The Challenges of Algorithmically Assigning Fact-checks

A Sociotechnical Examination of Google's Reviewed Claims

by

Emma Wittenberg Lurie

Submitted in Partial Fulfillment of the Prerequisite for Honors
Prerequisite for Honors

in the
Computer Science Department
April 2019

Abstract

In the era of misinformation and machine learning, the fact-checking community is eager to develop automated fact-checking techniques that can detect misinformation and present fact-checks alongside problematic content. This thesis explores the technical elements and social context of one such "claim matching" system, Google's Reviewed Claims. The Reviewed Claims feature was one of the few user-facing interfaces in the complex socio-technical system between fact-checking organizations, news publishers, Google, and online information seekers. This thesis addresses the following research questions:

RQ1: How accurate was Google's Reviewed Claims feature?

RQ2: Is it possible to create a consensus definition for "relevant fact-checks" to enable the development of more successful automated fact-checking systems?

RQ3: How do different actors in the fact-checking ecosystem define relevance?

I investigate these research questions through a series of methods including qualitative coding, qualitative content analysis, quantitative data analysis, and user studies.

To answer RQ1, I qualitatively label the relevance of 118 algorithmically assigned fact-checks and find that 21% of fact-checks are not relevant to their assigned article.

To address RQ2, I find that three independent raters using a survey are only able to come to "fair-moderate agreement" about whether the algorithmically assigned fact-checks are relevant to the matched articles. A reconciliation process substantially raised their agreement. This indicates that further discussions may create a common understanding of relevance among information seekers. Using raters' open-ended justification responses, I generated 6 categories of justifications for their explanations. To further evaluate if information seekers shared a common definition of relevance, I asked Amazon Mechanical Turk workers to classify six different algorithmically assigned fact-checks and found that crowd workers were more likely to find the matched content relevant and were unable to agree on the justifications.

With regard to RQ3, a sociotechnical analysis finds that the fact-checking ecosystem is fraught with distrust and conflicting incentives between individual actors (news publishers distrust fact-checking organizations and platforms, fact-checking organizations distrust platforms, etc.). Given the distrust among actors, future systems should be interpretable and transparent about their definition of "relevance" as well as the ways in which the claim matching is performed.

Fact-checking is dependent on nuance and context, AI is not technically sophisticated enough to account for these variables. As such, human-in-the-loop models seem to be essential to future automated fact-checking approaches. However, the results of this thesis indicate untrained crowd workers may not be the ideal candidates for modeling complex values in sociotechnical systems.

Acknowledgments

This thesis was only possible because of the unwavering support of Eni Mustafaraj, who has served as my professor, major advisor, principal investigator, thesis advisor, mentor, and so much more during my time time at Wellesley. Thank you Eni for your endless enthusiasm, wisdom, honesty, patience, and book recommendations.

Thank you to Catherine Delcourt and Takis Metaxas who provided valuable feedback on the scoping and framing of this thesis as members of my committee. Thank you to Ismar Volić for graciously agreeing to serve as the honors visitor.

Thank you to Sarah Barbrow for all of her advice, kindness, and assistance. Sarah's knack for finding interesting articles and relevant resources made this thesis (and bibliography) better.

I am grateful to all of my computer science professors at Wellesley (Sravana Reddy, Sohie Lee, Takis Metaxas, Stella Kakavouli, Ben Wood, Jean Herbst, Eni Mustafaraj, Lyn Turbak, Randy Shull, Christine Bassem, and Scott Anderson) for their passion and encouragement.

Thank you to all of the study participants who gave their time and feedback. A special thank you to the three Wellesley Cred Lab student participants, who spent several hours with me sharing their ideas and perspectives.

I am fortunate to have wonderful friends who have listened to me talk endlessly about fake news, fact-checking, and research for the past three years. I have learned so much from them throughout my time at Wellesley.

I have also benefited from an extraordinarily supportive family who have supported me for my entire life. I would not be the person I am today without them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

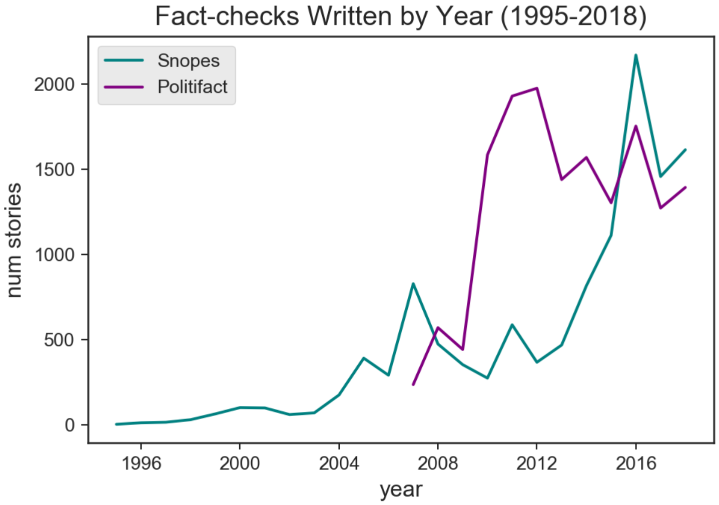
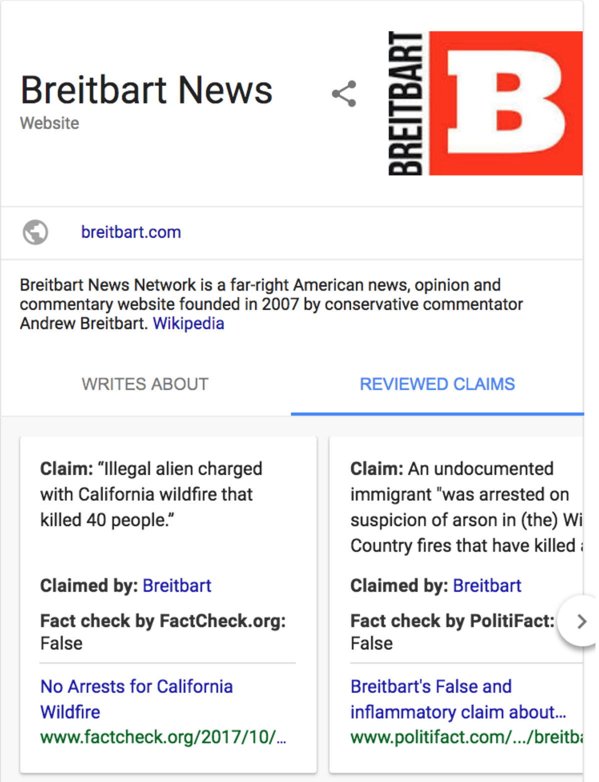


Figure 1.1: The number of fact-check articles written by the two largest fact-checking organizations from 1995-2018. There has been a sharp increase in the number of fact-checks produced by both Snopes and PolitiFact over the last several years.

1.1 Motivation

In response to fears about the spread of online misinformation, there has been a rapid growth and investment in fact-checking (see Figure 1.1). In August 2018, the Duke Reporters'

Lab identified 156 current global fact-checking initiatives. Of those organizations, 102 were established after 2013.[1] However, only a small percentage of people who are exposed to problematic online content are presented with corrective information like fact-checks. In the era of big data and artificial intelligence, a key agenda item for the fact-checking movement has been to develop automated fact-checking systems that leverage techniques like machine learning and natural language processing to limit peoples' exposure to misinformation by either having platforms 1) suppress stories that have an associated fact-check or 2) present relevant fact-checks alongside problematic content.



Breitbart News
Website

breitbart.com

Breitbart News Network is a far-right American news, opinion and commentary website founded in 2007 by conservative commentator Andrew Breitbart. Wikipedia

Claim: "Illegal alien charged with California wildfire that killed 40 people."

Claimed by: Breitbart

Fact check by FactCheck.org: False

No Arrests for California Wildfire
www.factcheck.org/2017/10/...

Claim: An undocumented immigrant "was arrested on suspicion of arson in (the) Wi Country fires that have killed

Claimed by: Breitbart

Fact check by PolitiFact: False

Breitbart's False and inflammatory claim about...
www.politifact.com/.../breitb

Figure 1.2: An example of the Knowledge Panel with the Reviewed Claims feature that was displayed in January 2018 with the query "breitbart." The Reviewed Claims component was visible on a subset of news publisher search engine results pages from November 2017 - January 2018.

In November 2017, as part of an expanded effort to provide users with context about news publishers, Google released Reviewed Claims (see Figure 1.2), a component of the

[1] https://reporterslab.org/the-number-of-fact-checkers-around-the-world-156-and-growing

Knowledge Panel that appears on many search engine results pages (SERP). The Reviewed Claims feaure displayed authoritative, third-party fact-checks about content produced by certain news publishers. This feature was publicized as a meaningful aid to users[2] in the fight against misinformation. However, without informing fact-checking organizations, news publishers, or information seekers, Google's Reviewed Claims "claim matched" over half of the fact-checks that appeared in the component. Claim matching is a process in which fact-check articles that do not list the original source of a fact-checked statement (i.e. the claimant) are algorithmically assigned to a piece of online content produced by a certain publisher (see Figure 2.2 and 2.3). Systems like Reviewed Claims that claim match existing fact-checks to online content have the ability to increase the number of stories that have been fact-checked and simplify the process of displaying fact-checks alongside problematic online content.

However, several news publishers complained about this feature in January 2018. They alleged that Google was displaying partisan bias and claimed that several of the fact-checks listed in their Reviewed Claims components were not relevant to the associated content. Soon after, Google removed the Reviewed Claims component from the Knowledge Panel, citing "bugs" in the feature's implementation. But what were the bugs? What can be learned from the mistakes of the Reviewed Claims feature to enable the development of better tools that combat misinformation?

To understand the failure of the Reviewed Claim feature requires 1) a technical understanding of the limitations of the Reviewed Claim system and 2) a sociotechnical perspective of the fact-checking ecosystem's stakeholders and incentives.

[2] https://www.blog.google/products/search/learn-more-about-publishers-google/

1.2 Contributions and Research Questions

The research questions this thesis addresses are:

RQ1: *How accurate was Google's Reviewed Claims feature?*

I find that the Reviewed Claims feature made a significant number of classification mistakes. In Chapter 4, I focus on how the designers of the claim matching system described in "Relevant Document Discovery for Fact-Checking Articles" [54] define and assess accuracy. This paper describes the system that likely generates the claim matches displayed in Reviewed Claims components.

The limitations described in "Relevant Document Discovery for Fact-Checking Articles" are present in the Reviewed Claims system and account for some of the classification errors detailed by news publishers. In Chapter 4.5, I outline some additional limitations of the system implementation, especially in regard to the authors' use of ClaimReview data.

In Chapter 5, I qualitatively label 118 algorithmically assigned fact-checks to assess the accuracy of the Reviewed Claims claim matches. Overall, I find that the 78% of the fact-checks are relevant to their assigned claimant documents. I assign three raters 30 claim matches, and they determine that 14 of the 30 are relevant.

RQ2: *Is it possible to create a consensus definition for "relevant fact-checks" to enable the development of more successful automated fact-checking systems?*

While it may be possible to create a common definition for "relevant fact-checks," I am unable to generate a satisfactory definition in this thesis.

In Chapter 5, I explore the results of qualitatively coding raters justifications for their assessment of the irrelevance of the fact-checks. The five categories I generate are:

(a) The fact-check concerns a minor detail in the article.

(b) The fact-check is verifying a different claim than what was in the article.

(c) The fact-check and article cover the same general topic but differ on details.

(d) The fact-check and article are framed differently.

(e) The fact-check is more general than the article.

In Chapter 6, I show that these rater justifications have significant overlap and differences with news publishers' criteria for relevance. However, crowd workers are not able to agree on which justifications apply to a given claim match. Thus, the justifications listed above are insufficient to define the multifaceted concept of relevance.

RQ3: *How do different actors in the fact-checking ecosystem define relevance?*

In Chapter 2.2, I explain that without a definition of a relevant fact-check, it is 1) impossible to determine the accuracy of algorithmically assigned claim-document models and 2) difficult to describe the feature to news publishers and online information seekers.

In Chapter 6, I use a sociotechnical lens to explore the different motivations and definitions of actors in the fact-checking ecosystem that inform their conception of relevance. In addition, I consider their influence within the fact-checking ecosystem. I find complex interactions exist between the different actors and some overlap between the various conceptions of relevance. If future automated fact-checking systems are to be more successful than Google's Reviewed Claims, system designers need to carefully consider how complex social values like relevance are currently modeled and should be modeled in the fact-checking ecosystem.

Through an exploration of these research questions and the findings discussed briefly above, this thesis offers the following primary contributions:

1. The first sociotechnical analysis of an automated fact-checking system. This sociotechnical lens builds understanding as to why similar systems will inevitably face pushback from a variety of actors. This analysis also delves into how "neutral" technical concepts like "accuracy" become entangled with challenging social and epistemological questions of "relevance."

2. Preliminary evidence suggests that untrained crowd workers may not be able to determine the nuances of "relevance" in fact-checking. This is significant since there is a push for platforms to increase reliance on crowd workers as quality control for technology platforms in fact-checking and related domains.

1.3 Thesis Outline

This thesis is organized in the following chapters:

- Chapter 2 provides the background needed for this thesis including an introduction to fact-checking, relevance, Google's Reviewed Claims, algorithm auditing, and the sociotechnical perspective.

- Chapter 3 explores the literature on misinformation detection, the benefits and critiques of fact-checking, automated fact-checking, and the sociotechnical perspective.

- Chapter 4 dissects the paper "Relevant Document Discovery for Fact-Checking Articles," which provides the technical specifications of the Reviewed Claims system. Then, I offer a critique of multiple elements of the paper through a quantitative analysis of the ClaimReview markup and a further analysis of the methods of the paper.

- Chapter 5 presents three qualitative studies that I designed and conducted to evaluate Google's Reviewed Claims' accuracy and information seekers' justifications of their accuracy assessments. The results paint a complicated picture of how information seekers define relevance but makes clear that information seekers did find not all of the claim matches to be relevant.

- Chapter 6 adopts a sociotechnical perspective of the fact-checking ecosystem's stakeholders' reactions to the Reviewed Claims feature. The sociotechnical lens reveals the complex relationship between stakeholders that make modeling relevance extremely challenging. I rely on a sociotechnical framework to analyze the individual actors and content analysis of various media reports of the feature from early 2018.
- Chapter 7 provides a roadmap for future work and offers a brief overview of the most significant findings of this thesis.

Chapter 2

Background

In this chapter, I outline key concepts relevant to the analysis I perform in this thesis, namely 1) fact-checking, 2) computational approaches to fact-checking, 3) the concept of “relevance” in claim matching, 4) the Reviewed Claims feature, 5) algorithm audits, and 6) the sociotechnical perspective.

2.1 Understanding Fact-checking

The American Press Institute (API) defines the purpose of fact-checking as the following:

> “Fact checkers and fact-checking organizations aim to increase knowledge by re-reporting and researching the purported facts in published/recorded statements made by politicians and anyone whose words impact others’ lives and livelihoods. Fact checkers investigate verifiable facts, and their work is free of partisanship, advocacy, and rhetoric.
>
> The goal of fact-checking should be to provide clear and rigorously vetted in-

> formation to consumers so that they may use the facts to make fully cognizant choices in voting elections and other essential decisions."[1]

As the API explains, fact-checking is different from traditional journalism because it is the "re-reporting" of verifiable, significant claims. A fact-check never produces new information, but rather selects a claim and uses existing reporting and research to assess the validity of that claim.

Each fact-checking organization has developed and followed its own process for fact-checking online information. An ethnographic study of PolitiFact's fact-checking process by Graves [21] describes the following stages:

1. Selecting claims to fact-check.
2. Contacting the claimants.
3. Tracing the origins of the claims.
4. Collaborating with domain experts to develop fact-check articles.
5. Ensuring the process and sources are properly documented in the article.

Even with these established processes, questions about the persuasiveness of fact-checks and concerns about fact-checking organizations' methodological rigor have plagued fact-checking initiatives. The various critiques of fact-checking will be further explored in Chapter 3.3.

Graves disagrees with many of these critiques and finds that similar to scientists and investigative journalists, fact-checkers deal "with controversies in which not just facts but rules for determining them are in question, and thus affords a view of the way material,

[1] https://www.americanpressinstitute.org/fact-checking-project/fact-checker-definition

social, and discursive contexts structure factual inquiry" [21]. This thesis aims to investigate the definition of accuracy in automated fact-checking systems by looking at the ways political and social contexts structure the rules of determining relevance.

2.2 Computational Approaches to Fact-checking

Some researchers believe that automated fact-checking approaches have the potential to effectively combat misinformation by addressing some of the limitations of manual fact-checking [24] with computational methods. Full Fact's 2016 report "The State of Automated Factchecking" [8] breaks down automated fact-checking approaches into four stages. End-to-end automated fact-checking systems will include all four stages in one system.

1. *Monitor claims*: Automated fact-checking systems need to monitor large amounts of raw content and identify the claimants and other information about the content. Metadata is essential for providing important context about the content, but is almost always insufficient.

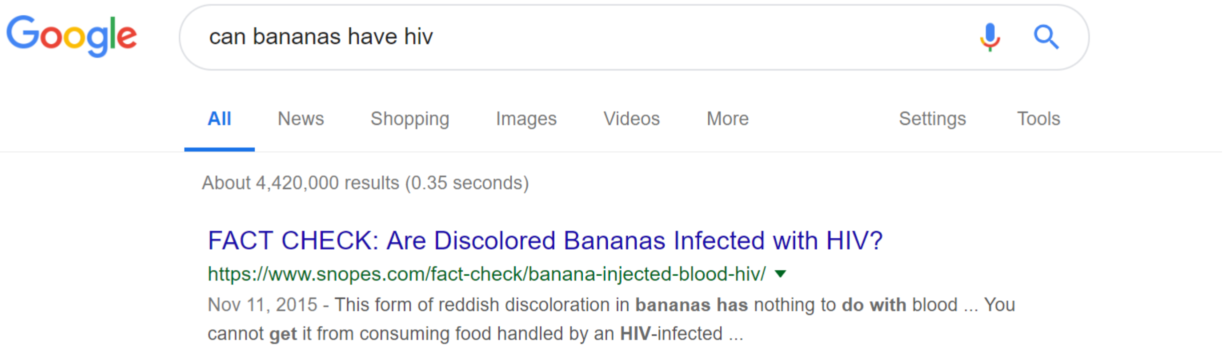
2. *Spot Claims*: Full Fact breaks down the tasks within this step as:

    - "Monitoring claims that have been fact-checked before in new text"
    - "Identifying new factual claims that have not been fact-checked before in new text"
    - "Making editorial judgments about the priority of different "
    - "Dealing with different phrasing for the same or similar claims"

3. *Check Claims*: See Chapter 3.1 and 3.4 for the different approaches currently being employed to verify claims.

4. *Create and Publish*: This step requires automated fact-checking systems to share the results of step 3 in a user readable format (fact-check article, tweet, etc.).

2.3 Defining Relevance



can bananas have hiv

About 4,420,000 results (0.35 seconds)

FACT CHECK: Are Discolored Bananas Infected with HIV?
https://www.snopes.com/fact-check/banana-injected-blood-hiv/
Nov 11, 2015 - This form of reddish discoloration in **bananas has** nothing to **do with** blood ... You cannot **get** it from consuming food handled by an **HIV**-infected ...

Figure 2.1: The first search result for the query “can bananas have HIV” is a fact-check. Google has started a number of initiatives to surface fact-checks at the top of the SERP (including Reviewed Claims, see Figure 2.2)

Fact-checking can take many different forms, but manual fact-checking (i.e. non-automated approaches) does not usually require thinking about relevance. In the simplest case, a fact-checker gets a tip from a reader about a possible false statement they came across, for example, bananas can have HIV. The fact-checker investigates that claim, identifies the origin of that claim, and writes up a fact-check. Soon after, when a user searches “can bananas have HIV” the fact-check article “Are Discolored Bananas Infected with HIV?” is displayed as a top search result (see Figure 2.1. In this example, the fact-check is *relevant* to the original claimant article (a social media post) by default, since it is a *response* to that original piece of content.

However, there are still ways in which the claimant can argue that the fact-check is not relevant to the original article. For instance, the fact-checker may have overlooked a piece of context necessary for understanding the claim or a missed a relevant detail in the article.

Other shortcomings of manual fact-checking itself will be discussed in Chapter 3.3.

Similar to many other machine learning tasks, automated fact-checking approaches (further explored in Chapter 3.4) must have a conception of accuracy to assess the quality of the model. In terms of the "claim-document discovery process," this means that if the fact-check "Are Discolored Bananas Infected with HIV? is algorithmically matched to an article about an earthquake in California, this should not be considered a relevant match. In fact, a model that algorithmically assigned such <fact-check, article> pairs would most likely have a low accuracy score. The determination of whether a fact-check is relevant or not seems to be a natural analog to traditional machine learning notions of accuracy.

Further examples reveal that the concept of relevance is actually quite complicated. If "Are Discolored Bananas Infected with HIV?" debunks claims about red discolorations of bananas, what if the algorithmically assigned article discusses blue discolorations of bananas infecting people with HIV? What about orange apples?

While this example does not appear in the Reviewed Claims example, the point remains: how can relevance be defined so that 1) the needs of information users are prioritized and 2) other stakeholders in the fact-checking ecosystem are treated fairly?

Even before Google, researchers recognized the complexities of defining relevance. In their 2000 paper, Introna and Nissenbaum write that "determining relevancy is an extraordinarily difficult task...Besides the engineering challenges, experts must struggle with the challenging of approximating a complex human value" [26].

Google researchers defined their conception of relevance in the claim-document discovery model as:

> Given a fact-checking article with claim c, a claim-relevant document is a related document that addresses c [54].

This definition leaves much to interpretation. Understanding and being able to communicate the definition of relevance is incredibly important to the success of automated fact-checking features. Without clarity as to what is and is not a relevant fact-check, it is impossible to determine the accuracy of algorithmically assigned claim-document models and accurately describe the feature to news publishers and online information seekers.

2.4 Introduction to Reviewed Claims

The Google search engine result page (SERP) no longer consists of ten blue links [17]. In the early days of Google, a user would have to click on one of the search result links to find the answer to their query; however, with the addition of the Knowledge Graph in 2012[2], there has been an increased emphasis on increasing content for informational queries (which differ from navigational and transactional queries [10]) on the SERP itself. The Knowledge Graph draws on information from a variety of sources including Freebase and Wikipedia. Previous studies have illustrated that Google relies on third-party sources, in particular, Wikipedia, to create high-quality SERPs [35]. Oftentimes, Knowledge Graph data is displayed at the top of the SERP in the Knowledge Panel.

In November 2017, in response to concerns that Google was not effectively combating misinformation on its platform, Google augmented its news publisher Knowledge Panels. The described use case of the expanded Knowledge Panels was as follows:

> “As tens of thousands of publishers of all sizes push out content every day, chances are you’ve come across a publication you’re not familiar with or one you wanted to learn more about.
>
> To help in this situation, publisher Knowledge Panels on Google will now show

[2] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

> the topics the publisher commonly covers, major awards the publisher has won, and claims the publisher has made that have been reviewed by third parties. These additions provide key pieces of information to help you understand the tone, expertise, and history of the publisher."[3]

The section of the Knowledge Panel that included "claims the publisher has made that have been reviewed by third parties" was named Reviewed Claims (see Figure 2.2). However, accusations of anti-conservative bias and errors in the classification algorithm caused Google to discontinue the Reviewed Claims feature in mid-January 2018 [45].

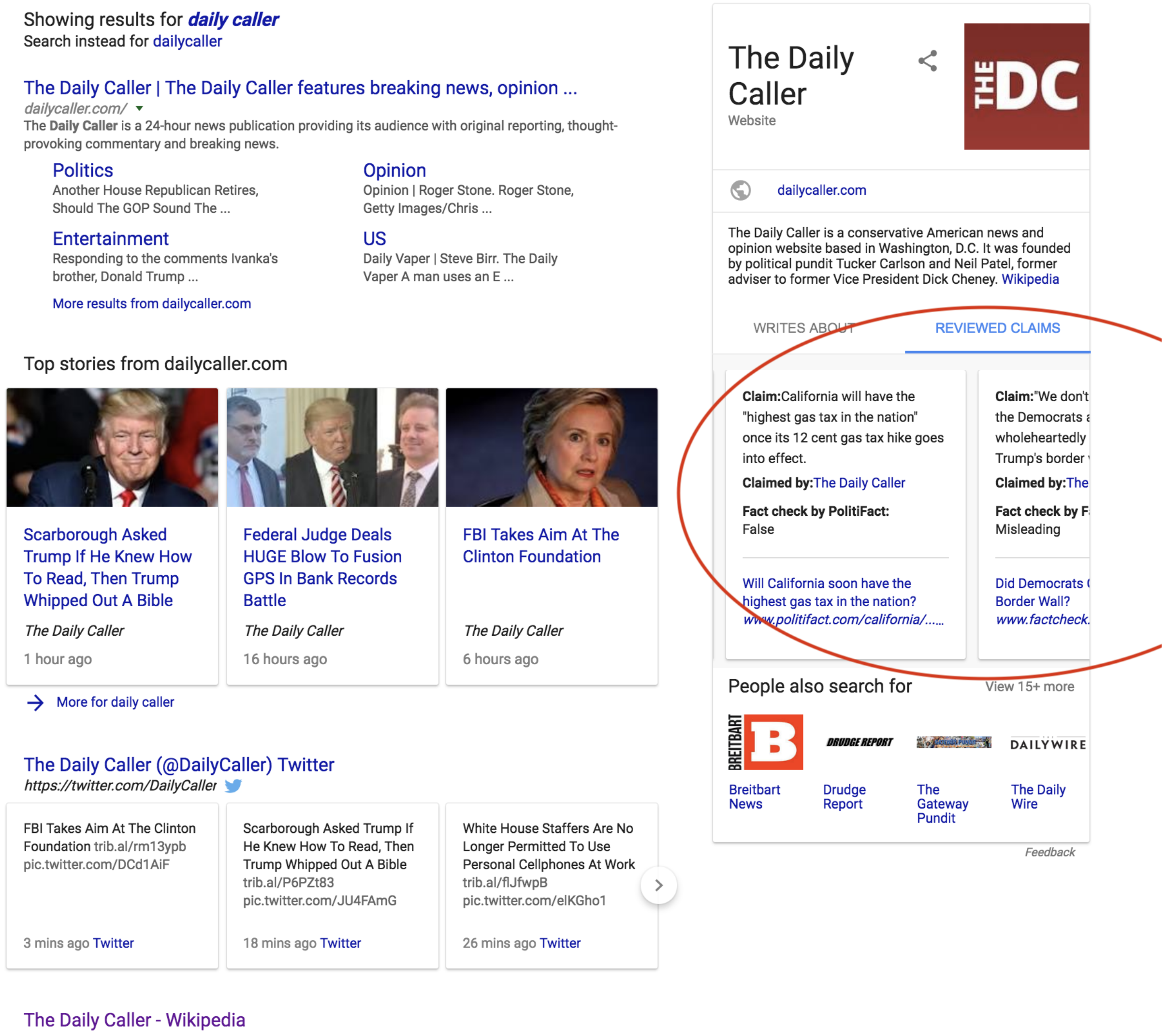
In particular, *The Daily Caller* complained that three of the ten fact-check articles displayed in the Reviewed Claims tab did not mention *The Daily Caller.* For example, in Figure 2.3, the "Claimed by" element of the interface explicitly states the claim comes from as *The Daily Caller.* However, the original fact-checking article from PolitiFact clearly identifies "Jim Patterson," a California politician, as the source of the claim. In fact, PolitiFact was unaware of *The Daily Caller's* article. Google was using algorithms (described in Chapter 4) to automatically assign fact-checks to news articles that were "relevant" to the fact-check. However, Google did not notify any stakeholder in the fact-checking ecosystem about the claim matching algorithm that comprised 57% of the analyzed fact-checks displayed in Reviewed Claims panels (detailed in Chapter 5).

In this thesis, I present the first analysis of the content in the Reviewed Claims feature. This analysis is based on a dataset of approximately 8,000 news publisher SERPs (search engine result pages), which were collected in early January 2018. In the dataset, I identified 59 news publisher SERPs that contained the Reviewed Claims tab. These 59 Reviewed Claims components contain a total of 221 fact-checks. Of the 221 fact-checks, I identified

[3] https://www.blog.google/products/search/learn-more-about-publishers-google

[4] https://www.poynter.org/fact-checking/2018/blame-bugs-not-partisanship-for-google-wrongly-appending-a-fact-check-to-the-daily-caller

[5] https://www.poynter.org/fact-checking/2018/google-suspends-fact-checking-feature-over-quality-concerns



Showing results for ***daily caller***
Search instead for dailycaller

The Daily Caller | The Daily Caller features breaking news, opinion ...
dailycaller.com/
The **Daily Caller** is a 24-hour news publication providing its audience with original reporting, thought-provoking commentary and breaking news.

Politics
Another House Republican Retires, Should The GOP Sound The ...

Opinion
Opinion | Roger Stone. Roger Stone, Getty Images/Chris ...

Entertainment
Responding to the comments Ivanka's brother, Donald Trump ...

US
Daily Vaper | Steve Birr. The Daily Vaper A man uses an E ...

More results from dailycaller.com

Top stories from dailycaller.com

Scarborough Asked Trump If He Knew How To Read, Then Trump Whipped Out A Bible
The Daily Caller
1 hour ago

Federal Judge Deals HUGE Blow To Fusion GPS In Bank Records Battle
The Daily Caller
16 hours ago

FBI Takes Aim At The Clinton Foundation
The Daily Caller
6 hours ago

More for daily caller

The Daily Caller (@DailyCaller) Twitter
https://twitter.com/DailyCaller

FBI Takes Aim At The Clinton Foundation trib.al/rm13ypb pic.twitter.com/DCd1AiF
3 mins ago Twitter

Scarborough Asked Trump If He Knew How To Read, Then Trump Whipped Out A Bible trib.al/P6PZt83 pic.twitter.com/JU4FAmG
18 mins ago Twitter

White House Staffers Are No Longer Permitted To Use Personal Cellphones At Work trib.al/flJfwpB pic.twitter.com/elKGho1
26 mins ago Twitter

The Daily Caller - Wikipedia

The Daily Caller
Website

dailycaller.com

The Daily Caller is a conservative American news and opinion website based in Washington, D.C. It was founded by political pundit Tucker Carlson and Neil Patel, former adviser to former Vice President Dick Cheney. Wikipedia

WRITES ABOUT REVIEWED CLAIMS

Claim: California will have the "highest gas tax in the nation" once its 12 cent gas tax hike goes into effect.
Claimed by: The Daily Caller
Fact check by PolitiFact:
False
Will California soon have the highest gas tax in the nation?
www.politifact.com/california/......

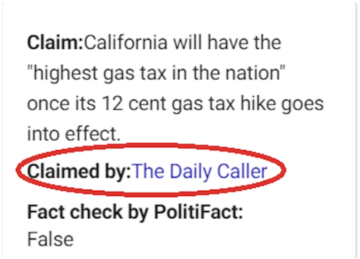
Claim:"We don't the Democrats wholeheartedly Trump's border
Claimed by: The
Fact check by F
Misleading
Did Democrats Border Wall?
www.factcheck.

People also search for View 15+ more

Breitbart News
Drudge Report
The Gateway Pundit
The Daily Wire

Feedback

Figure 2.2: The SERP of *The Daily Caller* with the Reviewed Claims feature circled as it appeared in January 2018. Reviewed Claims component contained fact-checks of articles produced by that news publisher.



Claim: California will have the "highest gas tax in the nation" once its 12 cent gas tax hike goes into effect.
Claimed by: The Daily Caller
Fact check by PolitiFact:
False

Figure 2.3: The Reviewed Claim feature indicates that the claim was made by *The Daily Caller*. However, the PolitiFact fact-check does not mention *Daily Caller*. This is an example of an algorithmically assigned fact-check discussed throughout this thesis.

119 as algorithmically assigned fact-checks.

I define an algorithmically assigned fact-check to be a fact-check article that does not reference the online document explicitly in the text or in the structured data of the fact-check.

2.5 Auditing Reviewed Claims

Technical researchers employ several methods to understand the workings of algorithmic systems. An increasingly common approach is the algorithm audit, which derives its approach from audits used to identify illicit financial transactions, housing, and hiring discrimination. Algorithm audits have been used to determine whether there is partisan bias on Google [44], sexism in online marketplaces [23], racism on platforms [38], and many other systems that are seemingly neutral or opaque without controlled experimentation.

Algorithm audits are performed in different ways. One option is to conduct a code audit (i.e. look at the code of the algorithm). However, as Burrell points out, even if one could decipher the code, machine learning models are difficult to decipher and it is impossible to understand the biases of input data from the code of the algorithm [14]. The second option is to conduct a scraping audit. As Sandvig describes it, a scraping audit is "when a researcher makes repeated queries to a platform and analyzes the results" [45].

In this thesis, I combine the two approaches. Google has not publicly released the code for the Reviewed Claims feature for analysis. However, at the 2018 Web Conference, Google Research published a paper titled "Relevant Document Discovery for Fact-Checking Articles in which they outline an automated claim matching system [54]. The paper presents a system that performs better than any other existing approach at the "claim-document discovery process." The claim-document discovery process matches existing fact-check articles to relevant documents that are not mentioned explicitly in the fact-check.

While the term "Reviewed Claims" does not appear in the paper, the objective, method, authorship, and cited limitations match up with what is known about the Reviewed Claims feature. Therefore, it seems likely that the model described in the paper is the backend of the Reviewed Claims feature. This paper allows me to perform a modified code audit, or "specifications audit," as I am able to examine the technical features of the system, but unable to examine the data that was fed into or produced by the model. As a result, this thesis does not contain any discussion of the

Although, I am unable to do a traditional "scraping audit" for this project where I would systematically test different inputs (fact-checking articles and news publisher SERPs), I have collected a set of news publisher SERPs. As described in Chapter 5.1, as part of a separate research project, I collected 59 Reviewed Claims tabs in the course of collecting the SERPs of approximately 8,000 news publishers in January 2018. I programmatically collected over 30,000 fact-check articles in early 2019 to examine the prevalence and structure of the ClaimReview markup, which Wang et al. [54] cited as essential to their system.

2.6 Introduction to the Sociotechnical Perspective

In the field of Science and Technology Studies (STS), systems consisting of both technical and social components are defined as "sociotechnical systems." In "Fairness and Abstraction in Sociotechnical Systems," the sociotechnical view is defined as requiring technical actors to engage with "different frameworks that provide guidance in identifying, articulating, and responding to fundamental tensions, uncertainties, and conflicts inherent in sociotechnical systems" [47].

Adopting a sociotechnical view to analyze the Reviewed Claims and discuss future automated fact-checking solutions is necessary because the success of the feature depends not only on the algorithms that Google researchers created to perform the claim matching, but also on how the various actors in the fact-checking ecosystem interact with the system. Moreover, as already mentioned in "Defining Relevance" if different actors define relevance differently, how can we assess the accuracy of "relevance classifiers?"

Thus, a sociotechnical lens is a useful perspective for analyzing the limitations of the Reviewed Claims feature. In Chapter 3.5, I detail two projects that have adopted a sociotechnical perspective to analyze shortcomings in technical tools or proposed solutions by analyzing the complex social interactions around technical systems.

Chapter 3

Related Research

In this chapter, I will discuss some of the relevant literature that informs automated fact-checking. This includes concerns about misinformation, benefits and critiques of fact-checking, computational approaches to fact-checking. I summarize two recent sociotechnical analyses that provide background on the insights produced by sociotechnical approaches as well as findings that inform this thesis.

3.1 Detecting and Limiting Misinformation

Online misinformation has created widespread concern. Combating problematic information is a major priority for many different research communities including computer science, psychology, and political science. While there has been widespread speculation that fake news impacted the 2016 U.S. presidential election, recent research casts doubt on those theories [22]. Much of the concern centered on the spread of fake news on platforms like Facebook, but a significant amount of traffic to fake news sites came via search engines [2].

Broadly, fake news detection efforts center on three different categories of approaches:

context-based, style-based, and knowledge-based [48].

Context-based approaches use social network analysis techniques to determine whether a story is credible [48]. These approaches often rely on propagation patterns and post metadata. See [52] [31] [11] [36] for more information. Style-based approaches rely solely on the linguistic features of text to determine the stance of an article or detect deceptive language. Work in this area of fake news detection includes the research presented in [42] [53] [43].

While these techniques hold promise, the remainder of my thesis discusses knowledge-based approaches, sometimes referred to as fact-checking. Knowledge-based systems create computational tools that build on existing human processes used to detect and combat misinformation. Additionally, many of the most heralded potential solutions to online misinformation are knowledge based approaches [34].

3.2 Benefits of Fact-checking

One benefit of fact-checking is that exposure to a fact-check can correct a person's misperceptions . Results from the Annenberg Public Policy Center's 2012 Institutions of Democracy Political Knowledge Survey found that people who were exposed to fact-checking were more accurate in their knowledge of campaign background as well as candidate's stance on issues, even after controlling for demographic information [20]. A longitudinal study of the 2014 election found that exposure to fact-checking significantly improved the accuracy of people's beliefs." This study emphasizes differences in the effect size for different demographic groups: fact-checking is substantially more effective for people with higher education levels and who identify as Democrats [41]. Additionally, social media posts with a link to a fact-check in the comments contain linguistic signals indicating that people begin to doubt that rumor once a fact-check link is added [28].

A second benefit to fact-checking is that it limits the spread of problematic information. A recent partnership between fact-checkers and Facebook resulted in fact-checked articles rated false to reach 80% fewer people;[1] however, further data about the reach and the relative coverage of fact-checks has not been publicly shared with researchers [6]. Nevertheless, research comparing Twitter, which has taken a less proactive stance towards limiting misinformation, to Facebook, after they made changes to limit misinformation on the platform, found that sharing of misinformation occurred 60% less on Facebook after their changes. This included but was not limited to the Facebook-fact checker partnership [3].

Additionally, since the rise of third-party online fact-checking, political candidates have been more cognizant of fact-checking. When state legislators believed that their statements were being fact-checked, they reduced the number of inaccurate statements they shared [40]. Mark Stencel, co-director of the Reporters' Lab at Duke University, categorized the way politicians interact with fact-checking.[2] These interactions include using fact-checking to validate politicians arguments and dedicating financial and personnel resources to responding to fact-checks. Stencel explains how fact-checks are weaponized by candidates to support their own viewpoints and undermine their opponents.

3.3 Critiques of Fact-checking

The first critique of fact-checking is that it is extremely difficult to change people's opinions about what is true [18]. While there is evidence of some corrective effects when people who are exposed to untrue information are presented with fact-checks (see the benefits section), the evidence is not conclusive [39]. Additionally, other studies find that people don't only

[1] https://www.blog.google/outreach-initiatives/google-news-initiative/building-trust-online-partnering-international-fact-checking-network/

[2] https://www.politico.com/magazine/story/2015/05/fact-checking-weaponization-117915

share information that they believe is true [5]. This means providing fact-checks may not limit the spread of online problematic information. There is an epistemological question as to whether it is even possible to objectively fact-check a claim. Amazeen argues that almost all statements have nuance and context that makes it extremely challenging to label the factual correctness of language [51]. Critics of fact-checking find empirical evidence of the epistemological and methodological limitations of fact-checking in Marrietta et al., which illustrates that fact-checkers, when they fact-check the same claim, often do not assign the same truthiness value to the statement [33] (although [4] finds the opposite). Studies also find that fact-checkers claims rarely overlap [30] [33], leading to bigger questions about the methods of story selection, discussed in [7]. This critique is amplified by perceptions of partisanship and bias of the fact-checking organizations themselves. This is especially important because fact-checking requires that readers trust the fact-checker.

3.4 Automated Fact-checking Systems

Beyond Reviewed Claims, there have been several other fact-checking claim matching system prototypes. One of the early systems that attempted automated fact-checking was Truth Goggles [46], a browser extension feature that would highlight the content on a webpage that was questionable. The interface was designed to encourage critical thinking about the claim. However, Truth Goggles was never able to implement an automated approach. The project has been revived and there are current efforts to fully automate claim matching.[3]

ClaimBuster [25] is another automated fact-checking system currently in development. The creators describe the project as “an end-to-end system for computer-assisted fact-checking.” Using natural language processing, machine learning, and database querying techniques, ClaimBuster matches statements to existing fact-checks in real time.

[3] http://www.niemanlab.org/2018/04/truth-goggles-are-back-and-ready-for-the-next-era-of-fact-checking/

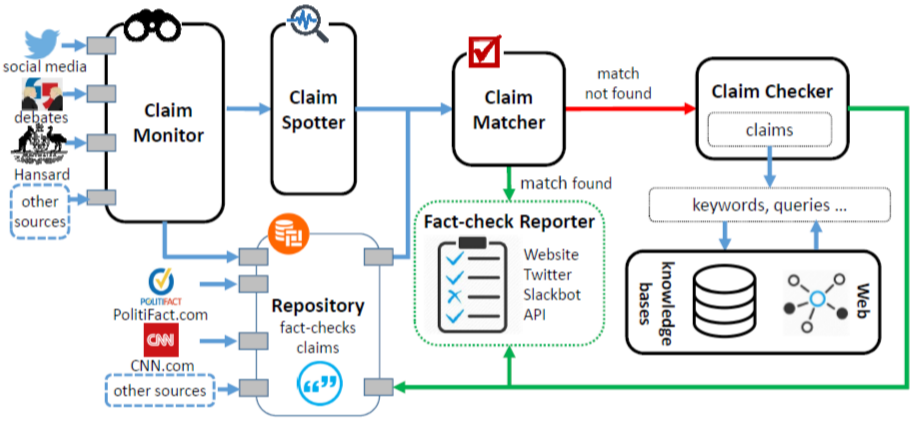


Figure 9: Major components of ClaimBuster (under continuous improvement and development)

Figure 3.1: An overview of the ClaimBuster system that highlights the 1) claim monitor, 2) claim spotter, 3) claim matcher, and 4) claim checker. Reprinted from [25].

ClaimBuster is a multi-part system, diagrammed in Figure 3.1:

- *Claim Monitor*: This element monitors texts from a variety of sources including the web, broadcast media, and social media.
- *Claim Spotter*: This system identifies “checkworthy factual statements” from the text provided from Claim Monitor using supervised learning. Some of the most important features include sentence sentiment, past tense verbs, and cardinal numbers.
- *Claim Matcher*: This is the part of the system that is relevant to the claim-document discovery process. Given a “checkworthy” factual statement, the claim matcher searches the database of known fact-checks for a relevant match. As for the technical implementation of this feature, all that is provided in the 2017 paper is that “the system uses two approaches to measuring the similarity between a claim and a fact-check. One is based on the similarity of tokens and the other is based on semantic similarity.” If the claim is not in the existing fact-check repository then it is processed by the Claim Checker. At Computation + Journalism 19, an update was given into the progress of

Claim Matcher. The authors came to the conclusion that the claim-document discovery problem required a human-in-the-loop [1].

- *Fact-check Reporter*: Synthesizes the results of Claim Matcher and Checker on the project website.[4]

This section illustrates 1) the structure of current automated fact-checking systems and 2) that claim matching is only one part of the automated fact-checking pipeline.

3.5 Sociotechnical Perspective

The term "sociotechnical systems" originated in organizational development, as a way to theorize about the interactions between workers and technology in English coal mines in the 1940s [50]. However, the term has been adopted by the Science and Technology Studies (STS) community to broadly discuss the complex relationships between social systems and technical components. In this subsection, I will focus on two papers that not only inform the methodology of this thesis but provide key insights in the related literature.

In "Fairness and Abstraction in Sociotechnical Systems," Selbst et al. contend that "to treat fairness and justice as terms that have meaningful applications to technology separate from a social context is...to make an abstraction error" [47]. Thus, they adopt a sociotechnical lens and identify five "traps" (common mistakes) of the fair machine learning (fair-ML) community. The authors explain that "each of these traps arises from failing to consider how social context is interlaced with technology in different forms, and thus remedies also require a deeper understanding of 'the social' to resolve problems." Below are the five traps:

1. *The Framing Trap*: "Failure to model the entire system over which a social criterion,

[4] idir.uta.edu/claimbuster

such as fairness, will be enforced."

2. *The Portability Trap*: "Failure to understand how repurposing algorithmic solutions design for one social context may be misleading, inaccurate or otherwise do harm when applied to a different context."

3. *The Formalism Trap*: "Failure to account for the full meaning of social concepts, such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms."

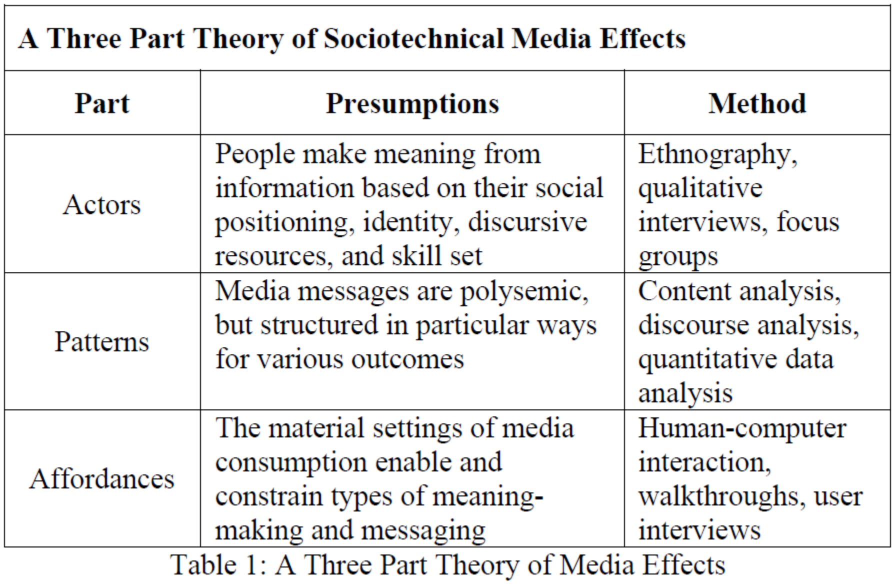
4. *The Ripple Effect Trap*: "Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system."

5. *The Solutionism Trap*: "Failure to recognize the possibility that the best solution to a problem may not involve technology."

The authors then proceed to describe case studies of each of the five traps and provide solutions. One such solution, targeted towards addressing the "framing trap," is by adopting a heterogeneous engineering approach [29] that includes inside "the boundaries of abstraction... people and social systems" as well as "local incentives and reward structures, institutional environments, decision-making cultures and regulatory systems."

In summary, Selbst et al. advocates that the "social must be considered alongside the technical in any design enterprise," rather than as an afterthought. They caution that technical systems that do not critically examine the social context and systems that exist are unlikely to be fair or just. They develop five categories common pitfalls listed above of current fair-ML models that can be used to discuss the shortcomings of other technical systems. Finally, the authors make some broad suggestions about how the fair-ML community can engage in more socially aware development in the future.

The second sociotechnical analysis "Why do People Share Fake News? A Sociotechnical Model of Media Effects" adopts a sociotechnical perspective of why people share problematic information online and finds that given the results, fact-checking and media literacy efforts are unlikely to succeed [34]. Marwick details her three-part sociotechnical model in the table and the corresponding methods in Table 3.2. From this table, it is clear that a wide variety of methods can be used to study sociotechnical effects.



A Three Part Theory of Sociotechnical Media Effects		
Part	**Presumptions**	**Method**
Actors	People make meaning from information based on their social positioning, identity, discursive resources, and skill set	Ethnography, qualitative interviews, focus groups
Patterns	Media messages are polysemic, but structured in particular ways for various outcomes	Content analysis, discourse analysis, quantitative data analysis
Affordances	The material settings of media consumption enable and constrain types of meaning-making and messaging	Human-computer interaction, walkthroughs, user interviews

Table 1: A Three Part Theory of Media Effects

Figure 3.2: An overview of the agents and methods described by Marwick. Reprinted from [34].

Marwick finds the combination of the following behaviors and conditions from the three outlined agents account for why people share fake news.

1. *Actors*: "Partisan Americans share fake news stories that support their pre-existing beliefs and signal their identity to like-minded others.

2. *Patterns*: "Successful problematic information builds on deep stories' found in mainstream conservative media or makes polysemic appeals that cross party lines."

3. *Affordances*: "Algorithmic visibility and social sharing massively increases the scale and spread of problematic information."

Marwick argues that these results call into question fact-checking and media literacy efforts. She cites *The Daily Caller's* reaction to Reviewed Claims, proposing that such fact-checking systems "may cause even more resentment and anger towards perceived liberal outlets, contributing to decreased trust." Marwick finds that her sociotechnical analysis calls for "a holistic approach" to thinking about solutions to sharing problematic information.

In summary, Marwick breaks down media effects into different categories to discuss why fake news spreads. She relies on a variety of different methods to understand why fake news spreads online and finds that fact-checking and media literacy efforts are unlikely to stem the spread of online problematic content.

3.6 Summary of Relevant Literature and Background

Listed below are the key takeaways from the past two chapters:

1. Online fact-checking has existed for over 25 years and is a growing field. Of the over 150 global organizations, there is a range of methodologies and objectives. Fact-checking is one of the key ways online information seekers are told to verify the credibility of sources, but research is mixed on fact-checking as a strategy to fight misinformation.

2. Automated fact-checking has the promise of 1) increasing the reach of fact-checking, 2) decreasing the time it takes for fact-checks to be applied to new online content. Fully automated fact-checking appears to be a distant dream. There are many attempts to create automated systems that either automate a piece of the fact-checking pipeline or have a "human-in-the-loop."

3. Reviewed Claims was a 2017-2018 feature deployed on Google search that incorporated an automated claim-document matching system. The feature was removed while under

substantial critique from conservative news outlets. Google cited “bugs” in the feature as the reason it was removed from the SERP.

4. The sociotechnical perspective involves the consideration of social actors and context in technical systems. Technical systems that do not consider social actors are unlikely to be successful.

Chapter 4

Technical Exploration of Reviewed Claims

The objective of this chapter is to understand the technical workings of the claim-document discovery process [54] that was most likely a key part of the Reviewed Claims system. What was the design the system? What limitations did Wang et al. [54] recognize in their system? What are other system design limitations?

4.1 Definitions

The following are definitions as provided in the Wang et al. paper, and I will adopt the terminology presented below for the remainder of my thesis.

Claim: the statement that is being fact-checked.

Claimant: the person or organization making the claim.

Verdict: the conclusion on the veracity of the claim according to the fact-checker.

Fact-checking article: an article that examines a single claim and produces a verdict.

Fact-checking article: an article that examines a single claim and produces a verdict.

Claim-relevant document: given a fact-checking article with claim c, a claim-relevant document is a related document that addresses c.

Supporting document: a relevant document that supports a claim c.

4.2 System Design

For claim matching to work, the system should identify with high confidence as many claim-relevant documents that support a claim c as possible.

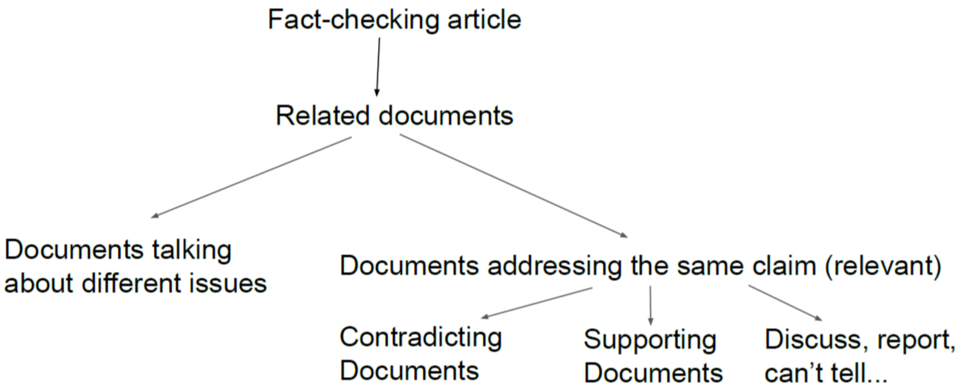

Fact-checking article
Related documents
Documents talking about different issues
Documents addressing the same claim (relevant)
Contradicting Documents
Supporting Documents
Discuss, report, can't tell...

Figure 4.1: An overview of the "claim-document discovery process." Reprinted from [54].

The system is summarized below:

> "We start from a set of fact-checking articles. For each article, we craft a set of queries and use a search engine to find a set of related articles. We then build a binary classifier to predict whether a related document is a relevant document. Finally, among all the relevant documents, we classify the stance of each relevant

document w.r.t. the fact-checking article and its claim" [54].

1. *Finding Related Documents* (i.e. candidate generation): The first task in the system is to find as many documents that "bear some topical or lexical similarity to the fact-checking article." Two approaches are taken to identify related documents.

   (a) Navigation approach: Wang et al. collects all outgoing links and citations from fact-check articles. This approach is not extremely valuable, as the authors find that most of these documents are related but not relevant to the fact-check.

   (b) Search approach: Wang et al. generates queries from the fact-check article and collects the Google search results that appear. Generating the ideal queries is a sizeable challenge. The authors list three different methods: 1) the text of the title and claim of the fact-checking article or associated ClaimReview data, 2) entity annotated representation of the title and claim, and 3) click graph queries of the 50 most popular search queries that led to the fact-check.

   The top-100 Google search results were collected for each query. In total, each fact-checking article generated 2,400 related documents for relevance classification. The click graph queries produced the best results of the three query generation strategies. Together, the three approaches had a recall[1] of 80%.

2. *Relevance Classification*: The authors construct a gradient boosted decision tree model[2] [55] to determine which of the generated candidate documents are relevant to the fact-check article. This classifier surpassed the accuracy of other relevance classifiers, with an accuracy of about 80%.

   The features of the model are all different measures of similarity between the fact-check article and the candidate document. They include:

[1] Recall is defined as the percentage of pertinent documents retrieved in terms of all pertinent documents.

[2] The gradient boosted decision tree is name of the machine learning model that the authors use to train and test their classifiers.

(a) Entity[3] similarity: To find the entity similarity, first, find the union of all the entities from the text of the claim (not of the entire fact-check article) and the entire text of the related document. Create two vectors, one for the fact-check article and one for the claimant document, with an associated confidence score for each entity. Then, find the cosine similarity[4] between the two vectors.

(b) Content similarity: In addition to the features that find the cosine similarity between the entire content of the fact-checking articles and the entire candidate document, there are many other features comparing the similarity of the claim text to the headline or title of the candidate document and comparing the claim text to individual sentences in the candidate document. This was identified as a particularly valuable feature. Claim text to candidate document paragraphs was attempted, as well as comparing "selected sentences" from the fact-check to sentences in the candidate document.

(c) Date published: Fact-checks and candidate documents that were published within a few days were more likely to be relevant.

3. *Stance Classification*: The authors create a gradient boosted decision tree model to determine whether a relevant document supports or contradicts the claim. Assuming that the fact-check article's verdict is false, relevant documents that support the claim are the documents that are the most important to discover as they are propagating problematic information.

   The authors adopt the strategy of attempting to identify contradicting relevant documents and pruning them out. To identify contradicting documents, Wang et al. compares "key textual evidence" (headline, title, important sentences) from the relevant document and the fact-check article claim. If any of the key textual evidence and article do not have a similarity score of greater than 0.8, the "key text" is not

[3] Entities include names, organizations, locations, etc.

[4] Cosine similarity is a common measure of similarity in text mining and information retrieval.

further considered. The remaining key text is concatenated with the sentence that comes before and after it. This allows for statements like “Does X happen? No, it does not” to be correctly analyzed.

Wang et al. constructed a lexicon of words that indicate a contradiction. They observe that headlines of fact-checks often contradict the claim. They extract the unigrams[5] and bigrams[6] from <claim, contradicting headline> pairs of approximately 3,000 fact-checking articles. They then constructed a vocabulary using the unigrams and bigrams with the greatest frequency. Some popular grams that indicated contradiction included: fake, purportedly, hoax, rumor, made up, fact check, not true, and no evidence. Then, for all the remaining key components (the key text + surrounding sentences), all the unigrams and bigrams are extracted and then a vector is constructed where each element of the vector is the frequency of a vocabulary word that was present in the key components.

4.3 Evaluation of Claim-Document Discovery System

Two different corpora are used to evaluate the claim-document discovery system. The first was a manual corpus, which was used to measure the performance of the candidate generation. Crowd workers[7] were recruited to discover as many supporting documents as possible through any means they wished. The recall of the candidate document generation was high (80%).

The second corpus, an unlabeled corpus was used to assess the accuracy of the classifier. The authors filtered out any fact-checks where the ClaimReview was missing a claimant,

[5] Unigrams of “to be or not to be”: to, be, or, not, to, be

[6] Bigrams of “to be or not to be”: to be, be or, or not, not to, to be

[7] Crowd workers are people who do crowd sourced tasks. Crowd sourced tasks often require a large pool of workers to perform a small task. Platforms like Amazon Mechanical Turk facilitate the interactions between those with tasks (like researchers) and crowd workers.

a claim, or a verdict. Additionally, duplicates and plagiarized fact-checks were excluded, in addition to fact-checks from organizations that were not part of the International Fact Checking Network. After filtering, approximately 15,000 fact-checking articles remained in the corpus. Then, to assess the relevance classifier, a sample of 8,000 of the 33 million (15k fact-check articles * 2.4K candidate documents) <fact-check, candidate document> pairs was presented to English-speaking crowd workers (with no additional qualifications). Workers were asked "Does the candidate document address the claim?" Three workers were asked about each claim (each worker could rate up to 6 claims). A <fact-check, candidate document> pair was considered relevant if at least two of the three workers found that the candidate document addressed the claim. The relevance classifier achieved 81.7% ± 1.8% accuracy (majority class label was 50%), which surpasses the accuracy of similar models. As for the misclassifications, the authors determine that they were primarily caused by either 1) bad inputs (e.g. erroneous publication dates, "poor" titles) or 2) not enough text (e.g. forum pages, online videos).

1,200 relevant documents were sampled to do stance detection. The same worker setup is adopted (English speaking, 3 workers per <fact-check, relevant document pair>, up to 6 claims), and workers were asked to classify the stance of each document as supporting, contradicting, neither, or can't tell. For 12% of pairs, two raters did not agree, these results were excluded. Their stance detection classifier achieved an accuracy of 91.6% (with 57% of documents being supporting documents) surpassing other models by approximately 2%.

4.4 Limitations Described in the Paper

The authors briefly discuss some examples where their system fell short. To explain why candidate generation did not include all related documents, the authors cite bad query construction and unfocused fact checks. They find that sometimes the article title and claim text are not good summaries for the content of the article. Moreover, they notice that when the fact-checking article covers several topics, the generated summaries may not be valuable.

As for the relevance classification, Wang et al. make two observations. The first is that sometimes the claim is a minor detail in the related document (ex. the claim is only mentioned once in passing). The second observation is that there is substantial variance in the number of relevant documents in a set of related documents.

The stance classification approach has shortcomings. The authors observe that some stance classifications require either an understanding of important context (i.e. domain knowledge) or an ability to interpret sarcasm. The authors note that current natural language processing techniques are not yet able to handle these challenging cases.

4.5 Additional Limitations

Defining Relevance in Fact-checking

RQ2 and RQ3 of this thesis focus on conceptions of relevance in fact-checking, and more specifically claim matching. In Wang et al. the definition of relevance is broadly construed. Recall that the authors define a claim-relevant document as: "a related document that addresses a claim, *c*." However, in their claim-document discovery process, not all "relevant documents" are relevant *matches*. The relevance of documents is determined in step two of their three step system. The stance classification of all relevant documents is a third step that determines whether the relevant *documents* are relevant *claim matches*. We could define relevant claim matches as relevant documents that support the claim (i.e. the stance classification model's determination), but this is never explicitly laid out in the paper. Moreover, supporting the claim is only defined by the model as not rejecting the claim. So, as long as words like “hoax”, “fake news”, etc. don’t appear, then the document supports the claim. So, “relevant documents that support the claim” is not a definition, but rather a post-hoc definition resulting from the reality of the designed system. However, Wang et al. does offer some clarification, explaining that “the claim-relevance discovery problem does not require the literal or precise claim to appear in the document but rather aims to find ones that seem to align in spirit.” Wang et al. do not delineate qualifications for aligning in spirit.

Thus, understanding if the “align in spirit” characterization of relevant matches is supported by information seekers or news publishers is a major focus of the rest of this thesis. In Chapter 5, I ask Wellesley undergraduate students and crowd workers to assess the relevance of some Reviewed Claims fact-checks. While I will save the majority of the analysis of those studies for 5.3, there does seem to be agreement among users and platforms that the exact claim does not have to be present in the document for there to be a relevant match. The challenge seems to be agreeing on what “aligns in spirit.” In Chapter 6, I extend this analysis to claimants (news publishers) and fact-checking organizations, which adds additional layers of complexity.

Examining the Claimants in ClaimReview

Bill Adair of PolitiFact believes that “ClaimReview is one of the untold success stories of the fact-checking movement.”[8] However, the Duke Reporters’ Lab estimates that approximately half of fact-checking organizations have not integrated the ClaimReview markup into their fact-checks.[9] As of October 2018, there are ongoing efforts to increase the adoption rates of the ClaimReview schema, including via the “Share the Facts” widget that has created an easy to use user interface for fact-check publishers.

While ClaimReview was developed in 2015, ideas about structured data being part of the

[8] https://www.poynter.org/fact-checking/2018/google-is-building-a-search-engine-for-fact-checks

[9] https://reporterslab.org/a-better-claimreview-to-grow-a-global-fact-check-database

web have existed since the earliest discussion of the Semantic Web in the early 2000s where it is described as:

> “The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users... [9]”

While the Semantic Web may not have come to fruition, structured data has become an increasingly important part of the web, especially on search. For example, robust fact-check structured data has always been viewed as a key prerequisite of automated fact-checking systems. The lack of such data was cited as a significant hindrance to the end game of fully automated real-time fact-checking [24]. However, since then, `schema.org`’s ClaimReview markup, as well as the Duke Reporter’s Lab/Jigsaw developed Share the Facts widget[10], have enabled fact-checks to appear on the SERP.

To create their corpus of fact-checks, Wang et al. depend on the ClaimReview markup that fact-checking organizations embed into fact-check articles. Not all fact-checks with ClaimReview are included in Wang et al.’s corpus: they filter out “any invalid markup where the ClaimReview fails to parse or is missing any of the three key fields (claimant, claim, verdict).” If there are variations in fact-checking organizations’ implementation of ClaimReview, this could limit the efficacy and reach of automated fact-checking systems. After observing inconsistencies in ClaimReview markup, I performed a systematic analysis of ClaimReview implementation.

I collected all of the fact-checks from the three most established fact-checking organizations: Snopes, PolitiFact, and Factcheck.org. This is the first publicly available detailed analysis of ClaimReview implementation of major fact-checking organizations (see Figure 4.2 and Table 4.1).

[10] http://www.sharethefacts.org

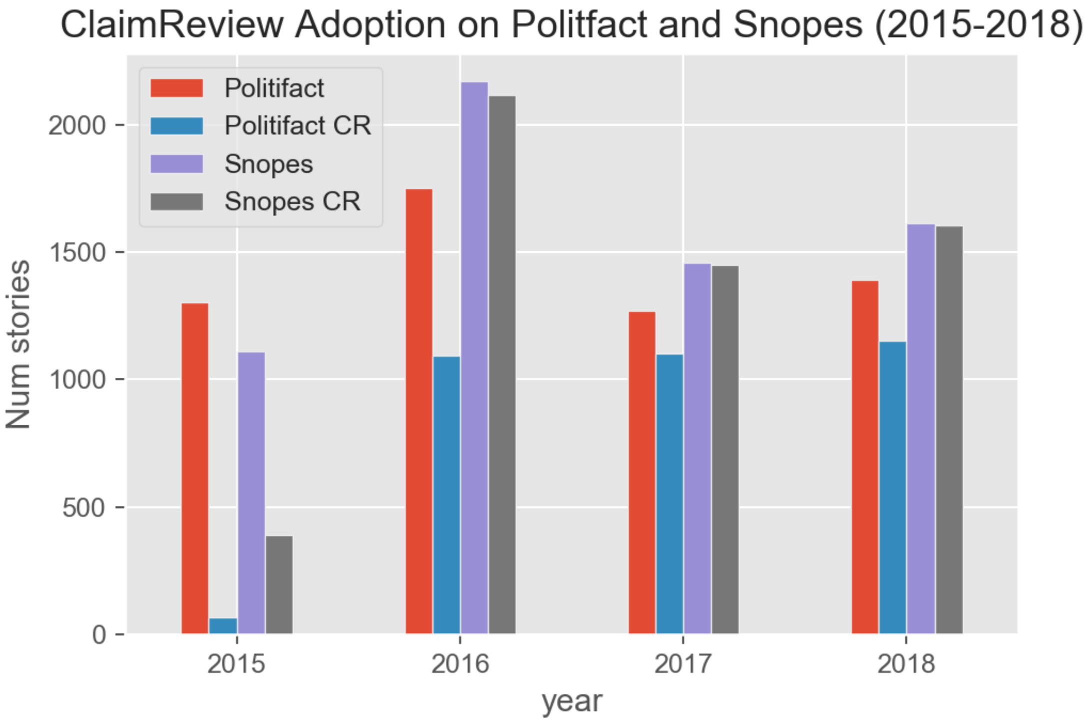


Figure 4.2: ClaimReview markup adoption rates from 2015-2018. Since, 2015, there has been a rapid adoption of ClaimReview markup. By 2018, almost all articles by Snopes and PolitiFact contained ClaimReview.

Snopes: Established in 1994, Snopes is arguably the oldest and most well-known online fact-checking organization. Snopes fact-checks are broken down into 45 categories[11] including racial rumors, science, and humor. Not all fact-check article have category tags, so I chose to use an automated Selenium browser[12] to navigate through the "Fact-check" section of the website. Snopes has a "News" section[13] of their website where they attempt to limit the spread of misinformation by providing a credible rundown of the day's news. These news articles do not meet the API's definition of fact-checking (see Chapter 2.1), so they have been excluded from my analysis. The metadata containing the ClaimReview metadata was in the form of JSON-LD (JSON for Linking Data).

Snopes published 1,613 fact-checks in 2018 (approximately 4 fact-checks per day). In

[11] https://www.snopes.com/archive

[12] Selenium is a tool that enables users to programmatically control a web browser.

[13] https://www.snopes.com/fact-check/

total, Snopes published 11,807 articles from 1995-2018. Of those, 6,680 contain ClaimReview metadata.

PolitiFact: Founded in 2007, PolitiFact[14] was originally owned by the *Tampa Bay Times* and is now controlled by *Poynter Institute for Media Studies.* Its mission is to fact-check political figures. Some pages have the JSON-LD metadata present on Snopes sites and others have microdata format present on FactCheck.org's pages. The microdata format is to be expected on the PolitiFact pages as it is the output of the user interface PolitiFact developed called the "Share the Facts" widget, as a way to increase metadata adoption rates.

In total, I collected 15,829 articles[15] from PolitiFact. In 2018, PolitiFact published 1,392 fact-checks. Of the 15,829 articles, 3,766 contained Claim Review markup.

Factcheck.org: Founded in December 2003, FactCheck.org has existed for 15 years as part of the Annenberg Public Policy Center and primarily fact-checks politicians. Initially, I collected 1,607 articles from FactCheck.org that were written between 2003 and 2018. However, like Snopes, FactCheck.org publishes other content besides fact-checks including a mailbag that contained quizzes and daily summaries. However, unlike the Snopes website, I was unable to easily differentiate from the website structure between fact-checks and other content. Closer examination of the headline structure of articles revealed that content with a date in the title followed by a colon (e.g. "April 24: Fact-Checking, Jobs, Climate Change") were not fact-checks. After filtering, 1,395 fact-checks remained. Of the 1,395 remaining fact-checks, only 135 (10%) contained the ClaimReview metadata. Because of the small number of articles with ClaimReview markup, I have only included PolitiFact and Snopes in further analysis.

Wang et al.'s method of only using fact-check articles that explicitly name a claimant excludes almost all the fact-check articles created by Snopes, which is the largest producer

[14] PolitiFact.com

[15] Not all were fact-checks and about 20 pages were no longer active URLs.

of fact-check articles. Of the 7,063 Snopes articles I scraped that contain the ClaimReview markup, only 1,297 (18%) contain a claimant name in the ClaimReview markup. Of those, 1,190 contain the source name "Multiple Sources." Thus, only 1.5% of Snopes fact-checks with ClaimReview name a specific claimant. Thus, there are additional questions to be asked about how the exclusion of so many fact-check articles impacted the training and testing of the Wang et al. model. Only eight claimants were listed more than once in ClaimReview markup by Snopes: Breitbart News (5), Last Line of Defense (4), Neon Nettle (4), 8shit.net (2), Ladies of Liberty (2), Freedom Daily (2), Fox News (2), and All News 4 Us (2).

However, every PolitiFact article has a claimant listed. Of the 404 claimants that are mentioned more than once (that comprise 74% of all claimants listed), only 28 (7%) are news publishers. 80% of the claimants are public figures, usually politicians (e.g. Donald Trump (516), Hillary Clinton (107)). Thus, by only including fact-checks with a claimant, Google's algorithm is "throwing away" most of the data. It is possible that Google has a shared database with fact-checkers or another algorithm to uncover more claimants, but there has been no media reports or other indication in the paper [54].

Variance in ClaimReview Implementation

From the analysis in ClaimReview in the previous subsection, it is clear that there is some variation in the way fact-checking organizations use the fields of ClaimReview markup. This is a common trade-off of decentralized data systems that Wang et al. recognize. Nevertheless, there has been no comparison of ClaimReview implementation between various fact-checking organizations. I have organized the prevalence of various features in the ClaimReview markup in Table 4.1.

Of the 32 fields, 18 fields appear in both ClaimReview markup. Only 14 fields appear in every ClaimReview for both of the largest fact-checkers. These substantial differences in

ClaimReview Fields	**PolitiFact**	**Snopes**	**Fields cont'd**	**PolitiFact**	**Snopes**
@context	3766	7063	itemReviewed/author/image	3766	0
@type	3766	7063	itemReviewed/author/jobTitle	3766	0
author	3766	7063	itemReviewed/author/name	3766	1297
author/@type	3766	7063	itemReviewed/author/sameAs	3766	0
author/logo	0	7063	itemReviewed/datePublished	3766	0
author/name	0	7063	itemReviewed/name	3766	0
author/sameAs	0	7063	reviewRating	3766	7063
author/twitter	28	0	reviewRating/@type	3766	7063
author/url	3766	0	reviewRating/alternateName	3766	7063
claimReviewed	3766	7063	reviewRating/bestRating	3766	7063
claimReviewSiteLogo	28	0	reviewRating/image	3766	7063
datePublished	3766	7063	reviewRating/Name	3766	7063
itemReviewed	3766	7063	reviewRating/name	28	0
ItemReviewed/@type	3766	7063	reviewRating/ratingValue	3766	7063
itemReviewed/author	3766	1297	reviewRating/worstRating	3738	7063
itemReviewed/author/@type	3766	0	url	3766	7063

Table 4.1: Fields present in ClaimReview markup of Snopes and PolitiFact. There are a total of 3,766 PolitFact articles and 7,063 Snopes articles. The differences in the implementation of the itemReviewed related fields are especially important given Wang et al.'s [54] approach for claim matching.

implementation reflect differences in the fact-checking organization approaches as well. For example, as mentioned in the previous subsection, PolitiFact primarily fact-checks people, not news publishers. Thus, there is the field jobTitle in ItemReviewed/author. More thought is needed about whether ClaimReview is truly standardizing fact-checks.

4.6 Summary

In this chapter, I provide a detailed description and analysis of the 2018 paper "Relevant Document Discovery for Fact-Checking Articles" which described the technical implementation of the claim-matching component of the Reviewed Claim system. There are several important takeaways from this chapter.

1. The objective of "Relevant Document Discovery for Fact-Checking Articles" is to identify a set of documents that are relevant and "align in spirit" to a given fact-check

article. The authors were able to beat the baseline performance of similar approaches by using a three step system: 1) candidate generation, 2) relevance classification, and 3) stance classification.

2. The outlined system relies upon existing ClaimReview data to identify and compare fact-check articles to candidate documents. However, ClaimReview implementation rate and style depends significantly on the fact-checking organization.

3. Wang et al. define relevance as: “given a fact-checking article with claim c, a claim-relevant document is a related document that addresses c.” And later that definition is supplemented with the “note that the claim-relevance discovery problem does not require the literal or precise claim to appear in the document but rather aims to find ones that seem to align in spirit.” As explored in future chapters, this definition does not fully model the complex idea of relevance.

In the following chapters, I explore how the claim-document discovery process described in this chapter models information seekers and other stakeholders’ notions of relevance and fairness.

Chapter 5

A User-Centered Examination of Reviewed Claims

The Reviewed Claims feature was designed for information seekers to better assess the credibility of news sources. Thus, the ways in which they perceive the Reviewed Claims feature should be of the utmost importance. This chapter describes how information seekers define relevance and generate justifications information seekers adopt to determine whether a claim is relevant to a document. Using a dataset of algorithmically assigned <fact-check, news article> pairs (i.e. <claim, document> pairs discussed in Chapter 4) and various populations of users' ratings and justifications of relevance, I analyze whether there is or can be a common conception of relevance among information seekers.

5.1 Data

As part of the Cred Lab's long-term research on web literacy and credibility of online sources, the Cred Lab[1] regularly monitors various Google search engine result pages (SERP). That is, for many query phrases, Google's search result pages are automatically collected and analyzed. In early January 2018, as part of a large data gathering about media organizations in the United States [32], I collected data from various lists about news publishers (newspapers, online news websites, fake and satire news, etc.)[2].

We collected the data of approximately 8,000 news publisher SERPs. Analyzing the content of over 8,000 collected SERPs, we identified 59 news publishers whose Knowledge Panels contained the Reviewed Claims tab. In this thesis, I examine 221 fact-checks that appeared in 59 Reviewed Claims tabs and focus my analysis on 118 algorithmically assigned fact-checks.

There were 119 algorithmically assigned fact-checks, but one claimant document only contained a video and that video was no longer accessible, so it was impossible to assess the relevance of the fact-check. 91 other fact-checks referred to the news publisher via the fact-check text, hyperlink, or metadata and 11 articles were excluded for analysis because I was unable to access the claimant document.

[1] I am extremely grateful to my fellow Cred Lab members and Professor Eni Mustafaraj for making all of this data available to me for this thesis.

[2] Details about the lists used for collection are available in the Appendix

5.2 Methods and Results

Author Labeling

I manually labelled 118 <claimant document, fact-check article> pairs (which did not explicitly mention the document in the fact-check) as "relevant" or "irrelevant," denoting the confidence level (low, medium, high) for each pairing. I defined a match as relevant if the fact-check applies to the central claim of the document. This process can be regarded as the human replication of the "relevance classifier" and assessment of whether a <claim, document> pair aligns in spirit as defined in [54].

	Confident	**Somewhat Confident**	**Not confident**
Relevant (n=93)	60	24	9
Not Relevant (n= 25)	3	12	10

Table 5.1: Labels and confidence levels for the 118 algorithmically assigned fact-checks. It was often more challenging to label a match as "not relevant" than "relevant."

Expanded Labeling Effort

Because I had low confidence in several of my relevance determinations, I selected 30 pairs that I was not confident about in the first pass of manual labelling. I recruited three undergraduate Wellesley College students to independently assess the relevance of these matches. I chose three raters because Google relied on three crowd workers to determine relevance (see Chapter 4). This was a labor-intensive task and took labelers between 5-10 minutes to evaluate each claim and provide justifications[3].

From the three raters, I ascertained an estimate of the accuracy of the overall Reviewed Claims classifier. The raters were instructed to read only the claim of the fact-check in the ClaimReview markup and the claimant document. If they were not confident in their

[3] Raters were compensated by the Cred Lab at their typical hourly rate (of $13/hr)

relevance assessment, they were then encouraged to read the entire fact-check article. In the majority of cases, the raters read the entire fact-check. Raters were instructed to not pass judgment on the veracity of the claim itself or to read sources other than the fact-check.

Raters were asked to assess matches as relevant or irrelevant, as well as provide a high, medium, or low confidence rating in their judgments. All of this data was labeled in a spreadsheet. Of the 30 selected claims that the three raters to analyzed, our raters had a fair agreement rate ($\kappa = 0.29$)[4], indicating the difficulty of this task. 16 of the 30 claims (53%) were rated as irrelevant to the assigned fact-check by at least two raters.

The raters were asked to provide a 1-3 sentence justification of their labels. Then, I generated five general justifications for *irrelevance* using a grounded theory approach to qualitatively code [16] the three undergraduate students raters justifications for why the documents were or were not relevant to their algorithmically assigned fact-check articles. From these responses, five justifications emerged that explain why a fact-check may not be relevant. Rater responses often included multiple justifications. The justifications included:

(a) The fact-check concerns a minor detail in the article.

(b) The fact-check verifies a different claim than what was in the article.

(c) The fact-check and article cover the same general topic but differ on details.

(d) The fact-check and article are framed differently.

(e) The fact-check is more general than the article.

The labelers and I then met for two hours to reconcile their findings. Research has found that a small number of knowledgeable raters can outperform a large number of crowd workers [37] and reconciliation of raters' judgments is commonly performed in HCI studies. The resulting agreement rate of agreement is $\kappa = 0.60$, which indicates moderate agreement

[4] https://en.wikipedia.org/wiki/Fleiss%27$_kappa$

between raters. The same 16 of the 30 claims (53%) were rated as irrelevant to the assigned fact-check by at least two raters.

Amazon Mechanical Turk Study

Then, I designed a Amazon Mechanical Turk (MTurk) study with IRB approval. Amazon Mechanical Turk[5] is an online crowd work platform that allows researchers to easily recruit and compensate study participants by having workers complete various tasks, known as HITs. There were two objectives in this phase:

1. Ascertain whether crowd workers had similar relevance determinations as the three raters.
2. Test if crowd workers selected the same justifications for a given claim.

Amazon Mechanical Turk studies are typically recognized to be substantially more demographically diverse than standard college student samples, and slightly more diverse than standard Internet samples [13]. As further detailed in the Limitations section, this sample was not representative of the U.S. adult population. Moreover, I recognize the potential for ethical abuses on Mechanical Turk and have followed academic best-practices of survey design and worker compensation[6].

Participants were asked for standard demographic information as well as about their political affiliation and news consumption habits. Following best practices for crowd worker payment, workers were compensated $5 per HIT for a survey that was designed to take around 30 minutes. The instructions mentioned that the task required close reading. To screen for high-quality results, I rejected HITs where Turkers who took less than 5 minutes

[5] https://www.mturk.com

[6] http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

to complete the survey, as it would be impossible to read three article pairs in such a short amount of time. In the first round, thirteen people completed the survey (three were rejected). In the second round, eleven people completed the survey (one was rejected). The mean completion time for the two rounds were 22 and 15 minutes respectively.

Compared to the crowd worker qualifications adopted by Wang et al. [54], stringent worker qualifications were placed on HITs. Specifically, I required 98% HIT approval rate, U.S. residency, >1,000 accepted HITs, and high political knowledge. To ensure a high level of political knowledge, I included political screening test that asked: 1) who is the Senate Majority Leader?; 2) Which amendment protects the right to keep and bear arms?; and 3) who was the U.S. Secretary of State in 2017? The qualifications and screening test are modeled after the experiment discussed in [12], which employed crowd workers to label the partisan and ideological leanings of various news articles.

Three pairs of <fact-check, claimant document> were shown to participants. Because I first implemented a pilot of showing three stories in a randomized order to 10 participants, I continued using the format of exposing the same three stories to participants in the second round of pairs. Similar to the rating scheme I employed with the three raters, participants were first asked to make a determination of relevance and then provide a confidence rating of their decision. This is an unconventional method to ascertain survey responses, a more traditional survey may have combined the two questions with Likert scale-based questions, but part of my objective was to assess the accuracy of the binary Reviewed Claims classifier. A Likert scale has the ambiguity of the neutral option, but the relevance classifier described in [54] does not have a neutral option.

Because one objective was to assess the relevance of a fact-check given the article's content, my goal was to remove all potential sources of bias from non-content related features, including source identifying content or references to other media organizations. Thus, I provided links to static versions of the text of the fact-check and article. I removed from these copies of the page news source identifying information and styling as well as all external hyperlinks. In addition, I changed all the references to other media organizations (e.g. *YourNewsWire*, *The Washington Post*) to generic sounding news organization names (e.g. *Denver Herald*, *Atlanta Post*).

One objective of this experiment was to assess whether information seekers can agree on the relevance justifications across individual <fact-check, news article> pairs. To measure this, instead of leaving the justifications open-ended, I created a multiple choice question bank with the five possible justifications described in the previous section as well as a sixth option for relevant matches: "the article and fact-check are a good match." Because discussions with the three student raters confirmed that not all of the justifications individually warranted a fact-check to be relevant or not relevant, but rather it affected raters' confidence scores, I decided to present the justifications to survey participants even when they had rated a story as relevant. Further discussion of the justifications will be included in Chapter 6.3.

Claim	**# of raters** who found the claim relevant (n= 3)	**# of Turkers** who found the claim relevant (n=10)
California will have the "highest gas tax in the nation" once its 12 cent gas tax hike goes into effect.	2	10
Says Ted Kennedy met "with the KGB in order to beat Ronald Reagan in 1984."	2	9
"The temperature is not rising nearly as fast as the alarmist computer models predicted. You know, it's much, much less, factors of 2 or 3 less."	2	9
Police in Charlottesville were issued a "stand down" order and told to let violence happen.	1	10
Bill Clinton was expelled from Oxford University for raping a British classmate named Eileen Wellstone.	0	10
A campaign ad attacking Republican Sen. Jeff Flake falsely says that he hasn't worked with President Donald Trump to 'repeal Obamacare.'	0	5

Table 5.2: Rater evaluations of the six claims shown to Mechanical Turkers. Only 1 claim was found to be not relevant by more than 1 crowd worker (see Table 5.3).

Of the six claims examined by the raters, five were found to be relevant by nine or more raters (see Table 5.2 for more details). On average, 3 different justifications appeared across raters for a given <fact-check, article> pair. Of the five claims that at least nine raters found relevant, only an average of 5.2 raters selected the justification that "The fact-check and article were a good match." See the breakdown of justifications of the claim: "A campaign ad attacking Republican Sen. Jeff Flake falsely says that he hasn't worked with President Donald Trump to repeal Obamacare." See Table 5.3 for more detail.

	Marked relevant	**Marked not relevant**
The fact-check and article cover the same general topic but differ on details.	4	1
The fact-check verifies a different claim than what is in the article.	1	3
The fact-check concerns a minor detail in the article.	1	1
The fact-check and article are framed differently.	1	0

Table 5.3: Justifications selected for the claim: A campaign ad attacking Republican Sen. Jeff Flake falsely says that he hasn't worked with President Donald Trump to "repeal Obamacare." There were a wide range of justifications used to explain raters relevance assessment. No crowd worker used the "good match" justification, despite five crowd workers rating the claim as relevant. Note: Crowd workers could select more than one justification.

5.3 Analysis

These three qualitative labeling exercises produced some novel findings and some new questions. First, it is significant that **at least 119[7] of the 221 fact-checks (57%) visible on Reviewed Claims, were algorithmically assigned**. This significant portion reveals that the conceptual challenges of determining relevance in terms of the algorithmic assignment of fact-checks is worthy of substantial consideration given the dependence of the Reviewed Claims system on such matches.

If the ability to agree on a common definition of relevance is a positive outcome, as I believe it is, the most promising finding of this chapter is the **significant improvement in the agreement rate after the three independent raters discussed their relevance assessments**. Note that the final agreement rate was 0.6, which is technically classified as moderate agreement, but 0.61 is classified as "substantial agreement." This indicates that

[7] I remark "at least 119" because 1) I was unable to access all claimant documents and 2) I was overly cautious in labeling something an algorithmically assigned fact-check. It was most likely unnecessary to exclude results that was only mentioned as an archive.is link to that site, with no mention of the source name or url in the text, source code, or ClaimReview markup.

at least among information seekers, there is the possibility for a common conception of a relevant fact-check.

The labeling exercise with crowd workers does not show any meaningful agreement with the Wellesley student raters. **At least half of the crowd workers found every claim to be relevant**, despite at least two Wellesley raters finding three of the six claims not relevant. One possible explanation for this discrepancy is that the Wellesley student raters gave this task more attention and thought than the crowd workers. However, comparing the mean completion times of the accepted HITs and the self-reported times for the Wellesley raters, there is not a substantial difference. Another possible explanation is that Wellesley students and crowd workers are two distinct populations who have substantial differences in their conception of relevance. Further experiments are needed to determine the cause of the discrepancy.

Before the reconciliation took place, I analyzed the independent raters' justifications and generated six justification responses to add to the Mechanical Turk multiple choice justification question (five responses for potentially not relevant matches and one for relevant matches). Before we began the reconciliation process, I presented the three independent raters with the justifications and solicited their suggestions for improvement. The raters were positive about the six justifications I generated, and as we went through the reconciliation process they were able to assign at least one of the justifications to the <article, fact-check> pair. **However, while all three raters agreed that while the justifications fit well, "there can be differences in the match without reaching the threshold for being a bad match."** This comment indicated that the justifications would be an insufficient tool for understanding relevance. I added an "Other" option to the justification options with an associated text box, in case that crowd workers had better suggestions, but no workers added their own justification. Sure enough, raters were not in agreement about which justifications to mark, even when labeling a pair as relevant (see Figure 5.3).

As mentioned in the Results section, of the five <fact-check, news article> pairs that at least 9 people rated as relevant, only an average of 5.2 selected "the article and fact-check are a good match.' This indicates that **it is not enough to define relevance, but also what is not relevant.** Similarly, 5 workers rated the claim: "a campaign ad attacking Republican Sen. Jeff Flake falsely says that he hasn't worked with President Donald Trump to repeal Obamacare" as relevant, but none selected the justification "the article and fact-check are a good match."

This reinforces the Wellesley raters' intuition about there being a threshold for marking something as a bad match. This threshold sounds similar to Wang et al.'s [54] comment about relevant documents not having to be precise matches, but rather aligning in spirit.

Given that at least half of all crowd workers found that each of the claims was relevant, it makes sense that the model described in [54] classified all of these documents as relevant, as Google used crowd workers to train its model. However, recall that Google's criteria for crowd workers was substantially less stringent than mine and that they only had three crowd workers assess each < fact-check, news article> pair requiring no other qualifications. While I do not aim to quantify the impact those choices had on their results, it is unlikely that Google's raters were more selective than mine. Thus, in both cases, it seems that having crowd workers assess relevance was not reflective of how the larger fact-checking ecosystem considers relevance. If it was, then there would not have been backlash due to the feature assigning irrelevant fact-checks to news publisher SERPs.

All three of Wellesley raters remarked that they struggled to understand some of the fact-check articles. They also noted that some of the matched news articles were difficult to understand, but this is less surprising as problematic information (fake news, extremely partisan articles, etc.) operate by misrepresenting events and are not held to the same journalistic standards as reputable sources. But, the fact-check articles should be elucidating, not confusing. When pressed further on why the articles were confusing,

raters elaborated that some of the fact-checks were extremely long, included many details, and required high levels of political knowledge. One rater admitted that she guessed on two ratings because she did not understand the fact-check. This should be a concern for fact-checkers as if people are struggling to understand the fact-checks as written, they are certainly not changing people's minds. There is other research on whether fact-checking is corrective (see Chapter 3.2 and 3.3), and this thesis does not aim to make a conclusion one way or another. Thus, I elected to ensure that the six examples I selected for the crowd workers to label did not include the articles the Wellesley raters had flagged as confusing.

5.4 Limitations

There are several limitations of this user-centered approach. One limitation is that I was only able to collect 59 Reviewed Claims panels. However, it seems likely that there were not hundreds of these panels as these 59 Reviewed Claims panels were incidentally collected from a sample of over 8,000 news publisher SERPs. Nevertheless, the data collection of Reviewed Claims feature was not systematic and likely excluded some news publisher SERPs that contained a Reviewed Claim component.

Moreover, we are only studying two questions in this chapter: 1) which "claim matches" are considered relevant by information seekers? and 2) what justifications explain relevance for information seekers? This means that we are ignoring another question: was the Reviewed Claims feature useful to information seekers? While this question is interesting, and similar to research questions I explored in previous work [32], I decided not to explore this in my thesis as the Reviewed Claims feature no longer exists. While the first two questions may be applicable to other "claim matching" fact-checking systems, the results about whether Reviewed Claims was useful would be difficult to separate from the feature itself.

Another limitation is that only one individual labeled the 118 claims. However, when compared to the expanded labeling effort which employed three independent, these ratings were the same as at least two of the three raters in 22 of 30 cases. Ideally, the three independent raters would have labeled the 118 claims, but time and financial resources were limiting factors. Therefore, author labeled data should not be considered the primary finding of this section, but rather a rough estimate of the accuracy of the Reviewed Claims system.

In terms of the expanded labeling effort, all three raters were first-year Wellesley College students who are members of Professor Eni Mustafaraj's Cred Lab. This makes it likely that they had significant exposure to concepts of credibility and trustworthiness before this experiment as well as similar age and education levels. There was no time limit on their labeling, while there was a provided time range for the Mechanical Turkers. Additionally, the Wellesley labelers were provided the actual URLs for the fact-check article and claimant document, so biases about the sources themselves may have come into play. The labelers also remarked that two fact-check articles were extremely challenging for them to comprehend, indicating that their assessments of at least two stories may not be accurate.

The Mechanical Turk study also has several limitations. One is a lack of demographic diversity of Mechanical Turk workers. Of the 20 workers, 12 were men and 17 were white. Like many other MTurk studies the mean age skews young and there was a similar bias towards more educated workers. There is also the issue about whether the crowd workers had the requisite background knowledge to be able to assess the relevance of the <fact-check, news article> pairs. I attempt to mitigate the effect of this limitation as Budak et al. [12] did with the political knowledge test, but knowing recent political history is different than being able to critically think about the <fact-check, news article> pairs.

5.5 Summary

As further explored in Chapter 6, information seekers needs were not taken into consideration as the Reviewed Claims feature was removed without meaningful explanation. In this chapter, I describe and analyze the results of multiple qualitative data labeling efforts. Overall, the takeaways from this chapter are:

1. There are mixed results on how many of the algorithmically assigned matches are "relevant matches." My labels and Wellesley College student raters labels indicate substantially lower raters of relevance than those of crowd workers.

2. I was unable to identify a clear set of criteria for what makes a match relevant, but generated several possible candidate criteria.

3. Crowd workers higher threshold for algorithmically assigned fact-check may help explain why Google's algorithm was so generous with its classification of relevance (as the model was trained on crowd worker labels).

Chapter 6

Reviewed Claims: A Sociotechnical Perspective

In this chapter, the social context of the Reviewed Claims system will be explored in depth by examining the interactions and sociopolitical position of actors in the Reviewed Claims ecosystem.

First, there is the **platform**. Fact-checks can be displayed in a variety of formats on platforms, and multiple platforms have prominently featured fact-checks in their content. Google developed an algorithm to match claims with fact-check articles. Fact-checks were featured not only in the organic search results, but also in a subset of news publishers' SERPs in the Reviewed Claims component. The focus of this analysis is Google's Reviewed Claims feature.

Second, there are the **claimants**. Claimants are the entities that either utter claims or produce online content that is fact-checked. These can be individuals in the public sphere (usually politicians), news organizations, or online news outlets. For the purposes of the Reviewed Claims feature, which only was displayed on a subset of news publisher SERPs, the focus will be on news publishers.

Third are the **online information seekers**, who can benefit from being shown that a specific claim is disputed or untrue. The public at large does not know much about fact-checking or the names of fact-check organizations. Web literacy best practices[15] encourage people investigating the veracity of claims to start by searching for previously written content about a claim. Thus, displaying fact-checks as part of a platform's interface may be an important way for the public to access additional information about source credibility.

Next are the **fact-checking organizations** like Snopes, PolitiFact, and *The Washington Post* Fact-check. They employ experienced human fact-checkers who select the claims that they will verify and then generate fact-check articles (sometimes) embedded with ClaimReview markup. These organizations are actively trying to increase the size of their audience, which is orders of magnitude smaller than that of problematic information. Simultaneously, these fact-checking organizations must be perceived as politically neutral and impartial to be able to correct misperceptions.

From these descriptions alone, it becomes clear that these agents have different agendas and motivations within the fact-checking ecosystem. I follow Marwick's technique in "Why do People Share Fake News?" [34] of using multiple methods to understand the sociotechnical perspective. I 1) conduct a content analysis that used the published writings of the various actors to explore how these differences played out with the Reviewed Claims feature as well as 2) supplement the lack of public writings from the information seeker perspective with those from the qualitative labeling discussed in Chapter 5.

6.1 Platform

When Google released the Reviewed Claims feature with great fanfare in November 2017, it never disclosed that it was algorithmically assigning fact-checks. As further explored in the Chapter 6.3, not even fact-checking organizations seemed aware of the way the Google was using claim matching to assign fact-checks to news publishers.

Google, the platform, not only designed and released the feature, but also removed it. After facing pushback to some of the algorithmically assigned fact-checks detailed in the next section, Google removed a few fact-checks. Facing continued public pressure from claimants, Google removed the feature in January 2018. No public explanation was provided to information seekers, only this comment to *Poynter* and *Daily Caller* journalists:

> “We launched the reviewed claims feature in our Knowledge Panel at the end of last year as an experiment with the aim of helping people quickly learn more about news publications...We said previously that we encountered challenges in our systems that maps fact checks to publishers, and on further examination it’s clear that we are unable to deliver the quality we’d like for users”[1].

Significantly, in this comment, the complaints of the claimants are not mentioned, only concerns about the quality of the feature for information seekers. From this statement, it appears that users may have complained about the quality of the feature, not news publishers. However, there is no indication that this was the case.

I discuss the details of the Reviewed Claims system in Chapters 2.4 and 4. In this remainder of this section, I will use the framework provided by Selbst et al.in “Fairness and abstraction in sociotechnical systems” [47] as a way to examine three ways the Reviewed

[1] https://www.poynter.org/fact-checking/2018/google-suspends-fact-checking-feature-over-quality-concerns

Claims system failed.

The Framing Trap: *"Failure to model the entire system over which a social criterion...can be enforced."*

The definition of relevance used by the claim matching algorithm that likely underpins the Reviewed Claims feature clarifies that relevance "...does not require the literal or precise claim to appear in the document but rather aims to find ones that seem to align in spirit. The formal definition Wang et al. provide for a relevant document is "given a fact-checking article with claim c, a claim-relevant document is a related document that addresses *c*."

The inability of information seekers to consistently agree on which claim matches are relevant (described in Chapter 5), exposes that relevance is difficult to model. News publishers' critiques (discussed further in the next session) of the algorithmically assigned fact-checks on Reviewed Claims further illustrates that Wang et al.'s definition, on which the whole feature is built, does not fully model relevance.

The Portability Trap: *"Failure to understand how repurposing algorithmic solutions design for one social context may be misleading, inaccurate or otherwise do harm when applied to a different context.*

While the Reviewed Claims feature was never applied to a context outside fact-checking, the claim-document discovery process outlined in "Relevant Document Discovery for Fact-checking Articles" [54] was incorporated into the Reviewed Claims system without resolving the limitations described in the paper. Having a less than perfect accuracy rate is acceptable for a conference paper, but those errors become consequential when news publishers become angry when algorithmically assigning fact-checks. In response to the irrelevant claim matched fact-checks, publishers alleged that fact-checking organizations and Google are biased against conservatives. This is especially concerning as public trust in U.S. media institutions and fact-checkers is dropping.[2] In Marwick's "Why People Fall for Fake News?" [34] she believes the Reviewed Claims feature may have caused "even more resentment and anger towards perceived liberal outlets, contributing to decreased trust."

The Solutionism Trap: *"Failure to recognize the possibility that the best solution to a problem may not involve technology."*

While I do not think it is fair to criticize Google for attempting to generate a technical solution, it seems unlikely that modeling relevance can be done without human participation (beyond labeling training data). Moreover, given the fragile fact-checking ecosystem, it may be that claim matching systems that rely on opaque machine learning algorithms are destined for failure in the fact-checking ecosystem.

6.2 Claimants

In the context of Reviewed Claims, the claimants are online news publishers. We can best understand their conceptions of relevance through their objections to the Reviewed Claims algorithmically assigned fact-checks.

Of the 59 claimants with Reviewed Claims panels, 12 published an article or post[3] about the Reviewed Claims feature. These news publishers had several objections to the Reviewed Claims feature.

The Daily Caller article about Reviewed Claims was the first piece of online content that was critical of the feature.[4] The 10 other articles about the Reviewed Claims feature

[2] https://www.washingtonpost.com/news/fact-checker/wp/2018/06/25/rapidly-expanding-fact-checking-movement-faces-growing-pains

[3] The claimants are: *Breitbart*, *Daily Caller*, *Daily Wire*, *Gateway Pundit*, *Federalist*, *Free Republic*, *Freedom Outpost*, *En Volve*, *Democratic Underground*, *Zero Hedge*, and *Natural News.*

[4] https://dailycaller.com/2018/01/09/googles-new-fact-check-feature-almost-exclusively-targets-conservative-sites

reference *The Daily Caller's* piece and several excerpt from it. However, *Daily Caller's* criticism of the feature is not limited to the algorithmically assigned fact-checks. The article instead focuses on three main issues:

1. Four of the fact-checks in the Reviewed Claims panel are not relevant to *Daily Caller's content*.
2. Google is a liberal company and Reviewed Claims reflects their political biases.
3. Comparable left-leaning sites do not have the Reviewed Claims component on their SERP.[5]

Exploring Claimant Justifications for Relevance

While only one of the four fact-check articles mentioned in the piece does not name *The Daily Caller* in the fact-check (which is how I define algorithmic fact-checks in my analysis), the critiques of the four fact-checks provide insight into how claimants conceptualize relevance. *The Daily Caller's* critique makes the following points:

1. **The fact-check was just reporting the facts, there was no broader claim.**

   For example, *The Daily Caller* writes that:

   > "the third-party 'fact-checking' organization says the "claim" in a DC article that special Counsel Robert Mueller is hiring people that "are all Hillary Clinton supporters is misleading, if not false. The problem is that TheDC's article makes no such claim. Their cited language doesn't even appear in the article. Worse yet, there was no language trying to make it seem that the

[5] While it's beyond the scope of the sociotechnical analysis of this thesis, see the Appendix for the partisan leaning of all 59 claimants with Reviewed Claims panels.

> investigation into the Trump administration and Russia is entirely comprised of Clinton donors. **The story simply contained the news**: Mueller hired a Hillary Clinton donor to aid the investigation into President Donald Trump."

In a second example, *The Daily Caller* provides a similar justification:

> "The 'claim' made, according to Snopes.com and Google, is 'a transgender woman raped a young girl in a women's bathroom because bills were passed' A quick read of the news piece shows that there was no mention of a bill or any form of legislation. The **story was merely a straightforward reporting** of a disturbing incident originally reported on by a local outlet."

Google must have found this argument persuasive, as it removed the second example from *The Daily Caller's* Reviewed Claims panel (this was the only algorithmically assigned fact-check of the four the *Daily Caller* referenced in its original article).

2. **The fact-check article did not evaluate the claim as false.** *The Daily Caller* took issue with a fact-check that had a rating of "mixture" from Snopes and was still displayed on the SERP, revealing their assumption that only claims rated false should appear on Reviewed Claims. Several claims that were not rated as false appeared on the Reviewed Claims component, as illustrates by Table A.1 in the Appendix.

3. **The article was satire, and therefore not deserving of a fact-check:** *The Daily Caller* claims that one of the assigned fact-checks was "obviously tongue-in-cheek" and does not warrant a fact-check.

Other claimants offer additional justifications. *Breitbart* claims that the **fact-check was assessing a different claim than the one in their article**.

> “the fact-checker claimed that *Breitbart* incorrectly reported that an illegal alien had been charged with starting a California wildfire, when in fact the story claimed that an illegal alien was arrested on suspicion of arson **a completely different claim**.”[6]

This argument is not unique to this *Breitbart* piece, note that *Daily Caller's* Mueller quote also said that “the problem is that TheDC's article makes no such claim.”

The Federalist also offers an example of such a claim.

> “Google shows a result from Snopes.com with regard to a Daily Wire story about Barack Obama praising Jay-Z while remaining publicly silent on the Congressional baseball shooting. Snopes.com suggests that the story was false, because Obama privately called Sen. Jeff Flake (R-AZ) an exchange reported only by Flake, not Obama. **But the entire premise of the story** was that Obama had remained publicly quiet on the shooting.”[7]

Whether these two examples actually illustrate <fact-check, news article> pairs that verify completely different claims, it is a common claimant justification for a fact-check not applying to their content.

Another justification that appears in *The Federalist* is that **the fact-check concerns a minor detail in the article.**

> “Consider the case of a woman named Eileen Wellstone. Out of many thousands of pieces published by *The Federalist* over the past four years, a single one mentions the name Eileen Wellstone. That article, detailing the sordid history

[6] https://www.breitbart.com/tech/2018/01/22/google-cancels-fact-check-program-fact-checked

[7] https://thefederalist.com/2018/01/10/googles-new-factchecker-is-partisan-garbage

> of Bill Clinton, **mentions her name exactly once**: 'Another woman, Eileen Wellstone, claimed Clinton raped her while he was at Oxford University in the late 1960s."'

Together, the 12 claimant articles produce the following justifications for when a fact-check may not be relevant to a piece of content:

1. The claimant document was simply reporting the facts, not making a broader political claim.
2. The fact-check did not evaluate the claim to be false.
3. The claimant article is satire.
4. The fact-check was assessing a different claim than the one in their article.
5. The fact-check concerns a minor detail in the article.

Claims Of Partisan Bias

The primary focus of claimant's reporting of the Reviewed Claims feature, was not the justifications of their content outlined above, but rather an attack on the partisan ship of Google fact-checks.

Of the twelve claimant articles about the Reviewed Claims feature, nine were articles and three (*Above Top Secret*, *Democratic Underground*, and *Free Republic* are forum posts). Of the nine articles, eight mentioned that Google was either biased towards liberals or anti-conservative. six claimants criticized fact-checking organizations for being partisan. All nine articles are from claimants that are conservative or conspiracy/pseudoscience sites.

As discussed in Chapter 3.3, fact-checking is only useful if the fact-checks are trusted. If the platform surfacing the fact-checks or the fact-checkers themselves are not trusted, then fact-checks will not be persuasive. Perhaps the distrust of fact-checks and Google make it impossible for Google to introduce successful fact-checking features.

In response to *Poynter's* article exploring the bugs in the Reviewed Claims algorithm, *The Daily Caller* publishes a follow-up article.[8]

> "Google's fact-checking program is a terrible idea," Peter Flaherty, president of the right-leaning nonprofit The National Legal and Policy Center... "It cannot be fixed. The problem is with the culture. Google has to fix the culture by hiring a greater diversity of people, especially conservatives and libertarians."

With this understanding, it is likely that no automated fact-checking feature could have been viewed positively by the entire fact-checking ecosystem, especially by conservative claimants.

The subsection will examine the influence of claimants and news publishers in the removal of the Reviewed Claims feature.

Claimants Role in the Reviewed Claims Ecosystem

News publishers cannot control the results displayed to users on Google search pages, including the Reviewed Claims panel. In the FAQ about the Reviewed Claims, publishers who disagree with a fact-check on the feature are instructed:

> "Reviewed Claims are made by publishers that fact check other publishers using the Fact Check markup and have been algorithmically determined to be author-

[8] https://dailycaller.com/2018/01/24/faulty-algorithms-liberal-bias-googles-fact-checking

> itative. If a publisher believes a reviewed claim is incorrect, Google recommends they contact the fact-checker that wrote the review. Publishers can also use the feedback link in the Knowledge Panel to report claims they believe are inaccurate."[9]

So, there is some evidence that news publishers, while not designing the feature or having any control over whether the feature appeared on SERPs about them, were an important part in having the feature removed.

6.3 Online Information Seekers

The Reviewed Claims feature was designed for online information seekers to help information seekers determine the credibility of a news source. The public at large does not know much about fact-checking, or the names of fact-check organizations, so displaying fact-checks as part of a Google's interface may be a useful public service that aligns with current web literacy approaches.

Online Information Seeker Justifications

As described in Chapter 5, I coded three undergraduate students' justifications for why thirty documents were or were not relevant to their algorithmically assigned fact-check articles. The articles selected include a subset of the articles from *The Daily Caller* and *The Federalist* that are mentioned above but include articles from other sources. From these responses, five justifications emerged that explain why a fact-check may not be relevant. Rater responses often included multiple justifications. Discussions with the raters confirmed that not all of

[9] https://support.google.com/websearch/answer/7568277?p=news$_publishers_kp$

the justifications on their own warranted a fact-check to be relevant or not relevant, but rather affected their confidence scores.

Below are the five justifications generated to describe when a claim match is not relevant:

1. **The fact-check concerns a minor detail in the article.** This justification appears in the claimant's justifications.

2. **The fact-check is verifying a different claim than what was in the article.** This justification appears in the news publishers' justifications.

3. **The fact-check and article covered the same general topic but differed on details.** This justification is used to describe when two a <fact-check, news article> are not a precise match. Raters often cited this justification in conjunction with another when labeling a claim match as not relevant.

4. **The fact-check and article are framed differently.** Fact-checks can be framed differently than the original topic by having different contextual details, by reporting on reporting of fact-checked events, or by using the fact-checked claim as the background for a new claim. Oftentimes, this justification was used to explain the raters' low confidence in their relevance assessment.

5. **The fact-check is more general than the article.** From discussions with the raters and their written justifications, one of the most difficult matches to evaluate was *The Gateway Pundit's* article about whether the DOJ leaked the name of an FBI informant, lobbyist William D. Campbell, who bribed the Clintons to give Russia Uranium. This multi-agent convoluted claim was difficult for raters to verify in the almost 2,000 word fact-check "What you need to know about Hillary Clinton, Russia, and uranium"[10]. The fact-check encompassed much more the specific claim about William Campbell,

[10] https://www.politifact.com/truth-o-meter/article/2017/oct/24/what-you-need-know-about-hillary-clinton-and-urani

who is not named in the fact-check. However, the idea that Secretary of State Hillary Clinton interfered in the process of Russia obtaining Uranium is disproven.

Online Information Seekers Reaction to Reviewed Claims

While Google may have conducted user studies to test the utility of the Reviewed Claims feature, there have been no peer-reviewed research or press releases detailing users' interaction with the feature. Thus, I cannot assess the usefulness of the feature. However, a previous study I conducted found that information seekers find the expanded news publisher Knowledge Panel beneficial to evaluate the credibility of news sources [32]. However, a *Gizmodo* article about the Reviewed Claims controversy is doubtful about the feature's utility:

> "...this fact-checking module isn't really all that useful—finding it requires clicking through to a submenu in a sidebar, and once the user's there the module's crowded with cut-off sentences and poorly contextualized information...It's not even clear having a fact-checking module appear alongside searches for a publication's name could noticeably impact the regular SEO referrals they're getting traffic from, which is what generates views."[11]

Again, no empirical research is publicly available that evaluates the ways information seekers use the feature.

While both the claimants and the platform played an important role in the removal of the Reviewed Claims feature, users, the intended audience of the feature, had no recourse to keep the feature or replacement mechanism for the feature. There is a small "Feedback" button that appears at the bottom of all Knowledge Panels, but users knowledge of its presence and the responsiveness of Google to that user feedback are both doubtful. Thus, I assess that

[11] https://gizmodo.com/conservatives-are-now-getting-angry-about-googles-fact-1821934885

users held limited ability to influence or even keep the Reviewed Claims feature and observe that the justification for a relevant fact-check have some, but not complete overlap with that of news publishers.

6.4 Fact-checking Organizations

There is more to learn about fact-checking organizations reactions to algorithmic fact-checking. The highly specific nature of some fact-checks reaffirm that many fact-checks are often written with a particular claimant and claim in mind, the reusability of fact-check articles may not be on the radar of many fact-checkers. However, some fact-checkers are more "algorithmically savvy" than others. For example, the British fact-checking organization Full Fact, is one of the leaders in developing automated fact-checking solutions.

In terms of the Reviewed Claims feature, the *Washington Post's* response to *The Daily Caller's* complaints about an algorithmically assigned Washington Post fact-check provides a preliminary indication of fact-checking organizations relationship to the feature.

When asked about the algorithmically assigned fact-check, *The Washington Post* responded with the following statement to The Daily Caller:

> "We went back and double-checked the story and the information submitted to Google, and The Daily Caller was not mentioned at all, even in links... We clearly labeled the source, so I cannot speak to how The Daily Caller ended up being erroneously listed as the source of the fact-checked quote in this case.[12]

The fact that *The Washington Post* believed that *The Daily Caller* was erroneously listed as the source, rather than understanding that their fact-check was algorithmically assigned

[12] http://archive.fo/vGAMx

by Google, illustrates the lack of transparency between platforms and fact-checking organizations. This is not a problem limited to Reviewed Claims, as fact-checking organizations have also been critical of Facebook's fact-checking efforts in terms of open data sharing[13].

While fact-checking organizations may not have been notified of the automated nature of the Reviewed Claims feature, they are certainly aware of the ClaimReview that they are adding to their articles. Adding ClaimReview clearly benefits Google, but recently fact-checking organizations have questioned if their efforts are beneficial to any stakeholder besides Google. In April 2019, The Vice President of Operations at Snopes, tweeted:

> "Last week we removed the ClaimReview markup that powers the fact check rich snippets in @Google from @snopes. The rest of the community should do the same and @Google should license the content or prove its [sic] actually effective."[14]

As of writing, Google has not public responded, nor has any other fact-checking organization publicly announced removing ClaimReview from their sites. However, this again shows the complex sociotechnical ecosystem of these actors.

Fact-checking organizations seemed unaware of the algorithmic assignments of their fact-checks on Reviewed Claims. Despite their lack of ability to contest Google's fact-check matching, fact-checking organizations still hold power within the fact-checking ecosystem as they 1) select claims to fact-check and 2) populate the structured data.

[13] https://techcrunch.com/2019/02/01/snopes-and-ap-leave-facebook-fact-checking-partnership

[14] https://twitter.com/vinnysgreen/status/1114188133099171841

6.5 Summary

Google's Reviewed Claims feature created a complex sociotechnical system built on the already fraught fact-checking ecosystem.

Google created the Reviewed Claims feature to address concerns that they were not actively tackling online misinformation on Google Search. Reviewed Claims was designed to provide contextual information about news publishers to assist information seekers to assess the credibility of online information and was released with much fanfare. However, Google failed to disclose a key piece of context: the feature was not simply relying on ClaimReview data supplied by third-party fact-checkers to identify relevant fact-checks, but frequently relied their own algorithm to assign fact-checks to news publisher content. Thus, this chapter focuses on how the various actors define relevance, as this definition is key to understanding the operation of Google's algorithm. I find that the definition in the technical description of a relevant document explored in Chapter 4, fails to account for the social differences and incentives when defining relevance ("the framing trap").

I identified 59 claimants (news publishers) who had Reviewed Claims panels. 12 of these claimants complained online and on TV about the feature. Several of these sources alleged that Google's anti-conservative bias was the reason Reviewed Claims was appearing on their SERP. They pointed out several instances that they felt that a fact-check did not belong in the Reviewed Claims. Some of the fact-checks were quickly removed by Google. While some of the claimant's concerns centered around issues of relevance, others emphasized that fact-checking and fact-checking organizations were not trustworthy or unbiased. I categorize and delineate the concerns raised by claimants to later compare with other actors definition of relevance.

As discussed in Chapter 5, the information seekers who labeled the relevancy of fact-checks had varying levels of agreement in defining and explaining what a relevant fact-check.

Nevertheless, I compared the justifications online information seekers provide to those of the claimants and found substantial overlap as well as discrepancies between the justifications of information seekers and claimants. There is no publicly available data that users complained about the feature. There is also no empirical data that they found the Reviewed Claims useful. Nevertheless, the feature was removed from the SERP two months after it was released.

Since 2015, a growing number of fact-checking organizations have added ClaimReview to their fact-checks. This allows Google to easily identify fact-checks and feature them in various ways on SERP. Fact-checking organizations have traditionally been supportive of such measures as they are always aiming for more information seekers to have more exposure to their fact-checks; however, recent frustrations with lack of licensure and transparency from platforms has led to increasing tensions and some fact-checking organizations to remove ClaimReview from their articles. With specific regard to the Reviewed Claims system, *The Washington Post* seemed unaware that their fact-checks were being algorithmically assigned to news articles not explicitly referenced in the text or ClaimReview of the fact-checks. It is unclear if other fact-checking organizations were aware of the Reviewed Claims feature.

The Reviewed Claims system designers failed to create an easily interpretable definition of relevance that could be applied and explained in the complex fact-checking ecosystem. Moreover, it seems that despite information seekers being the intended beneficiary of Reviewed Claims, there was little attention placed to how the feature affected their search experience or ability to evaluate online information.

Chapter 7

Conclusion

There is a large push to leverage machine learning to spread fact-checks further. This was the motivation of Wang et al.'s paper that I believe underpins the Reviewed Claims technical system. However, Wang et al.'s definition of relevance does not recognize that relevance is a complex human value. The algorithm described in Wang et al. fails to model (or even mention) how the sociotechnical system that is the fact-checking ecosystem affects their classifications. This limitation of their work is one common in machine learning, so common that Selbst et al. dubbed it "the framing trap" [47]. My detailed exploration of how different actors in the fact-checking ecosystem interacted with Reviewed Claims as well as their conceptions of relevance in "claim matching" attempts to qualitatively model relevance in the Reviewed Claims system. As Introna and Nissenbaum wrote in 1998, "determining relevancy is an extraordinarily difficult task...Besides the engineering challenges, experts must struggle with the challenging of approximating a complex human value" [26]. In Chapter 6, I illustrate that the Reviewed Claims feature was unable to successfully model the multi-faceted, multi-actor definitions of relevance.

I devote Chapter 5 to examining how users assess whether a claim match is irrelevant.

I label the 118 algorithmically assigned claim matches and find 25 matches (21%) to be "not relevant." Three raters independently label 30 <fact-check, news article> pairs that I struggled to label. I discover that raters had a fairly difficult time coming to the same relevance assessment. After a reconciliation process, the agreement rate between raters improved substantially, providing an indication that information seekers may be able to come to a common definition of relevance. Then I attempted to assess whether the justifications the raters used to explain whether a claim was relevant to a document were consistent among a larger population. In the process of running this experiment, I discovered that the crowd workers were far more generous in labeling claim matches as relevant than the independent raters or myself. As Wang et al. utilized crowd worker responses to train their data this is perhaps another reason that Wang et al.'s claim-document discovery process created a far from perfect implementation of relevance in the Reviewed Claims component.

This thesis is not arguing that humans should be excluded from automated fact-checking systems, instead, this thesis provides a compelling case study for why modeling complex social criterion like relevancy requires human input and discussion. Envisioning what such systems and discussion look like warrants further research, but it is concerning that Google, despite outlining the limitations of their "claim-document discovery approach" (summarized in Chapter 4.4), did not employ any human moderators to examine the outputs of their model. Of the over 8,000 SERPs collected, only 59 had Reviewed Claim panels. Even if 500 news publishers had the Reviewed Claims feature on their SERP, Google could have employed people to perform a final round check to ensure users found all algorithmically assigned fact-checks pertinent to the news publishers' content. Granted, I cannot conclusively say they did not employ human "moderators," but it seems unlikely given media reports of the feature.

Additionally, platforms like Google often argue that they are a neutral arbiter of information. Platforms make many meaningful judgments that illustrate they are in fact important gatekeepers of our information [19]. The Reviewed Claims feature is a perfect example of this. The Reviewed Claims feature only appeared on a very small subset of news publisher SERPs. How were the news publishers selected? This is certainly an editorial judgment, even if is done with an algorithm.

The lack of transparency in the Google Reviewed Claims feature is especially problematic. Despite the detailed public announcement of the Reviewed Claims feature[1], nowhere is it disclosed that Google will be algorithmically assigning fact-checks. As discussed in Chapter 6.2, in the FAQ about Reviewed Claims, publishers who disagree with a fact-check are told that "...if a publisher believes a reviewed claim is incorrect, Google recommends they contact the fact-checker that wrote the review"[2]. This is misleading considering that 57% of fact-checks in the sample of 221 claims were algorithmically assigned (see Chapter 5), so Google, not the fact-checking organizations, were responsible for the majority of fact-checks appearing in Reviewed Claims. Google's lack of transparency is a concern of fact-checkers. As explored in Chapter 6.4, Snopes, the largest fact-checking organization, removed all of the ClaimReview markup (the structured data that underpins the Reviewed Claims system) until Google 1) licenses the content created by fact-checking organizations and 2) provides evidence that their approach of incorporating fact-checks into the SERP are effective. As of writing, Google has not provided that data to Snopes and the ClaimReview data is no longer present on Snopes fact-checks.

This thesis finds that if platforms want to build automated tools that model a complex social value like relevance, they need to understand how all actors in the sociotechnical system define that value. Additionally, successful systems will provide meaningful and transparent explanations about their algorithms and data.

There are several promising avenues for future work, especially in regard to strengthening the findings related to **RQ2** and **RQ3**.

[1] https://www.blog.google/products/search/learn-more-about-publishers-google

[2] https://support.google.com/websearch/answer/7568277?p=news$_{p}ublishers_{k}p$

- Additional relevance assessment tasks for crowd workers that require the review of additional claims. This work should 1) recruit a more representative sample and 2) require workers to rate more than six claims.

- A more nuanced exploration of fact-checkers perceptions of how automated fact-checking affects their work. It seems that a follow-up ethnography to Graves' study [21] could be particularly valuable.

- Further examination of the reasons why crowd workers were less selective in their relevance assessments. Their broad interpretation of relevance does not model that of claimants (discussed in Chapter 6.2). This preliminary finding is worthy of future exploration as crowd workers are seen as a fix-all to the limitations of AI systems. Despite technology platforms touting AI as the solution to all problems, human labor is required (and often exploited [27]) to provide everything from platform content moderation to training data for supervised learning models (as in Wang et al. [54]).

- A broader analysis of fact-checking organizations. In this thesis, I have chosen to focus on examining fact-checking organizations that are not affiliated with a news organization, unlike *The Washington Post* Fact-checker, AP Fact-check, and several others. While Snopes and PolitiFact are frequently accused of liberal bias, charges of partisan seem to be especially fervent when it comes to mainstream news organizations. Looking at the differences in fact-checking process and perceptions of fact-checking organizations by institutional affiliation is another interesting area of study.

- An exploration of the differences between determining relevance and determining irrelevance. This thesis has primarily defined relevance by defining what is not irrelevant. There are possible implications to this approach that deserve further discussion and thought.

Without fact-checking, false claims stand with no direct confrontation. Fact-checks have the potential to take falsehoods head-on and correct people's misperceptions. Fact-check articles serve as a flag to online platforms that a claim being promoted by users or their algorithms may be problematic. Of course, there are complex questions surrounding fact-checking including: What content warrants a fact-check? How should fact-check articles be constructed? Currently, the question engrossing the fact-checking community on the front lines fighting fake news is: how can the reach of fact-checks be spread farther?

Chapter 8

Glossary

Algorithmically assigned fact-check: A fact-check article that has been matched to a claimant document through some technical system (e.g. the system described in Chapter 4).

Automated fact-checking: Fact-checking that relies on some kind of computational approach to identify, match, check, or publish fact-checks.

Candidate document: An article that bears some similarity to a fact-checked document.

Claim: A statement addressed in a fact-check article.

Claimant: the source of a claim that has been fact-checked.

Claim matching: The process by which fact-checks are algorithmically assigned to relevant documents.

Claim-document discovery process: Described in detail in Chapter 4, the claim-document discovery process describes the approach in [54].

Fact-check article: A fact-check never produces new information, but rather selects a claim and uses existing reporting and research to assess the validity of that claim.

Information Seeker: An individual using the web.

Knowledge Panel: A component on the Google search engine result page that provides contextual information about an entity, see Figure 1.2 for an example.

Machine Learning: a technical approach that leverages algorithms and statistical models to perform a specific task by relying on patterns.

Platform: An intermediary to the exchange of content (e.g. Google). See [19] for more.

Reviewed Claims: A component of the Knowledge Panel from November 2017 - January

2018 that surfaced fact-checks, some of which were algorithmically assigned.

Sociotechnical Perspective: a view of technical systems that incorporates social context, values, and actors.

Bibliography

[1] Adair, B., Li, C., Yang, J., and Yu, C. Automated pop up fact checking: Challenges & progress. In *Computation + Journalism* (2019).

[2] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*, 2 (2017), 211–236.

[3] Allcott, H., Gentzkow, M., and Yu, C. Trends in the diffusion of misinformation on social media. *NBER Working Paper Series*, w25500 (2019).

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

[5] Amazeen, M. A. Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism* (2017).

[6] Ananny, M. Checking in with the facebook fact-checking partnership. *Columbia Journalism Review* (2018).

[7] Babaei, M., Kulshrestha, J., Chakraborty, A., Redmiles, E. M., Cha, M., and Gummadi, K. P. Analyzing biases in perception of truth in news stories and their implications for fact checking. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), FAT* '19, ACM, p. 139.

[8] Babakar, M., and Moy, W. The state of automated factchecking. *Full Fact* (2016).

[9] Berners-Lee, T., Hendler, J., Lassila, O., et al. The semantic web. *Scientific American 284*, 5 (2001), 34–43.

[10] Broder, A. A taxonomy of web search. *ACM SIGIR Forum 36*, 2 (2002), 3–10.

[11] Budak, C., Agrawal, D., and El Abbadi, A. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web* (2011), WWW '11, ACM, pp. 665–674.

[12] Budak, C., Goel, S., and Rao, J. M. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly 80*, S1 (2016), 250–271.

[13] Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6*, 1 (2011), 3–5.

[14] Burrell, J. How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society 3*, 1 (2016), 1–12.

[15] Caulfield, M. *Web literacy for student fact-checkers.* Michael Arthur Caulfield, 2017.

[16] Charmaz, K. *Constructing grounded theory.* Sage, Thousand Oaks, CA, 2014.

[17] Chen, D., Chen, W., Wang, H., Chen, Z., and Yang, Q. Beyond ten blue links: Enabling user click modeling in federated web search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (2012), WSDM '12, ACM, pp. 463–472.

[18] Friesen, J. P., Campbell, T. H., and Kay, A. C. The psychological advantage of unfalsifiability: The appeal of untestable religious and political ideologies. *Journal of Personality and Social Psychology 108*, 3 (2015), 515–529.

[19] Gillespie, T. The politics of platforms. *New Media & Society 12*, 3 (2010), 347–364.

[20] Gottfried, J. A., Hardy, B. W., Winneg, K. M., and Jamieson, K. H. Did fact checking matter in the 2012 presidential campaign? *American Behavioral Scientist 57*, 11 (2013), 1558–1567.

[21] Graves, L. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique 10*, 3 (2017), 518–537.

[22] Guess, A., and Nyhan, B. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 u.s. presidential campaign, 2018.

[23] Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017), CSCW '17, ACM, pp. 1914–1933.

[24] Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., and Yu, C. The quest to automate fact-checking. In *Computation + Journalism* (2015), p. 5.

[25] Hassan, N., Arslan, F., Li, C., and Tremayne, M. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), ACM, pp. 1803–1812.

[26] Introna, L. D., and Nissenbaum, H. Shaping the web: Why the politics of search engines matters. *The information society 16*, 3 (2000), 169–185.

[27] Irani, L. C., and Silberman, M. S. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), CHI '13, ACM, pp. 611–620.

[28] Jiang, S., and Wilson, C. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 1–23.

[29] Law, J. Technology and heterogeneous engineering: The case of portuguese expansion. In *The social construction of technological systems: New directions in the sociology and history of technology*, W. E. Bijker, T. P. Hughes, and T. Pinch, Eds. The MIT Press, Cambridge, MA, 1987, pp. 1–134.

[30] Lim, C. Checking how fact-checkers check. *Research & Politics 5*, 3 (2018).

[31] Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (2017), Asian Federation of Natural Language Processing, pp. 252–256.

[32] Lurie, E., and Mustafaraj, E. Investigating the effects of google's search engine result page in evaluating the credibility of online news sources. In *Proceedings of the 10th ACM Conference on Web Science* (2018), ACM, pp. 107–116.

[33] Marietta, M., Barker, D. C., and Bowser, T. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum 13*, 4 (2015).

[34] Marwick, A. E. Why do people share fake news? A sociotechnical model of media effects. *Georgetown Law Technology Review 2* (2018), 474–512.

[35] McMahon, C., Johnson, I., and Hecht, B. The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies. In *Eleventh International AAAI Conference on Web and Social Media* (2017), AAAI Publications.

[36] Metaxas, P. T., Finn, S., and Mustafaraj, E. Using twittertrails. com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (2015), CSCW '15 Companion, ACM, pp. 69–72.

[37] Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., and Sigman, M. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour 2*, 2 (2018), 126–132.

[38] Noble, S. U. *Algorithms of oppression: How search engines reinforce racism.* NYU Press, New York, NY, 2018.

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

[40] Nyhan, B., and Reifler, J. The effect of fact-checking on elites: A field experiment on u.s. state legislators. *American Journal of Political Science 59*, 3 (2015), 628–640.

[41] Nyhan, B., and Reier, J. Estimating fact-checkings effects, 2017.

[42] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), Association for Computational Linguistics, pp. 231–240.

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

[44] Robertson, R. E., Lazer, D., and Wilson, C. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference* (2018), WWW '18, ACM, pp. 955–965.

[45] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a preconference at the 64th Annual Meeting of the International Communication Association* (2014).

[46] Schultz, D. Truth goggles : automatic incorporation of context and primary source for a critical media experience. Thesis, 2012.

[47] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), FAT* '19, ACM, pp. 59–68.

[48] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *arXiv:1708.01967* (2017).

[49] Starbird, K. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Eleventh International AAAI Conference on Web and Social Media* (2017), AAAI.

[50] Trist, E., and Murray, H., Eds. *The social engagement of social science, volume II: The socio-technical perspective.* University of Pennsylvania Press, Philadelphia, PA, 1993.

[51] Uscinski, J. E., and Butler, R. W. The epistemology of fact checking. *Critical Review 25*, 2 (2013), 162–180.

[52] Volkova, S., and Jang, J. Y. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018* (2018), WWW '18, ACM, pp. 575–583.

[53] Wang, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv:1705.00648* (2017).

[54] Wang, X., Yu, C., Baumgartner, S., and Korn, F. Relevant document discovery for fact-checking articles. In *Companion Proceedings of the The Web Conference 2018* (2018), WWW '18, ACM, pp. 525–533.

[55] Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), CIKM '09, ACM, pp. 2061–2064.

Appendix A

A.1 The Data Collection Process

As part of our long-term research on web literacy and credibility of online sources, the Cred Lab monitors the evolution of SERPs over time. That is, for various query phrases, we automatically collect the Google's search result pages and analyze changes in them. In early January 2018, as part of a large data gathering about media organizations in the United States [32], we collected data associated with two lists of online publishers that were made public by third-party journalists and researchers.

A.1.1 United States Newspaper List

We used the USNPL (United States Newspaper List) [1], which is a database of US-based local newspapers, TV, and radio stations broken down by state (n = 7269). It aggregates the newspaper and TV station lists from each state, which were scraped with express permission from the maintainers of the list. We chose to use this list instead of each state's Wikipedia list of newspapers, because we found the USNPL list to be more evenly distributed state-to-state and equally, if not more, complete.

A.1.2 List of Unreliable Websites

We aggregated several well-known lists of unreliable sources to create as comprehensive a list as possible (n = 585). Our list is comprised of the following sources:

- **Politifact's Fake News Alamanac** (n = 330) Last updated in November 2017, the Politifact Almanac lists sources that intentionally publish disinformation[2].

[1] http://www.usnpl.com

[2] https://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/

- **2016 and 2017 Buzzfeed List of Top Fake News Sites** (n = 96, 167) Buzzfeed used BuzzSumo to determine the most popular fake sites on Facebook in 2016 and 2017[3].

- **Zimdar's "False, Misleading, Clickbait-y, and Satirical 'News' Sources"**[4](n = 121) In November 2016, Melissa Zimdar collected and categorized unreliable sources. Not all sources on this list are fake, for example, some are satirical.

- **"Alternative Narrative"-affirming Sources** [49] explores the media ecosystem on Twitter after mass-shooting events. We selected the subset (n = 69) of that dataset that "affirmed" alternative narratives (e.g. rumors and conspiracies) after mass shootings.

The combined list of websites from the partisan dataset and the aggregated unreliable dataset amounts to 1150 website domain names (after removing duplicates).

A.2 The Composition of the 59 Reviewed Claims Panels

In Table A.1, we have aggregated the information for 15 out of 59 sites with a RC tab. This is due to space reasons, we cannot list all 59 sites in the table. However, we provide the names of the websites with the number of fact-checks assigned to them in Table A.2. Here is some information about these tables:

- The order of the websites (in both tables) is based on their global Alexa Rank[5] as of September 2018, with most popular websites shown first.

- The 2nd column, the bias value, is assigned based on the value present in the Buzzfeed dataset or by consulting the bias assigned by Media Bias/Fact Check (marked with a star). In total, we found 31 right-leaning websites, 14 left-leaning websites, and 14 conspiracy/pseudo-science/fake-news websites.

- The values for the column "Factual Accuracy" are also pulled from Media Bias/Fact Check (MBFC). The label "Questionable' according to MBFC indicates that a website "may be very untrustworthy and should be fact checked on a per article basis".

- The number of claims per website varies from 1 to 10. 22 websites have only one claim, and 11 websites have 10 claims. In average a tab had 3.7 claims (median was 2).

[3] https://github.com/BuzzFeedNews/2017-12-fake-news-top-50/tree/master/data

[4] http://d279m997dpfwgl.cloudfront.net/wp/2016/11/Resource-False-Misleading-Clickbait-y-and-Satirical-%E2%80%9CNews%E2%80%9D-Sources-1.pdf

[5] https://www.alexa.com/siteinfo

Website	**Bias (by BF)**	**Factual Accuracy (MBFC)**	**Snopes**	**Factcheck**	**Politifact**	**WaPo**	**Climate Feedback**	**Others**	**Total**
Breitbart	right	Questionable	False (2) Mixture (1)	False (2) Unsupported(1)	False (1) Mostly False (1) CR Unclear (1)		Inaccurate (1)		10
The Daily Caller	right	Mixed	False (2) Mixture (2)	Misleading (1)		3 Pinocchios (1)	Inaccurate (3)		10
Zero Hedge		Conspiracy/ Pseudoscience	False (3) Mixture (1) Unproven (1)	False (1)	Clintons role unclear (1)	Lacks Content (1)		CBS: False (1) Polygraph: False (1)	10
Daily Wire	right	Mixed	False (4) Mostly False (1) Mixture (1) Outdated (1)	No, he did not (1)	Mostly True (1)		Incorrect (1)		10
WND (Word Net Daily)	right	Mixed	False (3) Unproven (1) Mostly false (1) Mostly true (1)	False (1) Misleading (1)	False (1)		Incorrect (1)		10
Gateway Pundit	right	Questionable	False (2) Mixture (2)	False (2)	Pants on fire (1) False (1) CR Unclear (1)			Teyit.org: False (1)	10
The Federalist	right	High	Unproven (2)				Incorrect (1)		3
Upworthy	left	High	False (1)						1
Before It's News		Questionable/ Fake News	False (5) Misattributed (1)	False (2)	Pants on fire (2)				10
Palmer Report	left	Mixed	False (2) Unproven (4)						6
Natural News		Pseudoscience	False (3) True (1)	Probably at high doses (1)	Pants on fire (1)				5
Democratic Underground	left	Mixed	False (2)		Mostly false (1)			Fact/Myth: Fact (1)	4
Free Republic	right*	Mixed	False (2) Misattributed (1)	False (1) No, he did not (1)	Pants on fire (2) Half True (1)	4 Pinocchios (1)		Teyit.org: False (1)	10
The Conservative Tree House	right	Mixed	False (1)	False (1)					2
Above Top Secret		Conspiracy	False (3) Mixture (1) Unproven (1)			4 Pinocchios (1)			6

Table A.1: The top 15 (Alexa Rank) websites in our dataset that were assigned a Reviewed Claims (RC) tab by Google.

Bias or Accuracy	**List of websites to which Google assigned a Reviewed Claims tab (ordered by Alexa Rank)**
Right	Breitbart (10), The Daily Caller (10), The Daily Wire (10), WND [Word Net Daily] (10), The Gateway Pundit (10), The Federalist (3), Free Republic (10), The Conservative Tree House (2), OAN Network (2), Big League Politics (4), The Political Insider (2), Frontpage Magazine (2), American Greatness (2), American Renaissance (1), Bearing Arms (1), Red State Watcher (1), Truthfeed (6), 100 Percent Fed Up (2), Freedom Outpost (1), Commentary Magazine (1), The Millennium Report (1), VDARE (2), Sparta Report (1), En Volve (3), Conservative Fighters (4), Silence is Consent (2), America's Freedom Fighters (5), Freedom Daily (10), American News (1), American Conservative Herald (1), The New York Evening (1).
Left	Upworthy (1), Palmer Report (6), Democratic Underground (4), Counterpunch (3), Rightwingwatch (1), Bipartisan Report (3), True Activist (1), OpEd News (1), American Herald Tribune (1), Occupy Democrats (10), Egberto Willies (1), If You Only News (1), American News X (2), Resistance Report (1)
Conspiracy / Pseudoscience / Fake news	Zero Hedge (10), Before It's News (10), Natural News (6), Above Top Secret (6), Collective Evolution (3), Your News Wire (10), Investment Watch Blog (7), Awareness Act (1), Activist Post (2), Renegade Tribune (1), The Common Sense Show (1), Fellowship of the minds (4), Intellihub (2), 21st Century Wire (1)

Table A.2: The list of all 59 websites with a Reviewed Claims tab in our January 2018 dataset, grouped by political bias or factual accuracy. The labels come from BuzzFeed and Media Bias/Fact Check. The numbers in parentheses indicate the number of fact-checks that the RC tab contained.

- There are 10 different fact-check providers, Snopes, Factcheck.org, Politifact, Washington Post (WaPo), Climate Feedback (each of them as a column), and five others that we have put together in the before-last column: CBS News (1), Fact/Myth (1), Gossip Cop (3), Polygraph.info (2), and Teyit.org (2).

While the ratio 31 right-leaning vs. 14 left-leaning might be viewed as skewed, we have to consider that the partisan websites dataset composition is 490 (right) vs 177 (left). Thus, the findings are proportional to the size of the dataset.