2022-03-31

Examining the Perceptions of Fake News, Verification, and Notices on Twitter

Brendan Patrick Gwynn
Brigham Young University

Examining the Perceptions of Fake News, Verification, and Notices on Twitter

Brendan Patrick Gwynn

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Christopher Wilson, Chair
Jessica Zurcher
Scott Church

School of Communications

Brigham Young University

ABSTRACT

Examining the Perceptions of Fake News, Verification, and Notices on Twitter

Brendan Patrick Gwynn
School of Communications, BYU
Master of Arts

The rise of social media platforms has had a significant impact on the conventional model of gatekeeping. With increased access to information—as well as the ability to contribute to the public discourse—individuals no longer need to rely on the mass media for news. These realities have led to increased conversations surrounding credibility in the digital age. Although not a new concept, fake news has become increasingly common in recent years. The web—particularly social media outlets, like Twitter—have enhanced the spread of misinformation. To combat this, social media platforms have introduced gatekeeping features like verification marks and warning labels. However, questions remain regarding the credibility and effectiveness of these features. Furthermore, little information exists regarding the perceptions of these features. For this study, the researcher examined the perceptions of fake news, verification, and Notices (i.e., warning labels) as they relate to Twitter. These perceptions were captured through a survey that was distributed to Twitter users through MTurk. Results were examined generally as well as in the light of political orientation, ranging from very liberal to very conservative on a 4-point scale. Within the scope and limitations of this study, results indicate that the majority of Twitter users believe that fake news on the platform is a major problem. Additionally, results show that there is no significant difference between the effectiveness of verification and the effectiveness of Notices in slowing the spread of fake news, and neither feature is perceived as strongly credible or effective.

Keywords: gatekeeping, fake news, Twitter, verification mark, media credibility, online news, online information, internet as a news source, censorship, warning labels, social media

ACKNOWLEDGMENTS

When I made the decision to enroll in graduate school, I had very little idea of what was on the horizon. I must confess that the only reason I returned to school was because I felt like it was God's plan for me—and it took me a little while to accept His path and to stop dragging my feet. But as I look back on the last two years, I have an immense amount of gratitude in my heart for my experience and have been reminded that God can always make more out of us than we could ever make out of ourselves.

In addition to my appreciation to God, I would like to thank everyone who has put their confidence in me throughout this process, including my professors, parents, and peers. I would especially like to thank my thesis committee—Dr. Chris Wilson, Dr. Jessica Zurcher, and Dr. Scott Church—who shaped my experience not only as a graduate student at BYU but also as an undergraduate student. Each of you have had such a profound influence on me during my time at this university, and I feel indebted to you for your mentorship and friendship.

To Dr. Wilson, thank you for the countless hours you spent reading my papers, offering feedback, and answering my questions. Your willingness to help me as a student speaks volumes about who you are as a person. To Dr. Zurcher, thank you for your passion about teaching and all things related to communications. Your excitement was contagious and provided encouragement when life was particularly demanding. To Dr. Church, thank you for sharing your insights on all of my work and for expanding my knowledge in ways that nobody ever has. Your teaching has deepened my ability and desire to think critically.

To my parents, thank you for supporting me in all of my academic endeavors over the years. I appreciate your unconditional love and your growing interest in my work as a student. I know communications is not exactly your area of expertise, but I am glad that I have been able to

share my own passion with both of you. I hope that you have felt my gratitude over the years and also hope that I have been able to make you proud of what I have accomplished and who I have become in the process.

To my peers, thank you for befriending me and for taking on projects together as we have made our way through the program. I will always cherish the memories I have made over the last two years, including 1) the Marvel and video game studies completed with Abbie and Brad, 2) the various presentations prepared and given with Caleb, Megs, Izzy, and Ellice, 3) the paper about *WandaVision* written with Caleb and Caleb, and 4) the 1:00am texts with Ellice about communications theory our first semester.

I have been blessed with the best cohort I could have asked for. I must give a shoutout to Alycia for saving me when I had SPSS issues the night before our quantitative project was due. I also must give a shoutout to Caleb P. for answering all of my questions when trying to complete the thesis. I honestly could not have made it without the help of so many of my fellow students and friends, and I am so grateful that we were able to share this experience together.

TABLE OF CONTENTS

LIST OF TABLES

Examining the Perceptions of Fake News, Verification, and Notices on Twitter

Introduction

The launch of the internet near the turn of the century redefined not only how we consume information, but also how we interact with each other (Gil, 2018). Perhaps the most immediate contribution of this new global networking system was the increased access people had to a wealth of information (Singer, 2001). In previous years, individuals had to rely on the media to obtain their news (Messner & Garrison, 2009). But thanks to the internet, users can now seek information from innumerable sources, any time of the day, at the mere click of a mouse (Choi et al., 2006; Heinecke, 2019; Messner & Garrison, 2009). In addition, the smartphone has enabled individuals to access information virtually anywhere (Napoli & Obar, 2014).

In its mere quarter century of existence, the internet has evolved to include streaming services, blog sites, online forums, and social media platforms—all of which lend an immediate voice to just about anyone who wants one (Blank, 2013). This period of increased user-generated content is known as Web 2.0 (Allen, 2012). Cormode and Krishnamurthy (2008) explain that the main difference between Web 1.0 and Web 2.0 is the ease of content creation—a reality that can largely be attributed to the resources, aids, and platforms that have been introduced to maximize the potential to create content, including text, audio, and video. As a result, the internet is not simply a universal source for obtaining information; it is also a universal place for sharing information (Benham, 2020; Cormode & Krishnamurthy, 2008).

These realities have led to increased conversations surrounding credibility in the digital age. Although not a new concept, fake news has become increasingly common in recent years (Rubin, 2019; Sauer, 2017). The web—particularly social media outlets, like Twitter—has enhanced and quickened the spread of misinformation (Rubin, 2019), while the simplicity of creating content has also contributed to inaccuracies published and shared on the web—both intentionally and unintentionally.

To combat this, social media platforms have introduced features such as verification marks (Cohen & Sutton, 2018) and warning labels (Gadde & Beykpour, 2020). However, questions remain regarding the use of these features (Clayton et al., 2020; Pennycook et al., 2020), including how effective they are in slowing the spread of fake news, how credible they are, and how opinions about their use differ based on political orientation. Furthermore, limited information exists regarding the perceptions of these features.

Because of this, it is worth exploring in greater depth how the ongoing development of news creation and distribution has impacted the consumption of online news, particularly on social media networks. This is especially important with the increased frequency of fake news—a term that is often used synonymously with disinformation and misinformation (Kumar, 2020; Rubin et al., 2015), and which is defined by Allcott and Gentzcow (2017) as “news articles that are intentionally and verifiably false and could mislead readers” (p. 213).

The purpose of this particular study is to analyze the perceptions regarding fake news, verification, and Notices (i.e., warning labels) on Twitter. This research will contribute to the existing literature by offering insights into perceptions regarding efforts to combat the spread of misinformation online. It will also explore Twitter as a platform—an area where the research is less prevalent than other popular social media platforms. The overarching goal of this study is to treat Twitter as a type of case study to further shed light on the implementation and impact of verification marks and warning labels on the web, including if these features are perceived as effective and credible.

For this study, the researcher will capture perceptions by distributing a survey to Twitter users that asks questions related to fake news, verification marks, and warning labels. The survey will be created on Qualtrics and distributed through Amazon Mechanical Turk, targeting active Twitter users from the United States ages 18-49. Results will be examined generally and in the light of political orientation.

Literature Review

This literature review will rely on an assessment of gatekeeping as a practice, including how the media has typically applied the process and how gatekeeping has evolved and expanded as a result of the internet. Other related topics will also be examined, including credibility, fake news, social media platforms as a news source, Twitter as a social media platform, verification marks, warning labels, and censorship. Gatekeeping is a particularly important lens for this research, not only because of its several applications on social media platforms—such as top-down, bottom-up, and algorithmic gatekeeping—but also because of how verification marks and warning labels are unique applications of the practice.

Historical Background of Gatekeeping

The communications concept of gatekeeping—the process by which countless messages are reduced to only a few—has its origins in food (Shoemaker & Vos, 2009). Sociologist Kurt Lewin, who coined the term in 1943, was curious as to how an individual could alter the food habits of a population. This led him to create a model of how food travels through different channels before it ultimately ends up on the dinner table (Shoemaker & Vos, 2009). David Manning White (1950), a professor at the University of Iowa, then extended this model to communications messaging, focusing specifically on how the media determined which information to distribute.

In the gatekeeping model, gates are decision points while gatekeepers are people or policies that determine which messages come in and which messages stay out (White, 1950). Thus, media gatekeeping is impacted by numerous elements, including instincts, deadlines, rules, newsworthiness, competing stories, trends, goals, audiences, sources, job roles, government, and space (Shoemaker & Vos, 2009). However, gatekeeping does not end there. A new process begins when audience members receive information from the media and make their own decisions about what to allow through their individual gates (Shoemaker & Vos, 2009). Therefore, individual gatekeeping practices are key to understanding the theory in a modern context of the internet.

Gatekeeping in the Age of the Internet

Prior to the creation of the internet, individuals had limited access to news sources and were effectively required to rely on the messages that made it through the media's gates (Messner & Garrison, 2009). But as the internet became a universal source of information, consumers were less restricted by the effects of gatekeeping (Benham, 2020). The amount of information available online meant that the historical applications of media gatekeeping would be challenged, and traditional news sources would now have to compete with those publishing messages on the web (Gil, 2018; Heinecke, 2019; Messner & Garrison, 2009), including blog writers, social media influencers, religious leaders, and celebrities.

White could not have foreseen the impact that the internet would have on gatekeeping and the traditional distribution methods of the media. However, while the gatekeeping forces are continuing to evolve, the process itself is perhaps as universal as it has ever been, with each individual having great control not only over what they consume, but also over what they share, in what has been deemed a "two-step gatekeeping" process (Singer, 2014).

This change in the gatekeeping process has had far-reaching effects. Research has shown that the internet has significantly impacted the gatekeeping position formerly held by traditional journalists (Williams & Delli Carpini, 2004). The public now has power to force the media's hand, shape the news, and influence public discourse (Messner & Garrison, 2009; Poor, 2006). Digital channels—such as social media platforms, wikis, and blogs—provide the public with countless ways to contribute information and opinions (Heinecke, 2019; Messner & Garrison, 2009). This shift highlights the era of the internet known as Web 2.0, which is differentiated from the early days of the internet by the ease for anyone to create and share content (Cormode & Krishnamurthy, 2008).

Credibility

With the internet increasing the number of voices adding to public discourse—as well as allowing users to easily access limitless information—questions regarding news credibility have become more prominent. Credibility has long been associated with *trust* (Metzger & Flanagin, 2013) and *legitimacy* (McClymount & Sheppard, 2020), and academic definitions often include *believability* in their descriptions (Castillo et al., 2011; Metzger & Flanigan, 2013; Sikdar et al., 2013). Nevertheless, credibility is a broad and elusive concept that can be difficult to define.

As has been noted in the literature, credibility is often specific to its particular context (Keshavarz & Givi, 2020). In an online context, credibility can become quite complex because information is usually impacted by numerous sources as it is disseminated (Sundar, 2008). Some scholars have argued that online credibility consists of the source, the channel, and the message (Metzger et al., 2016; Appelman & Sundar, 2016). These aspects have frequently been used in communication scholarship (Keshavarz & Givi, 2020) and are evidence that multiple concepts impact perceived credibility, particularly within media.

Media Credibility

As defined by Chung et al. (2012), media credibility is “a perceived quality based on multiple factors, including trustworthiness and expertise” (p. 173). Although this definition is a helpful starting point, it does not specify all of the factors that contribute to media credibility—a challenge for researchers that is evident in the existing literature.

Other researchers have identified elements that contribute to media credibility, such as perceptions of bias, fairness, objectivity, accuracy, and believability (Sundar, 1999). Gaziano and McGrath (1986) put together a more comprehensive list of subcomponents, which included: fair, not biased, tells whole story, accurate, respects people’s privacy, watches after readers’/viewers’ interests, concerned about community’s well-being, separates fact and opinion, can be trusted, concerned with public interest, is factual, and has well-trained reporters.

Appelman and Sundar (2016) explain that another difficulty related to media credibility is the reality that communication is heavily integrated within the medium. For example, if the same content were published by two different entities, individuals may judge the credibility differently simply based on who published the information. In this instance, many factors may impact how credibility is perceived, including author, editor, publisher, content, and platform (Appelman & Sundar, 2016).

Further, measurement of credibility across online mediums is not consistent because it can’t be. Various elements contribute to the credibility within each situation. For example, blog credibility has been judged by elements such as authenticity, timeliness, and popularity (Kang & Yang, 2011), while web credibility has been judged by elements such as page layout, URL, and date, among other things (Dochterman & Stamp, 2010). These realities make it especially difficult—if not impossible—to analyze credibility based on a standard set of criteria.

Credibility of the Media

At the forefront of conversations surrounding media credibility are varying opinions concerning the roles of the media and online information providers in delivering credible news (Messner & Garrison, 2009). While some argue that media gatekeeping is dangerous because it allows the media to dictate what information we receive, others counter that information under such a model is much more likely to be vetted—which cannot always be said of information on the internet (Benham, 2020).

Theoretically, media gatekeeping *should* be a process that includes vetting and brings trust in standard news sources (Benham, 2020). Nevertheless, news outlets might push an agenda, mislead the public, or even publish inaccurate stories—unintentionally or not (Singer, 2001). This can have a negative impact on trust in the media (Watts et al., 1999) and push consumers to search for information elsewhere.

In addition, news outlets may withhold information from the public, leading non-traditional journalists and others—such as bloggers—to sidestep these traditional gatekeepers by publishing the information anyway (Poor, 2006; Williams & Delli Carpini, 2004). These instances can have unintended consequences for the traditional media by reducing trust and forcing media outlets to publish news that they otherwise could have kept from the public. One example of this occurred when a blogger published details about the Clinton-Lewinsky scandal that the media had largely ignored initially (Williams & Delli Carpini, 2004).

Timing is another challenge that the media has faced in the digital age (Smith & Sissons, 2016). With news outlets wanting to be the first to publish information, fact-checking is sometimes inhibited (Smith & Sissons, 2016). This type of media misconduct occurred in 2013 when San Francisco television station KTVU reported the names of the pilots involved in an

Asiana Airlines crash, stating that they had been verified by the National Transportation Safety Board (Laing, 2013). However, the reported names proved to be fake—and were also racially insensitive—leading to criticism of KTVU's verification process (Laing, 2013).

While this particular error was not directly tied to the internet, the omnipresence of online news from both traditional and unconventional sources has led organizations to sometimes cut corners, thus further harming people's confidence in the media (Smith & Sissons, 2016).

Credibility of Online Sources

On the other side of the debate, those who use the internet to consume and distribute news are not exempt from these challenges of credibility (Messner & Garrison, 2009). Information on the web is often quite credible (Messner & Garrison, 2009), but the growth of misinformation published on the internet by unvetted gatekeepers is becoming a threat to balanced journalism (Benham, 2020; Heinecke, 2019).

Interestingly, early studies measuring the accuracy of online information indicated that internet users considered online news more credible than traditional counterparts—a result of the perceived credibility that stemmed from the growing popularity of the internet (Johnson & Kaye, 2000). In a 2003 survey, Johnson and Kaye (2004) found that blogs were ranked as more credible than traditional news sources.

In recent years, however, individuals have begun to find less credibility in both traditional and online sources (Rubin, 2019). Results from the Edelman Trust Barometer (2020) indicated that 57% of individuals believed that the media they used was contaminated with untrustworthy information. In addition, 76% were worried about fake news being used as a weapon (Edelman, 2020). These growing concerns about credibility raise questions about what internet outlets and online gatekeepers are doing to slow or prevent the spread of fake news.

Overview of Fake News

Fake news is not a new concept, but the spread of inaccurate information has been enhanced by the internet, particularly social media platforms (Rubin, 2019). The surge of false information in recent years—as well as former United States President Donald Trump's emphasis on the topic—led Collins Dictionary to choose "fake news" as its word of the year in 2017 (Sauer, 2017).

The existing literature on fake news indicates that there are countless types and definitions. Rubin et al. (2015) identified three categories of fake news, namely: (a) serious fabrications, (b) large-scale hoaxes, and (c) humorous fakes. In addition, Tandoc et al. (2017) reviewed 34 academic papers on the subject and found six overarching definitions: (a) news satire, (b) news parody, (c) fabrication, (d) manipulation, (e) advertising, and (f) propaganda. After interviewing journalism professionals, Benham (2019) added "imbalance" as a seventh overarching definition of fake news—a category used when news is so unbalanced or infused with opinion that it can no longer be considered accurate.

Tandoc et al. (2017) noted the commonality across these definitions: that fake news assumes the appearance of real news, including how websites look, how articles are structured, and how images include photo credit. Moreover, bots imitate the pervasiveness of news by constructing a system of fake websites (Tandoc et al., 2017).

Although certain categories of fake news may not have significant negative impact, other types can have severe consequences. Some common topics of misinformation are science, politics, terrorism, natural disasters, and urban legends (Vosoughi et al., 2018). Fake news can therefore influence political campaigns, create mass hysteria, contribute to polarization, and impact markets, among other things (Vosoughi et al., 2018).

A research study conducted by Zakharov et al. (2019) examined the perceptions of the content, purpose, and sources of fake news among college students. The researchers found mixed opinions regarding these topics, but one notable discovery was that many students had difficulty distinguishing opinion pieces from fake news (Zakharov et al., 2019).

Additionally, while the motives behind fake news vary, researchers and the public alike suggest that a significant reason for spreading fake news is to push a personal or political agenda, or to receive financial or personal gain (Zakharov et al., 2019). Some scholars have included the term “intentional” in their definitions of fake news (Frank, 2015; Pubjabi, 2017), but other scholars feel that sharing fake news is not always done purposely or maliciously (Rubin, 2017). Zakharov and her colleagues (2019) also found that just under half of those who participated in their research study believed fake news was intentional.

For the purpose of this paper, fake news will be defined as *false information used to mislead the recipient by imitating credibility* (Allcott & Gentzcow, 2017; Benham, 2019; White, 2017). The topic will be further examined specifically in the context of social media platforms in the following sections of this literature review.

Social Media as a News Source

Results from a 2018 Pew Research Center survey showed that 68% of adults in the United States obtained news through social media (Shearer & Matsa, 2018). Newman et al. (2016) found that social media is the most important news source among those ages 18-24. The top three news-focused social media sites are Reddit, Twitter, and Facebook (Shearer & Matsa, 2018). Seventy-three percent of Reddit users utilize the platform for news, while 71% of Twitter users and 67% of Facebook users do the same with those respective platforms (Shearer & Matsa, 2018).

Consumers most often identified convenience as the biggest benefit of obtaining news this way (Shearer & Matsa, 2018). However, 57% of these consumers indicated that they expect news on social media channels to be predominantly inaccurate (Shearer & Matsa, 2018). In fact, inaccuracy was the top concern raised about obtaining information on social media (Shearer & Matsa, 2018).

In spite of these concerns, social media platforms continue to grow in popularity, making them an ideal location for individuals and organizations to post and share information (Shearer & Matsa, 2018). For this research, emphasis will be placed on Twitter, largely because Twitter is a uniquely public platform—meaning that most information can be accessed without needing an account—and because it was the first channel to implement a verification feature (Cohen & Sutton, 2018).

Twitter: Social Media Platform and Micro-Blogging Site

According to Statista, Twitter boasts approximately 330 million active monthly users across the globe (Clement, 2019) and is one of the most popular social media platforms in the United States, with nearly 70 million American users as of October 2020 (Clement, 2020). Like other channels, Twitter has features such as likes, re-posts, mentions, and replies that help track reach and engagement. Re-posts on Twitter are known as Retweets, which "are distinguished by the Retweet icon and the name of the person who Retweeted the Tweet" ("Retweet FAQs," n.d., para. 3). In addition to text, Tweets can include images, videos, and links.

One feature that differentiates Twitter from its competitors is the platform's 280-character limit for Tweets, which was expanded from 140 characters in 2017 (Rosen, 2017). In addition to being a social networking site, Twitter is considered a micro-blogging site since Tweet character limits led the platform to become a place for quick dissemination of news

(Hermida, 2010). Political news is among the most popular types (Vosoughi et al., 2018; Wojcik & Hughes, 2019), but the site is also known for sports, religion, economics, travel, health, and more. According to a 2019 survey, Twitter is the most preferred social media platform for news consumption, with 56% of respondents indicating that they used the platform for news (Statista Research Department, 2021b). The next closest platform was Facebook, with 38% (Statista Research Department, 2021b).

Another element that makes Twitter unique is its openness. While some accounts are private—and therefore have protected Tweets—research has shown that approximately 87% are public (Remy, 2019). When an individual creates an account, Tweets are set to public by default, meaning that anyone can view and interact with their Tweets ("About public and protected Tweets," n.d.).

Fake News on Twitter

While Twitter is not the only social media platform that deals with bots and fake quotes, the network is notorious for these issues, which contribute to the spread of false information (Kirner-Ludwig, 2019; Wojcik et al., 2018). Researchers from the Pew Research Center analyzed a random sample of 1.2 million Tweets and discovered that—among news and current event websites—66% of Tweeted links to popular websites were made by suspected bots (Wojcik et al., 2018).

Nevertheless, researchers at MIT discovered that bots spread both accurate and false information at the same rate (Vosoughi et al., 2018). They also found that fake news travels six times quicker on Twitter than true stories (Dizikes, 2018) and indicated that humans are largely responsible for the large scale spread of fake news despite testimony before congressional committees that concentrated on bots as a culprit (Vosoughi et al., 2018).

Gatekeeping on Social Media Platforms

Although users effectively control what information they receive on social platforms—through features like following, friending, liking, and subscribing—algorithms, trending topics, recommendations, and paid advertisements also play a role in what type of content users will be shown. In this sense, algorithms serve as gatekeepers (Gil, 2018).

Many of these features are designed to improve and enhance the user experience (Gil, 2018), and some even have the capability of limiting the spread of fake news. But as with most beneficial elements, there can also be negative consequences, such as removing quality content from the discussion or unintentionally silencing voices. The intricacies of gatekeeping on social media platforms continues to expand as algorithms, executives, and users further impact how information is seen and received.

Algorithmic Gatekeeping

Algorithmic gatekeeping has become increasingly complex and controversial in recent years because algorithms can go well beyond what users are self-selecting to amplify content that social platforms choose, which may be driven by money, popularity, or agendas. While algorithms certainly drive users to content which is related to their interests and even valuable, these processes also have the capability to unintentionally endorse controversial content and voices (Darcy, 2019).

On Twitter in particular, the platform’s algorithm often fills feeds with popular content that is not promoted nor sought out by users (Darcy, 2019). Twitter has stated that this decision was made to further expose users to content they might be interested in (Darcy, 2019). However, this effort brings with it not only the capability of sharing irrelevant content but also of spreading misinformation and amplifying fake news (Darcy, 2019). Furthermore, some Twitter users have been frustrated by this feature, emphasizing that if they wanted those Tweets to show up in their feed, they would follow those individuals (Darcy, 2019).

Top-down Gatekeeping

Connected with algorithmic gatekeeping is the concept of top-down gatekeeping. While not much specific information exists regarding how executives and others within a company may be influencing what content gets through the gates, platforms have certainly made known some of their more obvious efforts, such as the implementation of warning labels—an effort which will be discussed later in this literature review.

Nevertheless, some social media users have accused platforms of limiting voices (Darcy, 2019), and recent research suggests that there tends to be a political bias in censorship (Stjernfelt & Lauritzen, 2020). Specifically on Twitter, a growing number of individuals have accused the platform of "shadow banning," the practice of suspending a user without their knowledge and which prevents their content from being seen (Darcy, 2019). Although Twitter has denied this behavior (Darcy, 2019), it is clear that some users are skeptical of the possibilities associated with top-down gatekeeping.

Bottom-up Gatekeeping

Another reality impacting content consumption on social media platforms is that of user or bottom-up gatekeeping. Users on social media have long had the ability to report content for being offensive, but controversies of recent years have led to increased reporting and suspension for content being dangerous or misleading (Newberry, 2021). These reports are reviewed in relation to platform rules and may result in removal.

On Twitter in particular, accounts may be suspended for things such as impersonation, abusive behavior, hateful conduct, and dangerous rhetoric ("About suspended accounts," n.d.).

Twitter's Transparency site posts reports in six-month increments that provide numbers for the number of accounts reports, accounts suspended, and content removed. In its most recent report featuring January 2021 to June 2021, Twitter noted that 1.2 million of the 4.8 million accounts reported were suspended, and that 5.9 million pieces of content were removed for violating safety, privacy, or authenticity rules ("Rules Enforcement," 2022).

In 2021, Twitter introduced Birdwatch, a "pilot in the US of a new community-driven approach to help address misleading information on Twitter" (Coleman, 2021, para. 1). Although the concept is still being built out, the idea is that users have the capability to identify details of a Tweet that they find to be misleading and can then write notes that provide helpful context and information (Coleman, 2021). Other users will be able to see these notes and rate them based on their helpfulness (Coleman, 2021). In its pilot stage, Twitter is keeping notes on a separate site from Twitter but hopes to build the capability to view notes directly into Twitter in the future (Coleman, 2021). As such, this offering is another method for users to be involved in content moderation and the fight against fake news.

Combatting Fake News on Social Media Platforms

The importance of individual gatekeeping is especially relevant when obtaining news via social media. Fact-checking websites have emerged to help combat the spread of misinformation and fake news on social channels (Amazeen et al., 2019; Hameleers et al., 2020), but such sites are only beneficial if they are seen and utilized (Amazeen et al., 2019).

Recent research indicates that individuals who encounter digital misinformation are rarely provided with fact-check options (Guess et al., 2017). Sites like Google, Facebook, and Twitter have begun developing solutions that aim to address this lack of fact-checking ability (Gadde & Beykpour, 2020; Lewandowsky et al., 2017); nevertheless, individuals should be active consumers of news rather than passive recipients if they are to ensure that accurate information is passing through their own gates (Hameleers et al., 2020).

Furthermore, consumers in the digital age have a responsibility to fact-check information found online before sharing it on their channels (Hameleers et al., 2020). Fact-checking is growing in popularity (Amazeen et al., 2019; Hameleers et al., 2020), but due to the possibility of posts going viral on social channels—and because online vetting is often limited to personal gatekeeping efforts—fake news still thrives on sites like YouTube, Facebook, and Twitter (Vosoughi et al., 2018).

While older generations are more likely to seek out fact-checking websites (Calvillo et al., 2020), younger generations are becoming increasingly more skeptical. Marchi (2012) found that teenagers on average tend to believe that traditional media is corrupt, and that social media allows people to voice the truth. Younger individuals are inclined to desire opinionated news rather than objective news (Marchi, 2012).

Another challenge that social media users face is knowing how to judge the credibility of content. On Twitter in particular, Castillo et al. (2011) found that perceptions of news credibility can be impacted by aspects such as visual design and the perceived gender of the author. In a study about information on Twitter, Azer et al. (2021) discovered that factors including time, effort, informal language, and the limited size of Tweet length can make it difficult to judge credibility on the platform.

In response to difficulties like this, computer systems have been developed to analyze Tweets for credibility, looking at elements such as social tags and sentiment (Azer et al., 2021). These systems also compare Tweets against “trusted” content in the system (Azer et al., 2021), but these determinations are still vulnerable to error and can be impacted by human biases.

In recent years, Twitter and other social media sites have wrestled with how to handle fake news on their platforms and have begun seeking additional solutions (Lewandowsky et al., 2017). Twitter has attempted to combat the spread of misinformation in a number of ways—some of which were mentioned previously—but perhaps the two most recognized are the use of the verification mark (Cohen & Sutton, 2018) and the application of warning labels (Gadde & Beykpour, 2020).

History and Overview of the Twitter Verification Mark

Introduced by Twitter in 2009, the verification mark was designed to indicate authenticity (Cohen & Sutton, 2018). Twitter initially used the feature to verify prominent businesses, organizations, journalists, and politicians (O'Sullivan, 2020), but for a short time in 2016, Twitter opened verification to anybody who wanted to apply (Tsukayama, 2016). The feature was also adopted by other social media platforms, including Facebook, Instagram, and YouTube (Paul et al., 2019).

While social media platforms have long stressed that verification does not equal endorsement, research shows that accounts and individuals with a verified status receive enhanced credibility (Paul et al., 2019). Due to confusion and misinterpretation among users, social platforms have begun rethinking how they grant verification (Cohen & Sutton, 2018). In 2018, Twitter acknowledged that the use of the feature had led to uncertainty, posting on its platform, "Verification was meant to authenticate identity & voice but it is interpreted as an endorsement or an indicator of importance. We recognize that we have created this confusion and need to resolve it" (Twitter Support, 2017).

These acknowledgements and revisions have sometimes been the result of platform blunders or controversies. In 2017, Twitter received backlash after verifying an individual named

Jason Kessler, who some had labeled as “alt-right” (Sokol, 2017). Kessler’s verification mark was soon taken away, and others deemed to be in the far-right movement also had their marks removed (Sokol, 2017).

In response, Kessler accused Twitter of silencing conservative voices and condemned the platform for being one-sided (Sokol, 2017). Of the decision, Yair Rosenberg, a writer for Jewish publication *Tablet Magazine*, Tweeted, “Whoever advised Twitter to turn verification into an approbation of views rather than a confirmation of identity did not think this through. Now Twitter can be held accountable for every controversial thing said by a blue checkmark”[1] (Rosenberg, 2017; Sokol, 2017).

Following this controversy with Kessler, Twitter suspended its verification process for a time (Bowles, 2017). After the process was ultimately reinstated, the platform once again faced criticism after a high school student in upstate New York managed to get an account verified for a fictitious United States Senate candidate (O’Sullivan, 2020).

These examples pose serious questions regarding the use and purpose of the verification feature. If Twitter introduced the verification mark to authenticate identity and voice, why would the platform strip legitimate individuals of their verified status? In addition, does such an action contradict Twitter’s claim that verification marks do not equal endorsement? If most people interpret Twitter verification marks as a sign of importance, is this an example of censorship?

Recent Changes to Twitter’s Verification Program

Since these controversies, Twitter has updated its explanation of verification status to include additional information on the process (“Verified account FAQs,” n.d.). According to Twitter’s help site in late 2020, “An account may be verified if it is determined to be an account of public interest. Typically this includes accounts maintained by users in music, acting, fashion,

[1] Note: The Tweet appears to have since been deleted.

government, politics, religion, journalism, media, sports, business, and other key interest areas" ("About verified accounts," n.d., para. 2). And, as previously stated by the platform, verification does not indicate endorsement ("Verified account FAQs," n.d.).

Twitter also published details regarding the loss of verification status. The company reserved the right to remove verification for behaviors on and off Twitter, including misleading people; promoting hate and violence; attacking people based on race, ethnicity, orientation, gender, age, ideology, religious affiliation, etc.; supporting organizations who promote prejudiced attitudes; harassment; engaging in violent and dangerous behavior; posting disturbing, violent, or gruesome imagery; promoting terrorism; promoting suicide or self-harm; or violating the Twitter terms of service ("Verified account FAQs," n.d.).

While these pages clarified Twitter's definition of verification—as well as explained the platform's stance on granting and removing verification status—the feature underwent additional updates in following months. As of January 2022, Twitter indicates that "the blue verified badge on Twitter lets people know that an account of public interest is authentic" ("Verification FAQ," n.d.). The platform states that your account must be "notable and active" to qualify for verified status, the six types of which are: government; companies, brands, and non-profit organizations; news organizations and journalists; entertainment; sports and esports; activists, organizers, and other influential individuals ("Verification FAQ," n.d.).

Furthermore, Twitter relaunched the ability to apply for or request verification, which was reimplemented in May 2021 (Twitter Inc., 2021). This action came in response to feedback from Twitter users and led the platform to remove verification from accounts that no longer met the qualifications (Twitter Inc., 2021). The company indicated that the "application rollout marks the next milestone in our plans to give more transparency, credibility and clarity to verification on Twitter" (Twitter Inc., 2021). One of the main reasons Twitter updated its verification program was to encourage healthy dialogue and to allow users to better determine if the conversations they are having are trustworthy (Twitter Inc., 2021).

Twitter Verification and Credibility

According to a research study conducted by Edgerly and Vraga (2019), verification marks on Twitter do not increase perceived credibility of a post or account. The results of this study refuted the prevailing view that verification marks impact how people make credibility judgments on the platform. Instead, "account ambiguity and congruency were more powerful cues in assessing credibility" (Edgerly & Vraga, 2019, p. 286). Thus, preliminary research suggests that Twitter users rely on other cues to determine the credibility of a Tweet—not a verification mark.

Number of likes, number of Retweets, Tweet sentiment, and trusted URLs are features that have been identified as "credibility detectors" (Azer et al., 2021). Additionally, Zubiaga and Ji (2014) found that readily available features such as profile handle, profile picture, and images used in a Tweet can impact how consumers view news. These researchers also noted that poor grammar and spelling do not appear to make much of a difference on how people determine credibility (Zubiaga & Ji, 2014). However, making the extra click to view the profile associated with a Tweet can be the difference between correctly identifying a Tweet as accurate or as fake (Zubiaga & Ji, 2014).

Edgerly and Vraga (2019) suggest that the value of the verification mark may have been diminished and diluted due to Twitter's decision to offer verification to anyone. The number of verified accounts on the platform jumped from 150,000 (Kamps, 2015) to 300,000 in less than three years (Edgerly & Vraga, 2019). While this number still represents a relatively tiny portion of Twitter users (Kamps, 2015), the increased prevalence of Twitter verification—especially on more visible accounts—might reduce its usefulness, regardless of what Twitter's goal for the feature ultimately is.

Nevertheless, the existing literature regarding Twitter verification and credibility was written prior to the platform's recent updates to the program. It is clear that verification marks as they are currently utilized by Twitter are a form of gatekeeping designed to enable users to make better judgments regarding the credibility and trustworthiness of content (Twitter Inc., 2021). As such, these alterations are likely to influence perceptions regarding how verification impacts (a) the spread of fake news, and (b) the credibility of the feature itself.

Research on Warning Labels

Another prominent feature making its way to social media platforms in the fight against fake news is the warning label. Disclaimer and warning labels have been used over the years to address body image (Fardouly & Holland, 2018), alcohol (Lou & Alhabash, 2020), historically insensitive material (McGowan, 2018), and general information (Clayton et al., 2020; Pennycook et al., 2020), among other things. These disclaimers and labels have appeared in places such as traditional advertising (Fardouly & Holland, 2018; Lou & Alhabash, 2020), movies (McGowan, 2018), and social media outlets (Clayton et al., 2020; Lou & Alhabash, 2020; Pennycook et al., 2020).

Research has shown that although there has been a growing interest in disclaimers and warning labels, these efforts are often ineffective in addressing the concerns of public officials and society at large (Fardouly & Holland, 2018; Pennycook et al., 2020). Nevertheless, social media platforms have begun adopting this feature as a way to combat the spread of fake news (Clayton et al., 2020; Colliander, 2019).

Warning Labels on Social Media Platforms

Researchers from Dartmouth found that false headlines on social media platforms were perceived as less accurate by those who received general warnings about misinformation or when headlines include a tag that read “Disputed” or “Rated False” (Clayton et al., 2020). However, general warnings also lessened belief in the legitimacy of true headlines, indicating that warning labels also have the potential to decrease belief in true information (Clayton et al., 2020).

Additionally, Pennycook et al. (2020) discovered that although warning labels led to a modest reduction in perceived accuracy of false headlines, the existence of warning tags caused headlines that were untagged to be viewed as more accurate. These researchers concluded that because of these results, using warning labels to counter misinformation presents a potential challenge and concern, especially since producing misinformation is easier than exposing it (Pennycook et al., 2020).

Colliander (2019) found that comments made by other social media users had a greater likelihood than warning labels of dissuading the belief in and stopping the spread of fake news. Colliander’s research (2019) suggests that users opt to rely on other consumers as a guide when it comes to online disinformation rather than trusting the stage crew of social media platforms to tell them what is credible—a claim supported by results from a 2021 survey, which indicates that an astounding 75% of social media users do not trust platforms to make fair content moderation decisions (Kemp & Ekins, 2021). Colliander (2019) indicates that this might be the reason social media outlets are moving away from warning labels, but certainly in the case of Twitter, the platform has only been more vigorous in adding additional types of warning labels—or Notices, as they are known on Twitter (“Notices on Twitter,” n.d.).

History and Overview of Twitter Notices

Within the last few years, Twitter has been actively creating a variety of Notices that can be applied to Tweets to provide context for consumers (“Notices on Twitter,” n.d.). As of 2022, Twitter teams and systems can add Notices to restrict immediate access to Tweets for any of the following: graphic violence or adult content, violation of Twitter Rules, controversial content, disputed or misleading information, manipulated media, or suspended accounts (“Notices on Twitter,” n.d.).

In June 2019, the company introduced a new notice to provide clarity for situations where Tweets violated rules but were left on the platform (Twitter Safety, 2019). Twitter explained that it occasionally allows controversial content to remain online when there is a public interest in its availability but adds a notice about the violation of rules and limits the Tweet’s engagement (Twitter Safety, 2019).

The notice was seen as a response to complaints made by activists regarding exemptions that have been given to prominent leaders who have violated Twitter Rules (Ortutay, 2019). The company had previously stated that allowing controversial Tweets to be published by leaders holds them accountable and encourages discussion (Ortutay, 2018).

In 2020, Twitter introduced several new Notices, including warning labels for Tweets that contain synthetic and manipulated media, and warning labels for Tweets that contain unverified claims, disputed claims, or misleading information (Roth & Pickles, 2020). The platform’s new policy goes so far as to remove Tweets that contain “severe” misleading information (Roth & Pickles, 2020).

Twitter took additional steps in October 2020. Those who attempt to Retweet posts that have a misleading information label are given a prompt directing them to credible information about a topic before they can share it (Gadde & Beykpour, 2020). Twitter also added warnings and restrictions to Tweets from political figures that have a misleading information label (Gadde & Beykpour, 2020). Individuals who encounter these Tweets must now click through a warning and will not be able to Retweet, reply, or like (Gadde & Beykpour, 2020). The platform also de-amplifies other Tweets with labels by limiting engagement and ensuring that such Tweets do not show up in searches, notifications, recommendations, timelines, and feeds (Gadde & Beykpour, 2020).

Censorship Questions on Twitter

While verification and Notices are designed to limit the spread of misinformation, questions remain about their effectiveness and credibility. In addition, opinions differ on the use of features that intentionally silence voices. Some individuals felt the new Notice implemented in June 2019—which allowed rule-breaking content to stay on the platform when it was deemed to be in the interest of the public—was a step in the right direction but did not go far enough (Ortutay, 2019). Keegan Hankes, a research analyst for the Southern Poverty Law Center's Intelligence Project, felt that Twitter's decision to leave this content on the platform indicated that hate speech can be in the public interest—a point he disagreed with (Ortutay, 2019). But others have expressed opposing views, indicating that it is not in the interest of the people to allow social media platforms to prevent open expression (Daseler, 2019).

Twitter has long faced competing opinions on content restrictions, but as a private corporation, the organization has the right to prohibit what is posted on its platform (Daseler, 2019). Nevertheless, Supreme Court Justice Anthony Kennedy stated in a 2017 ruling that the internet is "the modern public square" (Daseler, 2019), which calls into question whether these platforms should still have the ability to censor opinions.

Censorship on social media is not a new trend, but it continues to become more and more relevant. A recent analysis of content that was removed from social platforms suggests that there is a political bias in censorship (Stjernfelt & Lauritzen, 2020). Moreover, concerns have been raised regarding who determines what should be censored, even with topics that are scientific or objective in nature (Niemiec, 2020).

Questions also remain regarding who should be allowed to define what information is inaccurate or harmful, as well as whether these individuals can be trusted (Niemiec, 2020). Niemiec (2020) points out that major social media platforms have cited the World Health Organization (WHO) as an authoritative voice during the COVID-19 pandemic, but that even an established organization such as this can make mistakes. For example, some individuals voiced concerns about how pharmaceutical companies were able to influence WHO's guidelines during the 2009 swine flu pandemic (Cohen & Carter, 2010).

Along the same lines, many news outlets and even fact-checking platforms—such as *PolitiFact*—initially disputed the claim that the COVID-19 pandemic began as an outbreak from a lab, labeling it as a conspiracy theory. However, *PolitiFact* and others later retracted statements as more information became available (Adams, 2021). An editor's note by Li-Men Yan on *PolitiFact* from May 2021 reads:

> When this fact-check was first published in September 2020, *PolitiFact*'s sources included researchers who asserted the SARS-CoV-2 virus could not have been manipulated. That assertion is now more widely disputed. For that reason, we are removing this fact-check from our database pending a more thorough review. Currently, we consider the claim to be unsupported by evidence and in dispute. The original fact-check in its entirety is preserved below for transparency and archival purposes. (para. 1)

Social media platforms also censored individuals for this claim, going so far as to remove user posts for claiming that COVID-19 was man-made or came from a lab (Lima, 2021; Rosen, 2021). In May 2021, Facebook updated its misinformation policy surrounding COVID-19 to no longer censor individuals for this claim (Rosen, 2021). The company said that the reason for this was due to renewed debate surrounding the origins of the virus (Lima, 2021).

These realities are important to note because they not only prove that fact-checking sites can be wrong but also highlight that warning labels can be misapplied. Furthermore, verification status on Twitter in particular can be impacted by the platform determining that a user has posted misleading information ("Verified account FAQs," n.d.), meaning that some verified individuals may have been censored or stripped of verification status for claims that were initially seen as inaccurate but which were later deemed debatable (Lima, 2021).

As mentioned previously, a 2021 CATO Institute poll found that 75% of Americans lack confidence that social media platforms will be fair in their content moderation decisions (Kemp & Ekins, 2021). Results from this same survey indicate that 54% of polled Americans were more concerned about social media platforms censoring truth than they were about the spread of fake news (Kemp & Ekins, 2021). A large reason for this may be due to the fact that social media companies are largely distrusted by Americans as a whole, regardless of political orientation (Kemp & Ekins, 2021). Nevertheless, it is important to understand how political identification may impact attitudes regarding efforts to combat fake news on Twitter, including censorship.

Political Identification in the United States

Research indicates that political orientation in the United States is nearly evenly split among three main groups—independent at 34%, Democrat at 33%, and Republican at 29% (Gramlich, 2020). Nearly half (48%) of registered voters in the United States are between the ages of 18 and 49, with 49% of those leaning or identifying as Democrat and only 42% leaning or identifying as Republican (Gramlich, 2020).

Notably, eligible voters between the ages of 18 and 49 have historically been less likely to vote than older generations. Millennials and Gen X both saw an increase in voting turnout in 2016, with 50.8% of Millennials and 62.6% of Gen X voting, up 4.4% and 1.2%, respectively (Krogstad & Lopez, 2017). However, these numbers pale in comparison with Boomers and Silent/Greatest, with both groups hovering around 70% (Krogstad & Lopez, 2017). Nevertheless, voter turnout among those ages 18-34 rose to 57% in 2020—an increase from 49% in 2016 according to the United States Census Bureau (Fabina, 2021).

Political Identification and Activity on Twitter

According to a 2019 Pew Research study, Twitter users are more likely to identify as Democrats than Republicans, with 60% leaning or being Democrat and 35% leaning or being Republican (Wojcik & Hughes, 2019). Among those ages 18 to 49, this figure is even larger, with nearly two-thirds (63%) leaning or being Democrat (Wojcik & Hughes, 2019).

In alignment with these findings, the platform tends to be much more liberal than conservative and is less conservative than the general United States population (Wojcik & Hughes, 2019). While approximately 25% of Americans identify as “very conservative,” Pew Research found that only 12% of Twitter users do (Wojcik & Hughes, 2019).

Additionally, it is worth mentioning that Twitter users are more likely to be politically active according to current research (Wojcik & Hughes, 2019). Sixty percent of Twitter users voted in the 2018 midterm elections while only 55% of United States adults could say the same (Wojcik & Hughes, 2019). Results from Pew Research Center indicate that 42% of Twitter users in the United States use the site to discuss politics at least occasionally (Hughes & Wojcik, 2019). Of the top 10 percent of Tweeters, 42% indicated that they had sent at least one political Tweet in the last 30 days (Hughes & Wojcik, 2019).

Political Identification and Fake News

Researchers have found that both liberals and conservatives associate left-leaning and right-leaning news platforms that oppose their political identification with the term "fake news" (van der Linden et al., 2020). While information regarding the relationship between acceptance of fake news and political orientation continues to develop, current research indicates that conservatives are more likely to believe misinformation than liberals (see Fessler et al., 2017; Miller et al., 2016). Jost (2017) suggests that this is because of differences in cognitive processes that exist between the two groups while Miller et al. (2016) and Fessler et al. (2017) suggest that the reason is because conservatives view the world as more complex and threatening and are more susceptible to uncertainty.

In a similar vein, Guess et al. (2017) found that those who supported Hillary Clinton in the 2016 United States election were more likely to fact check than those who supported Donald Trump. Other research has also indicated that liberal-leaning individuals and those who voted for Clinton were more likely to fact-check (Amazeen et al., 2019).

When it comes to the use of warning labels on social media posts, Mena (2019) found that warning labels reduced the intent to share fake news among Democrats, Independents, and Republicans. However, those who identified as Democrats and Independents were more likely than Republicans to share news posts that contained false information regardless of whether a warning label was present or not (Mena, 2019).

In contrast to this research, Grinberg et al. (2019) found that conservatives were more likely to share fake news during the 2016 United States election than their liberal counterparts.

Additionally, Pennycook and Rand (2019b) conducted a study that suggested liberal individuals were more likely to identify fake news than conservatives, and Calvillo et al. (2020) found that conservatives were less accurate in distinguishing between real and fake headlines.

Calvillo et al. (2020) suggest that political framing done by political leadership and the media can have a significant impact on the perceptions of fake news. These researchers indicate that once an issue becomes politicized, the way it is framed may impact the way it is perceived (Calvillo et al., 2020). But even though individuals on both sides of the political aisle are susceptible to fake news, Faragó and her colleagues (2020) discovered that the acceptance of pro-government and anti-government fake news was more driven by partisanship than it was by political orientation.

Political Identification and Social Media Gatekeeping Practices

Kemp and Ekins (2021) report that strong liberals are much more likely than strong conservatives to report content on social media platforms. Research shows that this behavior is heavily correlated with political identification, with 65% of strong liberals and 44% of moderate liberals having done so while 24% of strong conservatives and 21% of moderate conservatives have done so (Kemp & Ekins, 2021).

In relation to these statistics, conservatives are more likely to be censored on social media platforms and are more likely to have their accounts suspended (Kemp & Ekins, 2021; Stjernfelt & Lauritzen, 2020). Kemp and Ekins (2021) indicate that more than a third of conservatives have personally experienced a post being reported or removed, while only a fifth of liberals have. The reason for this may be the differing political opinions regarding the removal of content on social platforms, with 80% of strong conservatives saying platforms are going too far and 68% of strong liberals saying they're not doing enough (Kemp & Ekins, 2021).

Research Questions

While consumers still have access to significant amounts of information across the web, Twitter's use of verification marks and Notices limits what content users receive, sometimes at the expense of accurate information (Kemp & Ekins, 2021). This is particularly important as more and more individuals are using these outlets for news (Shearer & Matsa, 2018).

These applications of social media gatekeeping call into question not only their effectiveness in slowing the spread of false information, but also whether users consider them to be credible. More understanding is needed regarding the perceptions surrounding these specific Twitter features as well as how fake news is viewed on the platform. With this in mind, the following research questions have been proposed for further examination:

RQ_1: What are the perceptions of Twitter users surrounding fake news?

RQ_2: What are the perceptions of Twitter users surrounding the credibility of verification?

RQ_3: What are the perceptions of Twitter users surrounding the effectiveness of verification in slowing the spread of fake news?

RQ_4: What are the perceptions of Twitter users surrounding the credibility of Notices?

RQ_5: What are the perceptions of Twitter users surrounding the effectiveness of Notices in slowing the spread of fake news?

RQ_6: Do the perceptions surrounding fake news, verification, and Notices differ based on the political identification of the Twitter user?

Methods

This study was conducted through the distribution and quantitative analysis of a survey, which was created using Qualtrics, distributed through Amazon Mechanical Turk, and analyzed using SPSS Statistics. The research relied on allocated funds of $500 from the BYU School of Communications and resulted in 286 valid responses.

Population Description

The population for this research was active Twitter users, ages 18-49. As defined in the research, an active user was *any consumer who accesses the platform at least once a month*—a metric that Twitter previously reported, and which is still used by other data reporters (Statista Research Department, 2022).

The purpose of analyzing active users was two-fold: one, they were likely to use Twitter for news consumption (Statista Research Department, 2021b), and two, they were very likely to have encountered warning labels on the platform. The main reason the researcher surveyed and analyzed users ages 18-49 is because nearly 70% of Twitter users are between the ages of 18 to 49 (Statista Research Department, 2021a). Additionally, the proposed age range includes more than 50% of the registered voters in the United States as of 2019 (Gramlich, 2020).

Data Collection

This research relied on Amazon Mechanical Turk as its main method of data collection. Amazon Mechanical Turk—which is often referred to as MTurk—is a crowdsourcing platform where individuals are compensated for participating in surveys and other task-oriented projects. Screening questions helped narrow qualified participants to ensure that results reflected the intended population.

The survey took approximately five minutes to complete, and participants were given $0.60 as compensation. The decision regarding the amount of compensation was largely made because of a similar study conducted by Edgerly and Vraga (2019), who compensated MTurk participants $0.60 for a 6-minute survey.

Questionnaire Development

The questionnaire developed for this research study consisted of questions related to demographics, political identification, and perceptions surrounding three elements of Twitter: (a) fake news, (b) credibility and effectiveness of verification, and (c) credibility and effectiveness of Notices on Twitter. These perceptions were examined using (1) a series of one word text responses, (2) a series of 5-point semantic differentials, and (3) a series of 5-point Likert scales.

Measures

The scale used for political orientation was adapted from van der Linden et al. (2020) and was determined by self-placement according to the following options: very liberal, liberal, conservative, and very conservative.

The 5-point semantic differential scales, which were used to analyze perceptions of fake news and the credibility of verification and Notices, was taken from Edgerly and Vraga (2019), with the semantics being both replicated and adapted for this study. The scales were reduced from 7-point to 5-point in order to stay consistent with other measures. Some measurements from Gaziano and McGrath (1986) were also included.

- **Fake news on Twitter:** common/uncommon, dangerous/not dangerous, concerning/not concerning, easy to identify/hard to identify, controlled/uncontrolled, bot-generated/human-generated, minor problem/major problem
- **Verification on Twitter:** credible/not credible, accurate/inaccurate, biased/not biased, can be trusted/cannot be trusted, watches out for users' interests/does not watch out for users' interests, concerned with the community's well-being/not concerned with the community's well-being, concerned with public interest/not concerned with public interest, creates skepticism/creates confidence

- **Notices on Twitter:** credible/not credible, accurate/inaccurate, biased/unbiased, can be trusted/cannot be trusted, watches out for users' interests/does not watch out for users' interests, concerned with the community's well-being/not concerned with the community's well-being, concerned with public interest/not concerned with public interest, tell the whole story/do not tell the whole story, separate fact and opinion/do not separate fact and opinion, factual/not factual, prevent expression/do not prevent expression, censor/do not censor

Likert scales were used to evaluate the perceptions of fake news and the effectiveness of verification and Notices. These scales asked participates to rank responses about statements on a scale of 1 to 5, with 1 being *strongly disagree* and 5 being *strongly agree*. The measurement for this research was taken from Matthes (2012) and Schulz et al. (2020) and was adapted for this particular study.

- **Fake News**
    1. Fake news on Twitter is more common today than it was 5 years ago
    2. Fake news on Twitter should be monitored or addressed by Twitter
- **Verification**
    1. Verification on Twitter slows the spread of fake news
    2. Verification on Twitter eliminates the spread of fake news
    3. Verification on Twitter contributes to the spread of fake news
    4. Verification on Twitter is effective in reducing the spread of fake news
    5. Verification on Twitter leads people to find accurate information
    6. Verification on Twitter decreases belief in the accuracy of Tweets from unverified accounts

7. Verification on Twitter is dangerous
8. Verification on Twitter is beneficial

- **Notices**

    1. Notices on Twitter slow the spread of fake news
    2. Notices on Twitter eliminate the spread of fake news
    3. Notices on Twitter contribute to the spread of fake news
    4. Notices on Twitter are effective in reducing the spread of fake news
    5. Notices on Twitter lead people to find accurate information
    6. Notices on Twitter are dangerous
    7. Notices on Twitter are beneficial
    8. Notices on Twitter create confidence in the accuracy of news on the platform
    9. Notices on Twitter raise concerns related to censorship
    10. Notices on Twitter are against the First Amendment

Results

The survey was accessed by 347 individuals, but 48 did not qualify to participate due to their responses to screening questions related to age or frequency of Twitter use. An additional 13 responses were removed for quality control, leaving 286 completed surveys that met the population requirements of the research.

Of the sample, 34.6% (*n* = 99) were female, 64.7% (*n* = 185) were male, and 0.7% (*n* = 2) identified as other or preferred to not say. These results closely reflect the reported population on Twitter, which had approximately 38.4% female users and 61.6% male users as of January 2021 (Statista Research Department, 2021a). In addition, 27.3% (*n* = 78) of respondents were ages 18-29, 46.5% (*n* = 133) were 30-39, and 26.2% (*n* = 75) were 40-49 (see Table 1).

Table 1

Age (n = 286)

18-29	*30-39*	*40-49*
27.3% *n* = 78	46.5% *n* = 133	26.2% *n* = 75

While gender and age are reported in these results, the main demographic related to this research was political orientation (see Table 2). As reported by Wojcik and Hughes (2019), 63% of Twitter users ages 18-49 identify as leaning or being Democrat. The results of this survey are consistent with this statistic, with 63% of participants identifying as liberal or very liberal. This indicates that the results reported in this research are likely to closely reflect the perceptions of Twitter users overall.

Table 2

Political Orientation (n = 286)

Very liberal	*Liberal*	*Conservative*	*Very conservative*
21.7% *n* = 62	41.3% *n* = 118	27.6% *n* = 79	9.4% *n* = 27

Additionally, it is worth reporting the frequency of Twitter use among those qualified for the survey. As indicated in Table 3, approximately 95% of respondents access the platform at least weekly, with more than two-thirds accessing Twitter daily.

Table 3

Twitter Use (n = 286)

Daily	*At least once a week*	*At least once a month*
70.6% *n* = 202	24.1% *n* = 69	5.3% *n* = 15

Fake News on Twitter

RQ$_1$: What are the perceptions of Twitter users surrounding fake news?

The first question research participants were asked in the fake news section of the survey was, "What is the first word that comes to mind when you hear the term *fake news*?" By far, the most common response was "Trump" (*n* = 79), which was 67 mentions ahead the second most common, "Fox" (*n* = 12). The top 10 responses are listed in Table 4.

Table 4

First word that comes to mind when hearing the term "fake news"

	n =
1. Trump	79
2. Fox	12
3. Republicans	10
4. Fraud	10
5. CNN	9
6. Bot	9
7. False	7
8. Bad	7
9. Propaganda	6
10. Misleading	6

In spite of the huge gap between "Trump" and the rest of the responses, themes were identified that tightened the results. *Donald Trump* was still the most common theme (*n* = 84), but *words with similar meanings to "fake"*—fraud, false, dishonest, misinformation, misleading, deception, manipulation, lying, lies, fabrication, propaganda—was second (*n* = 51), *media* was third (*n* = 27), and *Republicans/conservatives* was fourth (*n* = 12).

Beyond these initial impressions, the survey sought to understand aspects such as impact, source, and frequency of fake news on Twitter, among other things. Table 5 provides descriptive statistics regarding these measures.

Table 5

Descriptive Statistics for Perceptions of Fake News on Twitter (n = 286)

	M	*SD*
1. Common/uncommon	2.18	1.13
2. Dangerous/not dangerous	2.07	1.17
3. Concerning/not concerning	2.02	1.20
4. Easy to identify/hard to identify	2.87	1.10
5. Controlled/uncontrolled	3.46	1.22
6. Bot-generated/human-generated	3.24	1.09

7. Minor problem/major problem	3.86	1.12
8. Is more common today than it was 5 years ago	4.18	1.06
9. Should be monitored or addressed by Twitter	4.04	1.10

Based on these results, Twitter users largely feel that fake news on the platform is common, dangerous, concerning, and a major problem. Furthermore, the majority of respondents indicated that fake news is more common on Twitter today than it was five years ago, suggesting that current efforts to combat misinformation might not have a significant impact in preventing its frequency, including its existence, impact, and spread. Nevertheless, respondents as a whole "agreed" that fake news on the platform should be monitored or addressed by Twitter (*M* = 4.04, *SD* = 1.10).

Verification on Twitter

General perceptions regarding verification marks were analyzed by asking respondents, "What is the first word that comes to mind when you hear the term *verification mark*?" The most common response was "check" (*n* = 19), with "Twitter" (*n* = 17) and "celebrity" (*n* = 17) tying for second. The top 10 responses are listed in Table 6.

Table 6

First word that comes to mind when hearing the term "verification mark"

	n =
1. Check	19
2. Twitter	17
3. Celebrity	17
4. Real	14
5. Trust	10
6. Good	10
7. Fact	9
8. Blue	8
9. Security	8
10. Verified	7

When analyzing these responses for common themes, it became apparent that Twitter users largely associate verification marks with *prominence* (*n* = 27), *checkmarks* (*n* = 25), *verified status* (*n* = 25), *authenticity* (*n* = 23), *accuracy* (*n* = 20), *Twitter* (*n* = 17), and *trustworthiness* (*n* = 15).

RQ$_2$: What are the perceptions of Twitter users surrounding the credibility of verification?

Respondents were specifically asked whether verification on Twitter was credible or not credible, with the results on a 5-point semantic scale indicating that users overall leaned toward the credibility of verification (*M* = 2.45, *SD* = 1.24). The remaining measures regarding credibility were captured in additional statements (see Table 7). Notably, each of the individual credibility measures were rated less credible than the credibility measure itself, with each mean value calculating at higher than 2.45.

Table 7

Descriptive Statistics for Perceptions of Credibility of Verification on Twitter (n = 286)

	M	*SD*
1. Credible/not credible	2.45	1.24
2. Accurate/inaccurate	2.46	1.17
3. Not biased/biased	2.96	1.30
4. Can be trusted/cannot be trusted	2.54	1.15
5. Watches out for users' interests/does not watch out for users' interests	2.72	1.22
6. Concerned with community's well-being/not concerned with community's well-being	2.70	1.50
7. Concerned with public interest/not concerned with public interest	2.62	1.22
8. Creates confidence/creates skepticism	2.51	1.20

The mean values for the remaining credibility measures suggest that Twitter users as a whole feel that verification on Twitter tends to meet the components of credibility. However, the mean value regarding bias sat almost directly in the middle of the scale (*M* = 2.96, *SD* = 1.30), the lowest of the individual credibility measures.

RQ$_3$: What are the perceptions of Twitter users surrounding the effectiveness of verification in slowing the spread of fake news?

Responses related to the effectiveness of verification on Twitter in slowing the spread of fake news generally show middling opinions, with a slight incline that supports that they do (see Table 8). Twitter users were specifically asked if they felt verification was effective in reducing the spread of fake news, and although opinions were not overwhelmingly strong, the sentiment fell on the side of agreeing with the statement (*M* = 3.20, *SD* = 1.23). Table 8 provides descriptive statistics regarding all statements.

Table 8

Descriptive Statistics for Perceptions of Effectiveness of Verification on Twitter (n = 286)

	M	*SD*
1. Slows the spread of fake news	3.18	1.18
2. Eliminates the spread of fake news	2.76	1.31
3. Contributes to the spread of fake news	2.82	1.21
4. Is effective in reducing the spread of fake news	3.20	1.23
5. Leads people to find accurate information	3.30	1.12
6. Decreases belief in the accuracy of Tweets from unverified accounts	3.13	1.18
7. Is dangerous	2.52	1.25
8. Is beneficial	3.49	1.13

These results indicate that verification on Twitter does not have a significant impact on eliminating, reducing, or slowing the spread of fake news, nor does it have a large influence on leading people to find accurate information. However, users overall do not consider the feature to be dangerous (*M* = 2.52, *SD* = 1.25), and they typically find the feature beneficial (*M* = 3.49, *SD* = 1.13).

Notices on Twitter

General perceptions regarding Notices on Twitter were analyzed by asking respondents, "What is the first word that comes to mind when you hear the term *social media warning label*?"

The most common response was "caution" (*n* = 19), with "danger" (*n* = 17) and "Facebook" (*n* = 9) rounding out the top three. The top 10 responses are listed in Table 9.

Table 9

First word that comes to mind when hearing the term "social media warning label"

	n =
1. Caution	19
2. Danger	17
3. Facebook	9
4. Warning	8
5. Report	7
6. Misinformation	7
7. Censorship	6
8. Bias	6
9. Useful	6
10. Alert	5

While responses on this question varied greatly, a few common themes emerged from the responses. The top five themes highlighted that social media warning labels are associated with *caution* (*n* = 52), *censorship* (*n* = 19), *misinformation* (*n* = 15), *beneficial* (*n* = 14), and *social media* (*n* = 11).

RQ$_4$: What are the perceptions of Twitter users surrounding the credibility of Notices?

Just as was done with verification, research participants were specifically asked if Notices on Twitter were credible or not credible, with the results on a 5-point semantic scale indicating that users overall leaned toward the credibility of Notices (*M* = 2.57, *SD* = 1.24). The remaining measures regarding credibility of Notices were captured in additional statements (see Table 10).

Table 10

Descriptive Statistics for Perceptions of Credibility of Notices on Twitter (n = 286)

	M	*SD*
1. Credible/not credible	2.57	1.25
2. Accurate/inaccurate	2.57	1.22

3. Not biased/biased	2.94	1.33
4. Can be trusted/cannot be trusted	2.67	1.27
5. Watch out for users' interests/do not watch out for users' interests	2.65	1.28
6. Concerned with community's well-being/not concerned with community's well-being	2.59	1.27
7. Concerned with public interest/not concerned with public interest	2.54	1.26
8. Tell the whole story/do not tell the whole story	3.11	1.25
9. Separate fact and opinion/do not separate fact and opinion	2.82	1.31
10. Factual/not factual	2.73	1.20
11. Do not prevent expression/prevent expression	3.01	1.32
12. Do not censor/censor	2.96	1.30

The majority of mean values for each of the individual credibility measures suggest that Twitter users overall tend to believe that Notices meet the components of credibility. Just as with verification, however, the mean value regarding bias was near the center of the scale (*M* = 2.94, *SD* = 1.33), as was expression (*M* = 3.01, *SD* = 1.32) and censoring (*M* = 2.96, *SD* = 1.30). In addition, respondents indicated by a slight margin that Notices do not tell the whole story (*M* = 3.11, *SD* = 1.25).

In comparing perceived credibility across variables, a t-test revealed that there was no statistically significant difference between the perceived credibility of verification on Twitter and the perceived credibility of Notices on Twitter, *t*(285) = -1.614, *p* = .108

RQ$_{5}$: What are the perceptions of Twitter users surrounding the effectiveness of Notices in slowing the spread of fake news?

Responses related to the effectiveness of Notices on Twitter in slowing the spread of fake news show a slight incline toward the belief that Notices aid in this effort (see Table 11). Table 11 provides descriptive statistics regarding all statements.

Table 11

Descriptive Statistics for Perceptions of Effectiveness of Notices on Twitter (n = 286)

	M	*SD*
1. Slow the spread of fake news	3.28	1.19
2. Eliminate the spread of fake news	2.78	1.29

3. Contribute to the spread of fake news	2.58	1.26
4. Are effective in reducing the spread of fake news	3.17	1.21
5. Lead people to find accurate information	3.38	1.14
6. Are dangerous	2.60	1.33
7. Are beneficial	3.53	1.17
8. Create confidence in the accuracy of news on the platform	3.40	1.16
9. Raise concerns related to censorship	3.16	1.29
10. Are against the First Amendment	2.69	1.34

Twitter users were specifically asked if they felt Notices were effective in reducing the spread of fake news, and the sentiment fell on the side of agreeing with the statement (*M* = 3.17, *SD* = 1.21). Notably, however, Notices were seen as slightly less effective in reducing the spread of fake news than verification (*M* = 3.20, *SD* = 1.23). Nevertheless, a t-test revealed that there was no statistically significant difference between the perceived effectiveness of verification on Twitter slowing the spread of fake news and the perceived effectiveness of Notices on Twitter in slowing the spread of fake news, *t*(285) = .327, *p* = .744.

The data reported suggests that Notices on Twitter have a minor influence on preventing the spread of fake news but that they are not necessarily viewed as a contributor in the spread of fake news. Overall, Notices are seen as beneficial (*M* = 3.53, *SD* = 1.17) and create confidence in the accuracy of news on the platform (*M* = 3.40, *SD* = 1.16).

Two notable measures in Table 11 are the last two statements, which relate to censorship and the First Amendment. The overall attitude of Twitter users is that Notices are not against the First Amendment (*M* = 2.69, *SD* = 1.34) but that they do raise some concerns related to censorship (*M* = 3.16, *SD* = 1.29). These results may suggest that individuals are concerned with social media companies silencing voices on their platforms but feel these companies are allowed to do so as private corporations.

Twitter Perceptions and Political Identification

Although the general perceptions surrounding fake news, verification, and Notices on Twitter offer valuable insights, one of the main goals of this research was to determine if these perceptions differed based on political identification. The existing literature has suggested that political identification may have an impact on attitudes surrounding fake news and social media warning labels, but more research is needed in this area. Additionally, not much information exists regarding political orientation and attitudes towards the credibility and effectiveness of verification marks on Twitter.

RQ$_6$: Do the perceptions surrounding fake news, verification, and Notices differ based on the political identification of the Twitter user?

Fake News and Political Identification. In analyzing the one-word responses about fake news, it is worth noting that those who identified as liberal or very liberal were more inclined to associate fake news with the opposite side of the political spectrum. Of the 180 participants on the left side of the spectrum, 52.2% (n = 94) referenced conservatism in some form while only 10.5% (n = 11) of those on the right referenced liberalism. More specifically of those on the left, 38.9% (n = 70) referenced *Donald Trump* and 6.7% (n =12) referenced *Republicans*, while only 0.3% (n = 2) of those on the right referenced *Democrats*. Interestingly, 13.2% (n = 14) of those on the right referenced *Donald Trump*—a greater percentage than the references to liberalism.

One-way ANOVA tests were performed to compare the effect of political orientation on perceptions of fake news. A complete overview of ANOVA tests can be found in Table 12 with mean values for the measures listed by political group in Table 13.

The results of the ANOVA revealed a significant main effect of the variable *dangerous/not dangerous* (F(3, 282) = 11.04, p < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive fake news as dangerous (*M* = 1.61, *SD* = 0.91; *M* = 1.90, *SD* = 0.95) than those who identify as conservative (*M* = 2.61, *SD* = 1.34), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *concerning/not concerning* (*F*(3, 282) = 11.24, *p* < 0.05. A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive fake news as concerning (*M* = 1.50, *SD* = 0.74; *M* = 1.88, *SD* = 1.10) than those who identify as conservative (*M* = 2.56, *SD* = 1.37), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *minor problem/major problem* (*F*(3, 280) = 8.21, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive fake news as a major problem (*M* = 4.29, *SD* = 0.89) than those who identify as conservative (*M* = 3.44, *SD* = 1.34) and very conservative (*M* = 3.59, *SD* = 1.19), *p* < 0.05. Additionally, those who identify as liberal (*M* = 3.97, *SD* = 0.94) are significantly more likely to perceive fake news as a major problem than those who identify as conservative (*M* = 3.44, *SD* = 1.34).

The results of the ANOVA revealed a significant main effect of the variable *is more common today than five years ago* (*F*(3, 282) = 5.35, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that fake news is more common today than it was five years ago (*M* = 4.50, *SD* = 0.94) than those who identify as very conservative (*M* = 3.74, *SD* = 1.23), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *should be monitored by Twitter* (*F*(3, 282) = 16.04, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that fake news should be monitored or controlled by Twitter (*M* = 4.52, *SD* = 0.72; *M* = 4.27, *SD* = 0.89) than those who identify as conservative (*M* = 3.53, *SD* = 1.22), and very conservative (*M* = 3.44, *SD* = 1.45), *p* < 0.05.

Table 12

Analysis of Variance for Perceptions of Fake News on Twitter Between Groups

Variable	*SS*	*df*	*MS*	*F*	*Sig.*
1. Common/uncommon	6.46	3	2.15	1.71	0.17
2. Dangerous/not dangerous	41.13	3	13.71	11.04	**< 0.01**
3. Concerning/not concerning	43.87	3	14.62	11.24	**< 0.01**
4. Easy to identify/hard to identify	6.05	3	2.02	1.68	0.17
5. Controlled/uncontrolled	8.62	3	2.87	1.94	0.12
6. Bot-generated/human-generated	4.65	3	1.55	1.32	0.27
7. Minor problem/major problem	28.66	3	9.55	8.21	**< 0.01**
8. Is more common today than 5 years ago	17.21	3	5.74	5.35	**< 0.01**
9. Should be monitored by Twitter	50.36	3	16.78	16.04	**< 0.01**

Table 13

Mean Values for Political Orientation and Perceptions of Fake News on Twitter

	V lib	*Lib*	*Con*	*V con*
1. Common/uncommon	2.00	2.17	2.39	2.00
2. Dangerous/not dangerous	1.61	1.90	2.61	2.33
3. Concerning/not concerning	1.50	1.88	2.56	2.30
4. Easy to identify/hard to identify	2.60	2.93	2.92	3.04
5. Controlled/uncontrolled	3.64	3.54	3.34	3.04
6. Bot-generated/human-generated	3.06	3.25	3.27	3.56
7. Minor problem/major problem	4.29	3.97	3.44	3.59
8. Is more common today than 5 years ago	4.50	4.27	3.94	3.74
9. Should be monitored by Twitter	4.52	4.27	3.53	3.44

The results of these tests indicate that those on the left side of the political spectrum are more likely to view fake news as dangerous, concerning, and a major problem. While this data is notable, the most significant variable related to support toward Twitter monitoring or addressing fake news on its platform. Those who identify as very liberal are extremely likely to support this effort (*M* = 4.52, *SD* = 0.72) while those who are very conservative are separated by more than an entire scale point (*M* = 3.44, *SD* = 1.45).

Effectiveness of Verification and Political Identification. In analyzing how political identification impacted one-word responses about verification, no patterns emerged. The results of this question offered no consistency in answers among political groups.

One-way ANOVA tests were performed to compare the effect of political orientation on perceptions of the effectiveness of verification on Twitter in slowing the spread of fake news. A complete overview of ANOVA tests can be found in Table 14 with mean values for the measures listed by political group in Table 15.

The results of the ANOVA revealed a significant main effect of the variable *eliminates the spread of fake news* (*F*(3, 282) = 3.21, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very conservative are significantly more likely to perceive that verification on Twitter eliminates the spread of fake news (*M* = 3.22, *SD* = 1.45) than those who identify as very liberal (*M* = 2.40, *SD* = 1.22), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *contributes to the spread of fake news* (*F*(3, 281) = 9.33, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very conservative and conservative are significantly more likely to perceive that verification on Twitter contributes to the spread of fake news (*M* = 3.56, *SD* = 1.41; *M* = 3.18, *SD* = 1.27) than those who identify as liberal (*M* = 2.51, *SD* = 1.13) and very liberal (*M* = 2.66, *SD* = 1.06), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *is dangerous* (*F*(3, 281) = 17.55, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as conservative are significantly more likely to perceive that verification on Twitter is dangerous

(*M* = 3.27, *SD* = 1.31) than those who identify as liberal (*M* = 2.21, *SD* = 1.06) and very liberal (*M* = 2.05, *SD* = 1.06), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *is beneficial* (*F*(3, 281) = 3.86, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that verification on Twitter is beneficial (*M* = 3.76, *SD* = 0.86) than those who identify as very conservative (*M* = 3.04, *SD* = 1.48), $p < 0.05$.

Table 14

Analysis of Variance for Effectiveness of Verification on Twitter Between Groups

Variable	*SS*	*df*	*MS*	*F*	*Sig.*
1. Slows the spread of fake news	0.61	3	0.20	0.14	0.93
2. Eliminates the spread of fake news	16.24	3	5.41	3.21	**0.02**
3. Contributes to the spread of fake news	37.78	3	12.59	9.33	**< 0.01**
4. Effective in reducing spread of fake news	4.81	3	1.60	1.05	0.37
5. Leads people to find accurate information	5.29	3	1.76	1.40	0.24
6. Decreases belief in unverified Tweets	0.70	3	0.23	0.17	0.92
7. Is dangerous	69.62	3	23.21	17.55	**< 0.01**
8. Is beneficial	14.28	3	4.76	3.86	**0.01**

Table 15

Mean Values for Political Orientation and Perceptions of Effectiveness of Verification on Twitter

	V lib	*Lib*	*Con*	*V con*
1. Slows the spread of fake news	3.11	3.22	3.19	3.11
2. Eliminates the spread of fake news	2.40	2.73	2.94	3.22
3. Contributes to the spread of fake news	2.66	2.51	3.18	3.56
4. Effective in reducing spread of fake news	2.95	3.28	3.25	3.22
5. Leads people to find accurate information	3.44	3.29	3.13	3.56
6. Decreases belief in the accuracy of Tweets from unverified accounts	3.19	3.09	3.10	3.22
7. Is dangerous	2.05	2.21	3.27	2.70
8. Is beneficial	3.76	3.59	3.29	3.04

Results from these tests indicate that those on the left side of the political spectrum are more likely to view verification on Twitter as beneficial while those on the right are more likely to view it as dangerous. Additionally, those on the right are more likely to believe that verification contributes to the spread of fake news.

Credibility of Verification and Political Orientation. One-way ANOVA tests were performed to compare the effect of political orientation on perceptions of the credibility of verification on Twitter. A complete overview of ANOVA tests can be found in Table 16 with mean values for the measures listed by political group in Table 17.

The results of the ANOVA revealed a significant main effect of the variable *credible/not credible* ($F(3, 282) = 5.14$, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that verification on Twitter is credible ($M = 2.26$, $SD = 1.10$; $M = 2.22$, $SD = 1.10$) than those who identify as conservative ($M = 2.85$, $SD = 1.32$), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *accurate/inaccurate* ($F(3, 281) = 5.20$, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that verification on Twitter is accurate ($M = 2.31$, $SD = 1.03$; $M = 2.26$, $M = 1.01$) than those who identify as conservative ($M = 2.89$, $SD = 1.32$), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *not biased/biased* ($F(3, 281) = 5.38$, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as liberal are significantly more likely to perceive that verification on Twitter is not biased ($M = 2.69$, $SD = 1.21$) than those who identify as conservative ($M = 3.33$, $SD = 1.25$) and very conservative ($M = 3.41$, $SD = 1.53$), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *can be trusted/cannot be trusted* ($F(3, 281) = 3.45$, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that verification on Twitter can be trusted (*M* = 2.32, *SD* = 0.95; *M* = 2.41, *SD* = 1.03) than those who identify as conservative (*M* = 2.86, *SD* = 1.30), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *creates confidence/creates skepticism* (*F*(3, 280) = 3.33, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as liberal are significantly more likely to perceive that verification on Twitter creates confidence (*M* = 2.34, *SD* = 1.09) than those who identify as conservative (*M* = 2.83, *SD* = 1.23), *p* < 0.05.

Table 16

Analysis of Variance for Credibility of Verification on Twitter Between Groups

Variable	*SS*	*df*	*MS*	*F*	*Sig.*
1. Credible/not credible	22.76	3	7.59	5.14	**< 0.01**
2. Accurate/inaccurate	20.33	3	6.78	5.20	**< 0.01**
3. Not biased/biased	26.01	3	8.67	5.38	**< 0.01**
4. Can be trusted/cannot be trusted	13.39	3	4.46	3.45	**0.02**
5. Users' interests/not users' interests	7.22	3	2.41	1.63	0.18
6. Community/not community	5.99	3	2.00	1.39	0.25
7. Public interest/not public interest	9.58	3	3.19	2.17	0.09
8. Creates confidence/creates skepticism	13.96	3	4.65	3.33	**0.02**

Table 17

Mean Values for Political Orientation and Perceptions of Credibility of Verification on Twitter

	V lib	*Lib*	*Con*	*V con*
1. Credible/not credible	2.26	2.22	2.85	2.70
2. Accurate/inaccurate	2.31	2.26	2.89	2.41
3. Not biased/biased	2.81	2.69	3.33	3.41
4. Can be trusted/cannot be trusted	2.32	2.41	2.86	2.67
5. Users' interests/not users' interests	2.71	2.56	2.95	2.74
6. Community/not community	2.79	2.53	2.85	2.78
7. Public interest/not public interest	2.48	2.50	2.91	2.59
8. Creates confidence/creates skepticism	2.34	2.35	2.83	2.69

While political groups differed on several points related to credibility of verification on Twitter as determined by the ANOVA and post-hoc tests, it is worth mentioning that all variable means across groups fell on the side of leaning towards credibility, with the exception of bias. In this variable, those who identified as conservative and very conservative were more likely to mark towards biased (*M* = 3.33, *SD* = 1.25; *M* = 3.41, *SD* = 1.53), an indication of a lack of credibility.

Effectiveness of Notices and Political Orientation. The one-word responses about Notices resulted in a few interesting patterns. The theme of *misinformation*—which consisted of misinformation, disinformation, and fake news—was entirely comprised of left leaning individuals (*n* = 15). Liberals were also more likely (*n* = 9) than conservatives (*n* = 2) to identify Notices as beneficial. Those who stated Notices were necessary were also all liberals (*n* = 4).

One-way ANOVA tests were performed to compare the effect of political orientation on perceptions of the effectiveness of Notices on Twitter in slowing the spread of fake news. A complete overview of ANOVA tests can be found in Table 18 with mean values for the measures listed by political group in Table 19.

The results of the ANOVA revealed a significant main effect of the variable *slow the spread of fake news* (*F*(3, 282) = 5.14, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as liberal are significantly more likely to perceive that Notices on Twitter slow the spread of fake news (*M* = 3.53, *SD* = 1.08) than those who identify as conservative (*M* = 2.91, *SD* = 1.22), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *contribute to the spread of fake news* (*F*(3, 281) = 5.20, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very conservative and conservative are significantly more likely to perceive that

Notices on Twitter contribute to the spread of fake news (*M* = 3.26, *SD* = 1.43; *M* = 3.06, *SD* = 1.33) than those who identify as liberal (*M* = 2.29, *SD* = 1.05) and very liberal (*M* = 2.19, *SD* = 1.16), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *lead people to find accurate information* (*F*(3, 281) = 5.38, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that Notices on Twitter lead people to find correct information (*M* = 3.65, *SD* = 1.03) than those who identify as conservative (*M* = 3.11, *SD* = 1.23) and very conservative (*M* = 2.85, *SD* = 1.38), *p* < 0.05. Additionally, those who identify as liberal are more likely to perceive that Notices on Twitter lead people to find correct information (*M* = 3.53, *SD* = 1.00) than those who identify as very conservative (*M* = 3.11, *SD* = 1.23), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *are dangerous* (*F*(3, 281) = 3.45, *p* < 0.05. A post-hoc Tukey test revealed that those who identify as very conservative and conservative are significantly more likely to perceive Notices on Twitter as dangerous (*M* = 3.30, *SD* = 1.23; *M* = 3.08, *SD* = 1.38) than those who identify as very liberal (*M* = 2.11, *SD* = 1.23), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *are beneficial* (*F*(3, 281) = 3.45, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that Notices on Twitter are beneficial (*M* = 3.95, *SD* = 0.86) than those who identify as conservative (*M* = 3.08, *SD* = 1.74) and very conservative (*M* = 3.19, *SD* = 2.23), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *create confidence in news on the platform* (*F*(3, 280) = 3.33, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter create confidence in news on the platform (*M* = 3.69, *SD* = 1.04; *M* = 3.62, *SD* = 0.95) than those who identify as conservative (*M* = 2.99, *SD* = 1.32), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *raise concerns related to censorship* (*F*(3, 281) = 3.45, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very conservative and conservative are significantly more likely to perceive that Notices on Twitter raise concerns related to censorship (*M* = 3.85, *SD* = 1.17; *M* = 3.72, *SD* = 1.14) than those who identify as liberal (*M* = 2.98, *SD* = 1.18) and very liberal (*M* = 2.50, *SD* = 1.32), *p* < 0.05.

The results of the ANOVA revealed a significant main effect of the variable *are against the First Amendment* (*F*(3, 281) = 3.45, *p* < 0.05). A post-hoc Tukey test revealed that those who identify as very conservative and conservative are significantly more likely to perceive that Notices on Twitter are against the First Amendment (*M* = 3.59, *SD* = 1.28; *M* = 3.27, *SD* = 1.27) than those who identify as liberal (*M* = 2.41, *SD* = 1.18) and very liberal (*M* = 2.10, *SD* = 1.30), *p* < 0.05.

Table 18

Analysis of Variance for Effectiveness of Notices on Twitter Between Groups

Variable	*SS*	*df*	*MS*	*F*	*Sig.*
1. Slow the spread of fake news	20.90	3	6.97	5.16	**< 0.01**
2. Eliminate the spread of fake news	2.41	3	0.80	0.48	0.70
3. Contribute to the spread of fake news	49.72	3	16.57	11.61	**< 0.01**
4. Effective in reducing spread of fake news	7.42	3	2.47	1.70	0.17
5. Lead people to find accurate information	19.99	3	6.66	5.41	**< 0.01**
6. Are dangerous	51.34	3	17.11	10.69	**< 0.01**
7. Are beneficial	33.58	3	11.19	8.85	**< 0.01**
8. Create confidence in news on the platform	27.51	3	9.17	7.28	**< 0.01**
9. Raise concerns related to censorship	68.53	3	22.84	15.92	**< 0.01**
10. Are against the First Amendment	79.45	3	26.48	17.24	**< 0.01**

Table 19

Mean Values for Political Orientation and Perceptions of Effectiveness of Notices on Twitter

	V lib	*Lib*	*Con*	*V con*
1. Slow the spread of fake news	3.40	3.53	2.91	3.00
2. Eliminate the spread of fake news	2.63	2.85	2.82	2.67
3. Contribute to the spread of fake news	2.19	2.29	3.06	3.26
4. Effective in reducing spread of fake news	3.29	3.29	2.92	3.11
5. Lead people to find accurate information	3.65	3.53	3.11	2.85
6. Are dangerous	2.11	2.38	3.08	3.30
7. Are beneficial	3.95	3.70	3.08	3.19
8. Create confidence in news on the platform	3.69	3.62	2.99	3.04
9. Raise concerns related to censorship	2.50	2.98	3.72	3.85
10. Are against the First Amendment	2.10	2.41	3.27	3.59

Notably, one of the two variables that did not return a statistically significant result between groups was *effective in reducing the spread of fake news*, suggesting that there is no difference between groups in this attitude. However, other measures returned significant results between groups related to the effectiveness of Notices in slowing the spread of fake news, such as *slow the spread of fake news*, indicating that the topic is more complex than one variable. Because groups differed on eight of the 10 measures, it appears that there is a significant difference between political groups.

Credibility of Notices and Political Identification. One-way ANOVA tests were performed to compare the effect of political orientation on perceptions surrounding the credibility of Notices on Twitter. A complete overview of ANOVA tests can be found in Table 20 with mean values for the measures listed by political group in Table 21.

The results of the ANOVA revealed a significant main effect of the variable *credible/not credible* ($F(3, 282) = 13.77, p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter are credible (*M* = 2.18, *SD* = 1.03; *M* = 2.26, *SD* = 1.02) than those who identify as conservative (*M* 3.20, *SD* = 2.96) and very conservative (*M* = 2.96, *SD* = 2.73), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *accurate/inaccurate* (*F*(3, 282) = 9.13, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter are accurate (*M* = 2.18, *SD* = 1.05; *M* = 2.36, *SD* = 1.06) than those who identify as conservative (*M* = 3.05, *SD* = 1.27), $p < 0.05$. Additionally, those who identify as very liberal are significantly more likely to perceive that Notices on Twitter are accurate (*M* = 2.18, *SD* = 1.05) than those who identify as very conservative (*M* = 2.96, *SD* = 1.53), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *not biased/biased* (*F*(3, 281) = 5.06, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter are not biased (*M* = 2.70, *SD* = 1.17; *M* = 2.71, *SD* = 1.21) than those who identify as conservative (*M* = 3.32, *SD* = 1.41), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *can be trusted/cannot be trusted* (*F*(3, 281) = 11.50, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter can be trusted (*M* = 2.18, *SD* = 1.06; *M* = 2.45, *SD* = 1.13) than those who identify as conservative (*M* = 3.22, *SD* = 1.27) and very conservative (*M* = 3.15, *SD* = 1.61), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *watch out for users' interests/do not watch out for users' interests* (*F*(3, 282) = 17.53, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal, liberal, and very conservative are significantly more likely to perceive that Notices on Twitter watch out for users' interests (*M* =

2.15, *SD* = 1.14; *M* = 2.36, *SD* = 1.10; *M* = 2.74, *SD* = 1.56) than those who identify as conservative (*M* = 3.43, *SD* = 1.19), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *concerned with community's well-being/not concerned with community's well-being* (*F*(3, 280) = 8.26, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter are concerned with the community's well-being (*M* = 2.19, *SD* = 1.14; *M* = 2.41, *SD* = 1.18) than those who identify as conservative (*M* = 3.13, *SD* = 1.23), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *concerned with public interest/not concerned with public interest* (*F*(3, 282) = 14.72, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter are concerned with public interest (*M* = 2.05, *SD* = 1.12; *M* = 2.26, *SD* = 1.11) than those who identify as conservative (*M* = 3.15, *SD* = 1.23) and very conservative (*M* = 3.07, *SD* = 1.41), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *tell the whole story/do not tell the whole story* (*F*(3, 282) = 5.02, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal and liberal are significantly more likely to perceive that Notices on Twitter tell the whole story (*M* = 2.81, *SD* = 1.16; *M* = 2.97, *SD* = 1.15) than those who identify as conservative (*M* = 3.53, *SD* = 1.25), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *separate fact and opinion/do not separate fact and opinion* (*F*(3, 281) = 12.31, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that Notices on Twitter separate fact and opinion (*M* = 2.35, *SD* = 1.20; *M* = 2.56, *SD* = 1.09) than those who identify as conservative (*M* = 3.46, *SD* = 1.36) and very conservative (*M* = 3.19, *SD* = 1.52), $p < 0.05$. Additionally, those who identify as liberal are significantly more likely to perceive that Notices on Twitter separate fact and opinion (*M* = 2.56, *SD* = 1.09) than those who identify as conservative (*M* = 3.46, *SD* = 1.36), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *factual/not factual* (*F*(3, 280) = 14.90, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that Notices on Twitter are factual (*M* = 2.16, *SD* = 0.98) than those who identify as conservative (*M* = 3.33, *SD* = 1.15) and very conservative (*M* = 3.11, *SD* = 1.50), $p < 0.05$. Additionally, those who identify as liberal are significantly more likely to perceive that Notices on Twitter are factual (*M* = 2.54, *SD* = 1.07) than those who identify as conservative (*M* = 3.33, *SD* = 1.15), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *do not prevent expression/prevent expression* (*F*(3, 281) = 10.60, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that Notices on Twitter do not prevent expression (*M* = 2.27, *SD* = 1.24) than those who identify as conservative (*M* = 3.39, *SD* = 1.30) and very conservative (*M* = 3.70, *SD* = 1.38), $p < 0.05$. Additionally, those who identify as liberal are significantly more likely to perceive that Notices on Twitter do not prevent expression (*M* = 2.94, *SD* = 1.21) than those who identify as very conservative (*M* = 3.70, *SD* = 1.38), $p < 0.05$.

The results of the ANOVA revealed a significant main effect of the variable *do not censor/censor* (*F*(3, 281) = 6.99, $p < 0.05$). A post-hoc Tukey test revealed that those who identify as very liberal are significantly more likely to perceive that Notices on Twitter do not censor (*M* = 2.45, *SD* = 1.24; *M* = 2.88, *SD* = 1.18) than those who identify as conservative (*M* = 3.35, *SD* = 1.32) and very conservative (*M* = 3.37, *SD* = 1.52), *p* < 0.05.

Table 20

Analysis of Variance for Credibility of Notices on Twitter Between Groups

Variable	*SS*	*df*	*MS*	*F*	*Sig.*
1. Credible/not credible	56.47	3	18.82	13.77	**< 0.01**
2. Accurate/inaccurate	37.38	3	12.46	9.13	**< 0.01**
3. Not biased/biased	25.71	3	8.57	5.06	**< 0.01**
4. Can be trusted/cannot be trusted	50.21	3	16.74	11.50	**< 0.01**
5. Users' interests/not users' interests	73.76	3	24.59	17.53	**< 0.01**
6. Community/not community	37.13	3	12.38	8.26	**< 0.01**
7. Public interest/not public interest	61.34	3	20.45	14.72	**< 0.01**
8. Tell whole story/do not tell whole story	22.54	3	7.51	5.02	**< 0.01**
9. Separate fact and opinion/do not	56.60	3	18.87	12.31	**< 0.01**
10. Factual/not factual	56.38	3	18.79	14.90	**< 0.01**
11. Do not prevent expression/prevent	50.43	3	16.81	10.60	**< 0.01**
12. Do not censor/censor	33.60	3	11.20	6.99	**< 0.01**

Table 21

Mean Values for Political Orientation and Perceptions of Credibility of Notices on Twitter

	V lib	*Lib*	*Con*	*V con*
1. Credible/not credible	2.18	2.26	3.20	2.96
2. Accurate/inaccurate	2.18	2.36	3.05	2.96
3. Not biased/biased	2.70	2.71	3.32	3.37
4. Can be trusted/cannot be trusted	2.18	2.45	3.22	3.15
5. Watch out for users' interests/do not watch out for users' interests	2.15	2.36	3.43	2.74
6. Concerned with community/not concerned with community	2.19	2.41	3.13	2.77
7. Concerned with public interest/not concerned with public interest	2.05	2.26	3.15	3.07
8. Tell the whole story/do not tell the whole story	2.81	2.97	3.53	3.22
9. Separate fact and opinion/do not separate fact and opinion	2.35	2.56	3.46	3.19
10. Factual/not factual	2.16	2.54	3.33	3.11
11. Do not prevent expression/prevent expression	2.27	2.94	3.39	3.70
12. Do not censor/censor	2.45	2.88	3.35	3.37

The results of these tests indicate that political orientation has a significant influence on the perceptions surrounding the credibility of Notices on Twitter. Generally speaking, the data suggests that those on the left side of the political spectrum are more likely to view Notices as credible while those on the right side of the scale are more likely to view Notices as not credible.

Discussion

This study analyzes the perceptions surrounding fake news, verification, and Notices on Twitter. Verification and Notices were examined to determine if they are viewed as credible and effective in slowing the spread of fake news on the platform. All variables were analyzed in light of political orientation to determine if differences existed between those who identify as very liberal, liberal, conservative, and very conservative. The research findings have important implications for politics, social media, communications theory, and Twitter specifically.

The Prevalence of Fake News

Results of this study indicate that Twitter users largely view fake news as dangerous, concerning, and a significant issue. These findings support the literature, which illustrates that the majority of individuals who use social media are concerned with fake news being used as a weapon (Edelman, 2020), wonder if the information they are accessing is accurate (Shearer & Matsa, 2018; Edelman, 2020), and question the credibility of news from these sources (Rubin, 2019).

The most likely reason for these beliefs is due to the prevalence of fake news in today's world, particularly on social media. Those who participated in this research consider fake news to be common on Twitter, and one of the highest mean values from survey results indicates that most respondents feel like fake news is more prevalent on the platform today than it was five years ago (*M* = 4.18). This finding could, in part, suggest that current efforts to combat the spread of misinformation on Twitter might not have as strong of an impact as desired. This subject will be explored further later in the discussion in connection with additional findings from this research, including the fact that users do not view warning labels as particularly effective in slowing the spread of fake news.

The increased amount of fake news on Twitter is also likely due to an increased number of users who intentionally or unintentionally contribute to the spread of misinformation. In just under five years—from Q1 of 2017 to Q3 of 2021—the number of daily active users on Twitter has practically doubled, increasing from 109 million to 211 million (Statista, 2022). While it is impossible to know how many of these users are sharing fake news in online conversations, it is undeniable that the sheer increase in daily active users would also increase the amount of misinformation being posted on the platform—an issue that stems from online realities, such as the ability to conceal one's identity and safely hide behind the screen.

Although social media platforms cannot vet every individual who creates an account, perhaps these companies should implement stronger requirements and features to better ensure the legitimacy of accounts. With the majority of research participants indicating that they believe Twitter should monitor or address fake news on the platform, it is clear that users at large expect more from platforms than what they are currently doing to combat the spread of misinformation. These findings thus have important implications for social media platforms who are attempting to connect with the needs of their users.

Fake News and Politics

Another important outcome of this research is its insights regarding fake news and politics. The results of this study support the finding by van der Linden et al. (2020) that liberals and conservatives associate fake news with outlets of the opposite political identification, with several right-wing individuals in the study mentioning CNN and several left-wing individuals mentioning Fox News. This is not surprising given the heated political landscape in the United

States and the growing skepticism that exists regarding news networks, particularly networks with the opposing perspective of the viewer.

Expanding upon this finding, results also indicate that the term “fake news” is largely associated with conservatism—particularly among liberals but also among some conservatives. It is important to note, however, that the association among conservatives in this research study is solely with Donald Trump. The connection to conservatism at large—and the significant amount of connection to Donald Trump—could be explained by the fact that Donald Trump popularized the term, which would also explain why several conservatives stated *Trump* in their response. Nevertheless, a deeper analysis suggests that there is more to this finding.

As noted in the results, more than 50% of left-leaning study participants associate fake news with *Donald Trump*, *Fox News*, *Republicans*, and *conservatives*, while just over 10% of right-leaning participants referenced words connected with liberalism. While the measures of this study do not fully explain the reasons for these results, the findings suggest that liberals are much more likely than conservatives to connect fake news with the opposing political wing. This could be due to any number of reasons, but one possible explanation is the way that liberals tend to feel about conservatives. Results from a 2020 Pew Research Center survey indicate that liberals are significantly less likely to date a conservative than a conservative is to date a liberal (Brown, 2020), and a 2014 Pew Research Center survey revealed that liberals are more likely to unfriend someone on a social network over differences in politics (Mitchell et al., 2014). Gramlich (2016) also noted that Clinton supporters had a more difficult time respecting Trump supporters than the other way around. Findings like these suggest liberals may tend to be skeptical of the character and integrity of conservatives as individuals, which could explain why the majority of left-leaning individuals in this study associated fake news with conservatism.

In contrast, conservatives were more inclined to identify characteristics of fake news, such as *fraud*. In some ways, this is surprising given the fact that Donald Trump often associated the term with left-leaning news networks such as CNN and MSNBC, which certainly could have influenced the way right-leaning participants responded in this survey. Interestingly, however, there was a larger percentage of conservative participants who referenced Donald Trump than those who referenced liberalism—13% to about 10% respectively.

These findings are important because they indicate that there are significant differences in the way people think of fake news, especially based on their political orientation. Results suggest that liberals are more likely to associate the term with the opposing political viewpoint, perhaps seeing the term through a cynical lens. At the same time, results suggest that conservatives are more inclined to connect fake news with other things that have become associated with the term, such as fraud, CNN, and Donald Trump. As shown by this study, the term "fake news" may be more politically charged for liberals than it is for conservatives.

Verification Inconsistencies

Perceptions surrounding verification as analyzed in this study suggest that the feature is generally viewed as credible; however, Twitter users overall don't feel that verification has much of an impact on reducing the spread of fake news and do not feel overly confident that the feature leads individuals to find accurate information. These results are not necessarily surprising given the fact that the main purpose of the verification mark as utilized by Twitter is to authenticate users (Cohen & Sutton, 2018; Twitter Support, 2017; "Verification FAQ," n.d.). Nevertheless, the platform hoped that updates made to the verification process in 2021 would also increase the credibility of content on the platform and would better allow users to determine if conversations are trustworthy (Twitter Inc., 2021). And although users may typically view news from verified accounts as more credible than not, the feature itself does not appear to ensure that users are receiving information that is any more accurate than that from unverified accounts.

These findings raise important points related to the concept of verification, including the reality that inconsistencies continue to exist in Twitter's explanation of the feature. The platform has long emphasized that verification does not equal endorsement, but this is contradicted by the explanation of its current process. After all, and as mentioned above, Twitter has suggested that verified status should help users better determine if content is trustworthy (Twitter Inc., 2021). This explanation implies that Twitter views verified accounts as better sources of credible information, even though any generic user could publish content just as credible—if not more credible—than someone with a verification mark. And since the platform limits verification to select groups, including government, journalists, entertainers, and athletes ("Verification FAQ," n.d.), the feature consequently and inevitably creates endorsement only for specific categories as judged by Twitter. Twitter thus appears to be suggesting that it has more confidence in its users with prominent societal status than it does in its everyday users. These realities should concern Twitter users, because it indicates that their opinions are automatically viewed as less credible than those with verified status.

Furthermore, research has shown that users with a verified status often receive enhanced credibility (Paul et al., 2019). This suggests that regardless of how Twitter defines the feature, it can be difficult to alter the way that individuals have been conditioned to think. It is quite likely that because verification on Twitter is represented by a checkmark, some users may associate the feature with credibility without consciously realizing that they are doing so.

These explanations are possibly some of the reasons that credibility of verification as analyzed in this study was viewed near the middle of the scale for the measure *biased*. With the platform limiting verification to select groups, many users may feel that their voices are viewed as less important—a point analyzed in previous paragraphs. Perhaps more frustrating to users is the knowledge that Twitter previously allowed all accounts to apply for verified status but later removed that option. While Twitter has justified their decision to alter the process of verification, it does not explain why an individual of prominence is any more qualified to be verified than a general user, nor does it explain why verified accounts are seen by Twitter as more likely to enhance the quality and credibility of conversations on the platform.

With all of this in mind, it is worth mentioning that Twitter's current application of verification might actually be harming the credibility of the feature—both perceived and real. The limitation of the feature to certain groups—and the fact that it is largely extended to only influential and well-known individuals—calls into question its ultimate function and purpose. Opening verification to all users may not solve the dilemma because it would dilute the value of the feature, as noted by Edgerly and Vraga (2019). Nevertheless, the platform continues to dance around the concept of endorsement while also discreetly suggesting that verified accounts should be viewed as more credible. Additionally, Twitter's frequent changes to the verification process over only a few years makes it difficult for users to keep up with the purposes behind the alterations, potentially causing skepticism regarding the feature for some users.

Twitter, as well as other platforms who utilize a verification mark, should thus consider both how and why the feature is being used and consider ways that they can ensure that verified status is meeting its purpose, fulfilling its definition, and avoiding bias. The challenge for Twitter and other platforms will be creating an environment where verification can be viewed as credible and unbiased while also avoiding the appearance of endorsement. This may require a complete overhaul of the verification system as it currently exists.

New Applications of Gatekeeping

As illustrated in the literature, the theory of gatekeeping has expanded due to new technologies and the increased simplicity of news distribution by everyday consumers (Benham, 2020; Cormode & Krishnamurthy, 2008). While traditional newsroom gatekeeping has long been accepted by consumers, this research suggests that individuals are less inclined to support new applications of gatekeeping, such as warning labels and other efforts implemented by social media platforms.

The findings of this particular study—in connection with the existing literature—call into question the use of top-down gatekeeping on social media. Research has suggested that bottom-up efforts used to combat the spread of fake news might be viewed as more credible and thus might be more effective (Colliander, 2019). It is no secret that users at large mistrust social media platforms, which means that top-down efforts utilized by these companies are likely to result in skepticism from consumers. Furthermore, the past failures of social media platforms to effectively implement and utilize features like verification and warning labels have likely led to cynicism.

As noted by Colliander (2019), some social media outlets are moving away from the use of warning labels, in part because users are more likely to rely on other consumers when it comes to determining what is credible. Additionally, 54% of Americans indicated that they were more concerned about social media platforms censoring truth than they were about fake news (Kemp & Ekins, 2021). Perhaps these are some of the reasons that Notices are overwhelmingly viewed as not especially credible or effective in reducing the spread of fake news. Furthermore, the fact that 75% of social media users do not trust platforms to make fair content moderation decisions (Kemp & Ekins, 2021) may also play a role in this attitude.

Gatekeeping vs. Censorship

Results from this study go so far as to suggest that top-down gatekeeping applications, including warning labels, raise concerns related to censorship. The most likely explanation for why these gatekeeping practices are more likely to be opposed—and are sometimes viewed as censoring—is because consumer voices are the ones being silenced. Although news published through traditional newsrooms can silence voices, there is no doubt that individuals will feel it more personally when *their* words on *their* accounts are flagged or removed. Online platforms are designed to offer a voice to all, which enables individuals to shape the news and influence public discourse (Messner & Garrison, 2009; Poor, 2006). Thus, when social media companies inhibit users from contributing to the conversation, some consumers will feel that there is an overreach in content moderation.

This is particularly relevant in the political climate that currently exists in the United States. As illustrated by this research, the attitudes towards Notices are likely to differ based on the political identification of the user, with those on the right viewing them as less credible than those on the left. This is not surprising given previous research, which shows that conservatives feel social media platforms are doing too much content moderation while liberals feel they are not doing enough (Kemp & Ekins, 2021; Stjernfelt & Lauritzen, 2020).

Additionally, the results of this research indicate that those on the right are more likely to view Notices as a method of censoring and preventing expression. This largely supports findings from the existing literature, which illustrate that some individuals feel that it is not in the best interest of the public to allow platforms to prevent expression (Daseler, 2019). One reason that those on the right may feel more passionate about censorship is because conservatives are more likely to be censored and have their accounts suspended (Stjernfelt & Lauritzen, 2020; Kemp &

Ekins, 2021). Furthermore, this could be a larger reflection of the difference in attitudes toward rules and regulation that exists between liberals and conservatives. Just as conservatives prefer small government, this research suggests that conservatives also prefer small tech. It is worth noting, therefore, that warning labels are likely to be less supported by those on the right because of both the political bias in content moderation and conservative attitudes toward regulation by establishments at large.

Interestingly, however, censorship concerns on Twitter still appear to exist to some extent in spite of political orientation. This may be connected with the perceived credibility of Notices, since survey variables such as *tell the whole story/do not tell the whole story* and *do not prevent expression/prevent expression* fell on the less credible side of the general scale (M = 3.11, M = 3.01). While those on the right ranked these measures as significantly less credible than those on the left, even those who identified as very liberal and liberal were not very confident that Notices tell the whole story (M = 2.81, M = 2.97).

On the other hand, perceptions surrounding verification differed in a few measures based on political orientation. The most notable distinction from survey results is that those on the right side of the spectrum were more likely to view verification as dangerous while those on the left were more likely to view it as beneficial. This finding supports the aforementioned reality that conservatives typically want less regulation and are more inclined to distrust establishments. The fact that those who identified as conservative and very conservative view verification on Twitter as biased (M = 3.33, M = 3.41) supports the idea that individuals on the right tend to be skeptical of most regulatory bodies.

These findings are important because of the implications they have on communications theory. As gatekeeping continues to expand and evolve, researchers should consider how these changes are impacting consumers and their attitudes toward gatekeeping efforts. As illustrated previously, individuals appear to be increasingly more skeptical of increased regulations, including company executives and policies that impact the ability to consume news more freely. These realities are likely to continue altering attitudes about gatekeeping, especially in the digital age where consumers have become accustomed to the ease of accessing information without restriction.

Connected with this is the impact that these efforts are having on attitudes toward social media platforms. These results suggest that social media platforms should be conscious of top-down gatekeeping efforts. Distrust in platforms is real, and gaining that trust back requires that companies are involved in efforts that communicate a willingness to meet consumer needs and listen to users rather than continuing to communicate executive power.

Notices vs. Verification

To further analyze top-down gatekeeping efforts on Twitter, results suggest that warning labels on the platform are viewed as slightly credible, though they are seen as less credible than verification. Furthermore, Notices are seen as slightly less effective in slowing the spread of fake news than verification. This is significant because Notices are largely viewed as the main feature currently utilized by Twitter to combat the spread of misinformation on the platform ("Notices on Twitter," n.d.).

As indicated in the existing literature, warning labels are largely controversial (Clayton et al., 2020), can be ineffective (Fardouly & Holland, 2018; Pennycook et al., 2020), and can have a negative impact on users (Pennycook et al., 2020). If Notices are viewed as less effective and less credible than verification—a feature on the platform which is primarily used to authenticate (Cohen & Sutton, 2018; Twitter Support, 2017; "Verification FAQ," n.d.) rather than combat fake news—then it is quite possible that Notices are not producing the results Twitter is expecting or desiring.

Freedom of Consumption

With both features being viewed essentially the same when it comes to combatting the spread of fake news, it is worth discussing if the use of Notices should be eliminated—at least for posts related to potential misinformation, since Twitter also utilizes the feature for offensive and graphic content. As identified earlier, it is quite possible that a few of the reasons for these attitudes toward warning labels are opposition to top-down gatekeeping, distrust in social media platforms, and concerns related to censorship. All of these reasons are interwoven into a larger tapestry that primarily comes down to the desire for "freedom of consumption."

Because this study analyzes Twitter users in the United States, it is important to highlight specific circumstances related to this audience that drive freedom of consumption. These include freedom of speech as protected by the First Amendment and the ease to instantly access information from just about anywhere at any time in the digital age. The main purpose of warning labels as utilized by Twitter and other social media platforms is to reduce, limit, or hide the content that users can consume. Not only does this compete with traditional American values that have largely enabled freedom of consumption, but it also expands and enhances the power of those in executive positions. Furthermore, social media platforms have long communicated that they connect individuals and lend everyone a voice—however, their efforts to moderate content contradict the supposed purpose of their existence.

Additionally, the internet has provided users with a path around traditional methods of gatekeeping. This has enabled them to access an unlimited amount of information and has also allowed them to determine what news they receive. With social media platforms implementing new methods of gatekeeping, individuals are once again facing the realities associated with people and policies that determine what they receive. Because consumers are no longer used to this in the digital age, it is likely that they find these realities frustrating—and ultimately may view efforts to limit what they receive as ineffective and not credible, as illustrated by the results of this research. With warning labels standing in the way of freedom of consumption, social media platforms may want to consider how they can respect the desires of users while also implementing ideas that will help address the concern surrounding fake news.

Conclusion

One limitation of this research was the limited number of participants who identified as very conservative. At times, it seemed as though the data for this group was not as accurate as expected, likely due to the small size of the group, which amounted to only 9.4% of the research participants (*n* = 27). Research has shown that this group makes up a relatively small percentage of Twitter users, with only 12% identifying as very conservative (Wojcik & Hughes, 2019). Nevertheless, future studies would benefit from ensuring that they have a reputable number of participants who identify as very conservative to be able to analyze this group in greater depth and with greater accuracy.

Additionally, this study only views perceptions of fake news, verification, and Notices (i.e., warning labels) on Twitter and by users of the platform. Future research should investigate if these results hold true across other social media sites that utilize similar features.

With the credibility of Notices being called into question by those on the political right, and with the effectiveness of Notices not being viewed as particularly strong by Twitter users at large, it is worth asking if Notices are the best approach to combat the spread of fake news. The existing literature has also raised questions about Notices as a method to slow the spread of fake news, indicating that regardless of what social media platforms believe, their users are not certain that this method is the right answer. Nevertheless, this study does suggest that there is strong support for Twitter monitoring or addressing fake news on the platform.

Because of the complexity associated with content moderation, Twitter simply won't be able to fully address the problem of fake news, especially in a way that is fully supported by all 330 million of its diverse users. However, the platform should certainly consider alternative methods to combat the spread of fake news. As indicated in the literature, the platform is currently testing a new feature known as Birdwatch, which is a community-based effort against misinformation (Coleman, 2021). This relies on bottom-up gatekeeping efforts, which may be viewed by users as more credible and effective than top-down efforts. Twitter should continue to refine this feature and work on implementation to see if it is viewed as more effective than warning labels. The platform could also seek ways to better help users become responsible media consumers rather than passive participants who rely on others to tell them what is true and what is not.

These recommendations also apply to other social media platforms who are likely facing similar challenges when it comes to combatting the spread of fake news. Consumers across these platforms are likely to have similar views to those on Twitter and may even be less supportive of content moderation due to political orientation differences that may exist on other channels. At a minimum, all platforms should consider the use of alternative methods in the fight against fake news.

References

About public and protected Tweets. (n.d.). Twitter. Retrieved November 7, 2020, from https://help.twitter.com/en/safety-and-security/public-and-protected-tweets

About suspended accounts. (n.d.). Twitter. Retrieved January 29, 2022, from https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts

About verified accounts. (n.d.). Twitter. Retrieved November 7, 2020, from https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Adams, B. (2021, May 22). PolitiFact retracts Wuhan lab theory 'fact-check.' *Washington Examiner*. https://www.washingtonexaminer.com/opinion/politifact-retracts-wuhan-lab-theory-fact-check

Allcott, H., & Gentzcow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, M. (2012). What was Web 2.0? Versions as the dominant mode of internet history. *New Media & Society*, *15*(2), 260-275. https://doi.org/10.1177/1461444812451567

Amazeen, M. A., Vargob, C. J., & Hopp, T. (2019). Reinforcing attitudes in a gatewatching news era: Individual level antecedents to sharing fact-checks on social media. *Communication Monographs*, *86*(1), 112–132. https://doi.org/10.1080/03637751.2018.1521984

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59-79. https://doi.org/10.1177/1077699015606057

Azer, M., Taha, M., Zayed, H. H., & Gadallah, M. (2021). Credibility detection on Twitter news using machine learning approach. *I.J. Intelligent Systems and Applications*, *3*, 1-10. https://www.doi.org/10.5815/ijisa.2021.03.01

Benham, J. (2020). Best practices for journalistic balance: Gatekeeping, imbalance and the fake news era. *Journalism Practice*, *14*(7), 791-811. https://doi.org/10.1080/17512786.2019.1658538

Blank, G. (2013). Who creates content? *Information, Communication & Society*, *16*(4), 590-612. https://doi.org/10.1080/1369118X.2013.777758

Bowles, N. (2017, November 10). Twitter, facing another uproar, pauses its verification process. *The New York Times*. https://www.nytimes.com/2017/11/09/technology/jason-kessler-twitter-verification.html

Brown, A. (2020, April 24). *Most Democrats who are looking for a relationship would not consider dating a Trump voter*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/04/24/most-democrats-who-are-looking-for-a-relationship-would-not-consider-dating-a-trump-voter/

Calvillo, D. P., Ross, B. J., Garcia, R. J., Smelter, T. J., & Rutchick, A. M. (2020). Political ideology predicts perceptions of the threat of COVID-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science*, *11*(8), 1119-1128. https://doi.org/10.1177/1948550620940539

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the 20th international conference on World Wide Web*, pp. 675–684. ACM Digital Library. https://doi.org/10.1145/1963405.1963500

Choi, J. H., Watt, J. H., & Watt, M. L. (2006). Perceptions of news credibility about the war in Iraq: Why war opponents perceived the internet as the most credible medium. *Journal of Computer-Mediated Communication*, *12*(1), 209-229. https://doi.org/10.1111/j.1083-6101.2006.00322.x

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolf, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*, 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Clement, J. (2019, August 14). *Twitter: Number of monthly active users 2010-2019*. Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Clement, J. (2020, October 29). *Leading countries based on number of Twitter users as of October 2020*. Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Cohen, D., & Carter, P. (2010). WHO and the pandemic flu "conspiracies." *British Medical Journal*, *340*(7759), 1274-1279. https://doi.org/10.1136/bmj.c2912

Cohen, D., & Sutton, K. (2019, October 21). Twitter and YouTube rethink verification, causing problems for content creators and brands. *ADWEEK*. https://www.adweek.com/digital/twitter-and-youtube-rethink-verification-causing-problems-for-content-creators-and-brands/

Coleman, K. (2021, January 25). *Introducing Birdwatch, a community-based approach to misinformation*. Twitter. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, *97*, 202-215. https://doi.org/10.1016/j.chb.2019.03.032

Cormode, G., & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday*, *13*(6). https://doi.org/10.5210/fm.v13i6.2125

Darcy, O. (2019, March 22). How Twitter's algorithm is amplifying extreme political rhetoric. *CNN Business*. https://www.cnn.com/2019/03/22/tech/twitter-algorithm-political-rhetoric/index.html

Daseler, G. (2019). Web of lies: The challenges of free speech in the age of social media. *The American Conservative*, *18*(4), 43-48.

Dizikes, P. (2018, March 8). *Study: On Twitter, false news travels faster than true stories*. MIT News. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

Dochterman, M., & Stamp, G. (2010). Part 1: The determination of web credibility: A thematic analysis of web user's judgments. *Qualitative Research Reports in Communication*, *11*, 37-43. https://doi.org/10.1080/17459430903514791

Edelman. (2020, January 19). *2020 Edelman Trust Barometer*. Retrieved November 9, 2020, from https://www.edelman.com/trustbarometer

Edgerly, S., & Vraga, E. K. (2019). The blue check of credibility: Does account verification matter when evaluating news on Twitter? *Cyberpsychology, Behavior, and Social Networking*, *22*(4), 283-287. https://doi.org/10.1089/cyber.2018.0475

Fabina, J. (2021, April 29). *Despite pandemic challenges, 2020 Election had largest increase in voting between presidential elections on record*. United States Census Bureau. https://www.census.gov/library/stories/2021/04/record-high-turnout-in-2020-general-election.html

Faragó, L., Kende, A., & Krekó, P. (2020). We only believe in news that we doctored ourselves: The connection between partisanship and political fake news. *Social Psychology*, *51*(2), 77–90. https://doi.org/10.1027/1864-9335/a000391

Fessler, D. M., Pisor, A. C., & Holbrook, C. (2017). Political orientation predicts credulity regarding putative hazards. *Psychological Science*, *28*, 651–660. https://doi.org/10.1177/0956797617692108

Frank, R. (2015). Caveat lector: Fake news as folklore. *The Journal of American Folklore*, *128*(509), 315-332. https://doi.org/10.5406/jamerfolk.128.509.0315

Gadde, V., & Beykpour, K. (2020, October 9). *Additional steps we're taking ahead of the 2020 US Election*. Twitter. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html

Gaziano, C., & McGrath, K. (1986). Measuring the concept of credibility. *Journalism Quarterly*, *63*, 451-462. https://doi.org/10.1177%2F107769908606300301

Gil, B. F. (2018). Gatekeeping changes in the new media age: The internet, values and practices of journalism. *Brazilian Journalism Research*, *14*(2), 486-505. https://doi.org/10.25200/BJR.v14n2.2018.1026

Gramlich, J. (2016, November 1). *It's harder for Clinton supporters to respect Trump backers than vice versa.* Pew Research Center. https://www.pewresearch.org/fact-tank/2016/11/01/its-harder-for-clinton-supporters-to-respect-trump-backers-than-vice-versa/

Gramlich, J. (2020, October 26). *What the 2020 electorate looks like by party, race and ethnicity, age, education and religion*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/sci ence.aau2706

Guess, A., Nyhan, B., & Reifler, J. (2017). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential election*. https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf

Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, *37*(2), 281-301. https://doi.org/10.1080/10584609.2019.1674979

Heinecke, S. (2019). The game of trust: Reflections on truth and trust in a shifting media ecosystem. In T. Osburg & S. Heinecke (Eds.), *Media trust in a digital world* (pp. 3-13). Springer.

Hermida, A. (2010). Twittering the news. *Journalism Practice*, *4*(3), 297–308. https://doi.org/10.1080/17512781003640703

Hughes, A., & Wojcik, S. (2019, August 2). *10 facts about Americans and Twitter*. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/

Johnson, T. J., & Kaye, B. K. (2000). Using is believing: The influence of reliance on the credibility of online political information among politically interested internet users. *Journalism & Mass Communication Quarterly*, *77*(4), 865-879. https://doi.org/10.1177%2F107769900007700409

Johnson, T. J., & Kaye, B. K. (2004). Wag the blog: How reliance on traditional media and the internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly*, *81*(3), 622-642. https://doi.org/10.1177%2F107769900408100310

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology*, *38*, 167–208. https://doi.org/10.1111/pops.12407

Kamps, H. J. (2015, May 25). Who are Twitter's verified users? *Medium*. Retrieved November 9, 2021, from https://medium.com/@Haje/who-are-twitter-s-verified-users-af976fc1b032

Kang, M., & Yang, S. (2011, May). *Measuring social media credibility: A study on a measure of blog credibility*. Paper presented at the 61st annual conference of the International

Communication Association, Boston, MA. https://instituteforpr.org/wp-content/uploads/BlogCredibility_101210.pdf

Kemp, D., Ekins, E. (2021, December 15). *Poll: 75% don't trust social media to make fair content moderation decisions, 60% want more control over posts they see*. CATO Institute. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more

Keshavarz, H., & Givi, M. E. (2020). A scale for credibility evaluation of scientific websites: findings from a cross-contextual approach. *Online Information Review*, *44*(7), 1369-1386. https://doi.org/10.1108/OIR-04-2020-0127

Kirner-Ludwig, M. (2019). Creation, dissemination and uptake of fake-quotes in lay political discourse on Facebook and Twitter. *Journal of Pragmatics*, *157*, 101-118. https://doi.org/10.1016/j.pragma.2019.07.009

Krogstad, J. M., & Lopez, M. H. (2017, May 12). *Black voter turnout fell in 2016, even as a record number of Americans cast ballots*. Pew Research Center. https://www.pewresearch.org/fact-tank/2017/05/12/black-voter-turnout-fell-in-2016-even-as-a-record-number-of-americans-cast-ballots/

Laing, K. (2013, July 15). NTSB: Fake Asiana pilot names originated with TV station. *The Hill*. https://thehill.com/policy/transportation/311075-ntsb-fake-asiana-pilot-names-originated-with-tv-station

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lima, C. (2021, May 26). Facebook no longer treating ‘man-made’ Covid as a crackpot idea. *Politico*. https://www.politico.com/news/2021/05/26/facebook-ban-covid-man-made-491053

Lou, C., & Alhabash, S. (2020). Alcohol brands being socially responsible on social media? When and how warning conspicuity and warning integration decrease the efficacy of alcohol brand posts among under-drinking-age youth. *Journal of Interactive Advertising*, *20*(2), 148-163. https://doi.org/10.1080/15252019.2020.1780651

Marchi, R. (2012). “With Facebook, blogs, and fake news, teens reject journalistic ‘objectivity,’” *Journal of Communication Inquiry*, *36*(3), 246–262. https://doi.org/10.1177%2F0196859912458700

Matthes, J. (2012). The affective underpinnings of hostile media perceptions. *Communication Research*, *40*, 360-387. https://doi.org/10.1177/0093650211420255

McClymont, K., & Sheppard, A. (2020). Credibility without legitimacy? Informal development in the highly regulated context of the United Kingdom. *Cities*, *97*, 1-10. https://doi.org/10.1016/j.cities.2019.102520

McGowan, D. (2018). Walt Disney treasures or Mickey Mouse DVDs? Animatophilia, nostalgia, and the competing representations of theatrical cartoon shorts on home video. *Animation*, *13*(1), 53-68. https://doi.org/10.1177%2F1746847717752585

Mena, P. (2019). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, *12*(2), 165-183. https://doi.org/10.1002/poi3.214

Messner, M., & Garrison, B. (2009). Internet communication. In D. W. Stacks & M. B. Salwen (Eds.), *An integrated approach to communication theory and research* (pp. 75-89). Routledge.

Metzger, M., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*(B), 210-220. https://doi.org/10.1016/j.pragma.2013.07.012

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D., & Mccann, R. (2016). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication Yearbook*, *27*(1), 293-335. https://doi.org/10.1080/23808985.2003.11679029

Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, *60*, 824–844. https://doi.org/10.1111/ajps.12234

Mitchell, A., Gottfried, J., Kiley, J., & Matsa, K. (2014, October 21). *Political polarization & media habits*. Pew Research Center. https://www.pewresearch.org/journalism/2014/10/21/political-polarization-media-habits/

Napoli, P. M., & Obar, J. A. (2014). The emerging mobile internet underclass: A critique of mobile internet access. *The Information Society*, *30*, 323-334. https://doi.org/10.1080/01972243.2014.944726

Newberry, C. (2021, February 3). *36 Twitter stats all marketers need to know in 2021*. Hootsuite. Retrieved on November 3, 2021, from https://blog.hootsuite.com/twitter-statistics/

Newman, N., Fletcher, R., Levy, D. A. L., & Nielsen, R. K. (2016). *Reuters Institute Digital News Report 2016*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/our-research/digital-news-report-2016

Niemiec, E. (2020). COVID-19 and misinformation. Is censorship of social media a remedy to the spread of medical misinformation? *EMBO Reports*, *21*(11). https://doi.org/10.15252/embr.202051420

Notices on Twitter and what they mean. (n.d.). Twitter. Retrieved November 11, 2020, from https://help.twitter.com/en/rules-and-policies/notices-on-twitter

Ortutay B. (2018, July 23). No, Twitter will not ban Trump, here's why. *AP News*. https://apnews.com/article/222ade5f1b4340a8b5d7246045ec31d7

Ortutay, B. (2019, June 27). Politicians' tweets could get slapped with warning labels. *AP News*. https://apnews.com/article/46c853b157f94a6a8bae0689ac555010

O'Sullivan, D. (2020, February 28). A high school student created a fake 2020 candidate. Twitter verified it. *CNN Wire.* https://www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html

Paul, I., Khattar, A., Kumaraguru, P., Gupta, M., & Chopra, S. (2019). Elites tweet? Characterizing the Twitter verified user network. *IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, pp. 278-285. https://doi.org/10.1109/ICDEW.2019.00006

Pennycook, G., Bear, A., Collins, E., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11), 4944-4957. http://doi.org/10.2139/ssrn.3035384

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–59. https://doi.org/10.1016/j.cognition.2018.06.011

Poor, N. (2006). Playing internet curveball with traditional media gatekeepers: Pitcher Curt Schilling and Boston Red Sox fans. *Convergence: The International Journal of Research into New Media Technologies*, *12*(1), 41-53. https://doi.org/10.1177%2F1354856506061553

Remy, E. (2019, July 15). *How public and private Twitter users in the U.S. compare — and why it might matter for your research*. Medium. https://medium.com/pew-research-center-decoded/how-public-and-private-twitter-users-in-the-u-s-d536ce2a41b3

Retweet FAQs. (n.d.). Twitter. Retrieved November 7, 2020, from https://help.twitter.com/en/using-twitter/retweet-faqs

Rosen, A. (2017, November 7). *Tweeting made easier*. Twitter. https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

Rosen, G. (2021, May 26). *An update on our work to keep People informed and limit misinformation about COVID-19*. Facebook. https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims

Rosenberg, Y. [@Yair_Rosenberg]. (2017, November 16). *Whoever advised Twitter to turn verification into an approbation of views rather than a confirmation of identity did not think* [Tweet]. Twitter.com.

Roth, Y., & Pickles, N. (2020, May 11). *Updating our approach to misleading information*. Twitter. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

Rubin, V. L. (2017). Deception detection and rumor debunking for social media. In Sloan, L. & Quan-Haase, A. (Eds.), *The SAGE handbook of social media research methods* (pp. 1-25). SAGE. https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-social-media-research-methods/book245370

Rubin, V. L. (2019). Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *Journal of Documentation*, *75*(5), 1013-1034. https://doi.org/10.1108/JD-12-2018-0209

Rubin, V. L., Chen, Y., & Conroy, N. K., (2015). Deception detection for news: Three types of fakes. *The Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4. https://doi.org/10.1002/pra2.2015.145052010083

Rules Enforcement. (2022, January 25). Twitter Transparency Center. Retrieved January 29, 2022, from https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jan-jun

Sauer, N. (2017, November 2). Collins Dictionary picks 'fake news' as word of the year. *Politico*. https://www.politico.eu/article/collins-dictionary-picks-fake-news-as-word-of-the-year/

Schulz, A., Wirth, W., & Müller, P. (2020). We are the people and you are fake news: A social identity approach to populist citizens' false consensus and hostile media perceptions. *Communication Research*, *47*(2), 201-226. https://doi.org/10.1177/0093650218794854

Shearer, E., & Matsa, K. E. (2018, September 10). *News use across social media platforms 2018*. Pew Research Center. https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

Shoemaker, P. J., & Vos, T. P. (2009). Media gatekeeping. In D. W. Stacks & M. B. Salwen (Eds.), *An integrated approach to communication theory and research* (pp. 75-89). Routledge.

Sikdar, S., Kang, B., O'Donovan, J., Hollerer, T., & Adah, S. (2013). Cutting through the noise: defining ground truth in information credibility on Twitter. *ASE HUMAN Journal*, *3*(1), 151-167.

Singer, J. B. (2001). The metro wide web: Changes in newspapers' gatekeeping role online. *Journalism & Mass Communication Quarterly*, *78*(1), 65-80. https://doi.org/10.1177/107769900107800105

Singer, J. B. (2014). User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society*, *16*(1), 55-73. https://doi.org/10.1177/1461444813477833

Smith, P. K., & Sissons, H. (2016). Social media and a case of mistaken identity: A newspaper's response to journalistic error. *Journalism*, *20*(3), 467-482. https://doi.org/10.1177/1464884916683551

Sokol, C. (2017, November 16). WSU's James Allsup stripped of 'verification' checkmark. *The Spokesman-Review*. https://www.spokesman.com/stories/2017/nov/16/wsus-james-allsup-stripped-of-verification-checkma

Statista Research Department. (2021a, April 14). *Percentage of U.S. adults who use Twitter as of February 2021, by age group.* Statista. https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/

Statista Research Department. (2021b, July 20). *Social media activities on select social networks by social media users in the United States in February 2019*. Statista. https://www.statista.com/statistics/200843/social-media-activities-by-platform-usa/

Statista Research Department. (2022, January 28). *Twitter: number of monthly active users 2010-2019*. Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Stjernfelt, F., & Lauritzen, A. M. (2020). *Your post has been removed: Tech giants and freedom of speech*. Springer.

Sundar, S. S. (1999). Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly*, *76*, 373-386. https://doi.org/10.1177/107769909907600213

Sundar, S. S. (2008). "The MAIN model: A heuristic approach to understanding technology effects on credibility", in M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73-100). The MIT Press.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

Tsukayama, H. (2016, July 19). Twitter's letting anyone apply to become verified. *The Washington Post.* https://www.washingtonpost.com/news/the-switch/wp/2016/07/19/twitters-letting-anyone-apply-to-become-verified/

Twitter Inc. (2021, May 20). *Relaunching verification and what's next*. Twitter. https://blog.twitter.com/en_us/topics/company/2021/relaunching-verification-and-whats-next

Twitter Safety. (2019, June 27). *Defining public interest on Twitter*. Twitter. https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html

Twitter Support [@TwitterSupport]. (2017, November 1). *Verification was meant to authenticate identity & voice but it is interpreted as an endorsement or an indicator of importance.* [Tweet]. Twitter.com. https://twitter.com/TwitterSupport/status/928654369771356162

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media, Culture & Society*, *42*(3), 460-470. https://doi.org/10.1177/0163443720906992

Verification FAQ. (n.d.). Twitter. Retrieved January 28, 2022, from https://help.twitter.com/en/managing-your-account/twitter-verified-accounts

Verified account FAQs. (n.d.). Twitter. Retrieved November 7, 2020, from https://help.twitter.com/en/managing-your-account/twitter-verified-accounts

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Watts, M. D., Domke, D., Dhavan, S. V., & Fan, D. P. (1999). Elite cues and media bias in presidential campaigns: Explaining public perceptions of a liberal press. *Communication Research*, *26*(2), 144-175. https://doi.org/10.1177/009365099026002003

White, A. (2017). *Fake news: It's not bad journalism, it's the business of communications in the digital age*. Ethical Journalism Network. https://ethicaljournalismnetwork.org/fake-news-bad-journalism-digital-age

White, D. M. (1950). The "gate keeper": A case study in the selection of news. *Journalism Quarterly*, *27*(4), 383-390. https://doi.org/10.1177/107769905002700403

Williams, B. A., & Delli Carpini, M. X. (2004). Monica and Bill all the time and everywhere: The collapse of gatekeeping and agenda setting in the new media environment. *American Behavioral Scientist*, *47*(9), 1208-1230. https://doi.org/10.1177%2F0002764203262344

Wojcik, S., & Hughes, A. (2019, April 24). *Sizing up Twitter users*. Pew Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Wojcik, S., Messing, S., Smith, A., Rainie, L., & Hitlin, P. (2018, April 9). *Bots in the Twittersphere*. Pew Research Center. https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/

Yan, L. (2021, May 17). *Archived fact-check: Tucker Carlson guest airs debunked conspiracy theory that COVID-19 was created in a lab*. PolitiFact. https://www.politifact.com/li-meng-yan-fact-check/

Zakharov, W., Li, H., Fosmire, M. (2019). Undergraduates' news consumption and perceptions of fake news in science. *portal: Libraries and the Academy*, *19*(4), 653-665. https://doi.org/10.1353/pla.2019.0040

Zubiaga, A., & Ji, H. (2014). Tweet, but verify: Epistemic study of information verification on Twitter. *Social Network Analysis and Mining*, *4*(1), 163. https://doi.org/10.1007/s13278-014-0163-y