verifi interface: Account View (A), Social Network View (B), Tweet Panel (C), Map View (D), and Entity Word Cloud (E). The interface can be accessed at [Verifi.Herokuapp.com](http://Verifi.Herokuapp.com).

for misinformation and we use the term “real news” as its antithesis to characterize legitimate information.

Several systems have been introduced to (semi-) automatically detect misinformation, disinformation, or propaganda in Twitter, including FactWatcher [73], TwitterTrails [124], RumorLens [155], and Hoaxy [165]. These systems allow users to explore and monitor detected misinformation via interactive dashboards. They focus on identifying misinformation and the dashboards are designed to present analysis results from the proposed models. Instead, Verifi aims to provide an overview of dimensions that distinguish real vs. fake news accounts for a general audience.

Our work is thus situated at the intersection of these research areas and focuses on studying users’ decision making about misinformation in the context of visual analytics.

### 3.3 Verifi: A Visual Analytic System for Investigating Misinformation

Verifi is a visual analytic system that presents multiple dimensions related to misinformation on Twitter. Our design process is informed by both prior research in distinguishing real and fake news as well as our analysis based on the data selected for our study to identify meaningful features.

A major inspiration for Verifi’s design is based on the findings of [191], who created a predictive model to distinguish between four types of fake news accounts. They find that attributes such as *social network interactions* (e.g., mention or retweet network), *linguistic features*, and *temporal trends* are the most informative factors for predicting the veracity of Twitter news accounts. Our design of Verifi is inspired by these findings: (i) we included a *social network view* that shows a visualization of account mentions (which includes retweets) as a primary view to allow users to investigate relationships between accounts; (ii) we developed an accounts view with *account-level temporal (daily) trends* as well as the most predictive linguistic features to facilitate users’ account-level investigation into the rhetoric and timing of each account’s tweets; and (iii) to choose the most effective *linguistic features*, we created a model to predict which linguistic features most accurately can predict the veracity of different accounts.

In addition to three different analytical cues inspired by Volkova et al. and our predictive model, we included visualizations and data filtering functions to allow participants to qualitatively examine and compare accounts. Based on existing research conducted on the ways news can be slanted and the diffusion of misinformation

Type	Real	Propaganda	Clickbait	Hoax	Satire
Account	31	30	18	2	2

Table 1: Distribution of types of news outlets

[18, 20, 51, 52], we included visual representation of three types of extracted *entities* (*places, people, and organizations*) to enable exploration through filtering.

### 3.3.1 Dataset

To create our dataset, we started with a list of 147 Twitter accounts annotated as propaganda, hoax, clickbait, or satire by [191] based on public sources. We then augmented this list with 31 mainstream news accounts [172] that are considered trustworthy by independent third-parties.<sup>3</sup> We collected 103,248 tweets posted by these 178 accounts along with account metadata from May 23, 2017 to June 6, 2017 using the Twitter public API.<sup>4</sup>

We then filtered the 178 accounts using the following criteria indicating that the account is relatively less active: (i) low tweet activity during our data collection period; (ii) recent account creation date; and (iii) low friends to follower ratio. In addition to these three criteria, we asked two trained annotators to perform a qualitative assessment of the tweets published by the accounts and exclude extreme accounts (e.g., highly satirical) or non-English accounts. After these exclusions, we had a total of 82 accounts, distributed along the categories shown in Table 1.

<sup>3</sup><https://tinyurl.com/yctvve9h> and <https://tinyurl.com/k3z9w2b>

<sup>4</sup>The Verifi interface relies on a public Twitter feed collected by the University of North Carolina Charlotte.

## Data processing and analysis

To analyze our tweet data, we extracted various linguistic features, named entities, and social network structures. The role of the computational analysis in our approach is to support hypothesis testing based on social data driven by social science theories [196].

**Language features:** Language features can characterize the style, emotion, and sentiment of news media posts.

Informed by prior research that identified multiple language features for distinguishing real versus fake news [191], we consider five language features, including *bias language* [153], *subjectivity* [206], *emotion* [190], *sentiment* [109], and *moral foundations* [65, 71]. For example, *moral foundations* is a dictionary of words categorized along eleven dimensions, including care, fairness, and loyalty. Table 2 provides an overview of the features we used to characterize the language of the tweets, with each feature containing multiple dimensions.

In total, we test 68 different dimensions (i.e., 34 different language feature dimensions and each with two different normalization methods – either by number of tweets or number of words) using a supervised machine-learning algorithm (Random Forest) with a 70/30 training/validation split. We eliminated highly correlated (redundant) features (see supplemental materials). Figure 4 provides the ranking of the top 20 predictive language features.<sup>5</sup> Using this ranking, we decided to include eight language features within Verifi: *Bias*, *Fairness (as a virtue)*, *Loyalty (as a virtue)*, *Negative*

---

<sup>5</sup>This model had a 100% validation accuracy (24 out of 24) on the 30% validation dataset.

Source	Features	Example
Bias Language Lexicon-driven	6	Bias, Factives, Implicatives, Hedges, Assertives, Reports
Moral Foundation Lexicon-driven	11	Fairness, Loyalty, Authority, Care
Subjectivity Lexicon-driven	8	Strong Subjective, Strong Negative Subjective, Weak Neutral Subjective
Sentiment Model-driven	3	Positive, Negative, Neutral
Emotions Model-driven	5	Anger, Disgust, Fear, Joy, Sadness, Surprise

Table 2: 34 candidate language features from five sources.

*sentiment*, *Positive sentiment*, *Fear*, and *Subjectivity* to assist users in distinguishing fake and real news.<sup>6</sup>

**Entity Extraction and Geocoding:** Verifi includes a word cloud to display the top mentioned entities and enable the comparison of how different media outlets talk about entities of interest. We extract people, organization, and location entities from the tweets.

**Social Network Construction:** To present the interactions between the accounts on Twitter, we construct an undirected social network. Edges are mentions or retweets between accounts. Nodes represent Twitter news accounts (82 nodes) as well as the top ten most frequently mentioned Twitter accounts by our selected accounts.

<sup>6</sup>We averaged Strong-Weak subjectivity measures into one single measure.

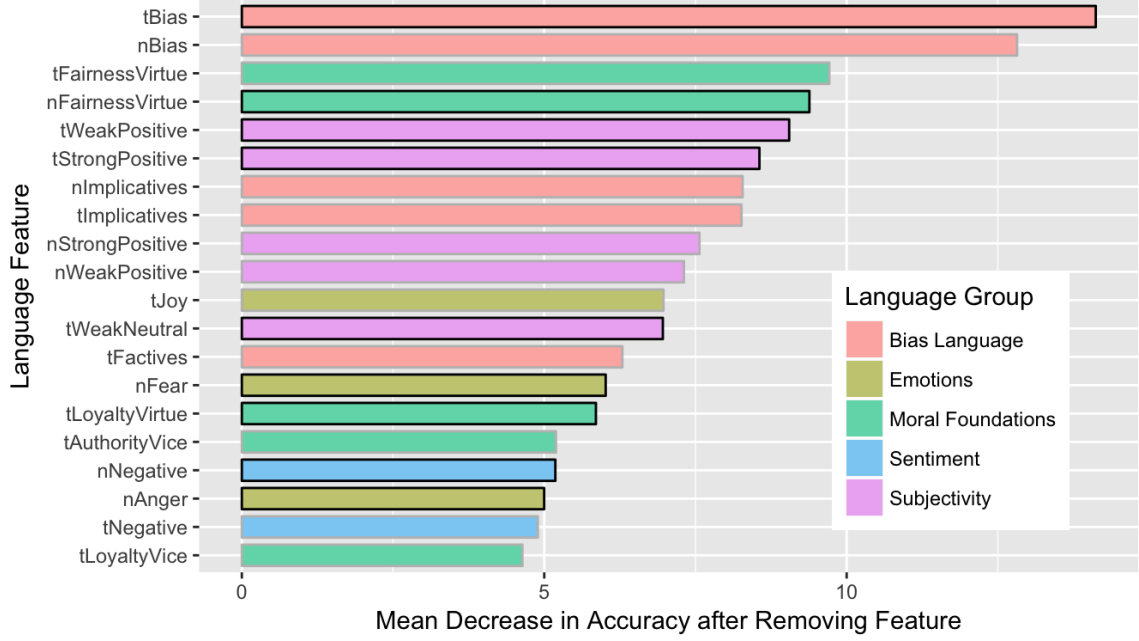


Figure 4: Top 20 most predictive language features of Fake and Real news outlets as measured by each feature’s average effect on Accuracy. ‘t’ prefix indicates the feature is normalized by the account’s tweet count and ‘n’ indicates normalization by the account’s word count (summed across all tweets). Features with borders are included in Verifi.

### 3.3.2 The Verifi User Interface

The Verifi user interface is developed using D3.js, Leaflet, and Node.js. The interface consists of six fully coordinated views that allow users to explore and make decisions regarding the veracity of news accounts (Figure 3).

**The Accounts View** (Figure 3A) provides account-level information including tweet timeline and language features. The circular button for each account is color coded to denote whether the account is considered real (green) or fake (red). The accounts colored in gray are the ones participants are tasked to investigate in our experiment. The timeline shows the number of tweets per day. The array of donut charts shows the eight selected language features (scaled from 0-100) that charac-



terize the linguistic content. For example, a score of 100 for fairness means that an account exhibits the highest amount of fairness in its tweets compared to the other accounts. Users can sort the accounts based on any language feature. The Account View provides an overview of real and fake accounts and enables analysis based on language features and temporal trends.

**The Social Network View** (Figure 3B) presents connections among news accounts (nodes) based on mentions and retweets (edges). The color coding of the nodes is consistent with the Accounts View (i.e., green for real, red for fake, gray for unknown). To increase the connectivity of the news accounts, we included ten additional Twitter accounts. These ten accounts (colored white) are the top-ranked Twitter accounts by mention from the 82 news accounts over the two week period. The Social Network View allows users to understand how a specific account is connected to fake or real news accounts on the social network.

**Entity Views:** The people and organization word clouds (Figure 3E) present an overview of the most frequently mentioned people and organization entities. The word clouds support the filtering of tweets mentioning certain entities of interest, thus enabling comparison across accounts. For example, by clicking on the word “American,” accounts that mention this entity would be highlighted in both the Accounts View and the Social Network View. In addition, tweets mentioning “American” will appear in the Tweet Panel View.

**The Map View** provides a summary of the location entities (Figure 3D). When zooming in and out, the color and count of the cluster updates to show the tweets in each region. Users can click on clusters and read associated tweets. Users can also

filter data based on a geographic boundary.

**The Tweet Panel View** (Figure 3C) provides drill-down capability to the tweet level. Users can use filtering to inspect aggregate patterns found in other views. Within the tweet content, detected entities are highlighted to assist users in finding information in text. This view is similar to how Twitter users typically consume tweets on mobile devices.

### 3.4 Experiment Design

We designed a user experiment to study how people make decisions regarding misinformation and the veracity of new accounts on Twitter with the help of the Verifi system.

#### 3.4.1 Research Questions

Situated in the context of decision-making with visual analytics, we organized our research focus on the following research questions:

**RQ1:** Would individuals make decisions differently about the veracity of news media sources, when *explicitly asked to confirm or disconfirm* a given hypothesis?

**RQ2:** How does uncertainty (conflicting information) of cues affect performance on identifying accounts that post misinformation?

#### Experiment Stimuli

After developing the Verifi interface, we loaded data from all 82 accounts (Table 1) into the system. To minimize the effect of preconceived notions, all news outlet names were anonymized by assigning them integer identifiers. Given the in-lab nature of the user studies and time limitations, we selected eight accounts that participants would

Table 3: Eight accounts with masked account names. Background colors indicate real (green) and fake (red).

Mask Name	Description
@XYZ	A news division of a major broadcasting company
@GothamPost	An American newspaper with worldwide influence and readership
@MOMENT	An American weekly news magazine
@Williams	An international news agency
@ThirtyPrevent	A financial blog with aggregated news and editorial opinions
@ViralDataInc	An anti right-wing news blog and aggregator
@NationalFist	An alternative media magazine and online news aggregator
@BYZBrief	Anti corporate propaganda outlet with exclusive content and interviews

investigate and would label as either real or fake based on their own judgments. The accounts were chosen to cover a range of different cues and degrees of uncertainty. We based our selection of experiment stimuli on classic studies in confirmation bias [202, 151].

Due to institutional concerns, we have masked the names of those accounts while preserving the nature of their naming. The eight selected accounts (4 real and 4 fake) with their masked names and description are shown in Table 3. The source of the description is Wikipedia and identifying information was removed to anonymize the accounts. Our goal in selecting the experiment stimuli was to enable participants to make decisions about a wide range of content with the aid of varied, sometimes conflicting, cues.

### 3.4.2 Experiment Tasks

To test the effect of confirmation bias, we designed an experiment with three experimental conditions: Confirm, Disconfirm, and Control. In the Confirm condition, participants were given a set of six cues about the grayed out accounts (i.e., the eight

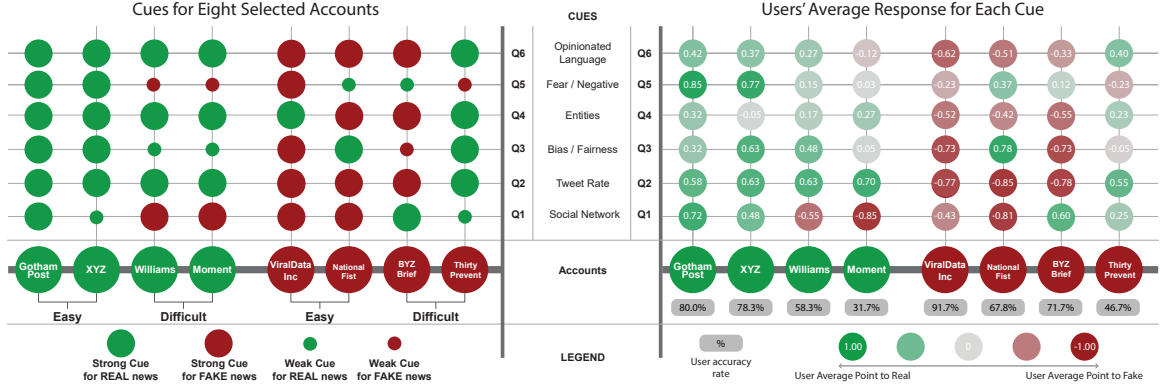


Figure 5: Available cues for selected accounts (column) and users’ response regarding the importance of these cues (row, Q1-Q6). Left: Shows each of the eight selected accounts as well as the cues available for each of them. Right: Shows average of importance for each cue per account based on participants’ responses. Values in gray circles below each account name show average accuracy for predicting that account correctly. The left figure is purely based on the (conflicting) information presented in the cues and is independent from user responses. The right figure based on the user responses on the importance of each cue coincides with the information in the left table.

selected accounts shown in Table 3) and were explicitly asked to *confirm* a given hypothesis that all grayed-out accounts were fake accounts. Similarly, in the Disconfirm condition, participants were explicitly asked to *disconfirm* the given hypothesis that all gray accounts were fake. Our third experiment condition was the Control, where the participants were simply asked to judge the veracity of the accounts; they were given neither the initial hypothesis nor the set of six cues. Following classic psychology studies in confirmation bias [132] where the information presented to the participants has inherent uncertainty, we added the element of uncertainty to the cues. We provided six cues (Q1-Q6) to the participant, of which three cues pointed to the account being real and three cues pointing to the account being fake. Each cue corresponds to a view in the Verifi interface.

The decisions that participants needed to make for the gray accounts involved an-

swering (True/False/Did Not Investigate) for each of the six statements listed below. Each statement is the same as the cue presented to the participant in the confirm and disconfirm condition; the purpose of the statements is to gather information on which cues the participants relied on when making decisions for a certain account.

**Q1** This account is predominantly connected to real news accounts in the **social network** graph. This characteristic is typically associated with known real news accounts.

**Q2** The average **rate of tweeting** from this account is relatively low (less than 70 tweets per day).

**Q3** On the language measures, this account tends to show a higher ranking in **bias measure and fairness measure**. This characteristic is typically associated with known real news accounts.

**Q4** This account tends to focus on a subset of polarizing **entities** (people, organizations, locations) such as Barack Obama or Muslims as compared to focusing on a diverse range of entities.

**Q5** On the language measures, this account tends to show a low ranking in **fear and negative** language measures.

**Q6** The tweets from this account contain **opinionated language**. This characteristic is typically associated with known fake news accounts.

These statements and the cues given to the participants at the beginning of the experiment (along with the hypothesis) are the same. Based on our data collection

and analysis, statements Q1, Q3, and Q5 point to an account while being a real account and while the rest of the statements (Q2, Q4, and Q6) point to the account being fake. For certain statements, we explicitly included information characterizing whether the cue pointed to the account being real or fake (as shown in Q1, Q3, and Q6). The presentation of these cues were deliberately chosen to add to the uncertainty of information presented to the users. In addition to asking participants about their decision-making process on each statement listed above, we also asked the users to rate the importance of each view in the Verifi interface in making those decisions (the Accounts View, Social Network View, Tweet Panel View, and Entity View) on a scale of 1 to 7. Additionally, we asked participants to indicate the confidence of their decision on a scale of 1 to 7 for each account, as well as an optional, free-form response section where participants could provide any additional information as a part of their analysis. All these questions were part of a pop-up form that was displayed when the participants clicked the “Choice” button shown alongside the account number in the Accounts View (Figure 3A). The responses to this form were captured in a database upon submitting the form during the task.

The information regarding each gray account and its cues is summarized in Figure 5 (Left). For simplicity of presentation, green circles indicate a cue pointing towards account being real, red circles indicate a cue pointing towards account being fake. The overview of how they score on cues demonstrate how the accounts exhibit different levels of difficulty for decision-making. For example, all evidence pointed to the *@GothamPost* account being real, which means that ideally, upon investigation, a participant would answer True for Q1, Q3 and Q5 and False for Q2, Q4 and Q6

when making their decision for that account. However, other real accounts chosen for investigation had more uncertainty in the cues. Notably, the *@MOMENT* account was chosen as one of the difficult accounts since it had a misleading social network cue (Q1) in that it had only one connection to a fake news account. For the fake news accounts chosen, *@ViralDataInc* had all evidence pointing towards the account being fake (except Q4, which means that the tweets from this account covered a diverse range of entities). This would make *@ViralDataInc* easier to judge as fake than, for instance, *@ThirtyPrevent*, which exhibits many more misleading cues.

### 3.4.3 Experiment Procedure and Participants

We recruited participants via in class recruitment, email to listservs, and the psychology research pool at our institution. Once signed up, participants came to the lab for a duration of one hour. After the informed consent procedure, participants viewed two training videos designed for this experiment. The first video introduced the interface and explained the different views. The second video provided a task example to determine the veracity of a sample account not used in the study. Both videos were identical across all conditions. After this training, participants completed a pre-questionnaire consisting of questions related to their demographics (age, gender, education), familiarity with visual analytics, social media, and Big-5 personality questions [85]. The participants were then assigned the task and asked to complete the task in 30 minutes. After completing their task, participants completed a post-test questionnaire which included six vignettes to assess participant’s propensity to confirmation bias in general [132].

Sixty participants completed the study, evenly split into three treatment groups. Participant ages were between 18 and 41 (mean=24.7). The gender distribution was 45% male and 55% female. A majority of the participants were undergraduates (65%), followed by Master’s (16.7%), Ph.D. (8.3%), and others (10%). The distributions of the participants between computing (48.3%) and non-computing majors (51.7%) was relatively even.

### 3.5 Data Analysis Methods

In this section, we introduce the analysis methods applied to our experiment data to answer the two research questions.

**To address RQ1**, namely, “are there significant differences in the way participants interact with the data and their resulting judgments based on the experiment condition?”, we use one-way analysis of variance (ANOVA) for testing and post-hoc Tukey’s honest significant difference (HSD) test to determine significance ( $\alpha=0.05$ ). Our experiment design is a between-subjects design with one level: the experimental condition.

**To address RQ2** regarding the effects of uncertainty, we used two logistic regressions to explore the effects of uncertainty (in cues, accounts, confidence, and treatment groups) had on users’ decision-making. Each regression included a different dependent variable: users’ accuracy (1 = correct decision, 0 = incorrect decision) and fake determination (1 = fake prediction, 0 = real prediction). This analysis allows us to determine which factors were most important and aligned with our expectation in terms of direction. For example, as mentioned in the Experiment Stimuli section,



cues Q1, Q3, and Q5 were selected to point to real accounts, suggesting a negative relationship with fake prediction (or less than 1 log odds ratios). Alternatively, cues Q2, Q4, and Q6 were selected to point to fake accounts (i.e., positive relationship or greater than 1 log odds ratios). In addition, we can also identify which cue was most important in decision-making as the one with the largest (in absolute magnitude) coefficient. In addition to the cues, we also include dummy variables for the account-level (using @XYZ as the reference level) as well as include users' confidence level and treatment group (Control group is the reference level) to understand if these factors played an additional role in the users' decisions.

### 3.6 Analyses Results

In this section, we describe our findings and results. The detailed discussion about the implications of these findings is in the Discussion section.

#### RQ1: Testing the Effects of Confirmation Bias

Table 4 shows the user accuracy rate and fake prediction rate across all three experiment conditions. We found no significant differences between the experimental conditions, on a diverse range of factors. Participants in all three conditions did not differ on the number of accounts labeled as fake and the number of accounts labeled as real ( $p > 0.05$  for both). We tested the accuracy rate and found no significant difference in the rate of accuracy across experimental conditions ( $p > 0.05$ ). In addition, we tested whether the participants interacted differently with the data, depending upon the experiment condition. To test this hypothesis, we computed the total time spent for participants in each condition, including time spent interacting with the data

	Control	Confirm	Disconfirm
Accuracy	60.4%	73.8%	63.1%
Fake Prediction	54.1%	55.0%	51.9%

Table 4: User accuracy and Fake prediction across conditions.

presented in each view in Verifi (e.g., Social Network View, Accounts View). We found no significant differences in the amount of time spent overall or in any specific panel on the interface across the three conditions.

### 3.6.1 RQ2: Measuring the Impact of Uncertainty

While we did not find significant differences in users’ decisions (e.g., accuracy) between experiment conditions, we expect differences in accuracy and fake prediction given uncertainty in cues for each account. Based on the cues in Figure 5 Left, we categorize accounts into two types: Easy and Difficult. These categories are based on how each account scores on the six cues and are independent from users’ responses. In this section, we describe regression analysis to analyze the effect of cues and account on users’ decision-making. We then present thematic analysis of users’ comments regarding their decisions.

**Regression Analysis:** Our results provide evidence that the prevalent factors in users’ decision-making were the cues and the accounts. Table 5 provides the log odds ratios for the independent variables by each regression. We observe three findings. First, in general cues have a significant effect on users’ fake prediction and accuracy. For the cues, we recoded the responses to indicate whether the cue was used consistent or not (e.g., depending on the direction of the cue relative to fake or real accounts).

Independent Variable	Dependent Variable	
	Accuracy	Fake
(Intercept)	0.18**	0.21**
Social Network Cue (Q1)	2.03***	0.99
Tweet Rate Cue (Q2)	1.24	1.06
Fairness Cue (Q3)	1.30*	0.74**
Entity Cue (Q4)	1.43*	1.77***
Fear Cue (Q5)	1.53***	0.90
Opinionated Cue (Q6)	2.78***	2.74***
@ViralDataInc (Fake Easy)	9.86***	117.96***
@NationalFist (Fake Easy)	1.90	9.7***
@GothamPost (Real Easy)	0.95	0.84
@Williams (Real Difficult)	0.90	2.13*
@MOMENT (Real Difficult)	0.36**	5.70***
@BYZBrief (Fake Difficult)	4.91***	24.21***
@ThirtyPrevent (Fake Difficult)	3.70**	18.89***
User Confidence	1.14	0.86
Confirm Group	1.97**	0.91
Disconfirm Group	1.16	0.75

\*\*\* = 99%, \*\* = 95%, \* = 90% confidence

Table 5: Log odds ratios for each independent variable in two logistic regressions. The Accuracy column is 1 = Correct, 0 = Incorrect Decision. The Fake column is the user’s prediction: 1 = Fake, 0 = Real. The @accounts variables use @XYZ as the reference level and the Group variables use the Control Group as the reference level.

We find that the opinionated, fear, and social network cues were the most important in explaining correct decisions when used consistently. Alternatively for explaining Fake decisions, we find that log odds ratios align to the cue direction as mentioned in Figure 5. For example, cues Q2, Q4, and Q6 point to the account being fake and we find the log odds ratios above one, although only Q4 and Q6 are statistically significant.

Second, we find that certain accounts had a significant effect on both users’ accuracy and fake prediction. This observation implies that some accounts were more difficult and systematically over or under predicted as fake. For example, @MOMENT has

Correct?	Type	Group	Comment	category
1	Yes	real	easy Several language features are consistent with predominantly real accounts	quantitative
2			News appears more factual reporting rather than opinionated discussion of events, which leads me to believe it is a real news account.	quantitative
3			This account does not seem to deal much with controversial topics, and although it has a lower loyalty score, it has a high fairness score and high bias, which are normally indicative of real accounts.	quantitative + qualitative
4			While this account only has one connection and it's to a fake account. I didn't notice anything suspicious in the tweets. The People and Organizations view only showed topics that are normally discussed in the news and nothing overly controversial.	quantitative + qualitative
5		fake	easy A lot of the tweets were not even news but simply them stating their opinions about a variety of issues.	qualitative
6			Only follows one account, tends to only tweet about one topic (Trump), and it's all negative and uses opinionated language.	quantitative
7			High fairness but low loyalty. Little amount of tweets (seemed inconsistent). Very high anger. When looking at the network, it was associated with a wide range of different accounts.	qualitative + qualitative
8			difficult This account is 100% angry, with a low tweet amount. This user also doesn't focus on that many people within their tweets.	quantitative + qualitative
9	No	real	easy Compared timeline of tweets as other tweets. The timeline and tweet content about taking Mosul for this account do not match with other "real" news.	quantitative
10			For this account, language within the tweets tipped me to believing this is a fake news account, or at least an extremely conservative or right-leaning (with high bias) news account. Wordage like "marxist left mainstream media" for instance.	qualitative
11			difficult Contains a lot of opinionated language in it's tweets.	qualitative
12			Despite the high tweet rate, their bias and subjectivity scores were high, which tends to relate to fake accounts. That added to the fact that it's only linked to another fake account and some verified accounts led me to believe this is a fake.	quantitative
13		fake	easy Though this is very opinionated, it leans towards an overall criticism of America, as opposed to an organization attempting to sway a constituency.	qualitative
14			Admittedly, personal bias played a role in deciding the "real"ness of this account as the information in the tweets, though not seemingly produced by big media, appears real, though not unbiased.	qualitative
15			difficult Connected to real accounts and has lower subjectivity.	quantitative
16			Although there was a high rating of anger, it seems as though none of the tweets expressed any anger or high bias.	quantitative + qualitative

Figure 6: A sample of users' comments about their decisions. Highlighted text shows users' mention of either a qualitative or quantitative reason. Green denotes reasons/cues pointing to the account being real while red pointing to being fake.

a very low log odds ratio for users' accuracy as users overwhelmingly incorrectly predicted @MOMENT, a real-difficult account, as fake (as indicated by its high log odds ratio for fake prediction).

Last, we find that confidence has no significant relationship in explaining accuracy or fake decisions. While there may be a univariate relationship between confidence and user decisions, this may likely be explained through the account level dummy variables as confidence also varied by accounts. Also, we find the Confirm condition maintains a weakly significant effect on accuracy relative to the Control group (reference level for treatments).

**Thematic Analysis of Comments:** Our regression analysis revealed that cues played an important role in users' decision making on misinformation. When cues point to conflicting directions of an account being real or fake, users are more likely to arrive at inaccurate decisions. In each decision, users had the option to leave

comments in regards to their decisions. These comments are extremely valuable in helping us decipher users’ rationales. We examined all comments (95 total) and thematically categorized users’ strategies. Our analysis focuses on how different usage on all or a subset of the cues affect their decision making. Similar to our quantitative analysis, we evaluate these themes through the lens of cue uncertainty and account difficulties.

Our thematic analysis classified comments into three categories: *Quantitative (32 comments)*, *Qualitative (37)*, and *Qualitative + Quantitative (26)*. We categorized mentions of social network connection, language feature score, and tweet timeline as quantitative. Any mention related to entities and users’ understanding of the text of tweets such as “opinionated language,” “news-like text,” and “style of text” were considered qualitative. The quantitative and qualitative dimensions extracted from the comments aligned well with the six cues provided to the participants.

**Easy Accounts:** Easy accounts (column 1, 2, 5, 6 in Figure 5 left) are the ones with most cues pointing to the accounts being either real or fake; thus leading many users to correct decisions. Fifteen comments for the easy accounts mentioned quantitative cues such as language features scores (Figure 6, row 1) and social network connections (Figure 6, row 6) as the basis of their decisions. 12 of these comments led to correct decisions. Seventeen comments focused on the qualitative cues such as opinionated language or entities, e.g., one real account decision based on “factual reporting” and a fake account decision due to seeming “too opinionated” (Figure 6, rows 2 and 5).

**Difficult Accounts:** Difficult accounts (column 3, 4, 7, 8 in Figure 5 left) are the ones with the cues pointing to contradicting directions, resulting in more uncertainties

in decision making. Seventeen comments focusing on quantitative cues such as fewer social network connections to other real news accounts for some real-difficult accounts yielded eleven inaccurate decisions (Figure 6, rows 12 and 15). Furthermore, twenty comments focused on qualitative cues such as users’ notion of opinionated language, in which seven cases it drove them to wrong decisions (Figure 6, row 11). Finally, fourteen comments focused on both quantitative and qualitative cues with only three of them yielding wrong decisions. In two of these cases, users decided to disregard the account’s anger ranking (Figure 6, row 16).

We observe that when users leverage both quantitative and qualitative cues with a thorough analysis of an account, they are more likely to make an accurate decision. Most comments contained a mix of qualitative and quantitative analysis (including language features, social network connections, and opinionated language) helped users to come to the correct decisions (Figure 6, rows 3, 4, 7 and 8).

### 3.7 Discussion and Future Work

Our goal was to assess the effect of confirmation bias and uncertainties on the investigation of misinformation using visual analytics systems. Although our post-questionnaire vignette, based on prior psychology research [132] showed that most of our users demonstrated a high level of confirmation bias, our experiment did not find significant differences between the experiment conditions. One explanation would be the hypothesis (all eight accounts are fake) we gave the participants did not resonate with them. If we had asked the participants to form their own hypothesis of the eight accounts being either real or fake by going through an example account, they

may have been more invested in the hypothesis and inclined to confirm or disconfirm it. Another explanation involves the use of Verifi, the visual analytics system that empowers users’ decision-making by allowing users to iteratively analyze multiple aspects of the news accounts. Often, people are instructed to ‘slow down’ and inspect information more critically [88] as an antidote to falling for confirmation bias. The Verifi interface could have played a role of somewhat mitigating confirmation bias in our experiments. This will be the subject of our follow-up studies.

We observe that participants’ responses to the cues were consistent with the account uncertainties/difficulties. Figure 5-Right shows how users’ average cue responses matched our original understanding of these accounts. Moreover, our regression analysis shows that certain cues significantly affected our users’ decisions (Q4-Q6) more than others. Opinionated language which had the strongest effect on users fake prediction stands out as an important lesson learned for future attempts to address misinformation. The fact that we allowed the opinionated cue to be purely based on users’ understanding of tweet texts, opens a whole new research question: How can we help users’ to more objectively identify/quantify opinionated language?

Furthermore, we find that uncertainty affected our users’ prediction accuracy. Our research shows that when a combination of quantitative and qualitative cues are presented clearly and with minimal uncertainty, users are successful in correctly differentiating between fake and real news accounts. In order to be resilient to these uncertainties, it is essential to take effective measures to communicate these uncertainties, motivate users to not be anchored on specific cues, and to holistically focus on a combination of qualitative and quantitative evidence. We plan to conduct a

followup experiment with adding uncertainty of the cues to the visual analytic system to test this hypothesis. One limitation of our study was the number of accounts chosen. Due to the time duration of our study (one hour), we decided to ask each participant to make decisions about eight accounts with varying difficulties. In order to test whether our results can be generalized, we plan to conduct a follow-up study that focuses on annotating a larger number of randomized accounts. The current study provides guidance on how we would instruct human coders to categorize all accounts based on the cues into different difficulty levels.

### 3.8 Conclusion

This chapter introduces a visual analytics system, Verifi, along with an experiment to investigate how individuals make decisions on misinformation from Twitter news accounts. We found that the account difficulty as mixed cues indicating real versus fakeness has a significant impact on users' decisions. The Verifi system is the first visual analytics interface designed to empower people in identifying misinformation. Findings from our experiment inform the design of future studies related to decision-making around misinformation aided by visual analytics systems.



## CHAPTER 4: VISUAL ANALYTICS FOR MULTIMODAL EXPLORATION OF MISINFORMATION SOURCE BEHAVIOR

This chapter is a slight modification of our paper with my co-authors Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin Arendt, Samira Shaikh, Wenwen Dou titled "Vulnerable to misinformation? Verifi!" published at the Proceedings of the 24th International Conference on Intelligent User Interfaces [91].

### 4.1 Introduction

The rise of misinformation online has a far-reaching impact on the lives of individuals and on our society as a whole. Researchers and practitioners from multiple disciplines including political science, psychology and computer science are grappling with the effects of misinformation and devising means to combat it [34, 140, 166, 168, 207].

Although hardly a new phenomenon, both anecdotal evidence and systematic studies that emerged recently have demonstrated real consequences of misinformation on people's attitudes and actions [147, 10]. There are multiple factors contributing to the widespread prevalence of misinformation online. On the content generation end, misleading or fabricated content can be created by actors with malicious intent. In addition, the spread of such misinformation can be aggravated by social bots. On the receiving end, cognitive biases, including confirmation bias, that humans exhibit and the echo chamber effect on social media platforms make individuals more susceptible to misinformation.

Social media platforms have become a place for many to consume news. A Pew Research survey on “Social Media Use in 2018” [12] estimates a majority of Americans use Facebook (68%) and YouTube (73%). In fact, 88% younger adults (between the age of 18 and 29) indicate that they visit any form of social media daily. However, despite being social media savvy, younger adults are alarmingly vulnerable when it comes to determining the quality of information online. According to the study by the Stanford History Education Group, the ability of younger generations ranging from middle school to college students that grew up with the internet and social media in judging the credibility of information online is bleak [207]. The abundance of online information is a double-edged sword - *“[it] can either make us smarter and better informed or more ignorant and narrow-minded”*. The authors of the report further emphasized that the outcome *“depends on our awareness of the problem and our educational response to it”*.

Addressing the problem of misinformation requires raising awareness of a complex and diverse set of factors not always visible in a piece of news itself, but in its contextual information. To this aim, we propose Verifi2, a visual analytic system that enables users to explore news in an informed way by presenting a variety of factors that contribute to its veracity, including the source’s usage of language, social interactions, and topical focus. The design of Verifi2 is informed by recent studies on misinformation [207, 108], computational results on features contributing to the identification of misinformation on Twitter [191], and empirical results from user experiments [97].

Our work makes the following contributions:

- Verifi2 is one of the first visual interfaces that present features shown to separate real vs. suspicious news [191], thus raising awareness of multiple features that can inform the evaluation of news account veracity.
- To help users better study the social interaction of different news sources, Verifi2 introduces a new social network visualization that simultaneously presents accounts' direct interactions and their topical similarities.
- Verifi2 leverages state-of-the-art computational methods to support the investigation of misinformation by enabling the comparison of how real and suspicious news accounts differ on language use, social network connections, entity mentions, and image postings.
- We provide usage scenarios illustrating how Verifi2 supports reasoning about misinformation on Twitter. We also provide feedback from experts in education/library science, psychology, and communication studies as they discuss how they envision using such system within their work on battling misinformation and the issues they foresee.

## 4.2 Production and Identification of Misinformation

Misinformation and its many related concepts (disinformation, malinformation, falsehoods, propaganda, etc.) have become a topic of great interest due to its effect on political and societal processes of countries around the world in recent years [160, 126, 99]. However, it is hardly a new phenomenon. At various stages in history, societies have dealt with propaganda and misinformation coming from various sources

including governments, influential individuals, mainstream news media, and more recently, social media platforms [24]. It has been shown that misinformation indeed has some agenda setting power with respect to partisan or non-partisan mainstream news accounts or can even change political outcomes [185, 30]. Misinformation is a broad and loosely defined concept and it consists of various types of news with different goals and intents. Some of the identified types of misinformation include satire, parody, fabricated news, propaganda, click-baits, and advertisement-driven news [178, 176]. The common feature among these categories is that misinformation resembles the look and feel of real news in form and style but not in accuracy, legitimacy, and credibility [108, 178]. In addition to the concept of misinformation, there is a broader body of research on news bias and framing defined as “to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation” [50]. By studying how misinformation is produced and how audiences believe it, as well as carefully surveying different means of news framing, we can have a strong motivation and guidance on how to address misinformation from a visual analytic perspective.

In this chapter, we approach misinformation from three different perspectives. First, we discuss existing research on how misinformation is produced. More specifically, how news outlets and individuals use different methods to alter existing news or to fabricate completely new and false information. Next, we discuss what makes audiences believe and trust misinformation. Finally, we review over existing methods for identifying, detecting, and combating misinformation.

### Production of misinformation:

How and why misinformation is created is a complex topic, and studying it requires careful consideration of the type and intent of misinformation outlets, as well as different strategies they use to create and propagate “effective” articles. Several intents and reasons can be noted on why misinformation exists. These reasons and intents include economic (e.g., generating revenue through internet advertisements and clicks), ideological (e.g., partisan accounts focusing on ideology rather than truth), or political factors (e.g., government produced propaganda, meddling with other countries political processes) [176, 178, 23]. Driven by these intents, certain news outlets use different strategies to produce and diffuse misinformation.

News media that produce misinformation often follow topics and agendas of larger mainstream partisan news media, but in some instances, they have the potential to set the agenda for mainstream news accounts [185]. Furthermore, they have a tendency to have narrow focuses on specific types of information that can manipulate specific populations. Often, we can observe extreme ideological biases towards specific political parties, figures, or ideologies [171, 23]. Some outlets take advantage of the psychological climate of audiences by focusing on fearful topics or inciting anger and outrage in audiences [23]. Moreover, misinformation accounts tend to take advantage of cognitive framing of the audience by focusing on specific moral visions of certain groups of people adhere to [53, 23, 106]. They also filter information to focus on specific subsets of news, often focusing on specific political figures or places [18, 20].

Misinformation is not always present just in text. Biases and news framing are

often presented in images rather than text to convey messages [61]. The power of images is that they can contain implicit visual propositioning [17]. Images have been used to convey racial stereotypes and are powerful tools to embed ideological messages [123]. Different means of visual and framing have been used to depict differences in political candidates from different parties [64]. For example, misinformation outlets use images by taking photos out of context, adding text and messages to the images, and altering them digitally [115].

Another important aspect in the production and dissemination of misinformation, specifically in the age of social media, are social bots. Social bots are automatic accounts that impersonate real outlets and manipulate information in social media. They are sometimes harmless or useful, but are often created to deceive and manipulate users [56]. Social bots can like, share, connect, and produce misinformation [108], they have the potential to amplify the spread of misinformation, and exploit audiences cognitive biases [107]. It has been shown that bots are active in the early stages in the spread of misinformation, they target influential users and accounts through shares and likes, and might disguise their geographic locations [166].

#### 4.2.1 Why and how we believe misinformation

Misinformation is falsified and fabricated information taking the form of real news. Equally as important to how misinformation outlets slant or fabricate news is produced, is to understand why and how people believe certain falsified information or take part in propagating the information. There are various social and psychological factors that effect the way we accept or reject information. It has been shown that col-

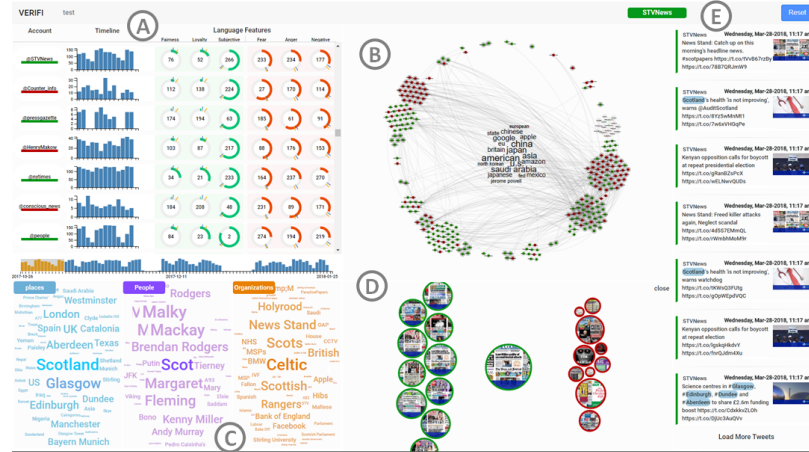


Figure 7: The Verifi2 interface is designed to support the investigation of misinformation sources on social media. The interface consists of 5 views: A) Account view, B) Social Network view, C) Entity Cloud view, D) Word/Image Comparison view and E) Tweet view.

lective means and social pressure affect our decision-making processes. Hence, we are likely to believe a news article to be true if our social group accepts it [169, 107, 49]. Prior exposure to a statement also increases the likelihood that individuals believe it to be true [140]. These factors form echo chambers in our real and digital societies and effect our ability to judge information accurately. These echo chambers are also amplified by algorithmically controlled news outlets that expose users to news they are more likely to interact with [174, 24, 58]. As individuals, we are also affected by confirmation bias and are more receptive of views that confirm our prior beliefs [128, 107]. In contrast, the ability to think analytically has a positive effect on the ability to differentiate misinformation from real news [141]. The uncertainty of misinformation and the ability to qualitatively and quantitatively analyze misinformation also has some effect on individual's ability to distinguish misinformation [?].

### 4.2.2 Battling misinformation

In order to combat the spread and influence of misinformation, we need to improve the structures and methods of curtailing misinformation to decrease audiences initial exposure, but we also need to empower individuals with the knowledge and tools to to better evaluate the veracity of the information [108].

The gravity of the issue has brought many researchers in computation to develop novel methods of detecting and identifying misinformation of various kinds. Many methods focus on detecting individual stories and articles. A number of fact-checking websites and organizations heavily rely on humans to detect misinformation [4, 11]. Researchers have developed a model to detect click-baits by comparing the stance of a headline in comparison to the body [34]. By focusing on sarcasm and satirical cues researchers were able to create a model that identifies satirical misinformation [158]. Images have also been used to distinguish misinformation on Twitter. It has been shown that both meta user sharing patterns and automatic classification of images can be used to identify fake images [70]. Using previously known rumors, classification models have been created to detect rumor propagation on social media in the early stages [208, 168]. Even though detecting misinformation on the early stages is extremely valuable, it has been noted that enabling individuals to verify the veracity and authenticity of sources might be more valuable [108].

Social media data has been used in visualizations related to various domains including planning and policy, emergency response, and event exploration. [90, 113, 112]. Several visualization based applications have been developed that allow users to



explore news on social media. Zubiaga et al. developed TweetGathering, a web-based dashboard designed for journalists to easily contextualize news-related tweets. TweetGathering automatically ranks newsworthy and trending tweets using confidence intervals derived from an SVM classifiers and uses entity extraction, several tweet features and statistics to allow users to curate and understand news on social media. [212]. Marcus et al. designed *twitInfo*, an application that combines streaming social media data several models including with sentiment analysis, automatic peak detection, and a geographic visualizations to allow users to study long-running events [116]. Diakopoulos and colleagues developed Seriously Rapid Source Review (SRSR) to assist journalists to find and evaluate sources of event-related news [47]. Using a classification model, SRSR categorizes sources to organizations, journalist/bloggers, and ordinary individuals. Furthermore, they use several features derived directly from Twitter as well as named entity extractions to allow journalists to understand the context and credibility of news sources. Even though these applications do not focus on misinformation, their lessons learned from dealing with social media are important inspirations for the design of Verifi.

There have been some efforts to classify and differentiate between verified and suspicious sources or to visualize specific types of misinformation such as rumors or click-baits. It has been shown that linguistic and stylistic features of news can be helpful in determining trustworthy news sources. Moreover, focus on different topics can be used to characterize news sources [127]. On Twitter, social network relationships between news accounts, as well as emotions extracted from tweets have been useful in differentiating suspicious accounts from verified sources [191]. Furthermore,

using different language attributes such as sentiment, URLs, lexicon-based features, and N-grams models have been created that can accurately classify rumors [107]. There are also a handful methods or systems that aim to enable humans to interact and understand veracity of sources or news articles. Besides the mentioned human curated fact-checking websites, There are systems that allow users to explore sources or (semi-) automatically detect news veracity on social media including FactWatcher [73], Rumorlens [155], and Hoaxy [165].

Most of these systems allow some exploration and fact-checking on specific misinformation related factors. As a recent survey of more than 2,000 teachers regarding the impacts of technology on their students' research ability shows, current technologies should encourage individual to focus on a wide variety of factors regarding sources of misinformation. Inspired by previous research in psychology, social sciences, and computation; and In line with recommendations that focusing on empowering individuals and enabling them to differentiate sources are the best strategies to battle misinformation [108], we developed Verifi2. Verifi2 employs an exploratory visual analytic approach and utilizes a combination of existing and new computational methods and is inspired by existing literature on misinformation, as well as a number of previous user studies on early prototypes to understand how users interact with misinformation [97].

### 4.3 The Design of Verifi2

In this section, we detail the inspirations for the Verifi2 system design, the tasks that Verifi2 aim to support, and distinguish Verifi2 from earlier prototypes that are

specifically developed for user experiments to evaluate cognitive biases [?].

#### 4.3.1 Task Characterization

Combating misinformation requires a multidisciplinary effort. On the one hand, the data mining and NLP research communities are producing new methods for detecting and modeling the spread of misinformation [168]. On the other hand, social scientists and psychologists are studying how misinformation is perceived and what intervention strategies might be effective in mitigating them[147]. More importantly, pioneers on misinformation research have been calling for an educational response to combating misinformation [207, 108]. Verifi2 aims to contribute to this cause by bringing together findings from various disciplines including communications, journalism, and computation and raise awareness about the different ways news accounts try to affect audiences’ judgment and reasoning. In addition, following Lazer et al.’s recommendation [108], Verifi2 aims to support the analysis of misinformation on the source/account level, in contrast to investigating one news story at a time.

Verifi2 is designed to support multiple tasks and enable users to evaluate the veracity of accounts at scale. The tasks are partially inspired by our comprehensive review and categorization of the recent literature on misinformation mentioned in section 2. The list of task are also inspired by our discussions and collaborations with researchers that have been working on computationally modeling misinformation. The tasks include:

- Task1: presenting and raising awareness of features that can computationally separate misinformation and real news;

- Task1A: presenting linguistic features of news accounts
- Task1B: presenting social network features of news accounts
- Task2: enabling comparison between real and suspicious news sources;
  - Task2A: presenting top mentioned entities to enable the comparison around an entity of interest by different news accounts
  - Task2B: enabling comparison of slants in usage of image postings and text between real and suspicious news accounts
- Task3: supporting flexible analysis paths to help users reason about misinformation by providing multiple aspects that contribute to this complex issue.

The tasks drove the design of the Verifi2 system, including the back-end computation and the front-end visualizations. Details of the Verifi2 system are provided in section 4.4.

#### 4.4 Verifi2

In this section, we provide detailed descriptions on the computational methods, the Twitter dataset we collected for misinformation investigation, and the visual interface. The overall system pipeline of Verifi2 is shown in Fig. 8.

##### History of Verifi

Verifi2 is a product of incorporating feedback from 150+ users from two experiments on decision-making on misinformation [?] and a comprehensive literature review on how misinformation is produced, believed, and battled. Verifi1<sup>7</sup> was is an interface

---

<sup>7</sup><https://verifi.herokuapp.com/>

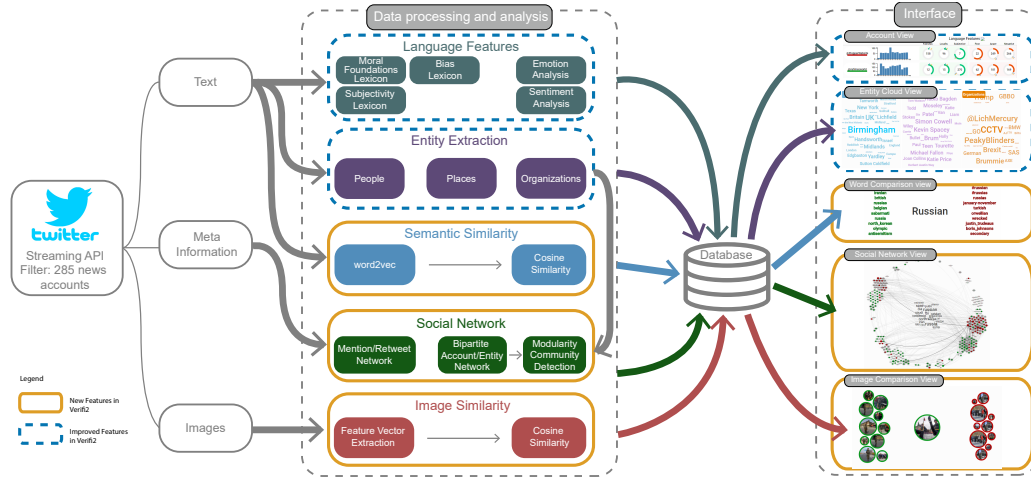


Figure 8: Verifi2 pipeline. Tweets, meta information, and images from 285 accounts are collected from Twitter streaming API. Language features, named entities, and word embeddings are extracted from the tweets. A social network is built based on the mention/retweet relationships. The accounts are clustered together using community detection of a bipartite account/named entity network. Image similarities are calculated based on feature vectors. All of these analysis are stored in a database and used for the visual interface.

that is built for conducting studies on cognitive biases using decision-making tasks around misinformation. It contains a small number of masked accounts and simple visualizations. Users would use Verifi1 to make decisions on whether an unknown account is real or fake based on given cues and information [?]. After conducting a thorough literature review and further analyzing the user studies, we realized that users are will not always have a visual analytics system at hand to make real-world decisions. Therefore, we developed Verifi2 with a new goal: Enabling users to explore news through a visual analytic system, learn about how news sources induce bias and slants in their stories, and ultimately transfer this knowledge in their real-world scenarios. However, The experiments and studies conducted using Verifi1 highlighted important points that guide the new features and goals of Verifi2:

- Users rely heavily on their qualitative analysis of news article in conjunction of

provided information. This result guided us to create completely new features that support qualitative analysis of news/context of news such as usage of visual imagery, topic similarity, and social network communities.

- Verifi1 included a simple force directed visualization of an undirected mention/retweet network. Even though these networks were shown to be computationally discriminating, they proved to be extremely misleading for some specific cases. In Verifi2 we provide a new network visualization with directed edges and topic communities.
- Verifi1 included a view for map visualization of place entities. This view was the least used view in Verifi1 and users suggested it was difficult to use. In Verifi2 we replace that view with a word-cloud visualization of the place entities.
- Verifi1 included linguistic features that proved to be vague or hard to interpret for users. We removed those features including non-neutral language, added new color coding metaphors and encodings for medians and averages of each score.

Finally, even though the design language of Verifi2 is influenced by Verifi1, Verifi2 incorporates new features or important improvements in every step within the pipeline (Figure 8), including more social media news accounts, different language features, new algorithms to support the comparison between real and suspicious accounts, and modified visual representations compared to Verifi1.

## Data

Verifi2 is designed to enable investigation of misinformation on a source-level. Therefore, our data collection starts with identifying a list of verified and suspicious news accounts. The list was obtained by cross-checking three independent third-party sources and the list has been used in prior research on identifying features that can predict misinformation [191]. The list includes 285 news accounts (166 real and 119 suspicious news accounts) from around the world. We then collected tweets from these accounts between October 25<sup>th</sup> 2017 to January 25<sup>th</sup> 2018 from the public Twitter Streaming API. As a result, more than 1 million tweets were included in our analysis. In addition to the account meta data and the tweet content, some tweets include links to images. Since images can play an important role in making a tweet more attractive and thus receiving more attention [123], we further collected all images for analysis. As a result, all tweets, account meta data, and images were stored in our database and were accessed using NodeJS. The tweets were then processed through a variety of different computational methods to support the analysis and comparison between sources of real news and misinformation.

### 4.4.1 Data Analysis and Computational Methods

This section will describe the data analysis and computational methods used to enable users to achieve the tasks Verifi2 aims to support.

#### 4.4.1.1 Data analysis to support Task 1

**Supporting Task 1A–Language use by news accounts:** Research on misinformation and framing suggests that suspicious accounts are likely to utilize anger or outrage-evoking language [23, 106, 171]. Verifi2 enables users to explore language use of news accounts. To achieve this task, we characterize language use by utilizing a series of dictionaries developed by linguists and a classification model for predicting emotions and sentiment. We started with a broad set of language features, including moral foundations [65, 71], subjectivity [206], and bias language [153], emotion [190], and sentiment [109] (Fig. 8 Language Features). Each of these features consists of various sub-features/dimensions. For example, the moral foundations includes five dimensions, namely fairness/cheating, loyalty/betrayal, care/harm, authority/subversion, and purity/degradation. We extract six different emotions (fear, anger, disgust, sadness, joy, and surprise) and three sentiment dimensions (positive, negative, and neutral). Overall, we collected over 30 dimensions that can be used to characterize language use. In [?], we performed a random forest to rank the dimensions based on their predictive power for a news account being real or suspicious. Among all of the calculated features, high score in fairness, loyalty, subjectivity was found to be predictive of real news accounts and fear, anger, and negative sentiment were found to be predictive of suspicious news accounts. The details for all language features can be found at [?], which presented a study leveraging Verifi1. For Verifi2, we re-examined the language features and removed the ones participants found



confusing. As a result, six language features were included in the Verifi2 interface.

### **Supporting Task 1B—Understanding relationships between news accounts:**

There are multiple ways to investigate the relationships among real and suspicious news accounts on Twitter. On the one hand, previous research [191] suggests that it is more likely for suspicious news accounts to mention or retweet real news accounts but not vice versa. On the other hand, literature on misinformation and framing shows that accounts have the tendency to focus on a subset (narrow set) of news topics compared to real news accounts [171, 23]. To present these two important aspects, we constructed a social network of all 285 news accounts based on both the retweet/mention relationship and the common entities they mention. To model the relationship among accounts based on their entity mentions, we constructed a bipartite network with two types of nodes: news accounts and entity mentions (people, places, and organizations). If an account mentions an entity there would be an edge between the two, weighted by the number of times the account mentioning that entity. We then applied maximum modularity community detection [130] on the bipartite network to identify communities of news accounts.

The community detection result contains 28 communities each with a number of accounts and entities with meaningful relationships. To briefly illustrate the community detection results, we describe three salient communities. The largest one contains 49 accounts (38 suspicious and 11 real). The top entities in this community include Russia, ISIS Iraq, Catalonia, Putin, and Catalonia which suggest an international focus of this community. Indeed, after examining self described locations of the accounts from their Twitter profiles, a vast majority were from Europe, Russia,

and the Middle East. The second largest community contains 41 nodes (31 real and 10 suspicious), with top entities in the community including Americans, Congress, Democrats, Republicans, and Manhattan which suggest a focus on American politics. Investigating the locations of these news accounts show that almost all of these accounts are from the United States. Another large community comprised entirely of suspicious accounts mention entities including Muslims, Islam, Lee Harvey Oswald, and Las Vegas which hints towards sensational topics related mostly to the United States.

#### 4.4.1.2 Data analysis to support Task 2

**NLP methods to support Task 2A—Understanding the top mentioned entities but real and suspicious news accounts:** Extant research shows that different accounts tend to have specific focus on topics such as places and political figures. By highlighting these entities from the text of tweets, we allow users to explore topical features of these news accounts. By extracting three types of named entities (people, places, and organizations) we can enable users to both easily explore topics of interest and compare how different accounts treat specific entities or topics. To extract named entities, we used python SpaCy.<sup>8</sup> After semi-automated clean up of the false positives and negatives, we saved the results in the database along with the tweets to be utilized in the interface.

Based on the top entities, Verifi2 further integrates computational methods to

---

<sup>8</sup><https://spacy.io/>

support comparing real and suspicious account postings around the same issue. Being able to compare the keywords usage around the same issue between real and suspicious news accounts contributes to the understanding of how different news media outlets frame their news. To this aim, we integrated a word embedding model [125] in Verifi2 to present the most semantically related words to entities selected interactively by an user. We first obtained vector representations of all words in our tweet collection using TensorFlow[13]. We then use the selected entity as the query word and identify the top keywords by calculating the cosine similarity between the query word and all words in the vocabulary. We performed this process separately for real news and suspicious news accounts. As a result, we were able to identify and display the most semantically similar keywords to a query word from both real news and suspicious news accounts’ postings.

**Computational methods to support task 2B—Compare real and suspicious account preference on images:** Research on misinformation and news framing shows that visual information including images are used heavily to convey different biased messages without explicitly mentioning in text [17, 123, 64]. To this aim, we utilized images from our 285 tweet accounts to enable users to compare and contrast the images posted by real and suspicious news accounts. We use the ResNet-18 model with pre-trained weights to extract a feature vector of 512 dimensions for every image from the “avgpool” layer<sup>9</sup> [75]. For any image that a user interactively selects as the query, we identify the top 10 similar images from both real news and suspicious news accounts measured by cosine similarity. Therefore, users can com-

---

<sup>9</sup><https://tinyurl.com/y9h7hex>

pare the images with accompanying tweets from real and suspicious accounts given an input image.

#### 4.4.2 Visual Interface

Verifi2 includes a visual interface with multiple coordinate views to support a variety of analytical tasks regarding misinformation described in section 4.4.1. The interactions are designed to support **Task 3** in that it enables flexible analysis strategies to make sense of the veracity of news accounts. The visual interface consists of five views: Account view, Social Network view, Entity Cloud view, Word/Image Embedding view and Tweet view. Through a series of interactions, each view enables users to focus specific aspects of misinformation while the rest of the views update accordingly. In this section, we will describe the features and affordances of each view while detailing how they are coordinated via users interactions.

##### 4.4.2.1 Account View

The Account view shows tweet timeline and language features of 285 Twitter news accounts (Fig. 7A ). Real news accounts are underlined by green lines and suspicious news accounts are underlined by red lines . The bar chart shows the daily tweet count of the account of the selected time range The six language features (fairness, loyalty, subjectivity, fear, anger and negativity) introduced in section 4.4.1.1 are represented by donut charts (scaled from 0-100). The numbers in the donut charts indicate the ranking of each language feature per account based on its score. For example, as seen in Fig. 7A the account @NYTimes ranks high on fairness (34<sup>th</sup> out of 285) and loyalty (21<sup>th</sup>) while ranking low on anger (237<sup>th</sup>) and negativity (270<sup>th</sup>). Two lines

in the donut chart indicate mean (orange) and median (blue) of the language feature to facilitate the understanding of how a particular account score given the mean and median of all accounts. The language features are displayed in two groups based on the correlations with the news accounts being real or suspicious. Overall, real news accounts have positive correlation with fairness, loyalty, and subjectivity while suspicious news accounts have positive correlation with fear, anger, and negativity. This view is connected to the analysis described in section 4.4.1.1 and supports Task1A.

Clicking on an account name filters the data for all views to only tweets from the selected account. Hovering the mouse over the name of each account invokes a tooltip showing the description of the hovered news account based on its Twitter profile. Users can sort the accounts based on any language features and explore how real and suspicious news accounts differ in language use.

### Social Network View

Based on the analysis introduced in section 4.4.1.1, the Social Network view present information on both retweet/mentions between news accounts, and the news account communities constructed based on the entities they mention (Fig. 7B). The visualization adopts a circular layout. Each node in the view represents a news account, colored in either red (suspicious) or green (real). The nodes are positioned at the perimeter of the circle. The directed links between accounts are established by retweet and mention, while the account communities are determined based on entity mention. To avoid adding more colors to denote communities, we alternated a light and darker shades of gray in the nodes background for adjacent communities to present

the clusters. This view supports Task1B.

Hovering on a node highlights all connections of the node to other nodes by with outgoing and incoming edges. An incoming edge represents the account is being mentioned or retweeted while an outgoing edge represents the account retweeted or mentioned another account. Moreover, hovering on a node shows a word cloud of named entities related to its community. Furthermore, users can pan and zoom in the Social Network view to switch focus between an overview of the social network and particular communities for analysis.

### Entity Cloud View

The Entity Cloud view presents three word cloud visualizations for designed for people, places and organizations respectively (Fig. 7C). Users can get a quick overview of the top mentioned named entities by all or selected accounts. We applied three colors - blue, purple and orange for each word cloud to distinguish mentions of these entities. The same color assignment is used in the Tweet view to highlight mentions of different entities in individual tweets. The Entity Cloud view supports Task2A.

Clicking on each entity in any of these word clouds filters the data to include tweets that contain these entities. It is important to note that the filters in Verifi stack, in that if an account is selected, then clicking on an entity shows tweets from that account containing that entity. Moreover, clicking on entities opens up the Word Comparison view to enable comparison of related terms from real and suspicious sources.

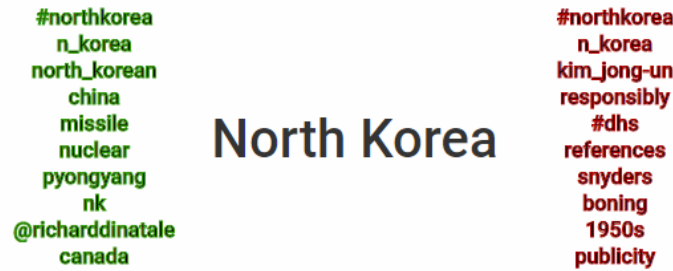


Figure 9: Word Comparison view. The query word is “North Korea”, the top related keywords from real news accounts are shown on the left while the keywords from suspicious accounts are displayed on the right.

#### 4.4.2.2 Word/Image Comparison View

The Word Comparison view (Fig. 9) is visible when users clicks on an entity from the Entity Cloud view. The view allows users to compare top 10 semantically similar words from real (Fig. 9 left) and suspicious news accounts (Fig. 9) right) to the selected entity. The user can find common and distinct words from either news sources. The user can further the comparison task by clicking on one of the semantic words. This action filters the tweets to those that include both selected entity and similar word. This view supports Task2A.

Enabled by the analysis described in section 4.4.1.2, the Image Comparison view (Fig. 7D) displays the top 10 similar images from real and suspicious news accounts given a query image. The central image circle of this view is the query image. To the left and right of this image are the top similar images from real and suspicious news accounts. The size of the circle encodes the cosine similarity value; with larger circles capturing the images being more similar. Hovering on an image shows the enlarged image as well as the associated tweet, account, and cosine similarity to the selected

image. Using this circular encoding, users can easily see whether images most similar to a selected image or mostly from real news accounts or suspicious accounts. This view supports Task2B.

#### 4.4.2.3 Tweet View

The Tweet view (Fig. 7E) provides details and allows users to read the actual tweets. This view enhances the reading experience by highlighting mentioned entities with the same color code as in the Entity Cloud view. The Tweet view also shows thumbnail images if a tweet links to an image. Hovering on an image thumbnail enlarges the hovered image for better legibility. Clicking on an image of interest opens the Image Comparison view to enable comparison of most similar images from real and suspicious news accounts.

### 4.5 Usage Scenario: exploring news related to an organization

In this scenario, the user is interested in exploring the news regarding the political climate in the United States. After launching Verifi2, The user starts by looking at entities in the Entity Cloud View (Fig. 10A). She immediately observes that one of the top mentioned organizations is GOP. By clicking the word in the Entity Cloud view, the interface updates to show tweets that mention GOP in the Tweet view. At the same time, the Word Comparison view gets activated so she can compare words most related to GOP from either real or suspicious news accounts (Fig. 10B). She observes that some of the closest words from real news (green colored words) are republicans, democratic, gops, bill, senate, and reform mostly “neutral” words that are directly related to the organization. She also observes that the related words from



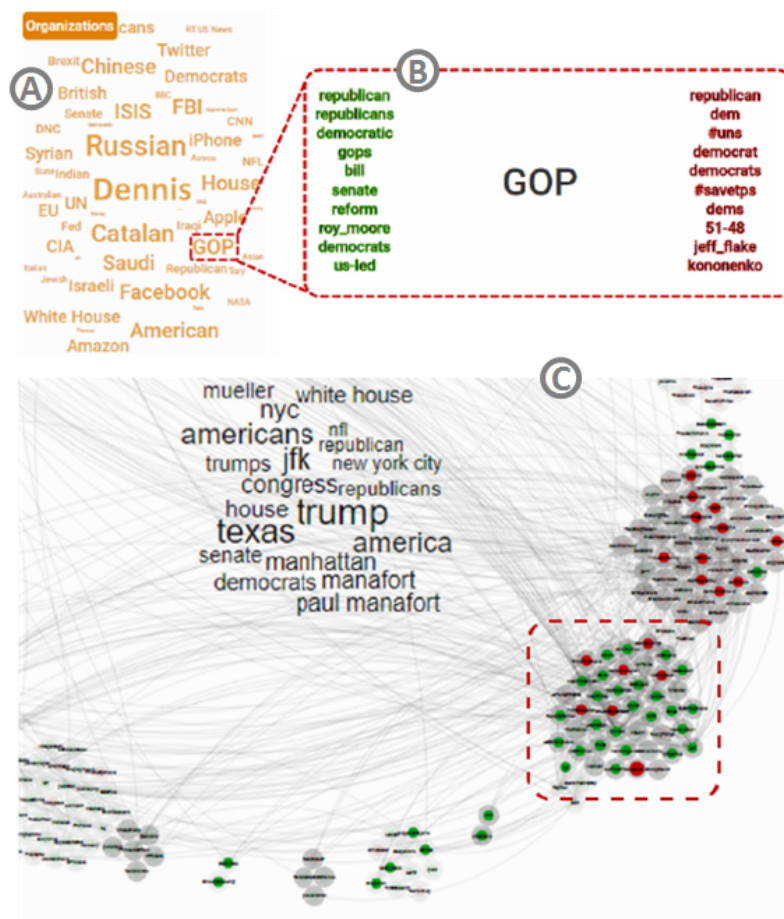


Figure 10: Top: Illustrating comparison between top semantically related words to the entity “GOP” (The Republican Party). Bottom: Community of news accounts in the social network that most mention the term GOP.

suspicious accounts (red colored words) are different. They include republican, dem, #UNS, democrat, #savetps, and jeff\_flake.

Through comparing the top related words to “GOP”, she finds the words dem, democrat, and democrats that are not among the top 10 words on the real news list. By clicking on the word “dem”, she can now cross-filter with the word “GOP” and “dems” and see how different accounts report their stories involving the two keywords. She finds out that the majority of the tweets that include both keywords in the Tweet view are from suspicious accounts. One of the tweets is a Russian account

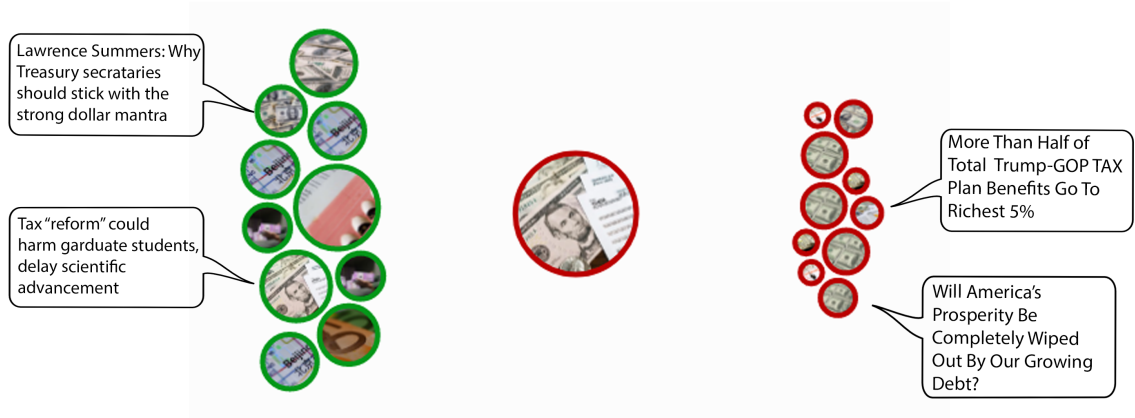


Figure 11: comparison of images/tweet pairs between real and suspicious news shows how these groups use images differently.

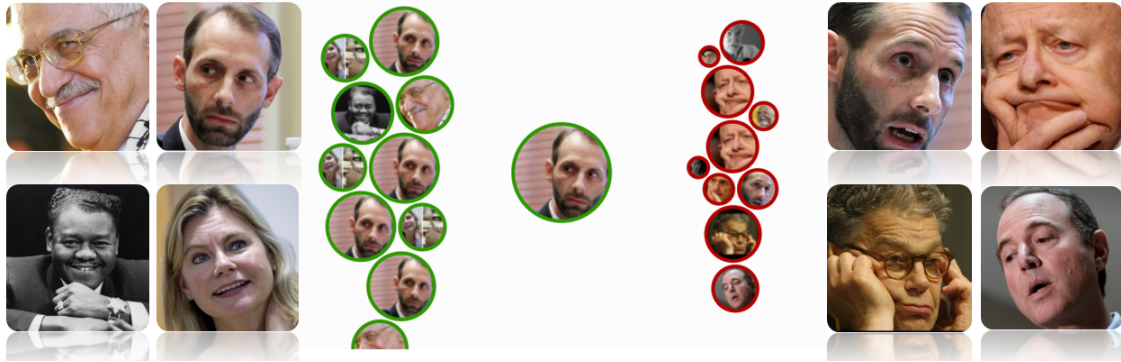


Figure 12: The Image Comparison view highlights how suspicious accounts frame images to convey messages.

retweeting an Iranian news account regarding a weapons ban proposal: *“GOP fails to back Dems’ weapons ban proposal”*. Other tweet looks extremely sensational and opinionated: *“BREAKING: Dems and GOP Traitors Unite Against Roy Moore”* and *“Enraged Liberals Snap After Senate Dems Get Treated Like Rag Dolls, Cave to GOP to End Govt...”*

The user then clicks on the word “bill” from the real news list which is not found on the misinformation list. Most of the tweets seems to be reporting events with neutral language about different bills: *“Rauner calls transparency bill ‘political manipulation,’*

*angering GOP”, “Property-tax deduction could help GOP reform bill” and “GOP leaders plan to hold votes on the bill early next week, first in the Senate and then the House.”*

Having learned from this comparison that the way suspicious news accounts frame news about GOP might be different than real news accounts, she continues to explore differences between accounts in the Social Network view. The view highlights the accounts that have used the term GOP. She notices that most of the accounts are concentrated in one community (See Fig. 10C). The top mentioned entities by this cluster shown in the graph include republicans, congress, house, and senate which are concepts related to American politics. After zooming into the cluster, she realizes that even though most of the accounts in this cluster are real accounts, there are a number of suspicious news accounts (in red). She then clicks on one of the red accounts to investigate how that account treats the GOP organization. She then goes to the Tweet view to examine tweets from these suspicious accounts. The first tweet is *“Democrats Were Fighting for the Rich in Opposing the GOP’s 401(k) Cut Proposal”*. The tweet includes an image that shows dollar bills and tax papers. By clicking on the image, she is able to compare the related images from real and suspicious news accounts (Fig. 11). She observes that all the images include pictures of money. She notices that one real news account is using the same image. By hovering on the images, she compares the tweets that link to these images from real and suspicious account. One tweet from a real news account is about tax reform: *“Tax ‘reform’ could harm graduate students, delay scientific advancement...”*. She then finds another image again with dollar bills and tax papers from a suspicious account: *“More Than Half of Total*

*Trump-GOP Tax Plan Benefits Go to Richest 5%*". She observes that this tweet has a suggestive tone in comparison to the more neutral tone of the tweet containing the similar images from real news.

#### Using Verifi2 to identify features of suspicious accounts

In this scenario, the user is interested in understanding features of suspicious accounts. The user starts with the Account view. He remembers that suspicious news accounts tend to rank high on the the Fear language feature. Therefore, he starts by sorting accounts based on the Fear language dimension. After sorting, his observation confirms that the majority of the top accounts with high fear ranking in the Account view are suspicious accounts. He then examines the names of different accounts. By hovering on the names, he is able to read the profile description of each account. One account that is ranked 39 in the fear score, describes itself as: *"You have found the tip of the spear in alternative media- Circumventing the dying dinosaur media systems of information suppression"*. The user finds this account interesting thus clicks on the name to start focusing on tweets posted by this account. Clicking on the name zooms the social network to center on this account. He finds that this account sits in a cluster that consists entirely of suspicious accounts. He also observes that this account has many network connections hinting that this is a prominent and active misinformation account. By exploring the directions of the links to other accounts, he finds that many of the suspicious accounts mention and retweet this account. However, he finds that none of the real news accounts have links going to this account. He interprets it as real news accounts are less likely to mention or retweet a suspicious

account.

The user then starts to read tweets from this account. He finds that many of the tweets are fear-mongering and highly opinionated. One of the tweets reads: “*Controversy has surrounded Ted Cruz’s father and his possible involvement in the Kennedy assassination...*”. He observes that the language does not have a regular reporting tone and the news is a conspiracy theory. Another tweet contains a black and white image of a elderly person with vampire teeth reads: “*Globalists Are Vampires That Must Be Destroyed*”. The user realizes that this image is obviously digitally altered, which could be a tactic that suspicious news accounts use. He clicks on the image to analyze whether other accounts use similar images. The image comparison view showed that none of the other tweets contain the same image. However, the most similar pictures shows black and white images from Russian accounts. A number of the images are pictures of a smiling Stalin and they talk about a ”massive new Stalin biography”.

#### 4.6 Expert Interviews

Our goal with Verifi2 is to help individuals learn about the different ways suspicious news accounts are showed to introduce slants, biases, and falseness into their stories. Moreover, we aim to produce a system to create help users make decisions in real-world scenarios. To assess our system on these grounds, we conducted five semi-structured interview sessions with experts from different disciplines who work on battling misinformation or conduct research on misinformation. These five individuals were university professors from education/library sciences, communications, digital

media, psychology, and political science. Naturally, each of our interviewees had a different approach to understanding and battling misinformation which resulted in valuable commentary on both the features of Verifi2, as well as the potentials for it to be used in real-world attempts to battle misinformation. The interviews were recorded, transcribed, and analyzed.

### Interview Setup

Interview sessions took between 45-60 minutes. Before showing the interface to the interviewees, we asked them to introduce themselves, their professions, and their specific concern with misinformation. Next, we asked general questions about misinformation and how technology can assist in attempts to battle it. We also inquired our interviewees about how they differentiate between misinformation and real information, suspicious and real news sources, and the reasons audiences chose to believe misinformation. After the initial questions, we opened Verifi2 and started thoroughly explaining the different parts of the interface. The interviewees made comments and asked questions regarding the features throughout the interface walk through. Finally, we asked to important closing questions. Namely if they thought of other features that needs to be added to the interface, and what potential uses they see for Verifi2. In this section, we will summarize the important lessons learned from these interviews. Direct quotes from interviewees are slightly edited for legibility.

#### 4.6.1 A spectrum of trustworthiness rather than binary classes

All of our interviewees acknowledged the challenge in defining misinformation. As well as the complexity of the issue which involves both news outlets with different

intents, as well as audiences with different biases. Their definitions of misinformation were different but with some common features. They made notes on our binary choice (real vs. suspicious) and how it is important to communicate the complexity of misinformation and move away from a simple binary definition thus not alienating the people who might be most prone to the dangers of misinformation.

The education expert highlighted the importance of sources, as well as factually false information: *“we always tell people [to] consider the source. So who is providing this information? [...] and even if the source is a reliable source, it can still be factually incorrect. I mean we have news networks that wouldn’t consider them to be reliable and even sources that are reliable [might publish inaccurate articles], [...] I think with the fake news or misinformation are things that are like blatantly not true.”*

The cognitive psychology expert described how misinformation is often bias, or sometimes completely factually wrong, but satire which has unique features, might not be considered as misinformation: *“misinformation would be information that is framed or biased or presented in such a way as to misdirect [...] it is basically only telling one side of the story with particular piece of information that when you take it in context, it seems very reasonable. But then when you take information out of context, [...], that form of misinformation is particularly dangerous.”*

The expert in digital media discussed the subtlety of the of the misinformation problem and hinted towards revisiting how we classify news accounts *you have maybe a factual article, but the headline is focusing on a particular part of that article and a kind of inflammatory way that actually has nothing to do with the constitution.* Our communications expert mentioned that misinformation could be in forms other

than news: : *“Misinformation could be in different settings, not just narrowly focused on news in particular, however, [completely false news] is probably one of the most problematic forms of misinformation because it has the potential to reach audiences”*. She then gave the example of satirical news: *“we do talk about how it’s different from satirical news meetings, news that is used to entertain. It’s fake anyway because it’s made to be fake, but may make people laugh. Right? And so there is difference between what you see on Onion network, let’s say versus you know, what you see on other websites that are purposefully designed to mislead people.”*

Our expert interviewee in cognitive psychology mentioned that Verifi would be a useful tool for conducting research on social media accounts, but she was concerned with the selection of the 285 accounts : *“...these two hundred and eighty five [accounts], how are they selected? How are they defined [in] that space? Or how were they selected to be included in the original Dataset? how are they representative? are they representative of [of the whole political spectrum]? What is the distribution across liberal and conservative we would need to know to make valid conclusions from the data. So for example, an account like Onion, is fake, but it is on the liberal side. This is really key for people looking at this and then being able to interpret the meaning.”*

The political science professor, emphasized the importance of minimizing the effects of users’ biases while dealing with misinformation: *“... people seek information that reinforces their biases. that bias is especially strong when in circumstances in which you cue a group [or] a social identity, So if you’re sending a group relevant message that cues you to think about [the other] group and distance yourself from an opposing group and pair that with misinformation about opposing or threatening group, that*



*can be a very powerful persuasive message.”*

All of the responses from our interviewees highlight a very important fact about misinformation. The type of article, and account (be it satire, click bait, or real news) is a very important factor. Furthermore, Two of the interviewees mentioned that the categorization of the real vs. suspicious accounts should be within a spectrum rather than a binary decision. Our digital media expert elaborated on this point: *“there’s a couple of websites or Twitter feeds that I’ve been following that are really, problematic because they’re basically funded by NATO and they’re there to track Russian misinformation and propaganda. But the really interesting thing about it is that they’re using language that’s very propagandistic themselves.”* She then continued about how the interface could respond to these gray areas: *“maybe rather than ascribing them [the accounts] as being real or [suspicious], they can be scored on a continuous spectrum taking into accounts all features in Verifi.”*

#### 4.6.2 Verifi2 as a potential tool for education on misinformation

Two out of five interviewees have been working on the “Digital Polarization” project, which is part of a national program to combat fake news <sup>10</sup>. Therefore, educating college students about misinformation has been on the top of their minds. After walking the interviewees through the usage of Verifi2, each had different thoughts on how the interface would be useful within their tasks.

The first point of discussion in our interviews was how they focus on misinformation as educators and what features they teach to their students. The communications

---

<sup>10</sup><http://www.aascu.org/AcademicAffairs/ADP/DigiPo/>

professor teaches a module on misinformation which she describes as: “[...] We do talk about different types of fake news. When talking about news, we’d talk about how we evaluate information. Just like when you write a paper, you were supposed to evaluate and cite your sources. [...] I also ] make them identify which news story, after reading it, is true or false. Sometimes it’s hard to tell. So we also play in the class some videos and some do discussions about why fake news exist and how fake news are hard to tell apart from satirical news and sometimes news in general”.

The education and library science expert emphasizes the importance of identifying sources: “... They [the students] should look and investigate who the author is. Basic things that we teach in terms of just general information literacy. Like generally [with] a .com website, you should be more investigative with because you know there’s a commercial purpose. [...] culturally there’s things that we consider reliable sources. So like a peer reviewed scholarly journal article is generally going to be a reliable source, a website blog article that can be produced by anyone that’s not vetted by anyone. You need to be a little bit more introspective and see where they’re coming from and look at the background of that person and things like that.”

When we asked our interviewees what features a technology or tool should have to enable their research and education tasks, they emphasized simplicity, ease of access, usage of visual information, and customizability. The education professor described a useful tool as: “I think something that uses plain language, easy to navigate, supporting a desktop and mobile version since so many students are accessing information from portable devices. [...] something that’s visual for students.”

After walking our interviewees through the usage of Verifi2. Each had different

thoughts on how the interface would be used within their educational tasks. The education and library sciences expert expressed interest in capitalizing on the comparison abilities of Verifi2 and suggested a new feature to enable users to actually filter the datasets to see retweets between accounts: *“I think it would be useful because it’s doing so much comparison between real news and fake news and I think that’s where we could potentially use that with our students. the [social network] graph is a little overwhelming. But I do like the idea of being able [to see who retweets whom] [...] It would be interesting if you could scale this data in a way someone could actually like look at what fake news accounts are saying about the real news. I wonder if they’re doing more than just re tweeting the story.”*

The communications professor who teaches a class with a module on misinformation mentioned that the interface could be useful for her students. However, she was concerned that the number of accounts in the interface, as well as some of the features might be too complex for an assignment and would need proper training: *“depending on the class that people teach and how much time you have and how complicated an exercise you want your students to do. I think just talking about the topic and fake news that takes a lot of time already. [...] but I think a simplified version with a recorded video explaining these features would be practical for an assignment.”*

The digital media professor discussed how Verifi2 would be a useful tool for teaching students about how to compare different sources. However, she was concerned about alienating individuals with different political preferences: *“... someone may subscribe to some fake news websites, and that’s what they think [is true]. And I think that’s all true if you presented this to them and then they saw that their preferred website came*

up as as fake information. Um, what happens is that'll just set their back up against a wall and they'll be like, well, no, this is totally wrong.". she then continued to describe how the interface would be useful for high school and college students if it allowed them to come to conclusions themselves: *"I think it would be such a great tool for teaching people how to evaluate information. I'm especially thinking of high schoolers [...] I think about some of my freshman students coming in and they're really good at parsing out information that's presented visually and they would be really good at navigating something like this and say, oh, OK, this is leaning towards true or this is leaning towards fake.."*

The political science professor emphasized the importance of creating a tool which discourages individuals from partisan motivated reasoning: *"one of the most important things that your tool would have to do is to help people disengage from a partisan motivated reasoning or predisposition based reasoning, which is actually really hard to do. I mean, you can cue people and try to reduce it, But we have an inherent tendency to shore up our existing beliefs, but a tool that, that somehow gets people to think about things in atypical ways [would be really beneficial]"*

There was also a discussion of using Verifi2 as an assignment for students through which they would apply their knowledge in media literacy to decide whether accounts are trustworthy or not. The education expert described this and how Verifi2 can be applied in a number of different educational settings: *"I'm thinking for a student assignment, we could [for example] pick 5 reliable sources and 5 unreliable ones. [...] I could even see this potentially being useful not just in these targeted classes [that we teach] but in some of the liberal studies courses, that are more Gen ed requirements.*

*I can see this, depending on the right instructor, if we knew about this tool that could educate people about [misinformation], they could be developing assignments for their writing classes or inquiry assignments. Overall, I see a lot of potential.”*

Our expert interviews highlight the potentials for Verifi2 to be used in scenarios where experts educate individuals about differences between trustworthy and suspicious sources of news. However, they highlighted the delicacy required to take on the complicated task of educating younger individuals on these concepts. Their recommendations included simpler interface and allowing users to decide the trustworthiness of the accounts based on their own assessment.

#### 4.6.3 Positive and negative usage of facial expressions in images

Biases and slants can be hidden not only in text, but also in the images. These hidden framings can activate our mental frames and biases without us noticing them [64, 123]. During our interviews with experts conducting research about misinformation, a professor in psychology who specializes in emotion analysis made observations about the differences between real and suspicious accounts on their usage of graphics/images. After receiving an introduction of the interface, the psychologist started exploring the image comparison feature in Verifi2. She was particularly interested in close-up images of individuals. Given her expertise in emotion contagion, her hypothesis was that real vs. suspicious news outlets may frame the visual information in tweets using subtle emotional encoding. She focused on the facial expressions in the images used in suspicious accounts compared to real news accounts.

Clicking on the picture of a person active in politics and examining the different

top related images from real or suspicious accounts, she said: *“These [pointing at the images from suspicious sources] are all more unattractive. you see this with [polarizing political figures], all the time whereas these [pointing at images from real news accounts] are not, they are all showing positive images, [...]. So these people [real news images] look, composed! these [suspicious news images] less-so. So fake news apparently is also much more likely to [use negative imagery].[...] That’s exactly what I do. When I started I did facial action coding. So you can actually do individual coding of facial muscles and they correspond with particular emotions and that is also something that you can [use the interface for]”*. Fig. 12 shows the pictures this expert was using as an example to show the way suspicious accounts frame their images using emotional encoding.

#### 4.7 Discussion

The final design of Verifi2 is inspired by three threads of research and experiments: data mining and NLP, social sciences, and direct user feedback from prior research on decision-making on misinformation [97]. With this in mind, Verifi2 supports calls to address misinformation from a multidisciplinary approach [108]. To evaluate our system, we interviewed experts from multiple disciplines who focus on the problem of misinformation, each who provided different feedback. First, when we are dealing with the ever-increasing but complex topic of misinformation, we need to pay attention to the different types of misinformation as well as the intents of such producers of misinformation. Additional consideration must be made as well on the vulnerabilities and biases of consumers of information. To better deal with this issue, our interview-

wees recommended allowing users to make modifications on the systems encoding on suspicious or real accounts based on the many qualitative and quantitative evidence available through the system. In addition, they emphasized the importance of moving from a binary representation of accounts (real vs. suspicious) to a spectrum. One solution to both these suggestions is to enable human annotations (e.g., adjust binary classifications). By allowing users to annotate, we will increase user trust. Human annotations will be critical for accurate and updated results. Our current data is based on third-party lists that, while are good proxies, may change over time. Human annotation could become critical to validate, filter, and override. Combined with Visual interactive labeling (VIL) like CHISSL [22], human annotators could provide real-time feedback on the accounts (e.g., cognitive psychologists who label extreme emotions).

Like many other researchers, our interviewees struggle with how to educate individuals on the spread of misinformation. They mentioned that Verifi would be a very useful tool to enhance traditional strategies such as media literacy courses. However, their responses made apparent the fact that the task of educating individuals requires much more work and rigor. It was mentioned that some simplifications and proper training needs to be done in order to make Verifi2 ready for education, especially for younger audiences. We are working with the researchers to modify, clarify, and prepare Verifi2 for such educational applications. Moreover, different educational settings require slightly different modifications. These modifications include enabling Verifi2 to focus only on a subset of the accounts or to enable users to observe completely new accounts, creating an infrastructure for Verifi2 to be used as course assignments,

and partnering with educational institutions to deploy Verifi2.

## 4.8 Conclusion

While not a new phenomenon, misinformation as an application of study is in its infancy [107]. In such an environment, visual analytics can play a critical role in connecting across disciplines [108]. In this chapter, we introduced the Verifi2 system that allows users to approach misinformation through text, social network, images, and language features. We described case studies to show some of the ways users can utilize Verifi2 to understand sources of misinformation. Finally, we interviewed a diverse set of experts who commonly focused on misinformation, but had different approaches. We learned that Verifi2 can be highly useful to be deployed as an educational tool. But most importantly, we learned that Visual Analytics can be an extremely useful tool to battle misinformation. This, however, requires more work and a common goal from the visualization community to address the complexities of such issues. We call for visualization researchers to take on this task, to make our democracies more resilient to the harms of misinformation.



## CHAPTER 5: IMPACT OF EMOTIONAL FACIAL EXPRESSIONS IN IMAGES ON USERS' JUDGEMENTS OF CONTENT BIAS AND SOURCE CREDIBILITY

### 5.1 Introduction

Following the turbulent 2020 United States presidential election, the storming of the capitol by rioters was a shock to many worldwide observers. Observing such an incident, we might ask what caused rioters to not believe the integrity of the election. More specifically, what causes audiences to find sources producing such misinformation as credible, trust these sources, and consequently believe their messages? In fact, recent work has shown that beliefs of content accuracy might not necessarily correlate with beliefs of source credibility and truthfulness [195]. This is a very complex problem with many interconnected aspects concerning the sources of news and their agendas, the specific strategies and characteristics they use to produce content, and the cognitive processes of content consumers [89, 201].

News sources curate their text and images with specific styles, tones, and emotions. Mainstream news media and misinformation sources take advantage of highly emotionalized content to influence their audiences [201, 23, 191, 156]. Various studies have shown that these strategies are often effective. People are more likely to be drawn to highly emotionalized news, click on articles associated with extreme emotions, and share headlines with more negative sentiments [35, 154]. Emotional content is also linked to virality on social media [31]. Furthermore, participants' self-reported

experience of heightened emotion such as elevated anger or sadness increases their perceived accuracy perceived accuracy of false news of false news but not of factually correct news [118].

In addition to emotional textual content, news and misinformation sources take advantage of social media’s highly visual nature by deliberately choosing images that amplify the persuasive power of the text content [201]. Examples of such visual content include images that communicate racist concepts not present within the text [123], politicians in the US and Europe [119], misleading images of historical diseases [87], and deceptive deep-fake images [184]. Images are also shown to impact how users judge the credibility of information sources in several ways. For example, after seeing an image of a brain, users are more likely to believe claims in certain scientific articles [163, 120] while images of smoking increases belief in messages in warning signals [167]. Articles accompanied by alarming images [103] or ones that depict victimization [211] are shown to increase users’ selective interaction with the articles. Emotional content in images also impacts the likelihood of people believing a statement. For example, being exposed to highly negative emotional images about different phenomena increases the believability of news content[188]. In comparison to positive imagery, being exposed to negative imagery has been associated with a greater perceived accuracy of false information[148].

This paper investigates how the accumulation of positive or negative facial expressions in images of social media posts affects users’ judgments about **content bias** and **source credibility**. Source credibility is a primary factor in the persuasiveness of news sources [77, 195]; while, content bias, defined as “having a perspective

that is skewed” is also a distinct element in determining source credibility [195]. We seek to answer the following questions: *1) how news content accompanied by images with happy (positive) or angry (negative) facial expressions impact users’ judgements about the credibility of content and news source; and 2) how users’ prior attitudes towards political identities impacts their judgements in light of emotional facial depictions of those identities in news content.* To answer these questions, we conducted two consecutive preregistered controlled experiments:

1. In study 1, we examined news accounts with highly emotionalized visual content without focusing on specific topics or politicians. We investigated how angry or happy facial expressions in images impact users’ judgements about multiple news accounts. Specifically, we evaluated if exposure to highly angry or happy facial expressions impacts users’ perception of anonymized right/left leaning and mainstream/misinformation accounts.
2. In study 2, we expanded our exploration of how emotions in images impact users’ judgments about sources. We focused on sources’ with highly emotionalized visual coverage of specific influential politicians such as Donald Trump or Angela Merkel. Furthermore, we investigated how users’ prior attitudes towards each politician interact with sources’ angry or happy portrayal of those politicians when making judgement about sources.

For both studies, in addition to quantitative analysis, we qualitatively analyzed users’ comments about their decisions to interpret the subtleties of users’ decisions outcomes under emotionalized visual information. In study one, we observe that users

find sources propagating tweets with angry facial expressions as less credible. However, we noticed that this effect is in general reduced by sources' political orientation. We suspected that this reduction could be due to differences in emotional facial expressions of different topics covered by sources in which a source might cover news about one person with angry imagery and others with happy. In study two, we found that when users are exposed to continuous angry coverage of specific politicians, they are more likely to perceive the content as systematically biased and the source as less credible. These studies highlight the importance of visual information in studying and combating misinformation on social media and the extent to which highly emotionalized visual coverage by sources might impact users' trust in both mainstream and misinformation media across the political spectrum.

## 5.2 Study motivation, design, and implementation

Considering users' perception of source credibility as a function of the content they consume, we formulate the overall structure of the studies as:

1. Users are instructed to go through multiple sets of curated social media posts from **anonymized** sources of news.
2. Users are asked to evaluate bias for multiple posts from one source before evaluating credibility.
3. In order to take individual differences in the amount of information a user needs to come to a decision into account, we allow users to form opinions after reading and evaluating a non-fixed number of posts. In other words, users are instructed

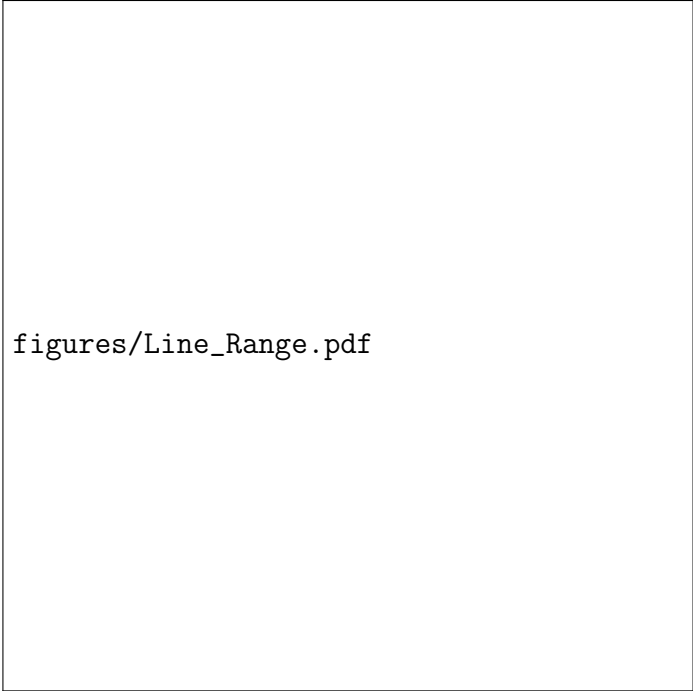
to view as many posts as they need.

4. After deciding they have viewed and evaluated enough posts from a source, users are instructed to provide a credibility rating of the source.

This structure is repeated for both studies 1 and 2. The randomized treatments are applied by altering the content users see from each source (details of each controlled experiments is provided in each respective section). Another important aspect of this study is also observing how users' uncertainty changes under different conditions. Tormala and Petty argue that more certain attitudes and judgements have important consequences including guiding behaviors, resistance to persuasion, and being persistent over time [181]. Thus, in addition to eliciting perceptions of content bias and source credibility, we ask users' to provide uncertainty about their judgement. Finally, We also collect users' ratings on sources' political orientations. Finally, we ask users to provide us with a short text description of how they arrived at their decision.

### 5.2.1 Dependent variables and elicitation method

As suggested by Karduni et al. [91], we aimed to allow users to evaluate bias and credibility on a spectrum as opposed to a binary choice. We elicit users' perceived bias as a continuous number between 0 (unbiased) and 1 (biased). We also elicit users' uncertainty around their decision as a confidence interval in the same range. We elicit users' perceived credibility of a source as a number between 0 (not credible) and 1 (credible). We elicit users' perceived political orientation of a source as a number between -1 (liberal) to 0 (center) and 1 (conservative). For both variables, we elicit



figures/Line\_Range.pdf

Figure 13: The Line + Range elicitation method.

users' confidence as a confidence interval range within each respective domain.

To elicit users' beliefs and uncertainty about the perceived bias, credibility, and political orientation variables as a continuous value rather than a binary choice. We adopted a modified version of the Line + Cone technique introduced by Karduni et al. [94] The Line + Cone technique was specifically designed to elicit users' beliefs about correlations between two variables. Within that study, users were instructed to first select a line that best represents their belief and then draw a range that encodes other plausible alternatives to their beliefs. In other words, this method first allows users' belief as a numerical value and a range denoting their uncertainty around the decision.

In this study, we introduce the Line + Range that adopts a similar design approach (See Figure 13). The Line + Range method enables eliciting users' choices in different

continuous spectrums such as bias, credibility, or political orientation. This method allows us to elicit users' perceptions as a continuous measure between bounded values while eliciting a range highlighting users' uncertainties in their decisions.

### 5.2.2 Dataset

In this study, we started with the same datasets used in [91, 97] which included a collection of tweets from multiple mainstream sources of news, as well as multiple accounts labeled by third-party sources as being likely to produce hoaxes, propaganda, or in general fake news [191]. To create a dataset of tweets that include facial expressions, we processed the dataset mentioned above through several computational methods. First, we used the python face-recognition library <sup>11</sup> to identify images that include faces. To identify images of different politicians, we used Google's FECNet [162] to extract feature vectors from all faces in the images and used the HDBScan clustering algorithm [121] to cluster the images of faces. We then manually identified prominent clusters of influential politicians.

To extract emotion scores from images, we utilized two readily available libraries, which provided us the ability to sort the images based on different emotion predictions [164] <sup>12</sup>. After generating a larger candidate dataset, we manually excluded low-resolution images, were inaccurately assigned to a cluster, or were mislabeled by the emotion prediction libraries.

---

<sup>11</sup><https://pypi.org/project/face-recognition/>

<sup>12</sup><https://github.com/thoughtworksarts/EmoPy> and <https://pypi.org/project/deepface/>

### 5.2.3 Study interface and procedures

We conducted Both studies 1 and 2 using a custom interactive interface developed with React JS and multiple javascript libraries such as D3.js for the Line + Range elicitation technique. The responses from users were collected in a MongoDB database through a NodeJS backend hosted on Heroku. The interface first prompts users with a consent form per IRB requirements. After clicking on consent, users are randomly assigned to their predefined conditions as specified by the study. Immediately after providing consent, we provide users with a brief demographic questionnaire. The next page provides an instruction to the Line + Range elicitation technique. After the elicitation technique instructions, the study interface provides instructions on the study's goals, i.e., describing Credibility and Bias and the choices users need to take. Consecutively, The interface routes users to the main study page. Based on their random condition assignment, they will go through pages resembling a social media page. Users go through tweets one by one, using the Line + Range method to provide a bias rating. Users are prompted to click on a "Make a decision" button that will open up a pop-up window with multiple questions about credibility, political orientation, and an open text box for comments on each users' decision. After providing ratings for one account, the study page is refreshed with a new round of tweets from a new account. This process is repeated until all accounts are evaluated. In the final stage, users are provided with a brief questionnaire to record their self-described political leanings, as well as two open-ended attention check questions. The final page of the study provides users with a debriefing on the study, as well as unique a unique token



for study incentive processing purposes.

#### 5.2.4 Study 1 & 2 models:

Our experimental design for both studies consisted of repeated measures for each user within two levels of responses including source level responses (credibility) and tweet level responses (bias). Furthermore, users' responses were continuous variables bounded between 0 and 1. For both studies, we used mixed-effects beta regressions to address the hierarchical design of our studies and the bounded dependant variables. Within the text, model coefficients as log of odds ratios are reported with corresponding confidence intervals. In the model figures, for ease of interpretation, we transformed the log-odds ratios to odds ratios. Odds ratios show the direction and strength of how each independent variable impacts the dependent variables. We used the normal approximation to calculate p-values of fixed effects and t-values produced by `lme4`.

In both studies, To measure the effects of our experimental design conditions, we built two mixed-effects models using R's `glmmTMB` for a beta regression.

#### 5.2.5 Topic modelling and qualitative analysis of topics:

First, we used Non-Negative Matrix Factorization (NMF) [42] to extract topics from the comments. We then qualitatively evaluated the topics through each topics' top-terms and top documents. Topics were then qualitatively categorized into themes based on their similarity. Each qualitative analysis section, includes samples of comments most related to each theme.

### 5.3 Study 1

In Study 1 we examined how exposure to positive (happy) and negative (angry) emotions in images influence perceived content bias (tweets) and source credibility (accounts). Participants completed a task in which they observed a series of tweets from eight different accounts. Text data (tweets) was collected from Twitter streaming API from October 25th to January 25th, 2018 from multiple news accounts [92]. The accounts are either labeled as **mainstream** (e.g., NYTimes, NY-Post, CNN, and Fox News) or known to produce **misinformation** (e.g., Breitbart, amLookout, investWatchBlog). The accounts are also categorized as being right-leaning or left-leaning<sup>13</sup>. Given these two dimensions of mainstream/misinformation and right/left political orientation, users evaluate a total of eight accounts (See table 6), two from each of four different categories of mainstream-left, mainstream-right, not misinformation-left, and misinformation-right. All images included facial expressions, and were scored on happiness and anger using multiple supervised machine learning algorithms (for more info refer to the methodology section). The content shown to users was manipulated in a between-subjects manner. Users' were randomly assigned to one of three conditions: **happy**, **angry**, and **mixed**. The happy condition receives tweets that contain images with the highest happy score. The angry condition receives tweets containing images with the highest angry image score. The mixed condition alternates between happy and angry images. Furthermore, one account from each category is randomly selected to include no images. For example, a

---

<sup>13</sup><https://www.allsides.com/unbiased-balanced-news> and <https://mediabiasfactcheck.com/>

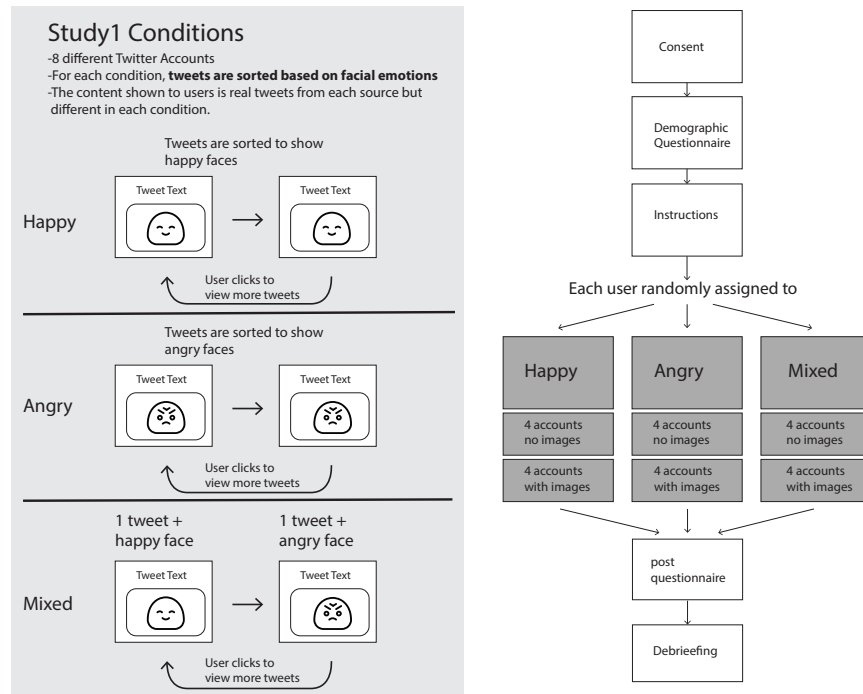


Figure 14: Study 1 conditions and process

participant assigned to a happy condition always views tweets sorted with the highest happy rated images but does not see images for four of the accounts (see Figure 14).

For each account, users first evaluate content bias in individual tweets. Participants record their belief and uncertainty about the bias of each tweet. By clicking on view more tweets, participants will view an extra tweet until they come to a decision about the source (a minimum of 5 tweets was enforced for each account). By clicking on “Make a Decision,” users see a pop-up view in which we elicit their perceived credibility and political orientation of each source. Users also have the option to use a text box to describe what influenced their decisions.

Table 6: 8 Twitter accounts that users evaluate in study 1. Each user goes through content sorted based on emotions in facial expressions collected from tweets’ images.

Source name	Type	Political Orientation
@veteranstoday	misinformation	left
@amlookout	misinformation	right
@opednews	misinformation	left
@InvestWatchBlog	misinformation	right
@MotherJones	mainstream	left
@nypost	mainstream	right
@cnnPolitics	mainstream	left
@Jeresulem_Post	mainstream	right

### 5.3.1 Hypotheses:

In Study 1, we hypothesize that in comparison to the tweets with happy images, users are more likely to assess tweets with angry images as biased<sup>14</sup>. Furthermore, we also hypothesize that as compared to the Mixed condition, users in the angry condition are more likely to assess sources as less credible. On the other hand, we hypothesize that users in the happy condition will be more likely to perceive tweets as less biased and sources as more credible. However, since it is likely for users’ judgment to be influenced by the inherent differences in the accounts, we also explore the effects of political orientation (right vs. left) and source type (mainstream vs. misinformation) on their judgements. In summary, we evaluate the potential effects of image emotion, source orientation, and source type on users’ uncertainty around their choices.

---

<sup>14</sup>pre-registration: <https://aspredicted.org/blind.php?x=9rn6i7>

### 5.3.2 Dependent & independent variables:

We considered four total dependent variables (DV): (1) content bias (bounded value between  $[0,1]$ ), (2) uncertainty range around content bias (bounded value between  $[0,1]$ ), (3) source credibility choice (bounded value between  $[0,1]$ ), (4) uncertainty around source credibility choice (bounded value between  $[0,1]$ ). For our independent variables (IV), we included the image emotion condition (angry, happy, or mixed), image shown (true or false), as well as political orientation (right or left) and source type (mainstream or misinformation). For models build with bias choice / uncertainty as the dependent variable, since the model is build on tweet level responses, there are only two image emotion conditions of happy or angry.

### 5.3.3 Model specification:

For each model, we included users' unique id and the source name as random effects. After comparing multiple model specifications using AIC, we also included interaction terms between sources' political orientation and image emotion. The reference conditions for credibility choice and uncertainty models are image emotion = mixed, source orientation = left, source type = mainstream, and image shown = False. The reference conditions for bias choice and uncertainty models are image emotion = happy, source orientation = left, source type = mainstream, and image shown = False.

### 5.3.4 participants

In study 1, we recruited a total of 81 (52 female, 28 male, and one other) university students with an average age of 21 years old. Per our pre-registration, we excluded responses from 9 participants who showed missing responses (due to unexpected technical difficulties), resulting in 72 accepted responses. 30 Participants were randomly assigned to the angry condition, 24 participants were assigned to the happy condition, and the remaining 18 were assigned to the mixed condition. Participants took an average of 26 minutes to complete the study. All participants either received either course extra credits or required research credits as incentives.

### 5.3.5 Results

#### **Beliefs and uncertainty of content bias:**

We used two mixed-effects beta regressions to study effects on users' beliefs about the bias of individual tweets (See Fig 16-left). Users found content from right-leaning accounts to be more biased ( $\beta = 0.487$  [0.31, 0.65]  $z = 5.63, p < .001$ ). Similarly, users judged misinformation sources to be more biased ( $\beta = 0.461$  [0.32, 0.59]  $z = 6.69, p < .001$ ). We also found that users rated tweets containing angry images as more biased ( $\beta = 0.540$  [0.38, 0.69]  $z = 6.991, p < .001$ ). This is in line with our hypothesis that angry imagery will lead to an increase in users' perceived bias. Interestingly, this effect was reduced for the interaction between right-leaning accounts and angry images ( $\beta = -0.568$  [-0.75, -0.37]  $z = -5.809, p < .001$ ). This interaction effect could be due to different topical focus and usage strategies of angry and happy imagery between right-leaning and left-leaning accounts. Surprisingly, even though the content

shown to users was solely sorted based on facial emotions in images, we did not find a significant effect of whether images were shown to users or not. One explanation could be that sources, in general, might choose images to match the overall tone of the text. Another explanation could be that the effect of images might be case-specific and users' judgements might be impacted by specific topics or tweets they are sensitive or knowledgeable about.

In our mixed-effects model on users' uncertainty around their beliefs (See Fig 16-right), we found that compared to left-leaning mainstream accounts, users were more likely to have larger uncertainty ranges for right-leaning accounts ( $\beta = 0.214 [0.09, 0.33]$   $z = 3.499$   $p < .001$ ). We also found a similar positive effect for misinformation accounts ( $\beta = 0.125 [0.039, 0.21]$ ,  $z = 2.853$ ,  $p < .005$ ). Furthermore, users were more likely to have higher uncertainty ranges when viewing tweets with angry images ( $\beta = 0.171 [0.033, 0.30]$ ,  $z = 2.430$ ,  $p < .05$ ). Finally, we observe that users are more likely to have smaller uncertainty ranges when evaluating angry tweets in right leaning accounts ( $\beta = -0.233 [-0.38 - 0.07]$   $z = -2.963$ ,  $p < .005$ ).

### **Beliefs and uncertainty of source credibility:**

We used mixed-effects beta regression to investigate the effects of study conditions on users' beliefs about source credibility. We found that in reference to left-leaning accounts, users are more likely to rate right-leaning accounts as less credible ( $\beta = -0.509 [-0.87, -0.14]$ ,  $z = -2.741$ ,  $p < .01$ ). Users were also more likely to rate misinformation sources as less credible ( $\beta = -0.515 [-0.69 - 0.33]$ ,  $z = -5.586$ ,  $p < .001$ ). We also see a similar negative effect where users are more likely to rate left-leaning sources as less credible when tweets contain images with angry facial expressions as

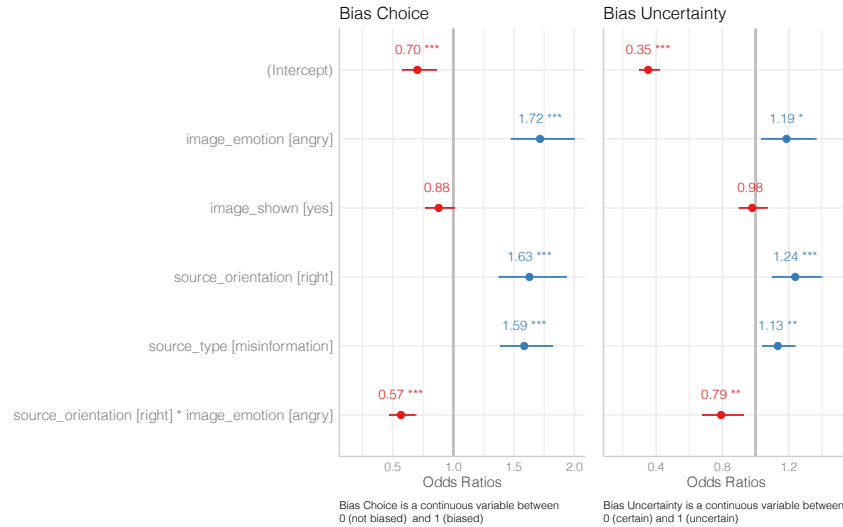


Figure 15: Study 1 fixed effects odds ratios for bias choice (left) and bias uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: \*\*\* 99.9%, \*\* 99%, \* 95%. For image\_emotion, the reference category is happy. For image\_shown, the reference category is no, left is the reference condition for source\_orientation, and mainstream is the reference condition for source\_type.

compared to the mixed condition ( $\beta = -0.440 [-0.78, -0.09], z = -2.494, p < .05$ ).

This finding is in line with our hypothesis that angry imagery will lead to a decrease in users' perceived credibility of sources. However, users' credibility rating is more likely to increase when considering right-leaning accounts in the angry image condition ( $\beta = 0.745 [0.281, 1.20], z = -2.494, p < .01$ ). We did not observe a significant effect of the happy image condition on users' perceived source credibility.

We also used a mixed-effects beta regression to study whether users' confidence around their decisions (represented through uncertainty ranges in the Line + Range technique). Interestingly, source type was the only factor that significantly affected users' uncertainty on source credibility. Users were more likely to be more confident in their decisions when source type was categorized as misinformation ( $\beta =$



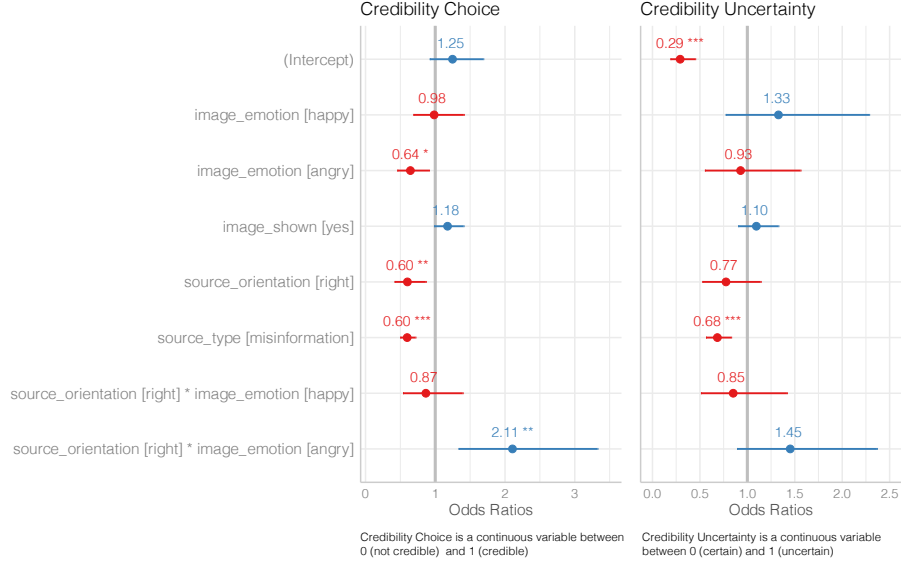


Figure 16: Study 1 fixed effects coefficients for credibility choice (left) and credibility uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: \*\*\* 99.9%, \*\* 99%, \* 95%. For image\_emotion, the reference category is mixed. For image\_shown, the reference category is no, left is the reference condition for source\_orientation, and mainstream is the reference condition for source\_type.

$-0.379 [-0.57, -0.18], z = -3.867, p < .001)$

### 5.3.6 Analysis of users' comments:

For study 1, each user had the option to answer one open-ended question for each account asking “please describe how the tweets (text and images) influenced your decisions about this account?.” Even though leaving comments was an optional part of the study, we received comments from all 72 participants and the majority of participants left comments for all 8 trials. In total, we collected 572 comments about users' decision-making influences. Since thematic analysis with a large number of comments is challenging, we used topic-modeling to facilitate the qualitative analysis of the comments and arrive at different themes of influences on users' judgments.

Table 7: Study 1 topic model of users’ comments. Shows a table of 20 extracted topics sorted based on number of unique users with comments assigned to each topic. [41].

Topic ID	Count users	Count Comments	Top Topic Terms	Theme
2	37	53	tweets , based , unbiased , tweets were biased , political	Facts vs. opinions
11	29	73	account , left , right , leaning , credible	Source attitude & political orientation
19	27	49	wing , right wing , right , left wing , clearly right	Source attitude & political orientation
1	27	45	political , orientation , political orientation , tweets were political , credibility	Source attitude & political orientation
6	24	35	source , credible source , credible , overall , opinionated with tweets	Facts vs. opinions
4	24	34	facts , stating facts , stating , titles , article titles	Facts vs. opinions
7	18	30	bias , credible , report , bias and direct , tell	Source attitude & political orientation
3	17	24	opinions , presented , facts , opinions presented , accounts	Facts vs. opinions
10	16	26	like a news , news , news source , like , source	Source attitude & political orientation
0	15	30	biased , tweets were biased , appear , tell , lean	Source attitude & political orientation
9	14	25	images , tell , provided , meme , use of images	Images and Pictures
15	12	17	opinionated , appear , articles appear , articles , opinionated with tweets	Facts vs. opinions
8	11	19	factual , factual tweets , report , pretty factual , gave were biased	Facts vs. opinions
5	10	19	pictures , tell , words , pictures influenced , think the pictures	Images and Pictures
17	6	8	clickbait language , language , zoomed in pictures , headshots , held	Language usage and tone
13	5	10	stance , political stance , nt , tweets didnt , issues	Source attitude & political orientation
16	4	9	associations , influence , helped , influence my decisions , decisions	Language usage and tone
18	4	6	variety , variety of tweets , wide variety , wide , misspelling	Language usage and tone
12	3	8	non biased , non , biased tweets , non biased tweets , pertaining to politics	Source attitude & political orientation
14	3	7	subjective , subjective language , language , tweet , highly subjective	Source attitude & political orientation

We categorized the extracted topics into four general themes: 1) source attitude or political orientation, 2) opinionated versus factual reporting, 3) specific language usage or tone, and 4) the effect of images. In this section, we will summarize each of these themes and offer a few example comments provided by participants.

**Attitude or political orientation:** The majority of comments in study 1 included mentions of general perceptions of source attitude such as political bias in sources or lack thereof. Topic 11 with a size of 79 documents from 29 unique users included such comments. For example, one user wrote : *“This account was talking about the right ruining everything so it has to be more left-leaning and it didn’t sound complete out there so I don’t it was 100% not credible but it definitely isn’t credible because it is biased.”* Another comment mentioned a similar comment but for right-leaning sources: *“These tweets made it clear that anyone on the left were out of their minds, and anyone on the right was perfectly fine, so it’s clear this is a right-leaning account. As for credibility, [...] I don’t know, some part of that sort of tone doesn’t seem too credible to me.”*

There were also more mentions of more specific source attitudes. Topic 19 with 49 comments from 27 users included such comments. For example, one comment took a repeated focus on Israel related topics as the basis for their judgement: *“lots about zionism and israel, usually doesn’t get talked about in right wing media.”*. Another user described harsh attitudes towards president Trump as their rationale: *“The Tweets were left-wing orientated with its expression on frustration and opposition to the Trump administration. The rhetoric was a bit harsh and seemed to be attacking Trump administration/right-wing ideology so it was bias and less credible.”*

**Opinionated versus factual reporting:** Another prevailing theme in users’ comments was a contrast between perceptions of opinionated vs factual reporting. Topics 2 (53 comments from 37 users), 6 (35 comments from 24 users), and 4 (34 comments from 24 users) included several comments related to this theme. For example, several comments mentioned how sources contained opinionated language: *“These tweets seem rather opinionated”* and *“Not a credible source, seems very opinionated with tweets.”*

On the other hand, sources that were deemed to report ‘facts’ were considered to be more credible as described by one of our participants: *“I found this to be a credible source because the tweets seemed factual.”* Another user mentioned the factual tone and a center-leaning political orientation, as the basis of their decision: *“This one seemed the most credible to me. The political orientation seemed to be pretty even and the article titles seemed to be based on facts.”*

**Specific language usage or tone:** There were a group of comments that highlighted how users sometimes take specific word usage, negative and positive sentiment,

and general tones as the basis for their decisions. These comments related mostly to writing style, word choice, and strong language as the basis for users' decisions. Topics 16 (9 comments from 4 users), 17 (8 comments from 3 users), and 14 (7 comments from 3 users) included such comments. Some comments mentioned specific words or phrases as cues for their decisions: *"Alarmist language "out of control homicides" "what you need to do to be safe" And, "They are using language such as "Looney Left" and the unflattering close-ups of democratic representatives"*. Another user noticed using all caps writing style as a rationale for their decision: *"Alarmist language, clear/strong opinions/writing in all caps"*. Finally, A group of comments also mentioned "clickbait" language as the basis for their decisions: *"...As for credibility, the way the tweets were written seemed "cheap" to me, like they were clickbait and just meant to draw anyone in based on shock value. I don't know, some part of that sort of tone doesn't seem too credible to me."*

**Effect of images:** Topics 9 (25 comments from 14 users) and 5 (19 comments from 10 users) included comments that mentioned images as an influence for their decision. These comments a wide range of observations from unflattering imagery, to facial-expressions and images being not serious. For example, a user mentioned how the images influenced their decision in an opposite direction of the text: *"Some of the tweets seemed to be about data rather than opinions. Some of the pictures seemed to take away merit."* Another comment explicitly mentioned facial expression of individuals helping them decide that the tweets have a left bias: *"Many of the tweets seemed to be credible as most were quotes by others. Some of the pictures had facial expressions that made the tweets seem left swinging."*

A number of comments cited comic or not serious imagery as the basis for their decisions. For example: *“Difficult to take the meme-like images seriously.”* And, *“The images were cartoonish and difficult to take it seriously. It was obviously making fun of trump.”*

A few comments mentioned unprofessional or unflattering images as the basis for their decisions: *“This account seemed biased towards Palestinians and Israelis. There were compliments of them and somewhat unflattering pictures of those who either commented against them or who may do (or not do) something against them.”* And, *“Some of the pictures were not professional and showed the president and alliances in a negative light.”*

These comments show that at least for a group of users, images sometimes serve as primary evidence for a sources’ lack of credibility. Although, majority of comments about pictures and images, did not specifically mention emotions in images as the basis for their decisions. This might be due to the fact that tweets in our dataset contained a diverse set of topics. Maybe, users look for a more systematic negative treatment of specific topics or individuals, rather than a combination of negative imagery from a wide ranging topics.

### 5.3.7 Discussion:

The main motivation behind this study was to assess how user judgements about bias in content and credibility of the source are affected by (the accumulation of) emotions in social media images. Users were randomly assigned to three groups of angry, happy, or mixed emotions. Each group saw content from 8 sources sorted based

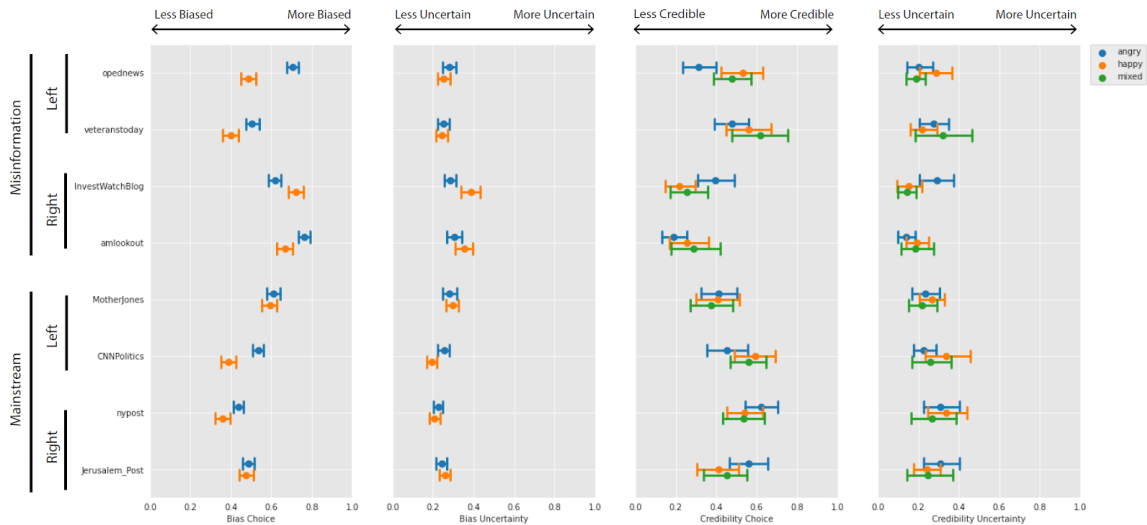


Figure 17: Mean and bootstrapped 95% confidence interval of users' responses for each account.

on the specified emotion. We hypothesized that the accumulation of tweets with angry images would increase the perceived content bias and reduce the perceived source credibility. We also hypothesized that happy images would lead to a reverse effect of a decrease in perceived bias and an increase in the perceived credibility of sources in comparison to a condition where happy and angry content is shown interchangeably. We found partial evidence for our hypotheses: we observed an increase in perceived bias and a decrease of perceived credibility of users in the angry image emotion condition. However, we did not find evidence of a reverse effect of happy emotion in comparison to a mix of content.

We also observed a noticeable reliance on information found from texts. The effect of angry emotion was somewhat reduced for right-leaning accounts for both content bias and source credibility (See figure 17). One explanation for this reduced effect could lie within the difference in tone, language, and topical focus of the sources between tweets with angry and happy imagery from these sources. Taking nypost as

an example, we can see from the content that the tweets that came with happy images mostly focused on celebrities. For example, one of the first tweets users reads in the happy condition has a picture of a smiling non-famous man and reads *“Teen sues over paramedic allegedly fondling her breasts after seizure”*. For nypost, the tweets in the angry condition are more politically charged. For example, one of the first tweets in this condition is about a republican politician, Roy Moore with a frowning picture of him looking down. The tweet reads: *“Trump [is] concerned about Roy Moore allegations: White House aide”*. Users’ comments help with reinforcing this observation. One comment about nypost from the happy condition mentioned *“It seems to be a celebrity tabloid account. I honestly think it is a parody of a tabloid account.”* while a comment from a user assigned to an angry condition for nypost paints a much different picture of the source: *“Headlines feel relatively neutral though maybe slightly conservative but no titles seem overly exaggerated or extreme.”*. This observation helps us hypothesize that the topic and the text content of news are likely to be of primary importance in affecting users’ judgements about source credibility and content bias.

Furthermore, we made another observation from Study 1. Although the content in study 1 was solely sorted based on emotions in images and we did observe the angry condition to significantly impact users’ judgements, we did not observe an effect of the presence of images on the outcomes (See Figure 16). The qualitative analysis of users’ comments might help us explain this observation. We can see that three of the four major themes were primarily related to the text content of the tweets. The themes highlight a series of heuristics utilized by users such as tone, attitude towards political

parties, and unflattering images. From these comments, we can see that users might heavily rely on cues from the tweet texts to judge bias in the content and credibility of sources. Moreover, it is possible that the emotional content in images is chosen by sources to match the textual content and that might result in not observing an effect. Furthermore, the text and topical focus from these conditions are different for every tweet. Our qualitative analysis hints that users might be sensitive to specific cues from specific tweets and that might overweight the effect of accumulation of emotions in images.

Finally, we observed that users' uncertainty around their judgement credibility was noticeably reduced for misinformation news sources. In other words, users were on average more confident in their decisions when rating the credibility of misinformation sources. One way to interpret this result is by considering the differences between misinformation and mainstream sources. Previous studies have shown that misinformation sources take use of more angry text, and are considered by users to be more opinionated [97, 191]. Given this difference between mainstream and misinformation sources, we can assert that content from misinformation sources might contain more cues in text and images for users to make more certain decisions. Of course, this needs to be further investigated in future studies.

## 5.4 Study 2

Motivated by the results of study 1, we developed study 2's experimental design to measure the impact of emotions in images on users' judgements, as well evidence that sources systematically portray different politicians with different emotional facial



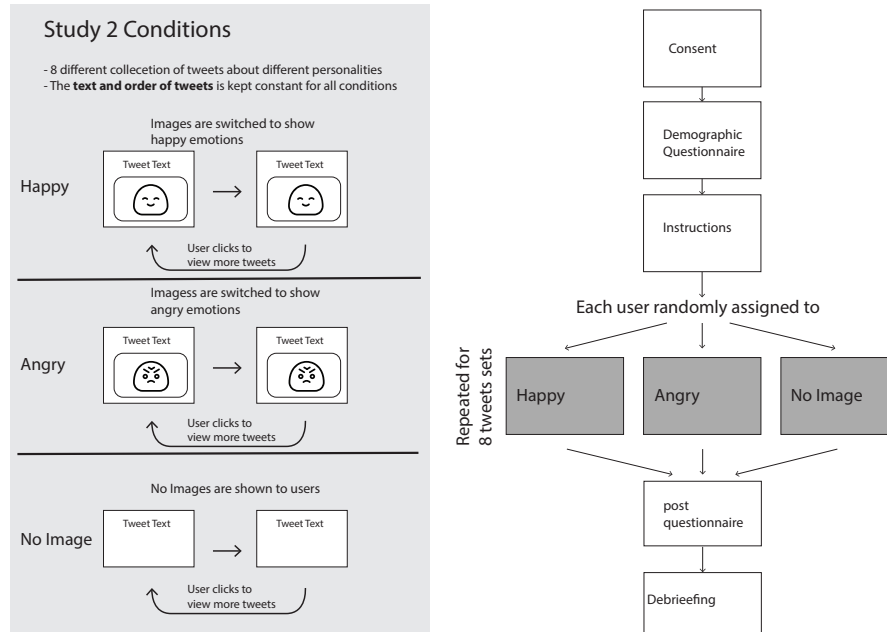


Figure 18: Study 2 conditions and process

expressions [137]. In study 2, we kept the text content shown to users same for all conditions and controlled images as either happy, angry, or no images. This requires each set of tweets to focus on a specific person and by switching angry and happy images of that person, we could measure the effect of systematic usage of negative (angry) or positive (happy) facial expressions on users judgements. Additionally, this experimental design allowed us to investigate the interactions between users' prior attitudes towards different politicians with the emotional content of images on their judgements about bias and credibility (see Figure 18).

#### 5.4.1 Experiment design:

We curated a dataset containing tweets on eight different politicians including Donald Trump, Hillary Clinton, Angela Merkel, and Emanuel Macron (see table 8). Users were instructed that all tweets mentioning each politician are from a unique

source. To control for perspective variance, we start by collecting the text or this study from mainstream sources. In order to limit the impact of text content on users' decisions, we downselected tweets with the following steps: we first conducted sentiment analysis on the tweet texts using Vader Sentiment [63]. Next, for each set of tweets, we selected tweets with the highest neutral sentiment scores. Finally, we manually evaluated the tweets and removed tweets with inaccurate scores from the sentiment analysis library. This resulted in tweets about 8 different politicians, from mainstream news sources, that were mostly of neutral tone. The images for this study are manipulated in a between-subjects manner in which users saw either happy, angry, or no images. For example, a user in the Happy condition, views tweets mentioning Hillary Clinton and are accompanied by happy images of her, while a user in the Angry condition evaluated the same tweets but with angry images of Hillary Clinton, and the control (no-image) condition will view no images.

Other than changes in study design, the procedures of the study were equivalent to Study 1 with one exception. For each set of tweets mentioning one of the eight politicians, users first answered two questions in a pop-up form about their familiarity and favorability of that person in a 5-level Likert scale.

#### 5.4.2 Hypotheses:

Since study 2 is a continuation of study 1, we expect to see a similar effect of angry emotions on bias and credibility scores. Moreover, we also hypothesize that users' favorability toward each politician interacts with the effects of image emotions such that if users are exposed to news focusing on specific figure, their perception of

Table 8: Politicians selected for study 2

Politician	Notes
Donald Trump	Former President of the US
Hillary Clinton	Former US Secretary of State
Barack Obama	Former president of the US
Theresa May	Former prime minister of the UK
Emanuel Macron	President of France
Angela Merkel	Chancellor of Germany
Kim Jong Un	Supreme Leader of North Korea
Vladimir Putin	President of Russia

bias and credibility is affected by both their prior favorability of that person and the emotion shown to the person<sup>15</sup>. More specifically, we hypothesize that favorability negatively interacts with angry emotion and positively interacts with happy emotion to predict users perceived bias and credibility. Furthermore, We will also investigate the impact of users' familiarity with each politician on their judgements.

#### 5.4.3 Dependent & independent variables:

The Dependent variables in study 2 are identical to study 1. We considered four total dependent variables(DV): (1) the tweet bias choice (bounded value between [0,1]), (2) uncertainty range around tweet bias (bounded value between [0,1]), (3) source credibility choice (bounded value between [0,1]), (4) uncertainty around source credibility choice (bounded value between [0,1]). For our independent variables (IV), we included the image emotion condition (angry, happy, or no image), as well as users' prior favorability and familiarity towards each politician in the form of 5 step Likert scales.

---

<sup>15</sup>pre-registration: <https://aspredicted.org/blind.php?x=9js6d5>

#### 5.4.4 Study 2 model specification:

For each model, we included users' unique id and the politician's name as random effects. After comparing multiple model specifications using AIC, we also included interaction terms between users' favorability and familiarity of each politician with the image emotion. The omitted reference conditions are image emotion = no image.

#### 5.4.5 Participants:

In study 2, we recruited a total of 126 (63 Female, 62 male, and one preferred not to say) participants. The average age of participants was 35 years old. 81 participants were recruited from Amazon Mechanical Turk and received a 2 dollar incentive. The rest of the participants were university students who received either research or extra credits for their participation. Per our pre-registration, we excluded responses from 12 participants who showed missing responses (due to unexpected technical difficulties) resulting in 114 accepted responses. 34 Participants were randomly assigned to the angry condition, 42 participants were assigned to the happy condition, and the remaining 38 were assigned to the no image condition. Participants took an average of 21 minutes to complete the study.

#### 5.4.6 Results

##### 5.4.6.1 Belief and uncertainty of tweet bias:

We used two mixed-effects beta regressions to study the effects of experimental conditions on users' beliefs and uncertainty of bias of individual tweets. We found that users viewing tweets with angry images rated tweets as more biased in comparison

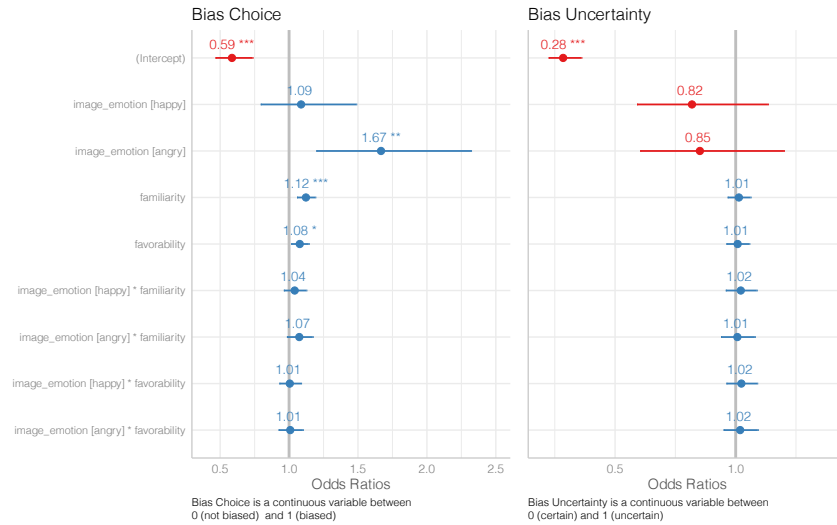


Figure 19: Study 2 fixed effects Odds Ratios for bias choice (left) and bias uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: \*\*\* 99.9%, \*\* 99%, \* 95%. For image\_emotion, the reference category is no image.

to when no images are shown to users ( $\beta = 0.511 [0.18, 0.84], z = 3.031, p < .01$ ). This is in line with our hypothesis that when the content of messages is the same, angry imagery leads to an increase in users' perceived bias. Similar to study 1, we did not observe an effect in the happy condition. Moreover, we did not find any evidence for an interaction between favorability/familiarity and image emotion. We did however, observe an overall positive effect of familiarity on users' perceived bias ( $\beta = 0.115 [0.05, 0.17], z = 3.82, p < .001$ ). We also observe a small positive effect of favorability on users bias choice ( $\beta = 0.074 [0.010, 0.13], z = 2.471, p < .05$ ). We did not observe a noticeable effect of any of the independent variables on users' uncertainty around their choices (See Fig 20).

**Belief and uncertainty of source credibility:** We used mixed-effects beta regression to investigate the effects of study conditions on users' beliefs about source

credibility. We found that users rated sources as less credible when tweets are accompanied with angry facial expressions ( $\beta = -0.397 [-0.73, -0.05], z = -2.308, p < .05$ ). This finding is in line with our hypothesis that angry facial expressions will lead to a decrease in the perceived credibility of sources. We also find that familiarity also decreases users' perceived credibility of sources ( $\beta = -0.153 [-0.245, -0.06], z = -3.256, p < .01$ ). Again, here we did not observe an interaction effect between favorability and the image emotion.

For our mixed effect model of users' uncertainty around their decisions, we did not observe a significant effect of image emotion on users' uncertainty around their decisions. However, we did observe that in cases that users' familiarity with subjects are higher, users are more likely to be more certain in their decisions ( $\beta = -0.178 [-0.29 - 0.06], z = -3.00, p < .01$ ). We also observe a positive effect on the interaction between happy emotion and users' familiarity ( $\beta = 0.191 [0.0350.34], z = 2.401, p < .05$ ).

#### 5.4.7 Qualitative analysis of users' comments:

In study 2, we collected a total of 881 comments from 116 users. Similar to study 1, we analyzed users' descriptions of the rationale behind their decisions using NMF topic modeling [42] and thematic analysis of the documents most representative of each topic. This helped us categorize the 20 extracted topics into 5 higher-order themes. Three of the themes were similar to the ones we extracted from study 1, while two are themes that are unique to this study. Since the goal of this qualitative analysis is to get an overview of how users conduct their decisions, we will mostly focus on the strategies and comments that are new and unique to this study (figure

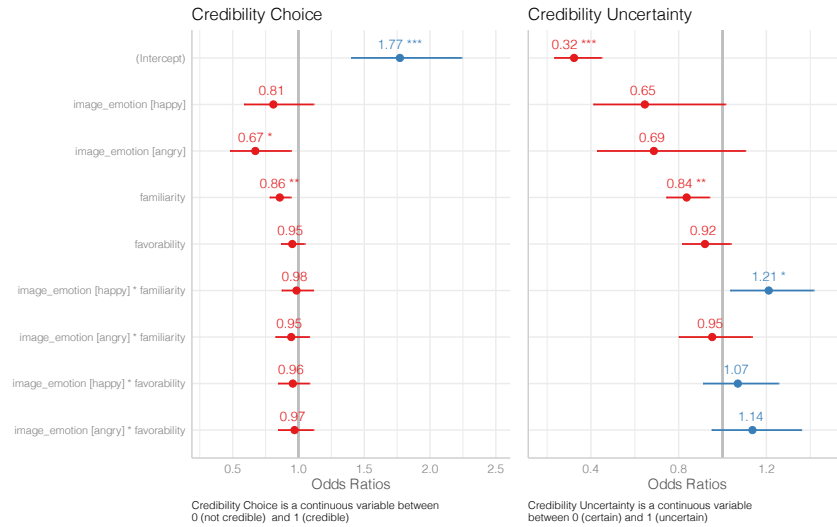


Figure 20: Study 2 fixed effects Odds Ratios for credibility choice (left) and credibility uncertainty (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: \*\*\* 99.9%, \*\* 99%, \* 95%. For image\_emotion, the reference category is no image.

9 provides an overview of all the extracted themes and 20 topics).

**Negative vs. positive attitude towards politicians:** This theme is mostly represented in Topic 3 (with 46 comments from 33 users) with descriptions about each set of tweets' attitudes towards the individual politicians. Many of the comments were about how a source is mostly covering negative or positive news about a politician. Often these comments contained a mention of both texts and images.

For example, a user mentioned how the source was mostly neutral but had some negative tweets and images about Barack Obama: *“There were a couple that seemed to be negative about [Obama] and images were a little negative. Not all were biased but some credibility was lost.”* Another user provided a similar comment about Donald Trump. *“There were a few tweets that had no opinion but the ones that did had some negative connotations toward Trump.”*

Table 9: Study 2 topic model of users’ comments. Top shows a table of 20 extracted topics sorted based on number of unique users who had comments associated with each of the topics. [41].

Topic ID	Count users	Count Comments	Top Topic Terms	Theme
10	51	92	tweets, factual, opinion, information, little	Facts vs. opinions
0	40	67	left, leaning, left leaning, right, news	Source attitude & political orientation
8	35	51	like, feel, feel like, sound, sound like	Source attitude & political orientation
19	34	71	facial expressions, expressions, facial, angry, angry facial	Images and Pictures
3	33	46	negative, positive, negative light, ones, light	Negative vs. positive attitude towards personalities
16	32	66	images, given, negative bias, bad, text	Images and Pictures
4	31	46	bias, given, news, way, statements	Source attitude & political orientation
6	31	50	political, orientation, source, political orientation, credibility	Source attitude & political orientation
9	31	47	neutral, pretty neutral, fairly neutral, pretty, fairly	Language usage and tone
2	27	38	biased, wording, read, rest, tweets	Source attitude & political orientation
5	27	46	facts, stating, stating facts, reporting facts, reporting	Facts vs. opinions
17	26	45	account, flattering, things, think, felt	Source attitude & political orientation
14	24	47	pictures, unflattering, headlines, unflattering pictures, stories	Images and Pictures
12	22	39	credible, unbiased and credible, tweet, source, tweets seemed credible	Source attitude & political orientation
15	22	28	know, know this person, person, nt know, nt	User was not sure
13	19	28	based, fact, fact based, opinion, matter	Facts vs. opinions
18	19	27	unbiased, tweets were unbiased, unbiased and credible, neutral tone, explain	Source attitude & political orientation
11	13	21	straight, reporting, straight up reporting, factual reporting, point	Source attitude & political orientation
1	12	18	sure, tweet, explain, little bit	User was not sure
7	3	8	tweets were worded, worded, pleasant or unpleasant, depended, unpleasant	Language usage and tone

Some of the comments actually noticed images as negative and text as more neutral towards the person. For example, one of the participants provided this comment about Donald Trump: *“The images paint [Donald Trump] in a negative light but the text wasn’t actually negative.”* Some users also mentioned that they did not find any negative attitudes towards a politician. For example, one of the users found the tweets about Kim-Jong Un to be not negative: *“There was nothing too negative and everything seemed legit. Images were not bad towards him.”*

**Neutral source attitude:** In study 1, we saw that there were many comments about sources being biased or having negative tones. In this study, we observed an emerging theme about the source being more neutral towards a person. Topic 9 (with 47 comments from 31 users) included many comments related to this theme. For example, one of the participants found tweets about Donald Trump as mostly neutral: *“The general tone of the tweets were neutral, informational with little or no opinion.”* Another user found tweets about Barack Obama as mostly neutral with a slight liberal bias: *“Most of the tweets seem pretty neutral, although in some seem to*



*be more liberal.”*

Once again, users in some cases found a mismatch between text and images in terms of neutrality. This comment is a users’ perspective about the Vladimir Putin tweet set: *“Text was neutral. The images were a mix of neutral and deliberating unflattering.”*

**Effect of Images:** In this study, we observe an increase in image related comments. We identified three topics that contain descriptions from users related to visual information 9). Topic 19 with 71 comments from 34 users and Topic 14 with 47 comments from 24 users include mentions of facial expressions, angry emotions, and unflattering portrayals. For example, a user found the text of tweets as mostly unbiased, and explained how portrayed facial expressions of Hillary Clinton was the basis for her judgement: *“While the text didn’t involve much biased words, the usage of certain images of Hillary Clinton depicting her facial expressions in an array of negative emotions showed a biased view, in which the account may have wanted viewers to take that negative emotion they may perceive through her image and unconsciously use it to influence their perceptions of Hillary Clinton herself.”*

Another user provided more similar details about how the images influenced their decisions about tweets related to Emanuel Macron: *“The usage of images shows him in a negative light, with angry and frowning facial expressions. There were also phrases used like “pulling no punches” that suggested some bias.”*

Some users also found a combination of specific language usage with “weirdly close up” and “unflattering” images leading them to believe a source is less credible: *“The texts were mostly bland, except for “...what do you think”, which is a tabloid-like*

*phrase for me. Photos were weirdly close-up facial views that were generally unflattering, which makes me wonder a bit about credibility...”*

Another interesting set of comments were about users perceiving the images as not correlating with the tweets: *“Most of the tweets were very normal, but there were a couple that had angry Macron pictures that did not correlate with the headlines presented.”* And *“The majority of tweets were unbiased with their headlines. Some of the pictures might have been a bit questionable.”*

A group of comments included a specific description of how images did not influence their decision. Topic 16 with 66 comments from 32 users, includes many such comments. For example, for the collection about Kim-Jung Un, a user mentioned how images were neither negative nor positive: *“The images were not the best images nor were they the worst images of him, but there was still a negative bias.”* Another user mentioned how images and tweets related to Donald Trump were both neutral, and honest looking: *“The tweets seemed to be honest and state the honest news about what is happening while the images associated with them.”*

#### 5.4.8 Study 2 discussion:

We asked participants to rate “sources” that each focus on a specific politician. The text was constant for all users, while we manipulated the conditions to include either angry or happy facial expressions of those politicians, or to include no images. We found that in comparison to no images being present, users in the angry condition found the content to be more biased and the sources less credible. However, we did not find a significant effect of happy images on users’ judgement. The difference between

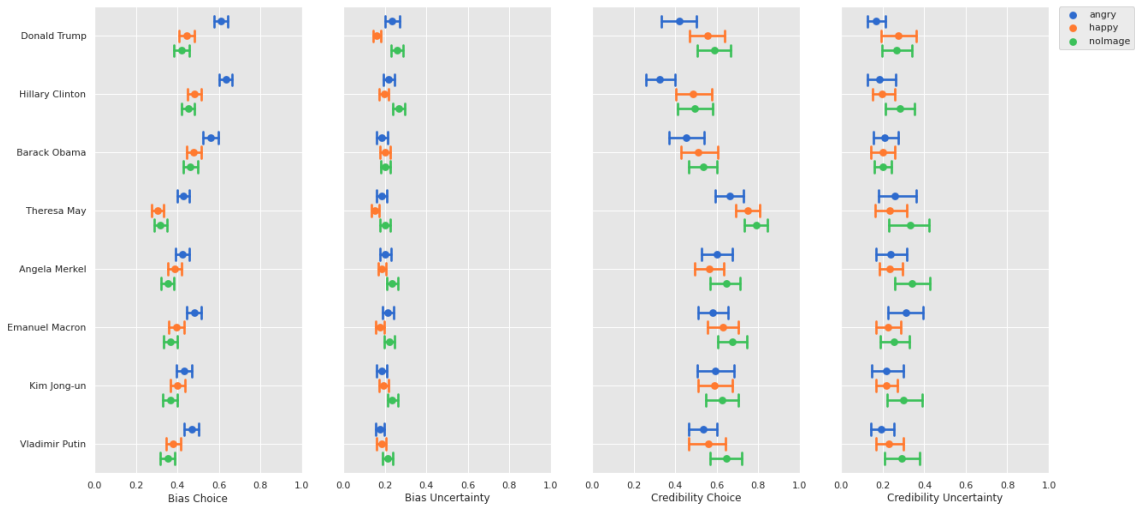


Figure 21: Mean and 95% bootstrapped confidence interval of choices for each condition/politician.

the happy and angry condition in our study could be better explained by a study on perceptions of negative or positive portrayals of politicians, in which Lubinger and Brantner found that participants mostly agreed on what constituted as negative, but perceptions of positive portrayals were wide varying [110]. This suggests that negative portrayals are likely more commonly agreed upon and thus might have a stronger effect on users' judgements.

Users' comments also included many mentions of angry, negative, or unflattering portrayals for these politicians. Comparing the results from studies 1 and 2, one might ask, why did we not observe a clear difference of happy images on users' judgments? First, it is worth reiterating that our quantitative and qualitative findings suggest that users' might put a stronger weight on the textual content of tweets when judging bias and credibility of choices. Second, the accumulation of angry emotions in tweets might signal a more systematic bias towards specific politicians and thus causing a stronger influence on users' judgements.

We observed that familiarity also impacts users' judgements of bias and credibility, as well as how certain they are in their decisions. Recent work on judgements on misinformation suggests that prior exposure and familiarity with misinformation increases the perceived accuracy of content [139]. We can assume that users' would have to rely on their memory to assess the accuracy of news headlines or articles. On the other hand, our work suggests that bias and credibility of sources might be a different dimension from the accuracy of content. Our comment analysis highlighted that users often rely on more analytical approaches and different kinds of heuristics such as negativity, word usage, or emotions in facial expressions to judge bias and credibility of sources. An explanation for the effect of familiarity on users' judgement about bias and credibility could be that when users are more familiar with politician, they could be more sensitive to the details of texts and images of content they view and therefore possibly less trusting of the source covering that person.

Finally, we observed a small overall effect of favorability on users' perception of tweet bias, and we did not observe an interaction effect between favorability and emotions in images. Assuming that users are engaged in deliberate reasoning by detecting cues that point towards sources being biased and not credible, these results do not provide evidence that users are engaged in motivated reasoning based on their favorability towards politicians. This result might yet be another evidence in the line of work pointing that motivated reasoning is not a primary factor in users' interaction with misinformation [143]. Moreover, We suspect that this interaction effect might be stronger if users are more invested in the topics covered in the study. In the future, we plan to repeat this study with identities and events that are more polarizing.

## 5.5 Discussion

Across two consecutive preregistered studies with a total of 207 participants, we find evidence that angry facial expressions in images accompanying social media news posts lead to an increase in users' perception of bias in content. We also found that users rate sources that show a systematic angry portrayal of different politicians as less credible. These findings provide evidence on the impact of emotional facial expressions on users' perceptions of source credibility and content bias. These results help paint a more detailed landscape of how users' trust in news sources are shaped by visual information such as images or videos.

Study 1, showed that angry facial emotions increases users' bias rating of content and decreases credibility rating of sources. We also found that political orientation of sources reduced the effect of angry emotions. Our qualitative analysis highlighted a wide range of heuristics that relate to both text and visual information employed by users. Users' perception of how opinionated a source is, choices of unusual or highly negative words, and "unflattering" portrayal of individuals in their images are among these heuristics. The combination of our qualitative studies and experimental results showed the impact of negative emotions on users' judgements, but also highlighted the complex interaction between topics covered by sources through a combination of text and images that was not explicitly considered in the study 1 design. In the design of study 2, we aimed to get a clearer picture of the impact of angry facial emotions by limiting users' choices to tweets mentioning specific politicians. In other words, through study 2, we investigated the systematic negative or positive visual

treatment of politicians on users' perceptions of credibility and bias. Our results show strong evidence towards part of our hypothesis, that a systematic negative treatment of politicians would lead to a decrease in users' rating of source credibility and content bias. However, we did not find evidence for our hypotheses around the interactions between users' favorability and emotional facial expressions.

Although these results provide clear evidence towards the impact of emotional facial expressions in images on users' judgements, there are many more aspects that remain to be studied. First, within each of angry or happy emotional categories, there are several finer levels of facial emotions ranging from extremely angry/happy to subtle frown/grin that might impact users' judgements. There are also other emotion dimensions such as sadness, surprise, fear, or disgust that might potentially impact users' judgements within this context. Finally, facial expressions rarely contain one unique emotion and subtle changes might communicate different meanings to individuals. We believe a natural next step for this study is to control for the amount and type of emotion in facial expressions by using Generative Adversarial Neural networks to produce image datasets with finer control on the facial expressions in images.

We also acknowledge some general limitations in the design and execution of our studies. First, the tweets used for both studies are approximately four years old and do not reflect current political events. Users might be more invested and impacted by current political events and make different decisions in light of more relevant news. Furthermore, the selected tweets for study 2 were selected from multiple mainstream sources instead of one source. A more coherent dataset from one source might yield clearer results. Finally, in order to limit the impact of users' preconceived notions

of sources, we masked all account names from our studies. Even though we believe that a scenario in which users encounter new and unknown sources is realistic and especially important in the context of misinformation, in many cases users might have a self-selected set of sources that they trust and refer to on social media. An important future step for our research is to investigate how users update their trust in sources they already know based on new content with different positive and negative emotional images. Such a study is significant in that it can open new ways of reducing trust in misinformation sources by identifying and highlighting content with highly emotional text and images.

Another important factor in understanding how news content impacts users' attitudes towards sources is their uncertainty around their decisions [181]. Through both studies, using a new elicitation technique, we asked users to provide their uncertainty ranges around their decisions (See methodology section) and explored the impacts of our experimental conditions on users' uncertainty ranges. For source credibility judgements, we found that some conditions significantly reduced users' uncertainty. In study one, only *source\_type = misinformation* showed significant reductions in users' uncertainty. One possible interpretation for this effect could be the inherent differences between misinformation and mainstream sources where misinformation sources use more extreme and suspicious language and images [189, 191, 97]. In study 2, we observed that familiarity reduced users' uncertainties. It is possible that in light of more familiarity with persons, users might be more likely to detect invalid / suspicious content and therefore make decisions with more confidence.

Results from our models on users' uncertainty around their bias choices were less

clear and interpretable. One possible reason could be that users' have much less information to inform their choices about bias for a single tweet, or that there are many different cues that influence their judgements. It is also important to note that several factors might impact users' uncertainty such as lack of knowledge, lack of clarity, lack of familiarity, or lack of correctness [146]. It is important to empirically clarify the meaning of our graphical elicitation uncertainty ranges for different judgements. Future work on uncertainty elicitation needs to address these subtleties in order to make such results more informative and useful.

Although this research was mostly motivated by the prevalence and impact of misinformation on our democracies and societies, we believe that our current globally politicized and extremely segregated political ecosystem calls for a more critical and holistic view on the whole media landscape. Although it is important to understand and mitigate users' trust in sources of misinformation, it is of equal importance to understand why individuals might elect not to trust more mainstream and generally trustworthy sources of information. Implicit, negative, and biased visual and verbal propositions of different politicians by mainstream sources might contribute to this lack of trust and lead to a question that remains mostly under-explored: How verbal and visual strategies by mainstream media might contribute to highly politically polarized societies, the likes of which we witnessed during the 2020 United States Presidential election?



## CHAPTER 6: CONCLUSION AND FUTURE DIRECTIONS

Combating misinformation using computation can be very effective in how we mitigate its dire effects on our society. We believe our efforts can be more effective if we move from solely detecting and flagging misinformation to developing user-centered systems based on a holistic understanding of how individuals and groups make decisions about misinformation. In the future, we plan to continue this work through two main overarching tracks: 1) identifying cognitive processing tendencies of users and creating models that can predict reactions to information (including misinformation) and 2) embedding these models into new systems, visualizations, and interaction techniques to encourage more rational reactions to data. Several specific research areas will contribute to these two future goals:

- uncertainty elicitation/visualization techniques

To effectively understand user decision-making and lead users towards more rational decisions, it is essential to take their individual beliefs into account and accurately and effectively elicit their prior beliefs and uncertainties [80, 94]. As a continuation of recent work on interactive techniques for eliciting and measuring the extent to which users update their beliefs using interactive visualizations [94, 100], we plan to develop new interactive techniques to measure how users update their beliefs when exposed to new information. This general technique

can be applied in various contexts. We plan to use this technique to understand how users update their beliefs about the credibility of misinformation sources.

- Understanding users' cognitive tendencies about misinformation:

Using the interactive technique developed in G1, I plan to study how different modes of information (text, image, social network, Images) affect users' beliefs and decisions about the trustworthiness and credibility of sources. In our series of controlled experiments on consumers' cognitive processes, we showed how *cognitive biases* and *uncertainty* played important roles in users' decisions about misinformation using visual analytic [97, 161, 40, 205, 92]. In addition to continuing our research on cognitive biases, Inspired by recent work on how users prior beliefs about a source affects their acceptance of news content [33] and other work on Bayesian cognition models in interactive visualizations [101], we plan to develop Bayesian hierarchical models to describe how users update their beliefs about sources in light of new information from different sources of news.

- Developing visual Analytics for understanding visual and verbal behavior of news and misinformation sources:

Research on misinformation suggests the importance of highlighting news sources' behavior instead of tagging individual news stories [108]. In collaboration with our research team at UNC-Charlotte, we have developed two visual analytic systems that use state of the art Machine Learning and Social Network analysis methods to highlight how misinformation sources differ from those of more trust-

worthy ones. Verifi features visualizations of models built to differentiate news sources based on linguistic, affective, and social network features [97]. Verifi2 builds upon Verifi by enabling users to explore semantic and visual similarities of content [92]. We plan to extend Verifi and develop a visual analytic system that 1) visualizes both text and image features most predictive in affecting users decisions and beliefs, 2) responds to users' prior belief and uncertainties about different news sources, and 3) helps users make more rational decisions about trustworthiness of sources.

Today, misinformation from various sources and with different intents is affecting our public health, democracies, and societies. Our work on visual Analytics can bring robust human-centered computational models to people who are making decisions about misinformation daily. Furthermore, future research on interaction techniques and models can contribute to research and practice in other disciplines such as Urban Planning [90, 93], social sciences [95, 204], and designing effective visualizations [96]. We sincerely hope that this work and our future endeavors can contribute to healing our highly polarized situations often fueled by misinformation and disinformation.

## REFERENCES

- [1] — facebook media and publisher help center. <https://www.facebook.com/help/publisher/182222309230722>. (Accessed on 01/22/2019).
- [2] Emergent. <http://www.emergent.info/>. (Accessed on 02/07/2019).
- [3] Fact- checking - duke reporters' lab. <https://reporterslab.org/fact-checking/>. (Accessed on 01/22/2019).
- [4] *Fact-checking U.S. politics*.
- [5] Fact-checking u.s. politics — politifact. <https://www.politifact.com/>. (Accessed on 01/22/2019).
- [6] Factcheck.org - a project of the annenberg public policy center. <https://www.factcheck.org/>. (Accessed on 01/22/2019).
- [7] Fiskkit.com discuss news that matters and find out what's true. <http://fiskkit.com/>. (Accessed on 01/22/2019).
- [8] Google, facebook and twitter agree to fight fake news in the eu - bloomberg. <https://www.bloomberg.com/news/articles/2018-09-25/google-facebook-and-twitter-agree-to-fight-fake-news-in-eu>. (Accessed on 01/22/2019).
- [9] Latest email hoaxes - current internet scams - hoax-slayer. <https://hoax-slayer.com/>. (Accessed on 01/22/2019).
- [10] *NC man told police he went to D.C. pizzeria with gun to investigate conspiracy theory*. The Washington Post. Accessed: 2018-03-31.
- [11] *Snopes.com*.
- [12] *Social Media Use in 2018*.
- [13] *TensorFlow*.
- [14] This is how facebook's news feed fact-checking will work in the uk — wired uk. <https://www.wired.co.uk/article/full-fact-facebook-fact-checking>. (Accessed on 01/22/2019).
- [15] Twitter trails: Tool for monitoring the propagation of rumors. <http://twittertrails.com/>. (Accessed on 02/07/2019).
- [16] Why the macron hacking attack landed with a thud in france - the new york times. <https://www.nytimes.com/2017/05/08/world/europe/macron-hacking-attack-france.html>. (Accessed on 12/02/2018).

- [17] L. Abraham and O. Appiah. Framing news stories: The role of visual imagery in priming racial stereotypes. *The Howard Journal of Communications*, 17(3):183–203, 2006.
- [18] W. C. Adams. Whose lives count? tv coverage of natural disasters. *Journal of Communication*, 36(2):113–122, 1986.
- [19] M. Allan. Information literacy and confirmation bias: You can lead a person to information, but can you make him think? 2017.
- [20] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.
- [21] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi. Sok: The evolution of sybil defense via social networks. In *Security and Privacy (SP), 2013 IEEE Symposium on*, pages 382–396. IEEE, 2013.
- [22] D. Arendt, E. Grace, R. Wesslen, S. Volkova, and W. Dou. Speed-accuracy tradeoffs for visual interactive labeling using a representation-free classifier. Submitted to VAST 2018, 2018.
- [23] A. Arsenault and M. Castells. Conquering the minds, conquering iraq: The social production of misinformation in the united states—a case study. *Information, Communication & Society*, 9(3):284–307, 2006.
- [24] V. Bakir and A. McStay. Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, pages 1–22, 2017.
- [25] V. Bakir and A. McStay. Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, 6(2):154–175, 2018.
- [26] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- [27] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.
- [28] G. Baym and J. P. Jones. *News parody and political satire across the globe*. Routledge, 2013.
- [29] Y. Benkler, R. Faris, and H. Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 2018.
- [30] Y. Benkler, R. Faris, H. Roberts, and E. Zuckerman. Study: Breitbart-led right-wing media ecosystem altered broader media agenda. *Columbia Journalism Review*, 3, 2017.

- [31] J. Berger and K. L. Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.
- [32] A. Bessi and E. Ferrara. Social bots distort the 2016 us presidential election online discussion. 2016.
- [33] R. Blom. Believing false political headlines and discrediting truthful political headlines: The interaction between news source trust and news content expectancy. *Journalism*, page 1464884918765316, 2018.
- [34] P. Bourgonje, J. M. Schneider, and G. Rehm. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, 2017.
- [35] N. D. Bowman and E. Cohen. 17 mental shortcuts, emotion, and social rewards: The challenges of detecting and resisting fake news. *Fake News: Understanding Media and Misinformation in the Digital Age*, page 223, 2020.
- [36] M. Carlson. The reality of a fake image news norms, photojournalistic craft, and brian walski’s fabricated photograph. *Journalism Practice*, 3(2):125–139, 2009.
- [37] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.
- [38] X. Chen, S.-C. J. Sin, Y.-L. Theng, and C. S. Lee. Why students share misinformation on social media: Motivation, gender, and study-level differences. *The Journal of Academic Librarianship*, 41(5):583–592, 2015.
- [39] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The anchoring effect in decision-making with visual analytics. In *Visual Analytics Science and Technology (VAST), 2017 IEEE Conference on*, Oct 2017.
- [40] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The anchoring effect in decision-making with visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.
- [41] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces*, pages 74–77, 2012.
- [42] A. Cichocki and A.-H. Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.
- [43] R. Coleman. Framing the pictures in our heads. *Doing news framing analysis: Empirical and theoretical perspectives*, pages 233–261, 2010.

- [44] A. Dang, A. Moh'd, E. Milios, and R. Minghim. What is in a rumour: Combined visual analysis of rumour flow and user activity. In *Proceedings of the 33rd Computer Graphics International*, pages 17–20. ACM, 2016.
- [45] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.
- [46] A. Day and E. Thompson. Live from new york, it's the fake news! saturday night live and the (non) politics of parody. *Popular Communication*, 10(1-2):170–182, 2012.
- [47] N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.
- [48] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmman, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM, 2014.
- [49] D. Eckles and E. Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*, 2017.
- [50] R. M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58, 1993.
- [51] R. M. Entman. *Projections of power: Framing news, public opinion, and US foreign policy*. University of Chicago Press, 2004.
- [52] R. M. Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173, 2007.
- [53] R. M. Entman. Media framing biases and political power: Explaining slant in news of campaign 2008. *Journalism*, 11(4):389–408, 2010.
- [54] J. S. B. Evans. *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates, Inc, 1989.
- [55] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.
- [56] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

- [57] Á. Figueira and L. Oliveira. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825, 2017.
- [58] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.
- [59] M. Flinthisham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran. Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 376. ACM, 2018.
- [60] D. Flynn, B. Nyhan, and J. Reifler. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150, 2017.
- [61] S. Frenkel. For russian ‘trolls,’ instagram’s pictures can spread wider than words. *The New York Times*, Dec 2017.
- [62] F. Gabbert, A. Memon, K. Allan, and D. B. Wright. Say it to my face: Examining the effects of socially encountered misinformation. *Legal and Criminological Psychology*, 9(2):215–227, 2004.
- [63] C. Gilbert and E. Hutto. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) [http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf](http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf), volume 81, page 82, 2014.
- [64] M. E. Grabe and E. P. Bucy. *Image bite politics: News and the visual framing of elections*. Oxford University Press, 2009.
- [65] J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.
- [66] L. Graves. Factsheet: Understanding the promise and limits of automated fact-checking - reuters institute digital news report. <http://www.digitalnewsreport.org/publications/2018/factsheet-understanding-promise-limits-automated-fact-checking/>. (Accessed on 01/24/2019).
- [67] L. Graves. Understanding the promise and limits of automated fact-checking. *Factsheet*, 2:2018–02, 2018.
- [68] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.



- [69] A. Gupta, H. Lamba, and P. Kumaraguru. prayforboston: Analyzing fake content on twitter. In *eCrime Researchers Summit (eCRS), 2013*, pages 1–12. IEEE, 2013.
- [70] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.
- [71] J. Haidt and J. Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116, 2007.
- [72] N. Hassan, F. Arslan, C. Li, and M. Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812. ACM, 2017.
- [73] N. Hassan, A. Sultana, Y. Wu, G. Zhang, C. Li, J. Yang, and C. Yu. Data in, fact out: automated monitoring of facts by factwatcher. *Proceedings of the VLDB Endowment*, 7(13):1557–1560, 2014.
- [74] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.
- [75] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [76] E. S. Herman and N. Chomsky. *Manufacturing consent: The political economy of the mass media*. Random House, 2010.
- [77] C. I. Hovland, I. L. Janis, and H. H. Kelley. Communication and persuasion. 1953.
- [78] L. Howell et al. Digital wildfires in a hyperconnected world. *WEF Report*, 3:15–94, 2013.
- [79] Y. L. Huang, K. Starbird, M. Orand, S. A. Stanek, and H. T. Pedersen. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 969–980. ACM, 2015.
- [80] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, 2018.

- [81] T. Hwang, I. Pearce, and M. Nanis. Socialbots: Voices from the fronts. *interactions*, 19(2):38–45, 2012.
- [82] C. Jack. Lexicon of lies: Terms for problematic information. *Data & Society*, 3, 2017.
- [83] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, pages 2972–2978, 2016.
- [84] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608, 2017.
- [85] O. P. John, E. M. Donahue, and R. L. Kentle. The big five inventory-versions 4a and 54, 1991.
- [86] E. Jonas, S. Schulz-Hardt, D. Frey, and N. Thelen. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology*, 80(4):557, 2001.
- [87] L. Jones and R. Nevell. Plagued by doubt and viral misinformation: the need for evidence-based use of historical disease images. *The Lancet Infectious Diseases*, 16(10):e235–e240, 2016.
- [88] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
- [89] A. Karduni. Human-misinformation interaction: Understanding the interdisciplinary approach needed to computationally combat false information. *arXiv preprint arXiv:1903.07136*, 2019.
- [90] A. Karduni, I. Cho, G. Wessel, W. Ribarsky, E. Sauda, and W. Dou. Urban space explorer: A visual analytics system for urban planning. *IEEE computer graphics and applications*, 37(5):50–60, 2017.
- [91] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. Arendt, S. Shaikh, and W. Dou. Vulnerable to misinformation? verifi! Submitted to VAST 2018, 2018.
- [92] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. L. Arendt, S. Shaikh, and W. Dou. Vulnerable to misinformation? verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 312–323, 2019.
- [93] A. Karduni, A. Kermanshah, and S. Derrible. A protocol to convert spatial polyline data to network formats and applications to world urban road networks. *Scientific data*, 3:160046, 2016.

- [94] A. Karduni, D. Markant, R. Wesslen, and W. Dou. A bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2020.
- [95] A. Karduni and E. Sauda. Anatomy of a protest: Spatial information, social media, and urban space. *Social Media+ Society*, 6(1):2056305119897320, 2020.
- [96] A. Karduni, R. Wesslen, I. Cho, and W. Dou. Du bois wrapped bar chart: Visualizing categorical data with disproportionate values. 2020.
- [97] A. Karduni, R. Wesslen, S. Santhanam, I. Cho, S. Volkova, D. Arendt, S. Shaikh, and W. Dou. Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. In *International Conference on Web and Social Media (ICWSM)*, 2018.
- [98] M. Kasra, C. Shen, and J. F. O’Brien. Seeing is believing: How people fail to identify fake images on the web. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page LBW516. ACM, 2018.
- [99] I. Khaldarova and M. Pantti. Fake news: The narrative battle over the ukrainian conflict. *Journalism Practice*, 10(7):891–901, 2016.
- [100] Y.-S. Kim, K. Reinecke, and J. Hullman. Data through others’ eyes: The impact of visualizing others’ expectations on visualization interpretation. *IEEE transactions on visualization and computer graphics*, 24(1):760–769, 2017.
- [101] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman. A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.
- [102] D. Klein and J. Wueller. Fake news: A legal perspective. 2017.
- [103] S. Knobloch, M. Hastall, D. Zillmann, and C. Callison. Imagery effects on the selective reading of internet newsmagazines. *Communication Research*, 30(1):3–29, 2003.
- [104] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM, 2004.
- [105] A. Kott, D. S. Alberts, and C. Wang. War of 2050: a battle for information, communications, and computer security. *arXiv preprint arXiv:1512.00360*, 2015.
- [106] G. Lakoff. *Moral politics: How liberals and conservatives think*. University of Chicago Press, 2010.

- [107] 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.
- [108] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
- [109] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.
- [110] K. Lobinger and C. Brantner. Likable, funny or ridiculous? a q-sort study on audience perceptions of visual portrayals of politicians. *Visual Communication*, 14(1):15–40, 2015.
- [111] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751–1754. ACM, 2015.
- [112] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual analytics science and technology (VAST), 2011 IEEE conference on*, pages 181–190. IEEE, 2011.
- [113] A. M. MacEachren, A. C. Robinson, A. Jaiswal, S. Pezanowski, A. Savelyev, J. Blanford, and P. Mitra. Geo-twitter analytics: Applications in crisis management. In *25th International Cartographic Conference*, pages 3–8, 2011.
- [114] A. Magdy and N. Wanas. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 103–110. ACM, 2010.
- [115] L. Mallonee. How photos fuel the spread of fake news, Jun 2017.
- [116] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM, 2011.
- [117] J. Marcus. *Mesoamerican writing systems: Propaganda, myth, and history in four ancient civilizations*. Princeton University Press Princeton, 1992.
- [118] C. Martel, G. Pennycook, and D. G. Rand. Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, 5(1):1–20, 2020.

- [119] L. Masch and O. W. Gabriel. How emotional displays of political leaders shape citizen attitudes: The case of german chancellor angela merkel. *German Politics*, 29(2):158–179, 2020.
- [120] D. P. McCabe and A. D. Castel. Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, 107(1):343–352, 2008.
- [121] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.
- [122] N. Mele, D. Lazer, M. Baum, N. Grinberg, L. Friedland, K. Joseph, W. Hobbs, and C. Mattsson. Combating fake news: An agenda for research and action, 2017.
- [123] P. Messaris and L. Abraham. The role of images in framing news stories. *Framing public life: Perspectives on media and our understanding of the social world*, pages 215–226, 2001.
- [124] P. T. Metaxas, S. Finn, and E. Mustafaraj. Using twittertrails. com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 69–72. ACM, 2015.
- [125] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [126] D. Miller. *Tell me lies: Propaganda and media distortion in the attack on Iraq*. Pluto Press, 2003.
- [127] S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 353–362. ACM, 2015.
- [128] C. R. Mynatt, M. E. Doherty, and R. D. Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The quarterly journal of experimental psychology*, 29(1):85–95, 1977.
- [129] V. Narwal, M. H. Salih, J. A. Lopez, A. Ortega, J. O’Donovan, T. Höllerer, and S. Savage. Automated assistants to identify and prompt action on visual news bias. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2796–2801. ACM, 2017.
- [130] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

- [131] N. Newman, R. Fletcher, A. Kalogeropoulos, D. A. Levy, and R. K. Nielsen. Reuters institute digital news report 2017. 2017.
- [132] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.
- [133] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.
- [134] S. Oraby, L. Reed, R. Compton, E. Riloff, M. Walker, and S. Whittaker. And that’s a fact: Distinguishing factual and emotional argumentation in online dialogue. *arXiv preprint arXiv:1709.05295*, 2017.
- [135] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.
- [136] E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.
- [137] Y. Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018.
- [138] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.
- [139] G. Pennycook, T. Cannon, and D. G. Rand. Prior exposure increases perceived accuracy of fake news. 2018.
- [140] G. Pennycook, T. D. Cannon, and D. G. Rand. Implausibility and illusory truth: Prior exposure increases perceived accuracy of fake news but has no effect on entirely implausible statements. *Available at SSRN*, 2017.
- [141] G. Pennycook and D. G. Rand. Cognitive reflection and the 2016 us presidential election. *Forthcoming in Personality and Social Psychology Bulletin*, 2018.
- [142] G. Pennycook and D. G. Rand. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. 2018.
- [143] G. Pennycook and D. G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.
- [144] D. M. Perlmutter. *Deep and surface structure constraints in syntax*. PhD thesis, Massachusetts Institute of Technology, 1968.

- [145] J. W. Peters. Wielding claims of ‘fake news,’ conservatives take aim at mainstream media - the new york times. [https://www.nytimes.com/2016/12/25/us/politics/fake-news-claims-conservatives-mainstream-media-.html?\\_r=0](https://www.nytimes.com/2016/12/25/us/politics/fake-news-claims-conservatives-mainstream-media-.html?_r=0). (Accessed on 12/01/2018).
- [146] J. V. Petrocelli, Z. L. Tormala, and D. D. Rucker. Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of personality and social psychology*, 92(1):30, 2007.
- [147] S. Pluviano, C. Watt, and S. Della Sala. Misinformation lingers in memory: Failure of three pro-vaccination strategies. *PLOS ONE*, 12(7):1–15, 07 2017.
- [148] S. Porter, S. Bellhouse, A. McDougall, L. Ten Brinke, and K. Wilson. A prospective investigation of the vulnerability of memory for positive and negative emotional scenes to the misinformation effect. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 42(1):55, 2010.
- [149] 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.
- [150] P. Pourghomi, F. Safieddine, W. Masri, and M. Dordevic. How to stop spread of misinformation on social media: Facebook plans vs. right-click authenticate approach. In *Engineering & MIS (ICEMIS), 2017 International Conference on*, pages 1–8. IEEE, 2017.
- [151] J. Rajsic, D. E. Wilson, and J. Pratt. Confirmation bias in visual search. *Journal of experimental psychology: human perception and performance*, 41(5):1353–1364, 2015.
- [152] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. *ICWSM*, 11:297–304, 2011.
- [153] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659, 2013.
- [154] J. Reis, F. Benevenuto, P. O. de Melo, R. Prates, H. Kwak, and J. An. Breaking the news: First impressions matter on online news. *arXiv preprint arXiv:1503.07921*, 2015.
- [155] P. Resnick, S. Carton, S. Park, Y. Shen, and N. Zeffer. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, 2014.
- [156] B. Richards. *Emotional governance: Politics, media and terror*. Springer, 2007.

- [157] A. Robb. Pizzagate: Anatomy of a fake news scandal – rolling stone. <https://www.rollingstone.com/politics/politics-news/anatomy-of-a-fake-news-scandal-125877/>. (Accessed on 12/02/2018).
- [158] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.
- [159] V. L. Rubin, Y. Chen, and N. J. Conroy. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 83. American Society for Information Science, 2015.
- [160] S. Sanovich. Computational propaganda in russia: The origins of digital misinformation. *Working Paper*, 2017.
- [161] S. Santhanam, A. Karduni, and S. Shaikh. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [162] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [163] N. Schweitzer, D. A. Baker, and E. F. Risko. Fooled by the brain: Re-examining the influence of neuroimages. *Cognition*, 129(3):501–511, 2013.
- [164] S. I. Serengil and A. Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020.
- [165] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.
- [166] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, pages 96–104, 2017.
- [167] Z. Shi, A.-L. Wang, L. F. Emery, K. M. Sheerin, and D. Romer. The importance of relevant emotional arousal in the efficacy of pictorial health warnings for cigarettes. *Nicotine & Tobacco Research*, 19(6):750–755, 2017.
- [168] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.



- [169] S. Sloman and P. Fernbach. *The knowledge illusion: Why we never think alone*. Penguin, 2018.
- [170] J. Soll. The long and brutal history of fake news. *POLITICO Magazine*, 2017.
- [171] D. Spohr. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150–160, 2017.
- [172] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, pages 230–239, 2017.
- [173] M. Sullivan. Why the term ‘fake news’ should be retired in 2018 - the washington post. [https://www.washingtonpost.com/lifestyle/style/its-time-to-retire-the-tainted-term-fake-news/2017/01/06/a5a7516c-d375-11e6-945a-76f69a399dd5\\_story.html?utm\\_term=.902df60602f3](https://www.washingtonpost.com/lifestyle/style/its-time-to-retire-the-tainted-term-fake-news/2017/01/06/a5a7516c-d375-11e6-945a-76f69a399dd5_story.html?utm_term=.902df60602f3). (Accessed on 12/01/2018).
- [174] C. R. Sunstein. *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press Princeton, NJ, 2001.
- [175] B. Swire, U. K. Ecker, and S. Lewandowsky. The role of familiarity in correcting inaccurate information. *Journal of experimental psychology: learning, memory, and cognition*, 43(12):1948, 2017.
- [176] D. Tambini. Fake news: public policy responses. 2017.
- [177] D. Tamibini. How advertising fuels fake news. *Media Policy Blog*, 2017.
- [178] E. C. Tandoc Jr, Z. W. Lim, and R. Ling. Defining “fake news“ a typology of scholarly definitions. *Digital Journalism*, pages 1–17, 2017.
- [179] E. C. Tandoc Jr, Z. W. Lim, and R. Ling. Defining “fake news” a typology of scholarly definitions. *Digital Journalism*, 6(2):137–153, 2018.
- [180] P. M. Taylor. *Munitions of the mind: A history of propaganda from the ancient world to the present era*. 2013.
- [181] Z. L. Tormala and R. E. Petty. Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, 14(4):427–442, 2004.
- [182] A. Tsang and K. Larson. The echo chamber: Strategic voting and homophily in social networks. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 368–375. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [183] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

- [184] C. Vaccari and A. Chadwick. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1):2056305120903408, 2020.
- [185] C. J. Vargo, L. Guo, and M. A. Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *new media & society*, page 1461444817712086, 2017.
- [186] C. J. Vargo, L. Guo, and M. A. Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *new media & society*, 20(5):2028–2049, 2018.
- [187] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An analysis of social network-based sybil defenses. *ACM SIGCOMM Computer Communication Review*, 41(4):363–374, 2011.
- [188] M. Vlasceanu, J. Goebel, and A. Coman. The emotion-induced belief amplification effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2020.
- [189] S. Volkova, E. Ayton, D. L. Arendt, Z. Huang, and B. Hutchinson. Explaining multimodal deceptive news prediction models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 659–662, 2019.
- [190] S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma. Inferring latent user properties from texts published in social media. In *AAAI*, pages 4296–4297, 2015.
- [191] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653, 2017.
- [192] A. Vrij, S. Mann, S. Kristen, and R. P. Fisher. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, 31(5):499–518, 2007.
- [193] M. Wagner, F. Fischer, R. Luh, A. Haberson, A. Rind, D. A. Keim, W. Aigner, R. Borgo, F. Ganovelli, and I. Viola. A survey of visualization systems for malware analysis. In *EG Conference on Visualization (EuroVis)-STARs*, pages 105–125, 2015.
- [194] E. Wall, L. Blaha, C. L. Paul, K. Cook, and A. Endert. Four perspectives on human bias in visual analytics. In *DECISIVE: Workshop on Dealing with Cognitive Biases in Visualizations*, 2017.

- [195] L. E. Wallace, D. T. Wegener, and R. E. Petty. When sources honestly provide their biased opinion: Bias as a distinct source perception with independent effects on credibility and persuasion. *Personality and Social Psychology Bulletin*, 46(3):439–453, 2020.
- [196] H. Wallach. Computational social science  $\neq$  computer science + social data. *Commun. ACM*, 61(3):42–44, Feb. 2018.
- [197] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*, 2012.
- [198] W. Y. Wang. ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.
- [199] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, D. A. Keim, K. E. Isaacs, A. Giménez, I. Jusufi, T. Gamblin, et al. State-of-the-art report of visual analysis for event detection in text data streams. In *Computer Graphics Forum*, volume 33, 2014.
- [200] C. Wardle. Fake news. it’s complicated. <https://firstdraftnews.org/fake-news-complicated/>. (Accessed on 12/01/2018).
- [201] C. Wardle and H. Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking. *Council of Europe report, DGI (2017)*, 9, 2017.
- [202] P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140, 1960.
- [203] J. Waterson, B. Esposito, and V. Sanusi. Here is all the fake news about the manchester terror attack. <https://www.buzzfeed.com/jimwaterson/manchester-arena-fake-news>. (Accessed on 12/02/2018).
- [204] G. Wessel, A. Karduni, and E. Sauda. The image of the digital city: revisiting lynch’s principles of urban legibility. *Journal of the American Planning Association*, 84(3-4):280–283, 2018.
- [205] R. Wesslen, S. Santhanam, A. Karduni, I. Cho, S. Shaikh, and W. Dou. Investigating effects of visual anchors on decision-making about misinformation. In *Computer Graphics Forum*, volume 38, pages 161–171. Wiley Online Library, 2019.
- [206] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

- [207] S. Wineburg, S. McGrew, J. Breakstone, and T. Ortega. Evaluating information: The cornerstone of civic online reasoning. In *Stanford Digital Repository*. Available at: <http://purl.stanford.edu/fv751yt5934>, 2016.
- [208] L. Wu, J. Li, X. Hu, and H. Liu. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 99–107. SIAM, 2017.
- [209] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 637–645. ACM, 2018.
- [210] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM, 2012.
- [211] D. Zillmann, S. Knobloch, and H.-s. Yu. Effects of photographs on the selective reading of news reports. *Media Psychology*, 3(4):301–324, 2001.
- [212] A. Zubiaga, H. Ji, and K. Knight. Curating and contextualizing twitter stories to assist with social newsgathering. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 213–224. ACM, 2013.
- [213] E. Zuckerman. Stop saying “fake news”. it’s not helping. — ... my heart’s in accra. <http://www.ethanzuckerman.com/blog/2017/01/30/stop-saying-fake-news-its-not-helping/>. (Accessed on 12/01/2018).
- [214] F. Zuiderveen Borgesius, D. Trilling, J. Moeller, B. Bodó, C. H. de Vreese, and N. Helberger. Should we worry about filter bubbles? 2016.