Bachelor thesis
Computing Science

Radboud University

Stop the infodemic

Designing and evaluating a misinformation warning mechanism based on expert voting

Author:
Jelte Smits
s1012014

First supervisor/assessor:
Dr. Hanna Schraffenberger
h.schraffenberger@ru.nl

Second assessor:
Dr. Eelco Herder
eelcoherder@cs.ru.nl

January 20, 2022

Abstract

The aim of this thesis is to investigate a new method to counteract the negative effects of fake news. The WHO states an "infodemic" is currently taking place, with consequences that include but are not limited to: efforts by state-actors to destabilize democracy, an increase in deadly Covid misinformation and polarization in politics. To identify possible counter measures, we first look at how misinformation spreads, how humans conduct truth assessments, what counter methods research currently agrees on and how this links with fact checkers and trust scores. Based on the obtained insights, we propose a new warning mechanism in which experts vote on the credibility of tweets in order to decrease the impact of tweets containing misinformation. A mock-up of this warning mechanism has been designed for Twitter, that shows an aggregated trust score combining the votes of different experts in their related fields. We conduct an experiment that compares our idea to a control condition without any warning and a current method Twitter uses for alerting users to misinformation. Initial results indicate that only our method works to decrease perceived credibility of tweets and that users often do not notice a warning tag Twitter currently employs. Finally, we discuss what these findings mean and give directions for further work.

Contents

Chapter 1

Introduction

> *"We're not just fighting an epidemic; we're fighting an infodemic. Fake news spreads faster and more easily than this coronavirus & is just as dangerous."*
>
> – Tedros Ghebreyesus, *Director-General of the WHO*

This quote emphasizes how fake news is becoming more and more prolific in today's society. In March of 2020 there were an average of 46,000 *daily* tweets which contained covid misinformation [33]. Countries that were most exposed to this misinformation, such as Iran, [33] had to deal with even bigger problems. Messages circulating on Iranian social media cited methanol as a potential Covid cure, causing many to overdose. As a direct result of the fake news in Iran, around 500 people died due to alcohol poisoning [10].

This example shows that fake news can have real, dangerous, consequences. But what is fake news? Axel Gelfert, professor of philosophy at the university of Berlin describes it as the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design [15]. This means satire is not fake news, since it is not misleading *by design*. Although the term has been around for a while, it was picked up heavily by the media during the 2016 presidential elections in the US. When Trump won these elections there were lots of claims saying that his victory had to do with fake news and online influence. Is the power of fake news overestimated, or do people still think too little of it? Can it even be stopped and how would that be possible? To answer these questions we must first zoom out to understand the problem from a broader perspective. We have to find out the origins of fake news, and people's motivation for spreading such content, before we can think of any sort of solution. This is what we address in the following chapter. However, first we specify what the research in this thesis will be about.

1.1 Research

This bachelor thesis was founded on a desire to not only highlight the problem of fake news, but also to find and study a new system that can possibly help. This new system is what the research and this thesis will be focused on. We aim to propose an accurate system based upon trust scores and warning messages. Trust scores can be automatically generated for content such as tweets or articles, or can be voted upon by actual people. Warning messages could explain why certain messages are seen as fake news and could help reduce the spread of such messages. We hope that by displaying a trust score together with a well-thought-out warning label, users are less likely to share and believe the misinformation displayed. This could potentially reduce the negative effects of fake news. If this one aspect of the big solution would be effective, we might no longer have to deal with all the terrible consequences of fake news. However, before we can design such a new system we must first understand much more about the subject of fake news. This is done in the following chapters, focusing on relevance and related work on misinformation.

In order to research the effectiveness of a new system, we use the following (personal) process. First, we have defined clear research questions, in order to show what the research is about. Secondly, we need to consume as much research and knowledge in this area as possible. This helps to shape our system in an effective way, by not making mistakes that have been shown not to work in previous research. Third, we have set up an experiment that helps to answer our research questions. Finally, we have drawn conclusions from our results. This section is dedicated to that first step, in which we outline our research by stating the important questions.

We have formulated several research questions which can help us to stay focused on our goal. Our main research question is more broad, while we focus on the specifics with two follow up sub-questions. The questions we ask ourselves are:

- Whether and to what extend would a new vote-based trust score system reduce the impact of misinformation?
  - Whether/how would the display of trust scores influence the spread of fake news?
  - Whether/how would trust indicators change the perceived credibility of a social media news post?

The following question can be used as a guideline to properly design the system we are trying to create. However, it is not a research question but instead a design question that we can keep in mind when coming up with the system.

- How do we display a trust score in such a way that it is most effective at reducing the impact of misinformation?

In the following chapter we now continue to the second step mentioned, consuming knowledge. In order to understand the problem from all possible perspectives, we look at the relevance the topic has in both the societal and scientific aspect. First we look at important examples, and highlight origins and motivations for spreading misinformation (Chapter 2). After we look at what is still missing in current research, showing the scientific relevance. In the next chapter we look at what is currently known about misinformation and all other relevant topics (Chapter 3). All of this helps us to create a system which can alleviate some of the troubles fake news is causing (Chapter 4). This system then needs to be tested by an experiment, in order to show its effectiveness (Chapter 5). The results of this experiment indicate if our system works or not, and shows some interesting findings (Chapter 6). These findings are then discussed to show its implications and to answer the research questions we set up (Chapter 7). Finally, we end with possible directions for future work, in which we show how this research can be continued in a useful manner (Chapter 8).

Chapter 2

Relevance

Everyone has dealt with fake news on some level. We have read stories about it, done research on it, or even consumed it without knowing. Some of the most influential fake news stories have swayed the world in one way or another. Starting this chapter, we will investigate some examples portraying the origins of misinformation, and motivations for spreading such content. After we look in detail at four different recent events where misinformation played a large role. The events we will look at are: Pizzagate, the US elections, the Russia operations, and the Covid-19 pandemic. Taking a good look at these different events will help us to introduce some important scientific concepts and gives us the necessary background information for the creation of our system.

2.1 Origins & motivations

Starting this section, we will look at a small town in Macedonia, called Veles. After the presidential elections in 2016 researchers found this one town where a lot of fake news originated from. A man called Mirko Ceselkoski lived there at the center of the Veles fake news industry [21]. This man learned many teenagers in Veles write engaging content, which was destined to be shared by many. The community did this for money. A woman living there earned about 30,000USD by writing a made-up story about the meaning of the color of toothpaste [21]. This shows that the fake news business is a lucrative industry. The town had the sentiment that if Americans did not make an effort in fact-checking their news, they might as well earn some money from it.

But not all misinformation enters the news cycle this way. We will now look at a second example, which is much more refined and sinister. This is the operation Russia is engaging in. A report by the CIA, NSA, and FBI states that the Russian president, Vladimir Putin, ordered an influence

campaign on the 2016 US elections [22]. The report states that the influence campaign was multifaceted. Not only did the campaign use social media, it included operations by state-funded media, third-party intermediaries, and paid social media users [22]. An article by The New York Times shows how Russia employed fake Americans, often bots, to tweet in accordance with the Russian agenda [43]. According to the American intelligence agencies the Russian purpose was clear: undermine the American faith in democracy, reduce the chances of Hillary Clinton winning the election, and helping Donald Trump win in the end. Since the report also states that Russian actors were not involved in vote tallying, the campaign was purely based around fake news and misinformation.

Finally, fake news can also come from the country it is describing itself. During the US elections researchers uncovered fake news from an American-based business. NPR, an independent non-profit news organization, launched an investigation into one particular popular fake news item titled: "FBI Agent Suspected In Hillary Email Leaks Found Dead In Apparent Murder-Suicide." This article was shared around 568,000 times and had more than 15.5 million impressions [17]. During the investigation NPR found that an American company called Disinfomedia was responsible for this fake news article [49]. In an interview with NPR the owner of Disinfomedia told the reporters that his company employs around 20 writers, all writing fake news stories. Although he makes around 10,000 to 30,000 USD from these stories, he claims that his company is about showing how easily fake news can spread.

These examples gave some information into the distribution of fake news. However, a large part of the problem is the people who share the misinformation. Scientific research has come up with several motivations for why people spread and re-share such content. One paper shows that people who share misinformation simply do not pay attention to the accuracy of the information presented [36]. They feel it is necessary to share important news, but since these people are absentmindedly sharing, they fail to check the accuracy of the content. Furthermore, fake news gets shared faster when people thought the information to be true, or had some pre-existing attitudes toward it [6]. This shows that people are more likely to share something that is already confirmed according to their worldview, even when the information is factually false.

What we showed in the previous paragraphs is that fake news is a broad problem. It has its roots not only in foreign, but also domestic agents. These agents have a broad variety of reasons for engaging in fake news distribution. Their motivation can be anything from destabilizing a democracy, to earning some money, to showing how fast such misinformation can spread. People sharing fake news do this primarily because they feel the news is important and do not check the accuracy, or have a pre-existing attitude towards the content.

2.2 Pizzagate

On 4 December 2016, a man entered a Washington DC based pizza place, in his hand an AR-15 assault rifle. He was not there to rob the place, the man had different intentions. He believed the restaurant, Comet Ping Pong, established a child trafficking ring for high-profile democrats. After shooting his assault rifle in the air he began looking for hidden children, secret tunnels, and more that would establish the theory of Pizzagate [14]. He did not find any of these things and was shortly later apprehended by the police. What started all this? It turned out to be a theory on the internet, called Pizzagate, that talks about high profile democrats running an underground child sex trafficking ring. The theory started on 4chan, an anonymous message board where anyone can post anything. On November 2nd, 2016 the first thread about Pizzagate started [53]. 4chan was at that time busy dissecting the Clinton emails, which had been leaked a few months prior. The users of 4chan believed that emails talking about the pizza restaurant, Comet Ping Pong, were actually code language. They soon started gathering "proof" in the many emails Hillary Clinton and her campaign had sent to each other. Two days after the initial thread on 4chan, the hashtag #Pizzagate goes viral on Twitter. With such a large audience the theory grew quickly. People from all over the internet started getting involved, each gathering more information about the supposed scandal. In the end, it grew so out of control that it concluded with an armed man walking into a completely normal pizza place in order to free children.

This is in essence a story about echo chambers. Online forums and communities where people with the same polarized opinion converse, without any outsider information coming in. The information, in this case the Pizzagate theory, gets more and more polarized until it eventually explodes. Echo chambers are one of the problems in today's internet, contributing to the spread of misinformation [54].

2.3 2016 US Elections

In 2016 "Post-truth" was selected as the word of the year, by Oxford Dictionaries [37]. This word was chosen because in 2016, especially in the months leading up to the elections, the word was significantly more searched on the Oxford dictionary website [37]. This indicates that interest has gone up for the term, most likely because fake news became quite prevalent in the 2016 US elections. The fact that fake news became a bigger problem during the 2016 elections is supported by research: a paper by Hunt Allcott and Matthew Gentzkow titled "Social Media and Fake News in the 2016 Election" shows that in the months before the elections there were approx-

imately 760 million clicks to fake news articles [4]. Furthermore, for the month before the 2016 elections, there were 159 million views on fake news websites [4]. When comparing these numbers to the number of American voters, the problem becomes clear: 760 million fake news clicks, vs 137 million voters.

But does fake news actually sway the opinion? This is something research is still divided on. The paper previously cited, by Allcott and Gentzkow, states that fake news did not have much of an influence. This is because the researchers compare fake news to television advertisements, which have been found to change vote shares by around 0.02 percentage points [4]. However, new research suggests that the influence could be bigger than earlier thought [35]. Researchers polled Obama voters on their belief in fake news articles, after letting the voters read 3 different articles. This revealed a strong correlation: Obama voters who believed in these articles were more likely to defect to Donald Trump in 2016 [35]. The conclusion from the cited paper state that people who believed in one or more fake news stories, were 3.3 times more likely to defect from Obama to Trump than people who did not believe a fake news story [35]. Because of the fact that the 2016 elections were really close, there is a real possibility fake news played an important role in the results. This might have been the first election won with misinformation, something that could happen more and more often.

2.4 Russia

As shown in the first section of this chapter, Russia is currently operating a highly controversial information campaign. The report by the FBI, CIA, and NSA, which we previously mentioned, has shown that the full US intelligence community believes Russia is actively interfering with the US in order to undermine its faith in the democratic process, among other reasons [22]. Most of this work is done by the so-called "Internet Research Agency" (IRA). This is a Russian company focused on online influence, heavily connected to the Russian government. A 100 page paper was requested by the US Committee on Intelligence, titled "The Tactics & Tropes of the Internet Research Agency" to take a closer look at the IRA [12]. The paper demonstrates that the main focus of the IRA was not on the US elections, but rather an effort to divide the American people [12]. Russian actors did this by posting highly dividing content across several social media platforms, mainly Twitter, Facebook, Instagram, and YouTube. These dividing posts had different controversial themes such as Black Lives Matter, LGBT culture, veteran's issues, and many more.

There is still much to be done to undermine these specific information campaigns. The authors of [12] advocate a deeper collaboration between researchers, tech platforms, and governments in order to detect foreign influence operations [12]. The report states that a steady, trustworthy collaboration between these entities is key in defeating this threat.

2.5 Covid

As illustrated by the opening quote of this thesis, the infodemic is seen as just as dangerous as the actual Covid-19 pandemic. This is because fake news can definitely kill. There are several specific examples of how fake news managed to wreak havoc during the pandemic. First off, the amount of supposed Covid cures floating around on the internet is ridiculously high. In the introduction we spoke about methanol being a Covid cure, but there are many more so-called "cures" which only worsened the pandemic. These cures were often advertised by people without an academic background in viruses and resulted in the spread and increase of fake news. In September 2021, the popular podcast host Joe Rogan contracted Covid and started advertising the drug Ivermectin as a great cure against the disease. Around the same time, the national poison data system in the US registered an increase of 245% in the number of exposures [41]. Other cures which circulated the internet were Chloroquine and Hydroxychloroquine. Again, these drugs did not slow a Covid infection but only resulted in deaths and poisonings [25].

Another second example takes place in the United States. Rumors floated around on the internet that there would be a nationwide lockdown. In response to these rumors, stores in the entire country had to deal with people panic-buying all the groceries [50]. In the Netherlands panic-buying happened as well, after the first corona restrictions. The people buying these groceries were scared of national shortages, possibly fueled by the fake news rumors spreading during that time.

Lastly, we take a look at a recent (Oct 2021) report by Amnesty International. Even Amnesty, an organization that normally focuses on human rights, sees the danger fake news poses in the current day and age. The report states that the Corona pandemic was the perfect breeding ground for misinformation and fake news [23]. This misinformation made it so that people disregarded the effect of face masks or tried experimental treatment. Amnesty believes that social-media companies should have a hard look at their actions and come up with better solutions in the future. Not only that, but countries have a larger role to play as well. These countries cannot simply apply censorship, or disable the internet, but should come up with good and accurate health campaigns.

During a pandemic news travels fast, but fake news travels faster [55]. Personally, we believe that our institutions should come out of this crisis with new tools and knowledge, better equipped to fight an infodemic. Especially during times when new, trustworthy information is scarce, we should have a system in place that effectively deals with information management. Average citizens should not have to feel the effects of a failed information system, and governments have a lot to learn from the past two years.

2.6 Summary of societal relevance

By now we have shown the damage fake news can do, where it comes from, and why people spread it. It can let people believe absolutely bogus information, potentially sway an election, disrupt democracy and even kill. It comes from many different sources aiming to make money or even influence another countries politics. These stories were just a few interesting examples of the real, obvious damage of misinformation. It is dangerous. Beneath the surface, there are many more invisible consequences of fake news such as declining trust in science, polarization in politics, and declining trust in others [29]. Although these consequences are harder to show in an example, research clearly shows that they exist. The many problems misinformation creates have become clear. We have shown the extent of the problem and how it connects to our modern-day society. We believe we should have an internet where stories and news are verifiable. Although a fully-fledged solution should be multifaceted and contain many more aspects, we will focus on the aspect of trust scores and warning messages by trying to develop a new system that could play a part in a better internet.

2.7 Scientific relevance

In the previous sections we showed the damage that misinformation does, and the dangers it still poses. However, aside from the societal relevance we showed, the topic also has a large scientific relevance. Research is constantly being carried out trying to answer the questions that still do not have an answer. These questions are called the knowledge gap. In this section we will identify this knowledge gap, showing what research does not yet know or still has to agree upon.

Almost all broad and influential papers looking into fake news specify that more research is definitely warranted. Researchers agree that we need more studies into why humans spread fake news. For example, research by [55], which looked at the spread of fake news on Twitter, states that we need to identify the factors of human judgement that drive the spread of true and false news. One of the research methods they mention is the use of a survey, which we have used in our own study. Our survey looks at how eager people are to share different kinds of messages, thus completing some of the future work of [55].

While [55] is more interested in looking into sharing of fake news, researchers from [29] are interested in the perceived credibility of messages. They mention in their future work that they would like to see more research on the effect of trustworthiness indicators [29]. In my thesis, these two questions regarding sharing and credibility are combined in a single experiment where we compare the effect of different trustworthiness indicators. Doing this allows us to work on both the future work of [55] and [29].

Broader questions are also asked, such as the question of how to design a new ecosystem that values and promotes truth [28]. This is also an interesting question, as my new idea which we mention in chapter 4 could be helpful for such a new ecosystem. This ecosystem would most likely consist of many different elements, but a large-scale vote-based trust system could potentially be a part of it.

Now that we have looked at some of the knowledge gaps that the broader research has pointed out, we will focus on research that is more similar to what we have done. [32] is such similar research, which looks at warning labels for Facebook posts. This research is done by letting participants look at different posts concerning international politics. However, the researchers mention that it might be interesting to test their findings against other topics than international politics [32]. This is something we tried to incorporate in order to address this knowledge gap.

To summarize: we will address several knowledge gaps stated by [55], [29], [28] and [32]. These knowledge gaps are:

- What are the factors of human judgement that drive the spread of true and false news? [55]
- What are the effects of trustworthiness indicators? [29]
- How do we design an ecosystem that values and promotes truth? [28]
- Do people react differently to misinformation correction concerning international politics than other kinds of misinformation? [32]

Everything we have seen points to the fact that this research field is still extremely active, with many open questions remaining. This means that our research in this area is warranted and we should try and help fill the knowledge gap.

Chapter 3

Related Work

In this section we will isolate the important topics of my thesis and take a closer look at the latest research in these areas. We will inspect influential papers in order to learn from them, and build on them. To design an effective experiment we need to know what is, and what is not known about our research questions. We do not only look at fake news research, but also cognitive science research into human judgment and credibility. These topics are extremely important to find out how warning labels could actually be effective. First, we will look at research into the spread of fake news. While some of the following research in this chapter is only done on social media platforms such as Facebook or Twitter (making it not entirely representative for the entire internet) we believe that the research is still extremely valid because a very large portion of the internet actively uses one or multiple of these platforms.

3.1 The spread of fake news

In this thesis we hope to find a new solution that can limit the impact of misinformation. Existing research will tell us more about how and why fake news spreads, specifically on social media platforms such as Twitter. Perhaps the most influential misinformation paper in recent years looks at the spread of fake news stories on Twitter, compared to actual verified news [55]. In comparing 126,000 news stories on Twitter, which were being tweeted about 4.5 million times, the researchers found that misinformation always spreads faster than actual verified news [55]. An example from this paper shows that the truth rarely diffused to more than 1000 people, whereas the top 1% of fake-news cascades routinely reached between 1000 and 100,000 people [55]. This tells us that Twitter is ripe with fake news, but sadly the problem is not only confined to Twitter. An analysis of Facebook data shows us the same, fake news spreads and outperforms actual news most of the time [46]. Other research tells us that during the 2016 US elections,

each American saw at least one fake news article [4]. This is likely a conservative estimate, the actual number could be higher since this research did not take every single source of fake news into account. The reason that all this misinformation spreads faster is simply because humans are more likely to interact with fake news than with actual news [55]. Researchers from [55] believe this is the case because people are more interested in novel information. When information is newer, it get interacted with more. Another reason the researchers mention is that false information inspires feelings as surprise and disgust, which could benefit the spread of such news [55].

3.2 Human judgment & perceived credibility

In order to find out how to reduce the spread of fake news, we first need to know how humans form their opinion. Not only that, we need to find out how we can help people reformulate people's opinions so that they are based on facts. This is important information to know, because if people will not accept our warning labels or do not believe them, we will not find out how to reduce the spread of fake news. Stephan Lewandowsky, a researcher in misinformation correction states that when people hear new information they tend to accept it more often than not [30]. This is because listeners are under the assumption that a speaker tries to be truthful and honest, although this is not always the case. When someone does try to assess the truthfulness of a statement, they do this in four steps. As stated by Stephan: "First, is this information compatible with other things I believe to be true? Second, is this information internally coherent?—do the pieces form a plausible story? Third, does it come from a credible source? Fourth, do other people believe it?" [30] This is important information that we can use to construct our warning labels. So when designing these labels, we make sure that the four questions mentioned will be answered by the label or a link on the label. However, even then the literature reveals problems.

Many studies mention the backfire effect, this is an effect in which when users are presented with statements that go against their beliefs, that this only reinforces their own believes [32] [27] [30]. Another effect that we must take into account is the continued influence effect. This states that people often keep replying on misinformation, even though the information has been disproven [13].

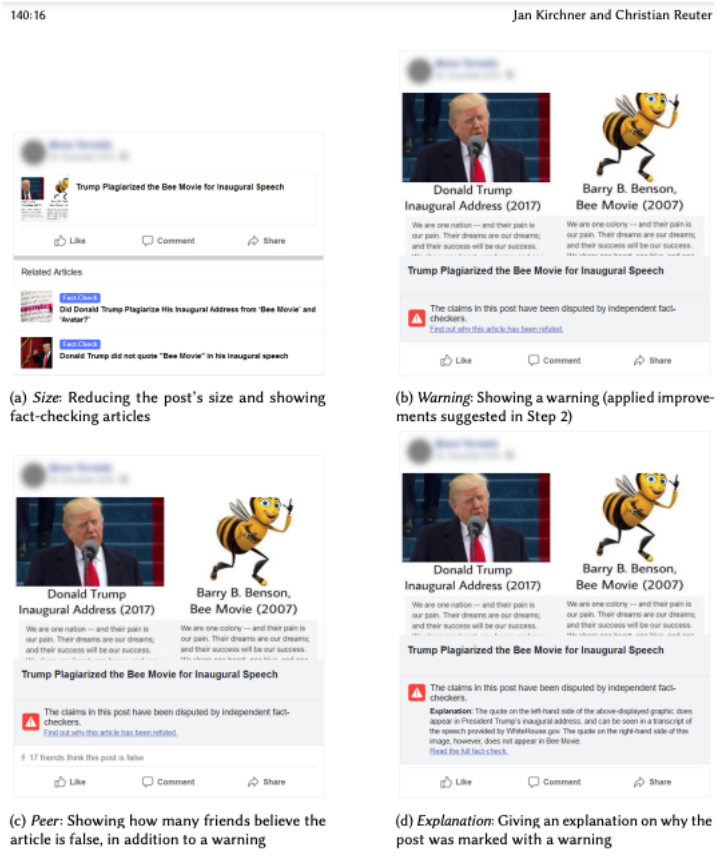
However, for both effects the literature offers solutions. First, the backfire effect will reduce significantly if certain conditions are met: A correction must not directly challenge someone's worldview, and corrections must contain an explanation [29]. Secondly, in order to reduce the continued influence effect we should: warn before initial exposure to misinformation (by using a warning cover), do repetitions of the warning, and let the warning label explain an alternative story by filling in the so called "coherence gap"[1] [30]. These are all important topics we keep in mind when designing warning labels.

[1] A missing piece of information that is necessary to understand a concept

3.3 Current research into solutions for stopping the spread and credibility of misinformation

My thesis tries to come up with a new and better way to reduce the impact of misinformation. We do this by looking extensively at previous research and solutions, and using this information to create something new. Existing research already covered several potential solutions, which we will now review.

A paper by Jan Kirchner and Christian Reuter tries to compare several design options for posts marked as misinformation. These options include: reducing the size of a post and adding related fact-checked articles, adding a warning label, showing how many friends believe this post is false, and showing a warning label together with an explanation [27]. On the following page is a screenshot from [27] showing the different options that have been researched.



(a) *Size*: Reducing the post's size and showing fact-checking articles

(b) *Warning*: Showing a warning (applied improvements suggested in Step 2)

(c) *Peer*: Showing how many friends believe the article is false, in addition to a warning

(d) *Explanation*: Giving an explanation on why the post was marked with a warning

Figure 3.1: Different conditions of warning label research by [27]

It turned out that the explanation option worked best. Similar research as [27] also used online tests in which they compared posts with and without warning labels [32]. The researchers found that the fake news posts with warning labels get believed less and get shared less [32]. This is in line with other research about warning labels vs social endorsement cues [51] and with research into warning labels for state-funded posts [34]. The last mentioned research states that a warning label has the ability to counteract the effects of state-funded misinformation.

Since the research from [27] has some similarities with our thesis, we will inspect the methods these researchers have used. Their research happened in three steps: first, the researchers conducted a survey to ask demographic information together with asking questions such as if you read an article before interacting with it. Secondly, the paper conducts semi-structured interviews to gain deeper insights into possible elements to be used to counter fake news. Lastly, based on the interviews four approaches (shown on the following page) are chosen which were then compared in effectiveness and user preference by means of an online experiment. This online experiment is something we have also used in order to stay true to actual research.

3.4 Fact-checkers

Fact-checking is the practice of independently checking certain statements, most often said by public figures such as politicians. Fact-checking is an important part of this thesis, since we will employ similar tactics and we need to know what works and what does not. First off, the literature agrees that fact-checking works to lower the belief in misinformation. Research shows that "the beliefs of the average individual becomes more accurate and factually consistent, even after a single exposure to a fact-checking message" [56]. Other researchers agree with this by stating in their papers that "fact-checkers do have the potential to correct attitude-congruent misinformation" [19] and that "the combination of news media literacy interventions and fact-checkers is most effective in lowering issue agreement and perceived accuracy of misinformation" [18]. This shows that fact-checking works, indicating that our own approach of using a new warning label containing expert votes can also work. However, the research also shows how misinformation is perceived as more credible when it confirms preexisting believes [18]. This is something we mentioned previously in the section on human judgment, where we discussed how a correction must not threaten a worldview since this can cause the backfire effect [30]. This is something we keep in mind in the design of our own solution.

Next, we will take a look at research on automated fact-checking. This can help us choose if we would like to automate our fact-checking process or focus on a more human-centered approach. There is currently much attention on automated fact-checking (AFC), as it could really help in the fight against misinformation. However, as of today these AFC systems are a long way off from being truly effective [16]. In research settings, these systems are good at identifying claims, but not so good at verifying and correcting those same claims [16]. This is because fact-checking requires a lot of context, something automated systems often lack. Efforts to create fully automated systems only show that we are still a long way off from truly AFC, with one paper stating that "Live, fully-automated fact-checking may remain an unattainable ideal" [20].

3.5 Trust scores

An important aspect that we have not yet touched upon is how tweets, or other content, can get labeled as misinformation. This is a complex topic with several different solutions. One interesting approach that we see in literature is trust scores: scores that indicate how trustworthy a tweet is. These scores can then be used to assign a warning message, or remove the content entirely. Trust scores are almost always calculated by algorithms, which will also be the focus of this part. These kinds of trust score algorithms always fall in one of three categories. Content-based, graph-based, or hybrid solutions [31]. Each one of these categories still has a variety of different solutions, in order to explain these categories better, we will take a look at an example for each of them.

3.5.1 Content-based trust scores

First off, content-based solutions look purely at hard metrics such as linguistics, retweets, or number of followers. An example of such a content-based method is a system that calculates the expertise score of users per topic [39]. The system first tries to find which topic a tweet is about, by using a method called topic modeling. This is a technique described in [38], in which an algorithm finds the topic of a tweet by using a statistical model and machine learning. This topic is then used in the actual algorithm, which calculates an expertise score of a user by looking at how often this user tweeted about the topic, compared to the person who tweeted most about it [39]. In the end, the algorithm is left with a table of users, topics, and trust scores which can be used for content moderation. But like mentioned before, even in content-based systems there are a lot of differences in the kind of systems used. For instance, there are also more AI-focused approaches that use active learning- and neural network models [26].

In this example from [26], researchers started by manually labeling a small set of tweets on trustworthiness and then training these models on the set. Once the models are fully trained and have achieved a certain accuracy, they are ready to be put into use.

3.5.2 Graph-based trust scores

Next, we have graph-based systems. These systems usually do not look too much at the content of tweets, but use graph-based solutions to come up with trustworthiness. This means that the algorithm looks at how connected users or tweets are to other (un)trustworthy users/tweets. The question then becomes how do we find out which user/tweet is trusted? One way this can be answered is by defining groups of trusted users, such as accounts from well-known people, news media, and US senators as [45] does. Next, the algorithm of [45] assigns attributes such as "politics" or "science" to each of these trusted users. Their graph-based algorithm then looks at the connections from these trusted users to users the algorithm does not yet know, and assigns trust scores in a propagating fashion in regards to the attributes. Simpler methods exist as well, such as automatically trusting users with a verified badge and only propagating trust based on these accounts instead of using attributes [42]. Furthermore, researchers are looking into ways to combine the graphs of users, tweets, and webpages into a single model [40]. Although this paper is still a proof-of-concept, it shows how a possible 3 layer graph could work, in which not only users and tweets are highlighted, but also the webpages that users are tweeting.

3.5.3 Hybrid trust scores

Finally, a system can use aspects from both content-based and graph-based systems: the hybrid approach. One such hybrid system first compares a tweet to a trustworthy news article to determine how trustworthy the tweet is, a content-based approach. But next, the system propagates the trust it calculated to other users interacting with the tweet, a graph-based method [58]. Another method works by using "coupled dual networks" [31]. These are two networks, a user network and a tweet network, that make connections to each other based on connections such as mentions or retweets. Next, all this data is stored in different matrices, that keep track of follower relations, mentions, retweets, and replies. All this data is then combined in large and complex mathematical equations to come up with a trust score.

All these systems are incredibly complex and the provided summary only shows a small part of a much larger field of research. We believe that when using these kinds of complex automated systems in practice, there are a lot of factors that can go wrong. Other than that, the current research is quite theoretical and still needs a long time before being actually implementable.

One of the factors that can go wrong with automated systems are the ethical concerns regarding bias in such systems. There are quite a few examples of automated algorithms having racist tendencies, hence we should thread carefully when employing these kinds of systems [59]. This is especially the case when using AI systems, which can operate as a "black box". This is a term researchers use when they do not really know what happens inside of the algorithm of an AI, but only know what they put in and what comes out. Not fully understanding a large-scale algorithm could have really bad consequences. That is another problem some of these trust score algorithms have, being too complex. All this complexity could lead to only a few people being able to understand the algorithm, making it so that they have the power to control and adapt it. If you do not understand something, you cannot shape it.

Finally, there are the problems of accuracy. An algorithm will always have a hard time understanding the subtleties of human text, making it unfit for questionable cases. Cultural differences and jokes could be hard to grasp but are just as important as normal text. There will always be edge cases that an algorithm will just not understand.

Because of all these different issues, we have decided to come up with our system, which can be simpler, with less bias, while working to decrease the impact of misinformation. This idea will be explained after the related work, in chapter 4. First, the next paragraph will quickly examine the most promising solutions we talked about in this chapter.

To summarize, from all the different visual design solutions we talked about in the previous section, having an explanation besides the warning message works best [27]. Thus we know that our design should have some sort of explanation. To counteract the continued influence effect, we should display the warning before the user sees the actual misinformation [30]. Other research agrees with this, by stating that using a cover instead of only a warning label works better to decrease perceived credibility [44]. Other remedies of the continued influence effect should also be kept in mind, such as repeating the warning or telling an alternative story. In order to not trigger the backfire effect, we let the explanation be subtle: it should not directly challenge someone's worldview.

Chapter 4

TrustIT: Proposing a new system based on privacy-friendly expert voting

This is the chapter in which we propose a new solution (working title "TrustIT") to the problem of misinformation. After reading promising and interesting research into misinformation correction, human judgment, fact-checkers and trust scores we have come up with a new system that hopefully has fewer flaws and a more practical implementation. The system is described in the context of Twitter. However, in theory it work for the entire internet if you were to swap tweets for web pages and make a few adjustments. In order to limit the scope of this idea and the research behind it we have chosen to only talk about Twitter. The text below is the first proposal of our idea. Ideally, this system will be refined based on experience and feedback.

TrustIT works by letting users vote on the trustworthiness of tweets. However, in order to be able to vote a user needs to be a verified expert in the topic of the tweet. These votes are then translated into a trust score which is publicly visible on the tweet, once enough experts have voted. We hope that in displaying such a trust score, combined with what we know about misinformation, we will reduce the perceived credibility and sharing intentions of tweets labeled as untrustworthy. Since this idea consists of several components, we will explain this in the different sections down below.

4.1 The experts

An important question in this system is, how do we decide who the experts are? An important distinction to make is that we do not appoint general experts, instead each expert can only say something about his or her own domain. This means that someone who studied medicine can only vote on tweets related to medicine and not on something different, such as politics. Furthermore, we differentiate between the level of experts. Someone who finished their masters in medicine is less capable than someone who has done a PhD. We propose two different levels of experts: Master experts and PhD experts. While master votes have a weight of 1, a PhD vote weighs as much as 4 master votes. This is because someone who has a doctorate in a certain area is much less likely to make mistakes and ideally knows better what they are doing. We chose a 4 to 1 ratio because this is the distribution of people with a masters vs people with a doctorate degree in the USA [7]. However, we do see the ethical concerns of this idea. This idea could create inequality between equal people, hence we will discuss the concerns in the limitations section in the discussion. To summarize: people can only vote on topics they have a masters or PhD degree in. (Other methods of choosing experts do exist, however we chose this method for an initial exploration of the concept.) Because authenticity can be faked on the internet, these experts need to prove that they are in the possession of such a degree. This is where attribute-based authentication comes in.

4.1.1 Attribute-based authentication

Attribute-based authentication (ABA) is a way of authenticating not all of yourself but only a part, namely an attribute. An attribute can be anything from age to a contract to an educational degree. Using ABA it is possible to keep a high level of privacy, while still proving certain facts about yourself. The privacy by design foundation, which started on the Radboud university, are working on a mobile app, IRMA, which can do exactly that [9]. IRMA is an ABA system with which users can (cryptographically) prove properties (attributes) about themselves, e.g., that they have a degree in medicine. We envision that in the future, Dutch degrees would be issued in the form of IRMA attributes by DUO[1]. This attribute is then stored in the mobile app from IRMA. It would then be possible to securely prove this degree, or any other attribute, to someone who asks for it such as Twitter or a potential employer. The person asking for the proof is called the *verifier*. What makes IRMA special is that proof can be given to the verifier, without the verifier actually learning anything else about the user (e.g., their name or university).

[1]The dutch organization responsible for executing education policy

How this works in practice is that any user can load their respective degrees on the IRMA app, proving that it is authentic. We assume that such a system can be used world-wide, and envision that degrees can be linked to Twitter.

Because we use IRMA in our system Twitter cannot learn any information from this degree, an official name or university all stay hidden. This greatly improves privacy, which is especially important when using documents such as official degrees. As mentioned earlier, it is possible to load in a degree using IRMA. The function is currently still a demo, but it shows that this solution is not far-fetched and can actually already be implemented with relatively little work [8]. This shows that such a system is highly possible and usable. Since the idea we are proposing is only a proof-of-concept, there are not yet ways to use ABA for degrees on a large, international scale. However, similar solutions are emerging world-wide and IRMA shows that ABA is highly possible.

4.1.2 Topic recognition

Another problem we need to solve is how a tweet gets turned into a topic and a degree linked to that topic. The first part of the problem, turning a tweet into a topic, has been previously researched. A paper by Twitter employees shows how this currently possible by using a large system of topic inference mechanisms [57]. These mechanisms are machine learning models that interact with each other, each carrying out their own task. The paper from [57] has 93% success rate in modeling topics on a large scale. The area of Twitter topic modeling is still discussed in research with other ideas being posed by [47] and [38]. We did not manage to find any papers on the linking of a scientific degree to a topic and consider this part out of scope for this thesis. Because we are more focused on decreasing perceived credibility and eagerness to share fake news content, we reserve this area for further research.

TrustIT in a whole starts when the topic of a tweet has been recognized. This topic is then linked to one of the research areas, which links to the experts that are also linked to a research area via a degree. These experts can then vote on the trustworthiness of the tweet. This now gives us a system in which experts use an attribute-based authentication system, like IRMA, to prove their degrees, which are then connected to Twitter who handles the voting. The next page contains a diagram that will summarize all of the connections between IRMA, Twitter, and the users.

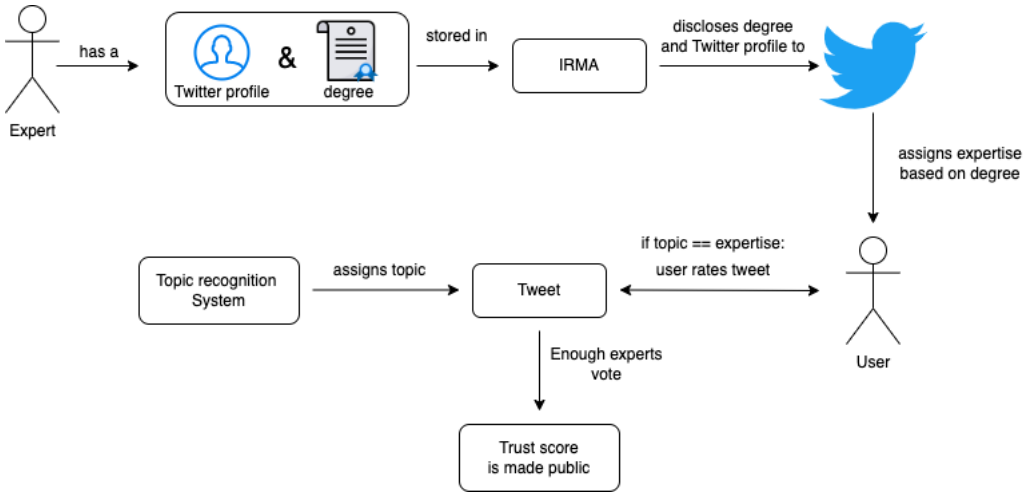


Figure 4.1: Diagram into the different components of TrustIT

4.2 The voting

The outlined idea is based less on automatic recognition, which is prone to errors, and more on human judgment. As explained, the experts will be able to vote on tweets that Twitter deems in coherence with their expertise. As the idea currently stands, the experts will have a separate timeline in which they see tweets they can vote on. These tweets will be displayed on basis of how popular they are. More popular tweets will get to the top, making it so that important tweets get verified quickly. Alternatively, people have the ability to suggest a tweet for a credibility check. If enough people indicate they doubt the credibility of a tweet, the system can add the tweet to the list of tweets meant for experts. Voting itself will work as simple and inviting as possible so that people are encouraged to vote. We envision a slider below the tweet, which can be moved by the correct experts. This slider will then converge on the average votes, which will be visible to everyone. This is displayed in figure 4.2 on the next page.

4.3 The display

Once a tweet reaches a certain threshold of votes, the trust score becomes live. This trust score should show in one glance how trustworthy a tweet is and how many experts gave their opinion. The display itself should be intuitive and simple. However, we should also keep in mind the research that we discussed in the related work chapter. In the section on human judgement & perceived credibility we showed research that can help to make a warning label more effective. We talked about what make people accept information and how to counteract the effects that diminish the warning labels, the

backfire and continued influence effect. We tried to make sure that people understand this message is coming from a credible source, together with showing how many other people accept this warning label. To counteract the backfire effect we tried to keep the label as neutral as possible, while also providing an explanation. To negate the continued influence effect we decided against using a warning cover, because this could be confusing and is already being tested in condition 2. Instead, we used a red outline and a "WARNING" text to make sure users will first see the warning label before the actual misinformation. All of this answers our design question, in which we asked ourselves: *"How do we display a trust score in such a way that it is most effective at reducing the impact of misinformation?"* After different iterations of the design, we found something we were happy with. This mock-up is what we used for our online experiment and is displayed in the figure below.

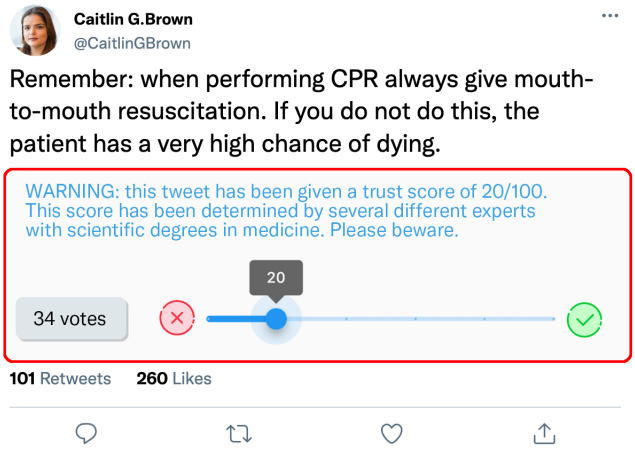


Figure 4.2: Visual of a tweet on TrustIT

A slider below the tweet is displayed for everyone, with a number indicating the trust score. The ends of the slider contain a simple red cross, and a green check, indicating a high or low trust. Furthermore, a label next to the slider displays how many experts have voted on the tweet. We choose a red outline for the trust score section to bring people's attention immediately to the important trust score.

4.4 Practical implementations

This idea is currently still a proof-of-concept, where we have shown how and with what techniques it can be turned into a reality. However, in order for TrustIT to truly work Twitter needs to play a big role. A large part of the system is concentrated on the Twitter platform: assigning expertise, voting and topic recognition. This is a big ask for Twitter. They would need to link IRMA to their services, program an entire voting system where only certain people can vote on certain tweets, create a topic recognition system and much more. A more realistic approach would be to develop this idea as a browser extension, in which much of the features are handled by this extension. This would still need to be programmed and maintained, but could be done by an independent foundation or organisation that sees the benefit in a system like this.

Chapter 5

Methodology

In this thesis we are interested in finding answers to the research questions previously mentioned. The main question we ask ourselves is whether and to what extend a vote-based trust score system would reduce the impact of misinformation. We also asked ourselves two sub research questions, in which we were wondering how our system would affect the spread and perceived credibility of misinformation. For the two sub research questions, we pose the following two hypotheses: *A correctly displayed vote-based trust score system can reduce the impact of misinformation by lowering eagerness to share* (H1). And: *A correctly displayed vote-based trust score system can reduce the impact of misinformation by lowering perceived credibility* (H2). We believe this is the case because we have seen both similar research in this area and general misinformation research that states well designed warning labels can have a real effect on eagerness to share and perceived credibility [32] [27] [30]. In order to accept or reject these hypotheses, we decided to use an online experiment and a questionnaire to gather the data necessary. This chapter demonstrates how we have done our research and have attempted to answer all research questions.

5.1 Design

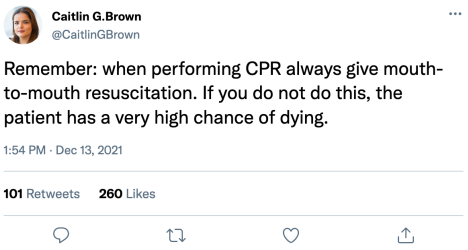
After considering several different approaches of doing our research, we decided on an experimental design that is between-groups and quantitative. This means we test different conditions on different groups of people, with the aim of gathering as much data as possible. The reason we choose a between-group experiment and not a within-subjects experiment is that we do not want any sort of learning or transferring to take place from one condition to the other. If we had chosen for a within-subjects design a participant would see the three different conditions we had prepared, which could possibly influence the answers a participant would give to the questions. Furthermore, a between-group experiment has the added benefits of

being simple (thus less room for error) and shorter (making more people willing to fill out the questionnaire). However, there are also downsides, such as the fact that you need more participants to have valid results. However, we decided that the advantages of between-groups study outweigh the disadvantages.

The reason that we used primarily quantitative data is that we felt that we could achieve better conclusions with this data. We did include a few optional questions in the questionnaire for participants to explain their choices, in the hopes of getting a better insight into how people think. In a qualitative analysis, such as with interviews, it is harder to generalize the data to a broader group of people. Finally, the reason that we choose an online experiment is the fact that this best mimics actual Twitter use: online. This way the experience of the experiment will be more similar to Twitter than if we were to do this in an offline setting.

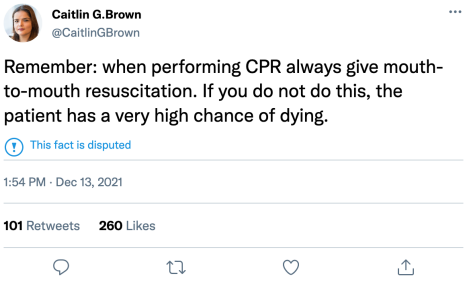
5.1.1 Research variables

The research variables are what a researcher changes, and what it measures. These are also called the independent, and dependent variables. As independent variables, we took three different approaches to the design of a trust score or warning label in combination with a tweet. First, a control condition. This condition just shows a normal tweet without anything added; this serves as our baseline to which we can compare new methods. Our second condition is a close copy of what Twitter currently uses to notify users of misinformation. This will show us how well current methods work. Finally, the third condition works with our new idea by letting people vote on topics they are an expert in. This condition shows a trust score, combined with an explanation as to how this trust score came to be. The tweet we used for this experiment is deliberately created to make people question its credibility, but not know an answer for sure. By making the content of the tweet not obvious fake news, we hope to more accurately measure the effectiveness of our conditions. Below we have outlined the different tweets, with all three conditions.

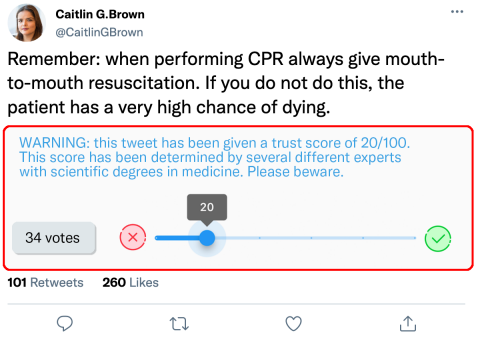


Control condition

Condition 2: Current measures



Condition 2: Current measures (After clicking view)



Condition 3: New idea

Next, as the dependent variable (the thing you measure), we measured both the message credibility and sharing intentions. To measure the credibility of a message, we used a scale outlined by [5] which was also used in previous similar research [32]. This credibility test asks how well the adjectives accurate, authentic, and believable describe the content presented on a 1 to 7 scale. These rating of these attributes can then be combined into a single credibility score which we can use for research. To measure sharing intention, we again used the questions in [32] which asks how likely you, and others, are in sharing this content. We tried to use the same questions in order to stay true to other research which we can then build upon. At the start of the survey we also measured the dependent demographic variables sex, age, education, and social media use. This information is necessary to know what kind of sample we have and how it compares to a general population. Another interesting section of questions we considered was asking about participants media literacy. This could have given insights into a possible correlation between how knowledgeable a participant considers his or herself to media and eagerness to share and perceived credibility of misinformation. However, media literacy is not a focus of this study. Fur-

thermore, research has already shown that media literacy does not help to identify fake news [24]. Although there is a chance that this could show a correlation in our data, we decided against adding this section. Another reason we decided this is because it adds multiple questions, adding time to our experiment.

5.1.2 Experiment setup

As stated before, we figured the best way to conduct quantitative research is by using an online experimental setup with a questionnaire. In this setup we had the ability to show participants a condition, and immediately gather data on this condition by letting the participant fill out questions. This section will explain some choices we took when designing this experiment. First off, we used LimeSurvey. Some advantages of this platform are that it is able to randomize participants, stores the data securely and since LimeSurvey is hosted by the university it does not rely on any 3rd party.

Another decision we took in designing the survey is letting participants know as little as possible before starting the study. All the participant knew at the start is what we wrote in the information letter at the beginning of the experiment, where we explained general information about consent, data use and possible discomforts of the study. We limited the information about study details, making sure that all a participant knew about the study before starting the experiment was that it had to do with Twitter. This made it so that we had less chance of priming, in which a participant will already think about the credibility of tweets before actually seeing the condition. Our hope is that in doing it this way the data will be cleaner. After asking the demographic questions previously described, the participants would see one of our three conditions. We used LimeSurvey with a random number generator to make it truly random which participant saw which condition.

On the same page where the participants could see the condition, we also displayed several questions about eagerness to share and credibility. As previously explained these questions were used by other research and could help us answer our research questions. To check whether the participants actually paid attention to the shown tweets, we added several attention tests after the research. These were questions that would be obvious to answer to anyone who looked at the tweet, but would not be so clear to anyone that just gave random responses. This is another method we used to make our data cleaner.

Finally, we used a debriefing after the experiment was over in which we explained what the study was about, along with giving participants the option to remove their data, ask (privacy) questions, or be notified when conclusions about this research were reached.

Chapter 6

Results

This chapter will detail the results we got from our online experiment. First, we will share statistics on our population to see how these statistics compare to a general population. After we will analyze the data from both the sharing- and credibility questions. Finally, we will discuss some remarks given by the participants in the last open question.

An important remark we must make before analyzing the results is that one of the attention tests was answered incorrectly a significant amount of times, in only one of the conditions. Attention test 3, which asked which of the three tweets the participants saw on the previous page, was answered incorrectly `12/23` times in the group of condition 2. We believe that these participants did in fact pay attention to the experiment but did not see the difference between the disputed tweet from condition 2 and the control tweet from condition 1. This is an interesting result which we will discuss in the next chapter. However, when analyzing these results we cannot be 100% sure that our explanation as to why so many participants answered incorrectly is correct and thus we will analyze the results both with and without these participants.

When performing statistics, it is important that you choose the correct statistical tests based on what data you have. We see the data from the sharing questions as ordinal data, since this was gathered using a Likert scale. Although there is some discussion on Likert data being ordinal or interval, we will follow the general advice on simple Likert data being ordinal [48]. The credibility score was also obtained using three Likert scale questions, however this data has then been transformed into a credibility score, making it interval data. As stated by [48], a good statistical test for ordinal Likert data is the Kruskal-Wallis test, which determines if the median for two groups is significantly different. This is the test we will use for data regarding the sharing questions. To analyze the data on credibility, we use

several statistical tests. First, one-way ANOVA will be applied to see if there is a significant difference between any two groups. One-way ANOVA works when your data is in interval, along with 5 other assumptions explained in [3]. If the result from ANOVA shows significance and the data adheres to all assumptions, we will use the Tukey post-hoc test to find out if one of the three groups is significantly different compared to another group. However, when there are outliers present in the data we first check if the outlier has an effect on the ANOVA results. If it does, we use the Kruskall-Wallis test, as is standard practice [3]. Kruskall-Wallis is a non-parametric statistical test that also works well with outliers. If this test shows significance we will again run a post-hoc test, but now we use the Mann Whitney U test. This non-parametric test tells us if two groups are significantly different from each other. We choose Mann Whitney U instead of Tukey because this test works best after the use of Kruskall-Wallis [1]. As alpha value for the statistical tests we choose `0.05`, this means that if the p-value of any test is lower than `0.05`, the result is significant. We choose this value because this is commonly used in social research.

6.1 Population

The experiment had a total number of `76` participants. However, there were several participants who had to be excluded because of different reasons. A total of `14` participants did not click "Yes" on the final question which asked for consent to use their data, instead they most likely clicked out of the survey thinking they had completed it. Furthermore, one participant answered incorrectly on attention test 1, asking what the tweet was about. Attention test 2, asking if the participant knew the author, also had a single incorrect answer. As mentioned before, attention test 3 had a lot of incorrect responses from participants in condition 2. Because we believe these participants did in fact pay attention we show the results with and without these participants. However, there was also another single participant that answered incorrectly on attention test 3, who was not part of condition 2, meaning we removed this incorrect result. This means that we are left with `59` useful responses. If we exclude the number of people from condition 2 answering incorrectly on attention test 3 we get a total of `47` useful responses.

To estimate what kind of population we were working with we asked questions regarding gender, age, education and social media use. Bar graphs of these results can be found in appendix A, showing all concrete numbers. The results we got regarding gender was slightly skewed towards males (Male = 35, Female = 22, Other = 2). When we removed the participants from condition 2 that answered incorrectly on attention test 3 this skew still exists (Male = 26, Female = 19, Other = 2). Next, we look at age. Our par-

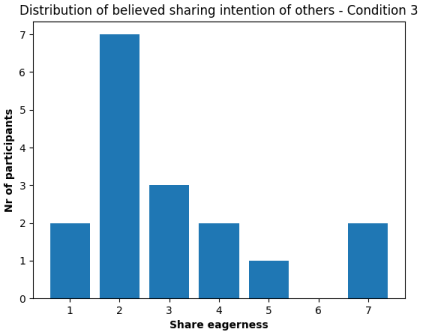
ticipants are mostly in the age group of 23-27 and 18-22. There are some participants in other age groups, but 88% of participants fall in one of the groups mentioned. On education, most participants fall into the category of having finished high school or bachelors. Finally, most participants use social media daily, with 1/5th of participants using hourly and 1/10th using weekly.

6.2 Sharing

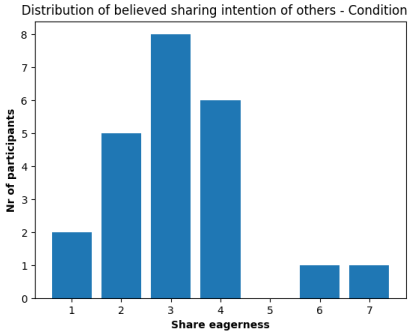
We asked two questions related to sharing: how likely the participant thinks it is that he would share the tweet and how likely the participant thinks it is that others would share the tweet. In regards to the first question, `52` of the `59` participants responded with `1 (Very unlikely)`. Because of the fact that the `7` remaining participants were not confined to one condition but spread out across all three conditions, we decided to not analyze the results of this question because the data does not show anything interesting. Furthermore, since `88%` of the participants answered the same significance cannot be reached. When removing the incorrect answers for attention test 3 we get `42` of `47` participants responding with `1 (Very unlikely)`, giving us a percentage of `89%` same answers.

The second question, "How likely do you think others are to share this tweet?" did not contain mostly similar answers and thus could yield significant results. In order to correctly analyze the results, we first present descriptive statistics in the form of bar graphs. The bar graphs on the next page will give us a quick look at the answers given for sharing question 2.

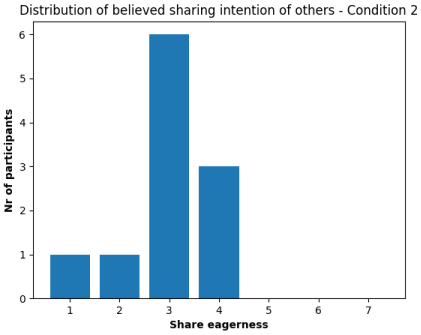
Control condition



Condition 3: new idea



Condition 2: All participants



Condition 2: Participants removed

According to the bar graphs all conditions are really similar, both with all participants and the participants from condition 2 who answered attention test 3 incorrectly removed. This indicates that the different conditions did not influence how people thought others would behave in sharing. However, to be completely sure whether or not there is a significant difference we have performed a Kruskall-Wallis test, as explained before.

We will first look at the results with *all participants*. Using all participants, the Kruskall-Wallis test showed no significant difference in sharing intentions of others between the three conditions, $H(2) = 0.898$, $p = 0.638$. When looking at the results with the *participants removed* from condition 2 who answered attention test 3 incorrectly, Kruskall-Wallis again showed no significance, $H(2) = 0.984$, $p = 0.611$.

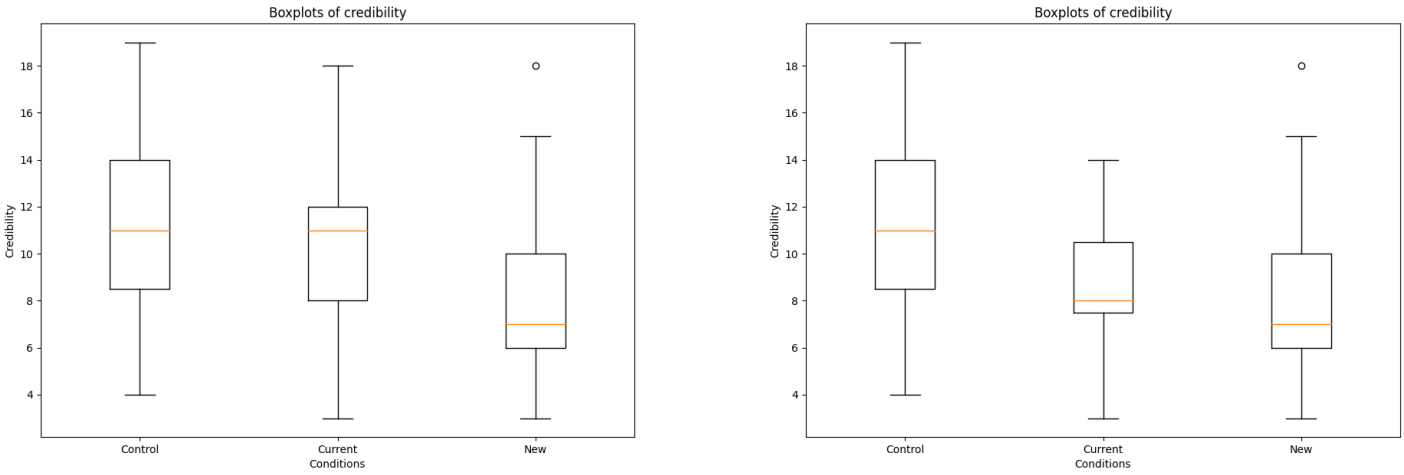
After we displayed the two share questions, we asked participants if they could share their motivation for their previous answers. The answers to these questions usually fall in one of these categories:

- Users stating that they do not share much
- Users stating that Twitter is not a good source
- Users stating that the warning put them off

This last category included answers such as: "A score of 20/100 is quite low, thus this tweet is not really credible", "The Information in the tweet doesn't seem to come from a professional by looking at the warning below it", "I probably wouldn't share a tweet about a topic I don't know anything about. Especially if experts gave it a low trust score". These comments were all placed by participants in condition three, indicating that some users do not want to share because of the warning.

6.3 Credibility

To measure the perceived credibility of the tweets in different conditions, we asked participants to rate on a scale from 1-7 how: accurate, authentic and believable they thought the content was. These three variables were then combined to create the credibility rating [5]. This time we use boxplots for descriptive statistics, making it so that we can quickly see an overview of the data.



(a) All participants

(b) Participants removed

Figure 6.1: Box plots of credibility ratings

These boxplots already show an interesting result: people who answered correctly on attention test 3, asking which tweet they saw, deemed the tweet less credible than people who answered incorrectly. Also, the boxplots seem to indicate that the tweet with the new system gets perceived as less credible, which would make sense since the tweet had a low trust score. A final interesting observation is that the boxplots show an outlier for condition 3. After examining all answers from the participant associated with the outlier we found no errors or other unusual observations in the participants answers. This means that the value is legit and we cannot simply remove it. Because ANOVA does not work well with outliers, we will run the test both with and without the outlier to see if there is a difference [3]. We will also look at the data with and without the participants from condition 2 that got attention test 3 wrong.

When we include *all participants*, we see that the one-way ANOVA revealed no statistically significant difference in credibility between at least two groups (F(2, 56) = 2.822, p = 0.068). Because we have an outlier we ran one-way ANOVA again, but with the outlier removed. This reveals that there is a significant difference in credibility between at least two groups (F(2, 55) = 4.348, p = 0.018). The results with and without the outlier are significantly different, this means we had to run a non-parametric test which is less sensitive to outliers [3]. The test we choose is the Kruskal-Wallis test, which can be used when all four of its assumptions are met. We followed advice from [2], checking the shape of the data in SPSS Statistics and found all assumptions are indeed met. The results from the Kruskal-Wallis test indicated that there is a significant difference in credibility between at least two groups (H(2) = 6.4598, p = .0396). Because we reached significance we ran a post-hoc test to find out in which groups the difference lies. The post-hoc test we can use for this is the Mann Whitney U test, in which we can compare all groups to each other [1]. Credibility scores of the tweet in group 1 (*Mdn*: 11) were not higher or lower than those in group 2 (*Mdn*: 11). A Mann Whitney test found no statistically significant difference (U = 191.5, p = 0.251, z = -0.669). However, credibility scores of the tweet in group 3 (*Mdn*: 7) were lower than those in group 1 (*Mdn*: 11). The Mann Whitney test found a statistically significant difference (U = 90, p = 0.012, z = 2.249). Finally, credibility scores of the tweet in group 3 (*Mdn*: 7) were also lower than those in group 2 (*Mdn*: 11). The Mann Whitney test again found a statistically significant difference (U = 119, p = .019, z = 2.079)

For completeness we also looked at the data with *participants removed*, these were the participants from condition 2 who answered attention test 3 incorrectly. The one-way ANOVA revealed that there was a statistically significant difference in credibility between at least two groups (F(2, 44) = 3.287, p = 0.047). Because we have an outlier we ran one-way ANOVA again, but with the outlier removed. We again found a statistically significant difference in credibility between at least two groups (F(2, 43) = 4.694, p = 0.014). Because in this case the outlier does not make a significant difference we do not have have to run other tests that are less sensitive to outliers. This means we can simply use the Tukey post-hoc test as planned. This test did however not find a statistically significant difference between groups 1 and 2 (p = 1.127), groups 1 and 3 (p = 0.096), and groups 2 and 3 (p = 0.989)

We previously mentioned two hypotheses to our main research question. Hypothesis H1 was: *A correctly displayed vote-based trust score system can reduce the impact of misinformation by lowering eagerness to share* and H2: *A correctly displayed vote-based trust score system can reduce the impact of misinformation by lowering perceived credibility.* As we have seen, the credibility is significantly lowered when we use a vote-based trust score system compared to when we do not. We have shown that group 3 is significantly different from groups 1 and 2 by using several statistical tests. However, this result was only present when using all participants, including the participants from condition 2 who answered incorrectly on attention test 3. We assume that all participants did in fact pay attention, but since we cannot know that for sure our conclusions are slightly less strong. We did not manage to reduce the eagerness to share when displaying our system. No significance was reached when performing tests on the sharing data. This means that we fail to accept hypothesis H1, but can accept hypothesis H2 on the assumption that all participants paid accurate attention.

6.4 Remarks

At the end of the experiment, we included an open question for people to give remarks or to let us know something. Almost all of the responses were not noteworthy, giving responses such as "No thanks" or "interesting". However, there was one theme that kept recurring. Some participants from condition 2 (warning label & disputed tweet) were having trouble answering one of the attention tests, in which they were asked to indicate which tweet they saw on the previous page. These participants indicated that they saw both the control condition and the warning label from condition 2. However, this is not the case. These participants were confusing the control condition with the second tweet of condition 2, which included a disputed tag.

Chapter 7

Discussion

This chapter will reflect back on our research questions and hypotheses, hopefully finding answers to the questions posed earlier. We will start with trying to answer our bigger questions and relating these answers to current literature, before diving more deeply into the interpretation of the results, coming back to our knowledge gaps and discussing limitations.

The main result of this thesis is that the use of a warning label based upon expert voting works to reduce perceived credibility, confirming H2. This reduction in credibility was only shown in condition 3, where we used our own idea and not in condition 2, which mimicked current solutions. None of the conditions worked to lower a participants own sharing intentions, or believed sharing intention of others. An unexpected, interesting result, is that half of the participants who saw the disputed tag Twitter currently uses to label misinformation did not notice this tag. The main research question we asked ourselves was: "Whether and to what extend would a new vote-based trust score system reduce the impact of misinformation?" We now know that such a system, in the way that we designed it, is able to reduce the perceived credibility of misinformation. When credibility of misinformation is lowered, the impact this misinformation has is consequently also lowered. However, because we did not manage to lower sharing intentions the impact of such a system is not as high as it can be. An assumption we make for these conclusions is that even the participants we removed paid attention, something we cannot be 100% sure about. Because of this our conclusions are not as strong as we would like them to be.

By doing this research we increased the body of knowledge into warning labels and misinformation. The results we obtained regarding credibility are in line with previous research by [27] and [32], who both found that labeled posts are perceived as less credible. In line with [27], we found that using an explanation (as we did in condition 3) worked best. However, while

both [27] and [32] also found lower sharing intentions, we did not find such results. The fact that condition 2 did not significantly reduce credibility or sharing intentions is not in line with current research by [30]. This could possibly be because participants could scroll past the first warning cover, and thus only saw the second tweet with the disputed tag.

7.1 Interpretation of results

Now that we have given an overview of the conclusions, we will dive deeper into the results we have gathered. First, we will discuss the issue regarding the many incorrect answers on the third attention test. This happened only with the group of condition 2, where we showed both a warning label and a second tweet with a disputed tag. `12/23` people in condition 2 indicated they saw condition 1 and there were four people mentioning in the remarks section that they saw both tweets (condition 1 & 2). This indicates that around 50% of the people who get shown a disputed tag do not actually register this. This is corroborated by the boxplots of the perceived credibility. Figure 6.1b shows the boxplots of participants who most likely actively noticed the disputed tag, while figure 6.1a contains all participants. We can clearly see that figure 6.1b shows lower perceived credibility on condition 2 (current), which would make sense if only the people from figure 6.1b actively saw the disputed tag. From these results we can conclude that a significant amount of people do not actively notice the disputed tag Twitter currently uses.

Next, we will take a closer look at the results from the sharing questions. First off, a high percentage of participants did indicate not wanting to share the tweet, regardless of which condition the participant was in. This could be due to the fact that Twitter has become a platform in which well-known people share tweets, but others rarely do. If someone usually does not share content, they would also not do that in this experiment. Secondly, people indicating how others likely they thought others were to share the tweet also did not result in differences between the conditions. A possible explanation is that people believe others will share content no matter what kind of warning is attached to it. However, another possible explanation is that the participants of this study simply did not know how Twitter users would react to the different conditions. Because of the fact that the Kruskal-Wallis test did not show any significance between any group we cannot accept nor reject these explanations.

Finally, we will discuss the results of the credibility questions. The boxplots start with an indication that condition 3 reduced the perceived credibility the most, while the perceived credibility in condition 2 only goes down when

participants actually noticed the disputed tag. The ANOVA and Kruskal-Wallis tests confirmed that there is indeed a significant difference between groups, both with all participants and with participants removed. The Matt Whitney U and Turkey post-hoc tests showed that this difference is only significant when we include all participants and look at the difference between the control condition and the condition with the new idea. This means that we have succeeded in coming up with a new idea that actually reduces perceived credibility. Condition 2 might have not reached significance because of the fact that with participants removed, this condition had a lot less data points than the other conditions.

But how is it that our own idea was so good in reducing significance, while something Twitter currently uses does not work? One possible explanation is the novelty effect: which state that a positive effect can be due to something being new and interesting, rather than actually effective [52]. It could be that our participants only perceived the credibility as lower because they saw this score for the first time, making it really interesting.

7.2 Knowledge gaps

Earlier in this thesis we mentioned several knowledge gaps that we would like to come back to. First off, we wanted to identify the factors that drive the spread of true and false news [55]. This is something our research could have potentially told us more about. The sharing questions we asked could have been linked to other factors such as social media use or perceived credibility. Sadly, because our sharing questions did not deliver significant results we could not gain any statistical insights from them. We did hear from some participants that they do not want to share something with our warning label, indicating that the TrustIT system could potentially be a driving factor to negate the spread of false news. Also, we have encountered some possible driving factors of spread in the literature we examined. One paper shows that people want to share content they consider important, but do not pay attention to the accuracy of the content [36]. Another driving factor is that people often share information they previously believed to be true, even if it is not [6].

Secondly, we asked what the effects are of trustworthiness indicators [29]. While we did not check all the different kinds of indicators, we have a clear view on the effect of our own trustworthiness indicator. We know that, based the results from the experiment, a trustworthiness indicator in the way that we designed it works to reduce perceived credibility. Furthermore, we know from our experiment that a simple warning tag below the tweet, which Twitter currently uses, most likely has a relatively small effect. This is because many people do not notice this warning tag.

Next, we stated a broader question, "How do we design an ecosystem that values and promotes truth?" [28] While an answer to this question most likely requires years of work from professionals from all fields, we feel that our thesis touched upon a part of the answer for this question. The system we designed is most definitively able to value and promote truth. We have shown that it works in the ecosystem of Twitter, but future research with other ecosystems could be interesting.

Finally, we asked ourselves if misinformation correction regarding international politics is any different from other kinds of misinformation [32]. This is something we originally planned to incorporate in our experiment, by letting participants see multiple tweets with different themes. In the end we decided against this because it would unnecessarily complicate the experiment. Thus we have not answered this question and recommend it as further work.

7.3 Limitations

7.3.1 Problems with the research

Although our experiment delivered some exciting results, there are limitations that we should mention. Most importantly, our sample of participants suffered from selection bias. This happens when the population used in an experiment does not represent a general population. As shown in the results section, and visually in appendix A, the population concentrated on several age and education groups. This happened because we shared the online experiment with my social groups, making the population skewed in certain categories. It would have been better to use a paid survey service such as Prolific, in which you can select your participants in such a way that the sample will represent a general population. This would help to make our conclusions more general and more scientifically sound. Furthermore, an increase in sample size would also help our research. Although the statistical tests showed that we obtained some statistically significant results, an increase in sample size could have led to more precise results. This could have potentially shown a correlation in credibility between the control condition and condition 2, something we did not manage to show.

7.3.2 Problems with the TrustIT

As mentioned previously, this system does have at least one ethical concern but there are more issues that we would like to address. First of all, the ethical concern of the votes of people with a PhD degree being worth more than masters degrees. We believe that it is nice for everyone to be equal, this is how a democracy functions and how most people like it. However, the fact

that on the internet everyone has an "equal voice" contributes to a lot of the harm that misinformation has done. People can easily spread misinformation or pretend to be someone else and because of this people are possibly quicker to believe false information. In our system we are most concerned about the truth, which is something that comes with years of study. It is no question that people who studied a subject for longer, know more about it. Since the primary goal of TrustIT is to show the truth, we believe that it is justified that people with a PhD degree have a vote that counts for more.

Another problem we foresee is that people who have a strong belief in conspiracy theories and distrust in authority might also distrust our system. Once people are stuck in such a mindset it is really hard to change their ideas, because people believe what fits in their worldview [29]. In order to reduce this effect as much as possible, we propose that the votes are as transparent as possible. If an expert would like to connect their name to a vote, this would be possible. This way people have the ability to converse and share opinions, which would hopefully take away some of the doubt people have. This would also reduce any doubts of the system being fake.

There is also the question of why experts would even take the time to rate tweets. Our hope is that since Twitter is such a large social network, enough people will feel the need to label tweets. Twitter is a platform with 211 million daily users, making it quite likely that enough people will vote on tweets [11]. Especially if Twitter could really show the importance of this system, we believe quite some experts will want their voices to be heard.

Finally, we are not sure if such a system as this could really stop the large misinformation campaigns other countries are currently participating in. State actors have the potential to discredit TrustIT and spread doubt about it. However, because we have shown that our method does work to decrease perceived credibility, we cannot say such a system will be completely powerless against state actors. Especially if you were to only use such a system during vulnerable times, such as elections, it could be a great help.

Chapter 8

Future work

We have answered some of our research questions, but are also still left with questions we did not find an answer for. The fact that we only got a statistically significant result using all participants already means that more research is definitely warranted. Although our conclusions are good, we cannot say with full certainty that all the participants from condition 2 who answered incorrectly on attention test 3 were paying full attention. That is why this final chapter will detail what possible directions further research could take.

First, we will suggest some ideas similar research could take as a follow-up. It would be interesting to repeat this research with other conditions. For example, we could find out if a positive, trustworthy expert score would increase perceived credibility. This is also important as it could promote truthful content that has less perceived credibility. Other conditions we could test are different designs of the trust score warning label, in order to find an optimal solution. Furthermore, in order to better validate the results we gathered, future work could repeat this experiment in a way where participants from each group see multiple tweets. Showing multiple tweets with different authors, retweet/like counts and themes could account for any influence of these and would help to complete the future work of [32]. If all different tweets show the same credibility results, there will be less chance that the content itself played a role in the perceived credibility but instead the warning label was the deciding factor. However, it would also be interesting to change the content of the tweet in order to find out how important the content is compared to a trust score. We also mentioned the novelty effect playing a potential role. In order to find out if this is indeed the case a longer experiment can be set up which can possibly find out if perceived credibility goes down over time with our method. Finally, the experiment can be repeated with more participants that better reflect a general population. This would help to better generalize the results.

We believe the idea of a fact-checking network with verified experts being able to vote has much more research potential. There are three interesting general directions further research could take when continuing this idea. First off, we looked at how we can design such a system for Twitter, but TrustIT does not have to be limited to Twitter. New research could find out if such a system could work for other social platforms such as Facebook, or even for general web pages. This research could follow a similar setup to ours, with an online experiment mimicking the actual environment. Improvements can be made by designing and programming an actual social media feed in order to truly let the participants be immersed.

A second direction we believe this research can take is of qualitative nature. Interviews can be conducted with experts in the field of misinformation and end-users of Twitter to gather valuable insights. Both researchers and people working in the private sector could share their opinion and make the idea more robust. It would also be interesting to conduct interviews with believers and sharers of misinformation, to try and gauge how such a system could help them. We have already received some interesting statements that we outlined in the results section. This shows that it would indeed be valuable to learn even more, by talking directly to those people.

Lastly, a third direction this research could possibly take is developing a proof-of-concept. This could be a browser extension displaying a trust score or slider for a certain website, or something that automatically adds the system to Twitter. Another proof-of-concept could be an international version of the attribute-based authentication system, where people from all countries can upload their degrees and other attributes.

An interesting finding of our study is that people often do not notice the disputed tag Twitter currently uses. Because researching this was not one of our initial goals we did gather some data on this, but the conditions were not designed to fully measure this effect. It would be interesting to set up a follow-up experiment in which this effect is fully tested. The experiment can be quite similar to what we did, using a control condition, a condition with a simple warning tag, and a condition with a warning tag combined with a cover-up label.

This thesis declared certain aspects out of scope. Something we did not fully research was the topic recognition system. Although we have shown some interesting papers ([57] [47] [38]) we did not fully dive into this aspect. Especially the part of the system that focuses on linking a scientific degree to a topic could be researched more. Future work could focus on this, as it is something we have deliberately left out of scope. The same is true for

echo chambers. While we mentioned their effect being potentially dangerous, we simply could not include this aspect in our research. Later research can find out if our method of warning against misinformation could work in the context of an echo chamber, where polarized content usually thrives. Also, we did not manage to address the knowledge gap regarding political misinformation compared to correction of other themes. Future research could look into this by repeating our experiment with multiple tweets with different themes, as suggested earlier.

8.1 Final thoughts

We started this thesis with a few broad and interesting questions we wanted to find answers to. Is the power of fake news overestimated, or do people still think too little of it? Can it even be stopped and how would that be possible? After studying this topic for several months, we now know that the dangers of fake news are most definitely not overestimated. We have seen the dangers it poses and the consequences of letting it go unchecked. This is symptom of an unhealthy internet that should be treated. Our goal was to develop a system that would help reduce the impact of misinformation. The first step in this has been achieved, by showing that our system has the potential to decrease perceived credibility. Although our solution cannot solve the misinformation crisis on its own, this thesis has shown a possible way forward and added to the body of knowledge regarding misinformation correction.

Bibliography

[1] Kruskal Wallis - Wikistatistiek.
`https://wikistatistiek.amc.nl/index.php/Kruskal_Wallis`.

[2] Kruskal-Wallis H Test in SPSS Statistics | Procedure, output and interpretation of the output using a relevant example. https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php.

[3] One-way ANOVA in SPSS Statistics - Step-by-step procedure including testing of assumptions. https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php.

[4] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017.

[5] Alyssa Appelman and S. Shyam Sundar. Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, 93(1):59–79, March 2016. Publisher: SAGE Publications Inc.

[6] Tom Buchanan. Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *PLOS ONE*, 15(10):e0239666, October 2020.

[7] United States US Census bureau.
about 13.1 Percent Have a Master Professional Degree or Doctorate. https://www.census.gov/library/stories/2019/02/number-of-people-with-masters-and-phd-degrees-double-since-2000.html.

[8] Privacy by Design Foundation. Toevoegen van diploma attributen. https://privacybydesign.foundation/uitgifte-diploma/.

[9] Privacy by Design Foundation. Website of the irma system and foundation. https://privacybydesign.foundation/.

[10] Mohammad Delirrad and Ali Banagozar Mohammadi. New Methanol Poisoning Outbreaks in Iran Following COVID-19 Pandemic. *Alcohol and Alcoholism*, 55(4):347–348, June 2020.

[11] Statista Research Department. Twitter global mDAU 2021. https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/.

[12] Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, and Robert Matney. The Tactics & Tropes of the Internet Research Agency. Technical report, New Knowledge, 2019.

[13] Ullrich K. H. Ecker and Luke M. Antonio. Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, 49(4):631–644, May 2021.

[14] Marc Fisher, John Woodrow Cox, and Peter Hermann. Real consequences of fake news leveled on a D.C. pizzeria and other nearby restaurants. *The Washington Post*, page 13, 2016.

[15] Axel Gelfert. Fake News: A Definition. *Informal Logic*, 38:35, 2018.

[16] Lucas Graves. Understanding the Promise and Limits of Automated Fact-Checking. page 8, 2018.

[17] Ryan Grenoble. Here Are Some Of Those Fake News Stories That Mark Zuckerberg Isn't Worried About, November 2016. `https://www.huffpost.com/entry/facebook-fake-news-stories-zuckerberg_n_5829f34ee`

[18] Michael Hameleers. Separating truth from lies: comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society*, pages 1–17, May 2020.

[19] Michael Hameleers and Toni G. L. A. van der Meer. Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research*, 47(2):227–250, March 2020. Publisher: SAGE Publications Inc.

[20] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. The Quest to Automate Fact-Checking. page 5, 2015.

[21] Heather C. Hughes and Israel Waismel-Manor. The Macedonian Fake News Industry and the 2016 US Election. *PS: Political Science & Politics*, 54(1):19–23, January 2021.

[22] Office Of The Director Of National Intelligence. Assessing russian activities and intentions in recent us elections. Technical report, The Central Intelligence Agency, The Federal Bureau of Investigation and The National Security Agency, 2017.

[23] Amnesty International. Silenced and misinformed, freedom of expression in danger during covid-19, 2021. https://www.amnesty.org/en/wp-content/uploads/2021/10/POL3047512021ENGLISH.pdf.

[24] S. Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, 65(2):371–388, February 2021. Publisher: SAGE Publications Inc.

[25] Berkeley Lovelace Jr. FDA issues warnings on chloroquine and hydroxychloroquine after deaths and poisonings reported, April 2020. https://www.cnbc.com/2020/04/24/fda-issues-warnings-on-chloroquine-and-hydroxychloroquine-after-serious-poisoning-and-death-reported.html.

[26] Tanveer Khan and Antonis Michalas. Trust and Believe - Should We? Evaluating the Trustworthiness of Twitter Users. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1791–1800, December 2020. ISSN: 2324-9013.

[27] Jan Kirchner and Christian Reuter. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):140:1–140:27, October 2020.

[28] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, March 2018.

[29] Stephan Lewandowsky, Ullrich K. H. Ecker, and John Cook. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369, December 2017.

[30] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the*

Public Interest, 13(3):106–131, December 2012. Publisher: SAGE Publications Inc.

[31] Peiyao Li, Weiliang Zhao, Jian Yang, and Jia Wu. CoTrRank: Trust Ranking on Twitter. *IEEE Intelligent Systems*, 36(1):35–45, January 2021. Conference Name: IEEE Intelligent Systems.

[32] Paul Mena. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, 12(2):165–183, 2020.

[33] Alexi Mostrous, Basia Cummings, and Ella Hollowood. the infodemic fake news coronavirus, March 2020. https://www.tortoisemedia.com/2020/03/23/the-infodemic-fake-news-coronavirus/.

[34] Jack Nassetta and Kimberly Gross. State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School Misinformation Review*, October 2020.

[35] Erik C. Nisbet, Paul Beck, and Richard Gunther. Fake news did have a significant impact on the vote in the 2016 election. http://theconversation.com/trump-may-owe-his-2016-victory-to-fake-news-new-study-suggests-91538.

[36] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, April 2021.

[37] Oxford University Press. Oxford Word of the Year 2016 | Oxford Languages. https://languages.oup.com/word-of-the-year/2016/.

[38] Siti Qomariyah, Nur Iriawan, and Kartika Fithriasari. Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. page 020093, Surakarta, Indonesia, 2019.

[39] Md.Habidur Rahman, Tabia Prama, and Md Anwar. Modeling Topic Specific Credibility in Twitter Based on Structural and Attribute Properties. pages 580–589. April 2021.

[40] Srijith Ravikumar, Raju Balakrishnan, and Subbarao Kambhampati. Ranking tweets considering trust and relevance. In *Proceedings of the Ninth International Workshop on Information Integration on the Web - IIWeb '12*, pages 1–4, Scottsdale, Arizona, 2012. ACM Press.

[41] Vanessa Romo. Poison Control Centers Are Fielding A Surge Of Ivermectin Overdose Calls. *NPR*, September 2021. https://www.npr.org/sections/coronavirus-live-updates/2021/09/04/1034217306/ivermectin-overdose-exposure-cases-poison-control-centers.

[42] Mourjo Sen and Evgeny Morozov. Analysing the Twitter social graph: Whom can we trust? page 24, 2014.

[43] Scott Shane. The Fake Americans Russia Created to Influence the Election. *The New York Times*, September 2017. https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html.

[44] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes. *arXiv:2104.00779 [cs]*, April 2021. arXiv: 2104.00779.

[45] Naveen Sharma, Saptarshi Ghosh, Fabrício Benevenuto, Niloy Ganguly, and Krishna P. Gummadi. Inferring Who-is-Who in the Twitter Social Network. *ACM SIGCOMM Computer Communication Review*, 42, September 2012.

[46] Craig Silverman. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

[47] Asbjørn Steinskog, Jonas Therkelsen, and Björn Gambäck. Twitter Topic Modeling by Tweet Aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 77–86, Gothenburg, Sweden, May 2017. Association for Computational Linguistics.

[48] Stephanie. Likert Scale Definition and Examples, August 2015. https://www.statisticshowto.com/likert-scale-definition-and-examples/.

[49] Laura Sydell. We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. *NPR*, November 2016. https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs.

[50] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, 53(3):171–174, May 2020.

[51] Eva Lermer Timo K. Koch, Lena Frischlich. The effects of warning labels and social endorsement cues on credibility perceptions of and engagement intentions with fake news. *PsyArXiv*, November 2021.

[52] Analytics Toolkit. What is a Novelty Effect? | Glossary of online controlled experiments. https://www.analytics-toolkit.com/glossary/novelty-effect/.

[53] Marc Tuters, Emilija Jokubauskaitė, and Daniel Bach. Post-Truth Protest: How 4chan Cooked Up the Pizzagate Bullshit. *M/C Journal*, 21(3), August 2018. Number: 3.

[54] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE*, 13(9):e0203958, September 2018.

[55] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018.

[56] Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, 37(3):350–375, May 2020.

[57] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916, New York New York USA, August 2014. ACM.

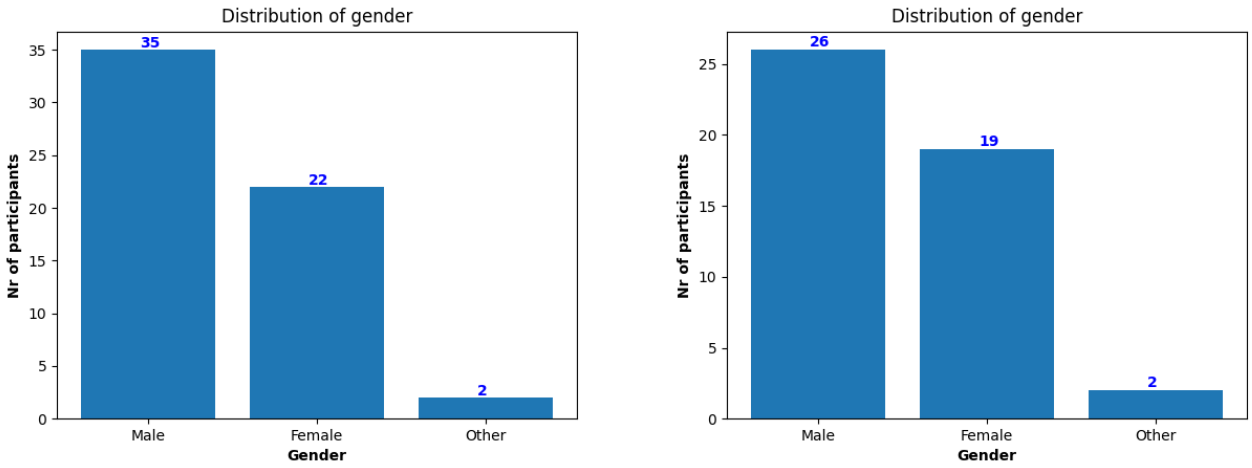
[58] Liang Zhao, Ting Hua, Chang-Tien Lu, and Ing-Ray Chen. A Topic-focused Trust Model for Twitter. *Computer Communications*, 14:38–1, October 2015.

[59] James Zou and Londa Schiebinger. AI can be sexist and racist — it's time to make it fair. *Nature*, 559(7714):324–326, July 2018. Bandiera_abtest: a Cg_type: Comment Number: 7714 Publisher: Nature Publishing Group Subject_term: Information technology, Society.

Appendix A

Appendix

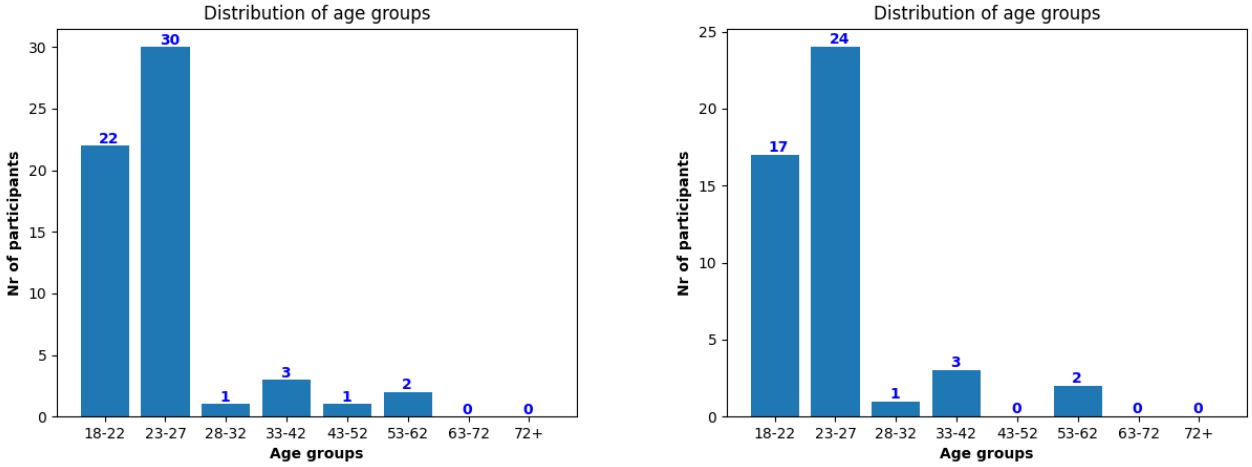
A.1 Appendix A - Bar graphs



(a) All participants

(b) Participants removed

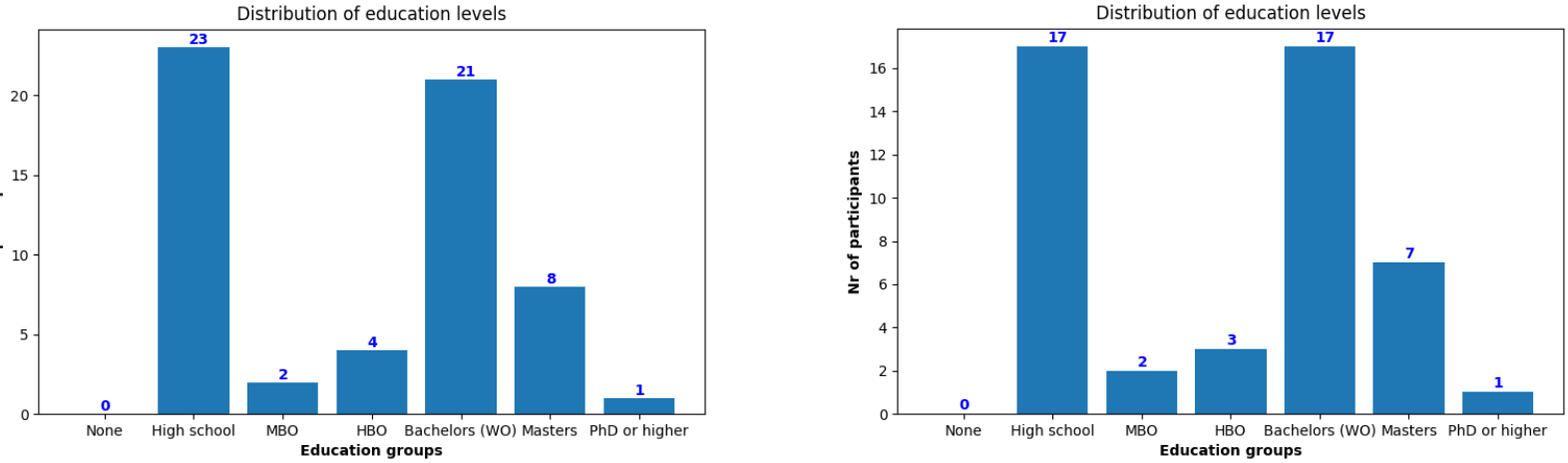
Figure A.1: Bar graph of gender distributions



(a) All participants

(b) Participants removed

Figure A.2: Bar graph of age distributions

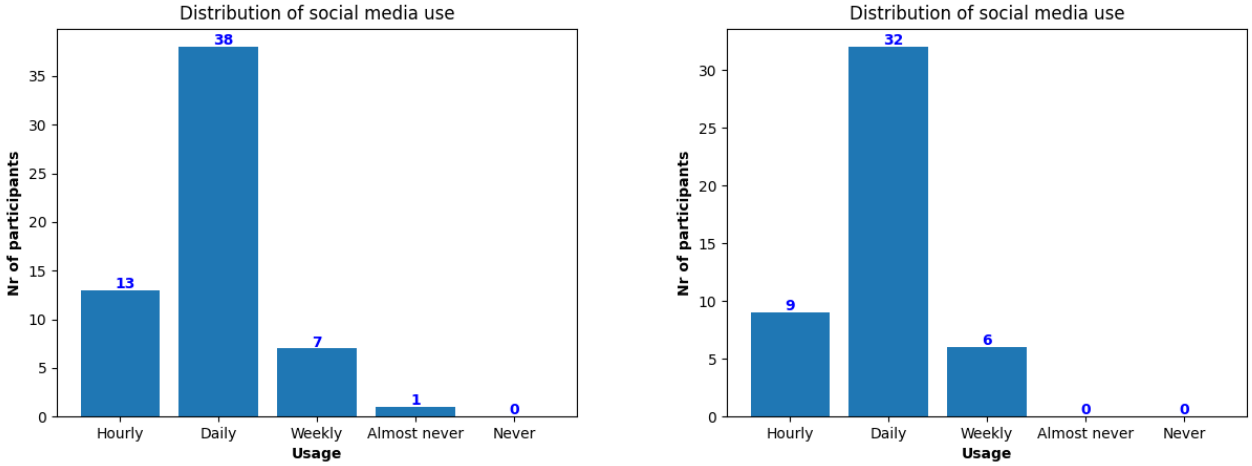


(a) All participants

(b) Participants removed

Figure A.3: Bar graph of education distributions

Social media use



(a) All participants

(b) Participants removed

Figure A.4: Bar graph of social media use distributions

A.2 Appendix B - Questionnaire

Bachelor thesis - Twitter research
By Jelte Smits
Introduction
Thank you for your interest in this study. Participation is voluntary. If you want to participate, we will ask you to give consent by clicking the "Agree" button on the next page. Please read the following information carefully. If something is not clear, or you would like more information, please contact me (jelte.smits@ru.nl) or my supervisor, Hanna Schraffenberger (hanna.schraffenberger@ru.nl). Please keep in mind that you have to be 18 years of age to participate in this study.

What to expect
Participation in my research will take a approximately 5 minutes. The research is about how you perceive posts on Twitter. First we will ask you a few demographic questions such as age, gender and education level. After this we will show you an image a tweet, followed by some questions about this tweet.

Voluntary participation
Your participation in this research is voluntary. You can decide to quit participating in the study at any time during the survey by simply closing the survey page. You do not have to give a reason for this. If you drop out before the end of the survey, we will delete your data. Your answers will

thus not be used it in the analysis.

What data is collected?
We will collect data on your personal background, such as age, gender, education, and social media use, as well as questionnaire answers. We will also collect timestamps (when a question group is answered) together with the answers. Since we do not ask for a name or any other identifying data item, your responses will be anonymous. We will publish research data and results in the bachelor thesis and share the data/results in presentations. None of the published data can be traced back to you.

What will happen to this data?
The data gathered during this study will be processed and stored anonymously. We store your consent (including timestamp) because we have a legal obligation to register your permission to participate in this study. Your consent information will be kept for 10 years upon completion of the research. Your anonymous research data will be stored for at least 10 years after the research has been completed.

Risks and discomfort
We do not expect any risks or discomfort.

More information?
If you have any questions about the study, please contact jelte.smits@ru.nl. This email can also be used for privacy questions specific to this study. You can also contact my supervisor, Hanna Schraffenberger at: hanna.schraffenberger@ru.nl. For general questions regarding privacy at the Radboud, please contact the office of the Data Protection Officer of Radboud University via privacy@ru.nl.

Consent

To be able to participate in the study we ask you to give your consent. Please read the following statements and indicate if you agree to it.

- I have been sufficiently informed about this research.
- I have read the information carefully.
- I have been granted the opportunity to ask questions about the research. If applicable: my questions have been answered satisfactorily.
- I have had the opportunity to carefully consider participation in this research.

- I have the right to stop participating in the research without giving a reason for this.
- I understand and agree to how the data of the research study will be stored and used.
- I participate in the study entirely on a voluntary basis.
- I am at least 18 years of age.
- I want to participate in the research.
- Agree
- I do not want to participate

General demographics

We will first ask a few demographic questions. These questions are ment to find out how my sample compares to a general population.
What is your gender?

- Male
- Female
- Other
- I do not wish to answer

What is your age?

- 18 - 22
- 23 - 27
- 28 - 32
- 33 - 42
- 43 - 52
- 53 - 62
- 63 - 72
- 73+

What is the highest degree or level of school you have completed?

- None

- Secondary education (VMBO, HAVO, VWO)
- Vocational education (MBO)
- Higher vocational education (HBO)
- Bachelors education (WO)
- Masters (WO)
- PhD or higher (WO)

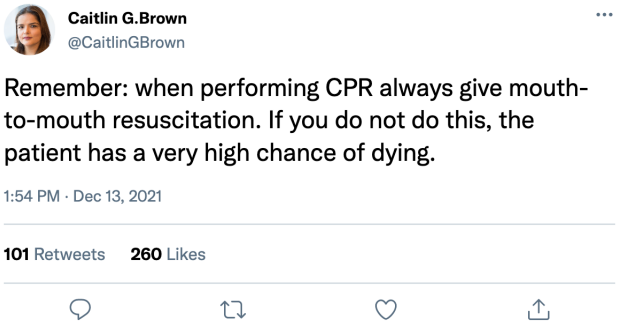
How often do you use social media?

- Several times an hour
- Daily
- A few times a week
- Almost never
- Never

Research

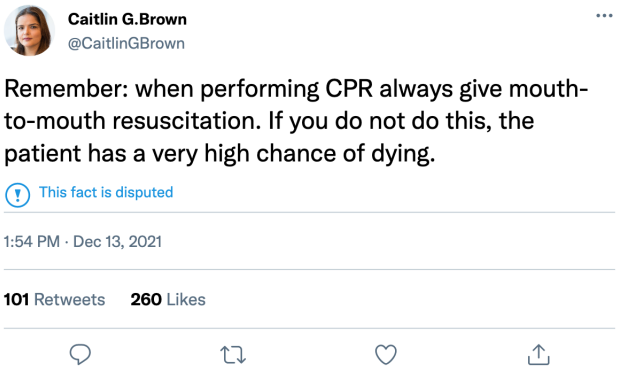
Now for the actual research! Please look carefully at the tweet(s) listed below and answer a few questions about it.

Only answer this question if the following conditions are met: (randnumber.NAOK == 1)

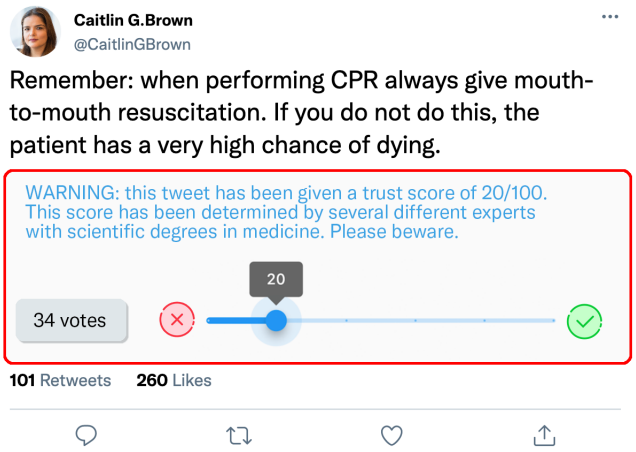


Only answer this question if the following conditions are met: (randnumber.NAOK == 2)

After clicking on the "View" button, you will see the following image:



Only answer this question if the following conditions are met: (randnumber.NAOK == 3)



Imagine you would user Twitter. How likely do you think ... * Please choose the appropriate response for each item:
You are to share this tweet?
Others are to share this tweet?
1 (Not at all) 2 3 4 5 6 7 (Very likely)

Could you please explain your motivation behind your answer on the sharing question?
To what extent do you find the content of the tweet...
Accurate, Authentic, Believable
Please choose the appropriate response for each item:
1 (Not at all) 2 3 4 5 6 7 (Very much)

Final questions

Please indicate what the tweet on the previous page was about

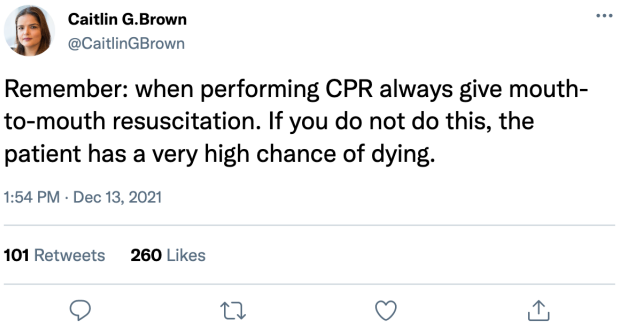
- Politics
- Corona
- Immigration
- Providing CPR
- I can't remember

Have you heard of this Twitter account before?

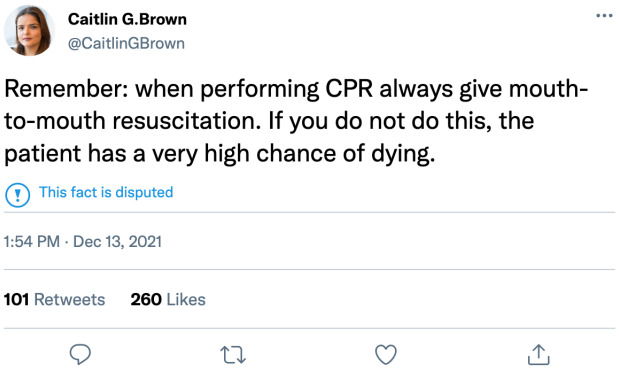
- Yes
- No
- I can't remember

Which of these 3 tweets did you see?

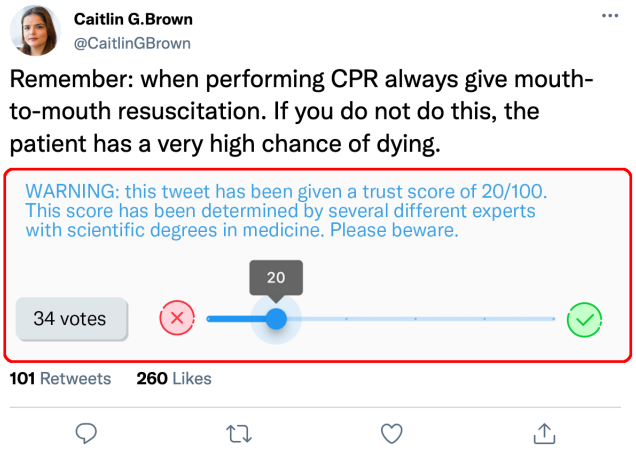
1:



2:



3:



- 1
- 2
- 3

Finally, would you like to give any remarks on this questionnaire or study?

Debriefing

Thank you so much for answering the questions! The study is about finding ways to counter misinformation. Currently social media platforms use some sort of warning message to warn if something is not correct. There has been much research on what works and what doesn't work. In my thesis I compared known methods with a new approach to combat misinformation. Some participants saw a normal tweet without any additional information.

Others particiants saw a tweet with a warning, and yet other participants saw a new method based on trust scores. The goal of this study is to find out how the envisioned trust scores would affect people's perception of tweets and how they compare to existing warning labels.

The tweet you saw in this experiment was not an actual tweet. It was designed by myself for the purpose of this study. The warning label was designed to mimic Twitter's existing warnings. The trust score was a mockup to investigate the potential effect of trust scores warnings. No medical experts have voted on the tweet's trustworthiness to establish the trust score.

If you would like to learn about the results about the outcome of this study, please send an email to: jeltesmits@ru.nl, with the subject "Thesis interest". This way I can filter all the emails and respond to those interested once my study is over.

If you would like to ask questions regarding the study to me, you can contact me on jelte.smits@ru.nl. If you would like to ask a question to my supervisor, you can contact her at H.Schraffenberger@cs.ru.nl.

If you do not want your data to be used after reading this debriefing, please answer the question down below with a "No". If you do not mind that your answers are used for this study, please click "Yes".

Do you agree to your answers being used in this study?

- Yes
- No

Once again, thank you for participating in this online experiment. If you could share this survey with all your friends and family, that would be greatly appreciated!

Link to share:
https://u1.survey.science.ru.nl/index.php/733662?lang=en