

BACHELOR THESIS
COMPUTING SCIENCE



RADBOD UNIVERSITY

Judging a Tweet's credibility

The effect of signature labels on perceived Tweet credibility

Author:

Marie-Sophie Simon
s1023848

First supervisor/assessor:

Dr. Hanna Schraffenberger
h.schraffenberger@ru.nl

Second assessor:

Dr. Eelco Herder
e.herder@cs.ru.nl

June 14, 2022

Abstract

Since social media became, popular the world has been getting more connected and new kinds of communities have developed. However, by now, the negative aspects have also become more apparent than ever. Whether it's cyberbullying, social media addiction or misinformation — social media's negative aspects can not be overlooked. Specifically, the problem of misinformation has recently been given a lot of attention and the societal impact has become visible. Both the 2016 US presidential election and the current COVID-19 outbreak have been heavily influenced by misinformation. In this research, we look into the Twid-project as one of the options that are currently being developed to address the problem of misinformation. The goal of the Twid-project is to help Twitter users judge the credibility of Tweets. With Twid, a user posting a Tweet would have the possibility to sign their Tweets with verified attributes. Signing a Tweet with a relevant label would give the user the chance to provide reliable background information about themselves. This research tries to answer two questions. The first question is whether people perceive Tweets with a relevant attribute-based verification label as more credible in terms of message and account credibility than regular Tweets. Our second question is whether an active indication of a missing signature reduces the perceived message and account credibility compared to both a signed and a normal Tweet. Our results are promising and show that a relevant attribute-based signature indeed does increase the credibility of a Tweet with true information. We did not find any evidence that the active indication of a missing signature decreases credibility compared to a Tweet without any label. This indicates that users understand the meaning of these signature labels. In an exploratory analysis, we further find an indication that credibility and sharing behaviour are related to each other. With these results, we show that the Twid-project has the potential to be part of a solution in the fight against misinformation. We suggest directions for future work to further investigate the potential and understand the effects that are at play.

Contents

1	Introduction	3
1.1	Misinformation on social media	4
1.2	Research	5
2	Preliminaries	7
2.1	Social Media	7
2.2	Misinformation	9
2.3	Authenticity	9
2.4	Integrity	10
2.5	Credibility	10
2.6	Attributes-Based Credentials (ABCs)	11
2.7	IRMA	12
2.8	Twid-Project	13
3	Related Work	14
3.1	Current countermeasures against misinformation	14
3.2	Credibility perception	15
3.2.1	Credibility perception on the web	16
3.2.2	Credibility perception on Twitter	19
3.3	Sharing Behaviour and Credibility	22
3.4	Absence of indicators	23
4	Method	24
4.1	Design of the Experiment	24
4.1.1	Conditions	24
4.1.2	Introduction of the labels	26
4.2	Apparatus and Material	26
4.2.1	The Tweet	27
4.2.2	General set-up and procedure	28
4.2.3	Measurement scales	29
4.3	Participants	30

5	Results	32
5.1	The effect of signature labels on credibility	32
5.1.1	Exploratory analysis: sharing behaviour	34
5.1.2	Expected plug-in use	35
5.1.3	Feedback about experiment	36
6	Discussion & Conclusions	37
6.1	Discussion	37
6.1.1	Relation to Previous Research	38
6.1.2	Limitations	39
6.1.3	Future research	41
6.1.4	Recent developments around Twitter	43
6.2	Conclusions	44
A	Demographics	51
B	Assumption Spearman's correlation	55
C	Questionnaire	57

Chapter 1

Introduction

In recent years, the popularity of social networks has been growing continuously and is expected to keep increasing in the near future (*Number of social network users worldwide from 2017 to 2025*, 2022). This has led to connections being strengthened and communities being formed leading to increases in life satisfaction and self-esteem of users (Tandoc Jr, Ferrucci, & Duffy, 2015; Valenzuela, Park, & Kee, 2009). However, social media has not only had positive impacts. Whether cyberbullying (O’Keeffe, Clarke-Pearson, on Communications, & Media, 2011), social media addiction (Hou, Xiong, Jiang, Song, & Wang, 2019), or misinformation (*Distribution of traffic sources for fake news in the United States in 2017*, 2019), the negative aspects of social media can not be overlooked. Specifically, the problem of misinformation has recently been given a lot of attention and the societal impact has become visible.

Misinformation has caused a lot of confusion and damage in the past years. Studies have shown that 64% of Americans think made-up news has caused them a great deal of confusion (Michael Barthel, 2016). According to Collins (2017), the Chair of the Culture, Media and Sport Committee of the UK, this increase in so-called fake news is “a threat to democracy and undermines the confidence in the media in general” (Collins, 2017, Chair’s comment section). One of the recent events where the spread of misinformation showed its impact, was the global COVID-19 pandemic. The impact here was of such significance that the World Health Organisation declared an ongoing *infodemic* next to the pandemic (Ghebreyesus, 2020). Due to misinformation claiming wrong treatments or diminishing the criticality of the virus, people risked their health and even lethal incidents occurred. There was, for example, the claim that drinking methanol¹ would cure the virus. This false claim led to at least 800 deaths, 5,876 hospitalizations and 60 cases of complete blindness (Islam et al., 2020).

¹Pure alcohol

Another critical example of misinformation playing a major role in society is the incident of the American elections in 2016. As a study done by the Ohio State University (Gunther, Beck, & Nisbet, 2018) shows, in this incident, fake news had a significant impact on the outcome of the United States election. As the study points out, different claims such as “Hillary Clinton is in very poor health due to a serious illness” (Gunther et al., 2018, p. 2) have been made and spread on social media which then may have lead to voters being swayed to vote for Donald Trump instead of Hillary Clinton. This demonstrates that the proliferation of false news can have a significant impact on our democracy.

The concept of fake news is not new (Burkhardt, 2017). As Burkhardt explains in his chapter “History of fake news”, rumours, gossip and misinformation have been spread by people for centuries. So why exactly is it, that the introduction of social media has increased the impact of misinformation by so much?

1.1 Misinformation on social media

Nowadays, in the age of social media, everyone can access all information and everyone can publish anything. Because of that, the reach of misinformation has increased enormously, meaning it spreads faster and further than ever (Vosoughi, Roy, & Aral, 2018). An additional factor in the spread of misinformation is the role that social media plays in acquiring news and information. A study done by the Pew research Center (Elisa Shearer, 2017) showed that in 2017, 67% of all Americans got at least some of their news from social media. This suggests that social media has a key role in the spread of misinformation. As a study from 2019 discovered (*Distribution of traffic sources for fake news in the United States in 2017*, 2019), social media platforms were responsible for spreading 42% of all false news in 2017.

The above-explained pressing issues around misinformation show, that there is a need for a practical solution to help users recognise misinformation. A realistic solution would need to maintain social media’s unique feature of being open to everyone while also being user-friendly enough to be used by a big enough number of people to have an impact. This is because, in the end, it is always the users of social media who decide whether they believe a piece of information to be true or not. So any practical solution, needs to be of help for the users to make a better and more accurate credibility judgment, than currently possible.

With this thesis, we are taking a first step on this journey, by starting

with an approach for the big social media platform Twitter. Twitter as a platform is very vulnerable to misinformation and already exploring options to fight this itself (Jacobs et al., 2021). By focusing on one platform, we have an easier start on developing and researching an approach. In case of promising results, this approach can then be extended more broadly on other platforms.

1.2 Research

The approach we will be researching is the Twid-project that is currently being developed at Radboud University, see *section 2.8 Twid-Project*. This project aims at using attribute-based credentials through IRMA, see *section 2.7 IRMA*, to allow users of Twitter to sign their Tweets with parts of their identity. With this, a doctor could sign a Tweet regarding a medical topic with the fact that they are a medical professional. In this way, the Twid-project hopes to provide verified background information to Tweet readers, allowing them to better judge a Tweet’s credibility than they currently can.

Until now, the Twid-project only exists as a prototype. This means we do not yet know whether it truly helps users judge credibility. This thesis is a first attempt to answer this question. More specifically, our research question is **whether people perceive Tweets with a relevant attribute-based verification badge as more credible in terms of message and account credibility than a normal Tweet**. We expect this to be the case since it gives the reader of a Tweet relevant and verified background information about the quality of the source, in this case, the account holder of the Tweet.

Aside from the effectiveness of a signature, another pressing open question is how to deal with unsigned Tweets. Thus, a second research question we address is **whether an active indication of a missing signature reduces the perceived message and account credibility compared to both the signed and the normal Tweet**. We expect it to be less credible than a normal Tweet because it might be perceived as too strong of a warning sign and decrease credibility more than necessary. The result of this question will be relevant for the future design of the Twid-project to optimally help users judge credibility.

Besides the effect of the labels on credibility, we are further interested in the way a user’s sharing behaviour is affected by the labels and the associated credibility judgment. This is because the project relies on the assumption, that a user would be more likely to share a Tweet if they perceive it as cred-

ible. To that end, we hope to find **whether the labels have an effect on a user's indicated sharing likelihood**.

The following thesis will be split into separate chapters. In *chapter 2 Preliminaries*, we will explore some topics that are critical to understand and define for the rest of this thesis. After that, in *chapter 3 Related Work*, we will put our research in the context of the current standard of knowledge. In *chapter 4 Method* we will explain the design and execution of our experiment and present the found results in *chapter 5 Results*. In the final part, *chapter 6 Discussion & Conclusions*, we will discuss the meaning of these results and advise on further research.

Chapter 2

Preliminaries

In this chapter, we will go over several concepts that are important to define for our research into judging a Tweet’s credibility. We will look at how these concepts are understood in other research and how we will use them in this research. These definitions will then be used throughout the rest of this thesis.

2.1 Social Media

Social media is not an easy concept to define. In trying to do so, there are two major issues that present themselves, according to a paper by Obar and Wildman (2015). The first challenge is the speed at which technology evolves. Because social media is an ever-changing idea, it is difficult to identify precise boundaries. Secondly, social media facilitates communication similarly to other technologies, such as the telephone or email. However, the term “social media” was coined through technologies that were developed later.

According to Obar and Wildman (2015), there are, however, four commonalities among social media services. The first one of these commonalities is the fact, that social media services are Web 2.0 internet-based applications. The development of Web 2.0 made the internet a more interactive space, which was a big milestone in the direction of the creation of social media. According to Obar and Wildman, the second similarity is that user-generated content is what makes social media what it is. Thirdly, the user-specific profiles are mentioned as a commonality. These profiles allow for connections to be made which leads to the fourth commonality, the development of social networks.

With all these commonalities combined, we have created a description of what social media is. In this thesis, we will be focusing on the social media

platform Twitter. Twitter is a so-called micro-blogging platform (Morris, Counts, Roseway, Hoff, & Schwarz, 2012). This means that registered users post text, optionally accompanied by images and videos, called “Tweets”, which are limited in size. In Twitter’s case, the length of the text is limited to 280 characters per Tweet (*How to Tweet*, n.d.). Besides posting yourself, a user can also read the posts of others. By “following” others, users create connections and personalise the content they view. Next to just the content of a Tweet, there is more information displayed with it. A user can always see a display name, the user name, a profile picture, the time and date of the posting, from what device (e.g. Twitter for iPhone, Twitter for Android, Twitter Web Client etc.) it was posted, the number of retweets¹, the number of quote Tweets² and the number of likes. Sometimes there is also a verification badge next to the name of a user, this means that an account of public interest is authentic (*About Verified Accounts*, n.d.) (see *section 2.3 Authenticity*). To receive the authentication badge, the account owner has to apply for it after which Twitter will check whether they meet the requirements of being authentic, notable and active. To check for authenticity, Twitter expects the account owner to prove their identity in a certain way, depending on what kind of account it is. In the case of an individual applying, an ID verification is required. To detect whether the account is notable, Twitter can choose to look at different criteria, such as a Wikipedia page about the account owner or being in the top .05% of follower or mention counts of their geographic location. The requirement of an active account includes a complete profile, active use, some security requirements and adherence to the rules of Twitter. There are further details about the application procedure, depending on the kind of account that applies. Important to note is, that the decision whether or not to give and also to keep the status of an authenticated account lies solely with Twitter.



Figure 2.1: An example Tweet, the first one ever posted

¹A retweet means it got re-posted by another user (*How to Retweet*, n.d.)

²A quote Tweet is like a retweet but you can add your own comment (*How to Retweet*, n.d.)

2.2 Misinformation

Besides the term misinformation, we can also find the terms disinformation and fake news. Often these are used interchangeably, although they are not exactly synonyms. Most commonly used are the definitions that misinformation is any wrong information (Caramancion, 2020; Guess & Lyons, 2020; Hilary & Dumebi, 2021). Disinformation, on the other hand, is a subset of this, that is, wrong information that is spread with the intent to deceive. So all disinformation is misinformation but not all misinformation is disinformation. Similarly, something can only be labelled as disinformation once it has been proven to be published with the intent to deceive. Fake news is the synonym for disinformation, however, through the mass use in media its connotation changed and can now also be perceived as a political tool to use against your opposition (Lazer et al., 2018).

As the phenomenon of misinformation has only gotten broad attention in recent years, the terms misinformation, disinformation and fake news are not always clearly defined and separated from each other. This is also clearly visible when looking up the definitions in the Cambridge dictionary. Here, misinformation is defined as “wrong information, or the fact that people are misinformed” and “information intended to deceive” (*Cambridge Dictionary Entry: misinformation*, n.d.). Disinformation, on the other hand, is only defined as “false information spread in order to deceive people” (*Cambridge Dictionary Entry: disinformation*, n.d.). For this paper we will be following the definition mentioned above, so misinformation is any false information, while disinformation assumes a malicious intent. In the used sources the difference between misinformation, disinformation and fake news is not always clearly defined. In that case, we will be referring to it in the way the cited source does.

2.3 Authenticity

Authenticity is defined as “a way to ensure that communication processes data, systems, or information are genuine” (de Oliveira Albuquerque, Villalba, Orozco, de Sousa Júnior, & Kim, 2016, p. 3743). To provide authenticity, one has to authenticate themselves. Through this process, a party proves that they are really who they claim to be by providing evidence. This evidence can be something you have (e.g. fingerprint), some information about yourself (e.g. name of the first pet), something you know about (e.g. password), or a combination of these (Kim & Hong, 2011). In this thesis, authenticity labels are added to Tweets. These authenticity labels provide viewers of Tweets with certainty about who (e.g. name) or what

(e.g. a doctor) the source of a message is. This is realized with the IRMA attributes, which the person posting needs to reveal when posting their Tweet.

2.4 Integrity

In the technical context of information security, integrity is defined as “the ability to guarantee accuracy and consistency of data and information during its entire life cycle” (de Oliveira Albuquerque et al., 2016, p. 3743). This means that the concept of integrity assures something to be in no way modified or damaged by an unauthorized party (de Oliveira Albuquerque et al., 2016). In the context of this thesis, that means, integrity guarantees that the Tweet a user is looking at, has not been modified by anyone, other than the original user tweeting it.

2.5 Credibility

In everyday language, credibility is understood as “the fact that someone can be believed or trusted” (*Cambridge Dictionary Entry: credibility*, n.d.). If we refer to it from a scientific standpoint, where we also want to measure it, this already gets more complicated. As discussed in the book “An integrated approach to communication theory and research” by Stacks and Salwen (2014) the concept of credibility is said to go back to the first discussions of rhetoric by the ancient Greek philosophers. It was generally understood as the idea that some sources are more reliable to speak the truth than others. The difficulty with the concept of credibility was determining where to place it and how to define it.

Nowadays, there is an extensive amount of literature on the concept of credibility, leading to ambiguity due to the differences in definition. With the rise of social media, where misinformation spreads faster than ever and reach more people than before (Vosoughi et al., 2018), the concept of credibility only got more complex. As Stacks and Salwen (2014) found, researchers have defined credibility as a multitude of concepts and studied it from different points of view. One thing most researchers agree on is that there is no one concept of credibility, but it relies on different pillars and aspects. However, as explained in this paper, most definitions differentiate between something similar to “message”, “source” and “channel” credibility. Furthermore, one thing that was found to consistently hold, is that, ultimately, credibility is always perceived by the message recipient. This is what makes credibility so interesting for this research. Because it is ultimately decided by the message recipient, they must have relevant and reliable information, to make a good credibility judgment. This is what the Twid-project tries

to achieve.

For this thesis, we distinguish only between message credibility and source credibility since our only medium of communication will be Twitter, and we will not be needing to differentiate between media.

For message credibility, we will be using the definition of Appelman and Sundar (2016) as their scale allows us to easily measure credibility. This scale is also used in other ongoing research at Radboud University (Schraffenberger, 2021-2022) so using the scale will allow for easier comparison of results in the future. In their paper, Appelman and Sundar (2016) refer to message credibility as “an individual’s judgement of the veracity of the content of communication” (Appelman & Sundar, 2016, p. 63). Because the judgment is only about the message, it is also said to only depend on “aspects of the message itself” (Appelman & Sundar, 2016, p. 63).

For source credibility, we apply the definition and scale of Metzger, Hartsell, and Flanagin (2020), for an easy assessment of source credibility. Likewise, it facilitates a future comparison with ongoing research at Radboud University (Schraffenberger, 2021-2022). In this paper, source credibility refers to how biased, professional and trustworthy a user finds a source to be. Because the only possible source in the context of Twitter are different accounts, we also refer to this as “account credibility”.

The concepts of authenticity and integrity are closely related to credibility. To judge credibility, it is important to know that the account holder is really who or what they say they are (authenticity) and that the post has not been modified (integrity). If, for example, someone were to post something and claim they were a doctor, while they are not, that could have fatal consequences. Similarly, if a doctor were to post something, but a hacker were to modify the message, this could have also fatal consequences. To this end, Twid provides as well authenticity as also integrity to give users a chance to hopefully make better judgements about credibility.

2.6 Attributes-Based Credentials (ABCs)

An attribute-based credential is a container that may contain multiple attributes of a user (Alpár & Jacobs, 2013). These attributes can be shown independently of each other so a user can authenticate themselves by only showing the necessary attributes while hiding others to maximise privacy. As Alpár and Jacobs (2013) state in their paper, attributes describe any property of a person. This can be things such as *My name is...*, *My nationality is...*, *My email address is...* or *My social security number is...*

An important differentiation they make here is between anonymous, so non-identifying, attributes (such as gender) and identifying attributes (such as bank account).

2.7 IRMA

IRMA is a mobile application that is similar to a passport on users' phones. With IRMA a user can collect attribute-based credentials and then authenticate themselves or certain attributes about themselves in a secure manner.

IRMA stands for "I reveal my attributes" (*IRMA in detail*, n.d.) and is a way of empowering users to disclose certain relevant attributes online while keeping irrelevant attributes private. Using IRMA to disclose these ABCs, a user can choose which attributes to use to authenticate themselves. This way they are giving away minimal information, protecting their privacy (Alpár et al., 2017). An example application would be a website that is only accessible to people above a certain age (like a chat room). In that case, you want to have the attribute *age*, revealed, but other than that you want to stay anonymous. So via these ABCs, a user can get access to things that require authentication while still protecting their privacy.

To receive attributes, a user needs to request them from an (attribute) *issuer*. This could be the official municipality for age, name and nationality, or the Privacy by Design Foundation for attributes like a phone number (*Issuance of mobile phone number attributes*, n.d.). These attributes are cryptographically signed by the issuer. The service provider that the user wants to get access to will then ask the user to disclose the relevant attributes. If the user does so, the service provider can do the cryptographic check and thus verify the attributes. Hence, the service provider is also called the *verifier*. Through this cryptographic check, they can see that the attribute is legitimate and has not been expired, manipulated, has been issued by the right issuer and indeed belongs to the right person (for more information see (*IRMA in detail*, n.d.)).

Besides attribute-based authentication, IRMA can also be used for attribute-based digital signatures (*IRMA in detail*, n.d.). A digital signature is an addition to, for example, a digital document that can only be created with the personal cryptographic key of the person signing it. These so-called private keys are bound to one individual via a certificate that contains the associated public key, with which the signature can be verified.

In an attribute-based signature, the attached signature contains at least one, possibly more, attributes of the signer. Everyone checking the signa-

ture can see those attributes. That way it can be verified that the signature was indeed from a certain person or that the signing person had a certain attribute, like being a medical doctor.

2.8 Twid-Project

The Twid-project is a proposed browser plugin for Twitter to provide certainty about who the source of a Tweet is (“authenticity”) and that this Tweet has not been modified since its upload (“integrity”) (Jacobs et al., 2021). The idea is to sign Tweets using IRMA in the above-explained way. By using IRMA, the attributes are already verified and a poster will be able to sign a Tweet using a number of these. This provides the reader of Tweets with valuable and verified background information about the poster. This will hopefully make it easier for the readers of Tweets to judge whether the information is credible and authentic.

For instance, through the Twid-project, users will have the possibility to sign Tweets using the domain address of their email, without revealing their real name. Users can also choose to sign with their real name via IRMA, which makes authentication accessible for everyone instead of only for accounts that have been verified by Twitter. Possible other attributes to use in a signature are a doctor’s license, city or country of residence or a teacher’s license. For example, in a post about COVID-19, the information that someone has a doctor’s license might give more credibility to their post. A user decides for every Tweet separately whether they attach a signature to a Tweet and also which attributes to include. This requires them to sign every Tweet separately which makes it possible to also provide integrity, alongside authenticity.

Chapter 3

Related Work

In this chapter, we will investigate how our thesis fits into the current state of research and what are important aspects to note for our study. Here, we look into current countermeasures against misinformation, to understand why those are not yet sufficient and what Twid would need to add. We furthermore look into the way users perceive credibility and the aspects they take into account there, first generally in an online environment, then specifically in the context of Twitter. Finally, we put the sharing behaviour of users into the context of credibility, because sharing is ultimately what spreads misinformation.

3.1 Current countermeasures against misinformation

The most common intervention to prevent the spread of misinformation is fact-checking. In this, a published piece of news gets checked to determine whether it is true or not. Then, if found incorrect, it either gets labelled as false or a reference to a source correcting the misinformation is placed next to it. This can either be done by individuals or organisations, although there is also research being done about automating fact-checking. However, developing algorithms to automate this process has not been successful yet and will likely also not get there in the foreseeable future, as Graves (2018) explains in his paper. As explained here, automated fact-checking tools can only be used to help human fact-checkers respond more quickly and effectively. One of the main problems found to automate this further is that algorithms cannot yet understand a claim well enough to fact-check it automatically. For this, the natural language processing in artificial intelligence needs to progress further. Additionally, it is found that the available data to check claims is not sufficient yet, especially in a way that it can be understood by software. Therefore, as stated by Graves (2018), algorithms are so far mainly used to flag claims that need to be fact-checked or discover

resurfaced claims that have been fact-checked by humans earlier.

Even if the process of fact-checking would be assumed to work perfectly, research suggests that it would still not be the ideal solution. As stated in an article published in the Science journal (Lazer et al., 2018), research has questioned the efficacy of fact-checking. As explained there, this is partly due to the reason that people tend to not question the credibility of information and accept it as true unless it violates their own beliefs. Similarly, the article also states, that people perceive information confirming their beliefs as more persuasive (confirmation bias) which makes it difficult to change peoples' opinions even when correcting misinformation. According to that article, fact-checking can even be counterproductive in certain settings. This is explained by the fact that people tend to remember information and how they feel about it, but forget the context in which they encounter it. That means, that repeating false information in a fact-checking context (i.e. labelled as false or corrected in an article) can lead to people only remembering the information, however, not the context of it being corrected.

So as can be seen, the process of fact-checking has its downsides and will probably never be effective enough to significantly reduce the amount of misinformation on social media. For that reason, science looks into additional solutions to prevent the spread of misinformation in a way that does not limit the possibilities of social media.

Any solution that will have a meaningful impact on the problem of misinformation must always keep the user in mind. The user must feel comfortable in using the solution and must be able to understand it easily. Only then will it be applicable to the broad audience and have a meaningful effect in the fight against misinformation. Further, it would be of advantage to keep the nature of social media intact, meaning everyone should be able to post and see everything. Taking away that freedom would not only change the essence of social media but also limit the freedom of speech, which is a basic human right as declared by the UN in article 19 (*Universal Declaration of Human Rights*, n.d.).

3.2 Credibility perception

Due to the before-mentioned rise of social media and the increase of misinformation, research in the field of misinformation and credibility judgments has increased. In *section 2.5 Credibility*, we have already touched upon the diversity of definitions of credibility and that almost every research has defined and measured it differently. Since we are mostly interested in the factors that influence a user's credibility judgement, we will only mention the

specific definition if relevant to us. We present related work on this matter in this section. Hereby we will first focus on the general factors that influence credibility judgments on the web and then focus on Twitter specifically.

3.2.1 Credibility perception on the web

It used to be the case that the recipient of a message could base their credibility judgment on the trust of the source through gatekeepers like, for example, newspapers (Jacobs et al., 2021). This is not the case anymore on the internet, where everyone can publish anything, without the need for review by any objective party. This means, that users needed to find new ways to judge the credibility of information on the web. In the following, we will examine the two main aspects that users can still rely on. These are the visual features or first impressions before any extensive cognitive processing has happened, and content-based features that the user will actively use to make a credibility judgment. Although we will exclusively focus on Twitter in our research, by first looking at credibility judgements from a broader perspective, we aim to discover elements that may also be significant for Twid.

Visual features

The first aspect a user can judge about a website, without needing to read the content, is the design and its visual features. This, of course, also plays a role in the credibility judgment a user has. Fogg et al. (2003) did a study with 2500 participants giving open-ended comments about website credibility. This study revealed that with just over 46% “Design Look” was the most often mentioned site feature a user considered. The second highest-rated feature was “Information Design/Structure” with nearly 29%. Only in positions three and four do people look beyond the surface elements of a website. That is, in place three users indicate to look at “Information Focus” (just over 25%) and in place four at the “Company Motive” (over 15%). So what visual features exactly are it, that make a website more or less credible?

In research conducted by Robins and Holmes (2008), it was discovered that when participants were asked to estimate a website’s trustworthiness based on the initial impression, they made a choice in an average of 2-4 seconds. As the authors of the study state, in this period “it is unlikely that any type of complex cognitive analysis and evaluation” (Robins & Holmes, 2008, p. 398) could have been possible, which means it was just decided by the design and looks of a website. The results of this study showed, that in almost all cases a website with a high aesthetic treatment was given a

higher credibility rating than a website with a low aesthetic treatment.

The MAIN-model described by Sundar (2008) provides a heuristic in which some of these surface-level characteristics are a determining factor for the youth to judge a website’s credibility¹. The important cues used in this heuristic relate to modality, agency, interactivity or navigability. Modality describes the structural features of content, so the way the content is presented. This can be, for example, via text, audio, images or videos. Agency cues relate to the source of information. These sources are not always clear on the internet, so the agency features can give at least some indication, like a website name or a friend who sent that message. As for interactivity cues, Sundar describes the interaction between user and source by, for example, dialogue boxes that ask for input. The navigability cues include features that bring you from one site to another by clicking e.g. hyperlinks. Sundar also mentions, that some of these cues are a “double-edged sword”, where, for example, more interactivity does not always have a positive correlation to the perceived credibility, but this depends very much on the context. In the next section, *section 3.2.1 Source features*, we will go into more detail about the agency cue relating to the source of something instead of simply the superficial design of a website.

Since this research focuses on Twitter, a website that has already been designed, we do not need to take all visual features into account for Twid. However, an important conclusion for us is that users have a first credibility judgement in a matter of seconds. So when designing our Twid label, we should make sure that it is easily visible and does not require a lot of cognitive processing.

Source features

As we have already seen in the study by Fogg et al. (2003), the fourth most-mentioned comment about credibility was “Company motive”. To consider a company’s motive, a user needs to know the company that provides the seen content, so the source. To provide a first indication of how the user judges the credibility of a source, we will look at the agency cues provided by Sundar (2008) in the already mentioned MAIN model. As described by Sundar, agency cues provide the user with information about the identity of the source. With agency is meant here, where the information came from, often “the agent itself is the source, at least psychologically” (Sundar, 2008, p. 83). As an example, Sundar mentions that in this way, it is, possible to name a “bot based news aggregator such as Google News” (Sundar, 2008,

¹Since the youth is one of the target groups of social media, we consider this model relevant to our research, although we do not know whether it holds for all age groups.

p. 83) as a source.

According to Sundar (2008), for agency cues, multiple heuristics play a role. The *bandwagon heuristic*, *authority heuristic*, *identity heuristic* and *machine heuristic* will be further explained here since these are most relevant for our experiment using the Twid-project.

The bandwagon heuristic implies that if other users select news stories to show to one user, then it is also rated as higher quality and more newsworthy than if it was selected by a news editor or by the user themselves (Sundar, 2008). As a consequence for Twitter, something might be perceived as more credible, if it has been shared before, liked or maybe even if the author has more followers, indicating they share stories that are worth following. Because the MAIN model does not involve Twitter in its research (to our knowledge), we do not know what consequences this will have for our research, but we will need to keep it in mind.

The authority heuristic is defined by Sundar (2008) as the notion that an official authority as a source is perceived as more reliable than other sources. This plays a role in the Twid-project because by signing a Tweet as, for example, a medical professional or even an official municipality, the signature would provide authority to a user and thus make their posts more credible.

The identity heuristic as part of Sundar (2008) MAIN-model is said to increase credibility, whenever a source is able to express part of their identity. Again, this is an important part of the Twid-project because the signatures can prove part of your identity, meaning you can express your true identity more than before. This is thus also expected to have consequences on our experiment.

The last relevant heuristic we will be looking at is the machine heuristic (Sundar, 2008). It says, that stories said to be chosen by a machine instead of an editor, were perceived as more credible by users. This is explained by the fact that a machine is thought to be more objective and thus closer to the truth. It is unclear how exactly this heuristic has an influence on Twitter since it is an algorithm choosing Tweets to show, however, they are often from people a user follows. While we do not know which effect it will have on our experiment, it is important to keep this heuristic in mind when evaluating our results.

In another study by Flanagin and Metzger (2007), they found that there is a difference in credibility judgment depending on the genre of a website. In this study, they distinguished credibility in three different aspects, namely

site credibility, *sponsor credibility* and *message credibility*. For sponsor and message credibility they found that news organization sites were perceived as the most credible. The special interest sites and electronic commerce sites were not significantly different from each other and the personal sites were perceived as the least credible. For site credibility the e-commerce site was rated equally credible as the news site, then the special interest site followed by the personal site. Besides the rating of sponsor credibility, there was no significant difference between actual and fictitious websites in these results. These results show that the perceived credibility judgment is affected by the genre of the source, even if the actual source is not known and thus no judgment can be made by earlier experiences with the source. While this refers to websites and not to accounts on Twitter, we do need to keep in mind that users can deduce some information from the user name of an account. In all the above aspects of credibility, personal sites are perceived as the least credible. This would mean the Twid labels are especially interesting for personal Twitter accounts so that these users can increase their credibility by supporting their posts with professional attributes.

The same study (Flanagin & Metzger, 2007) also looked into the correlation between self-reported information verification and observed information verification. Here it found, that individuals reporting that they verify information found online, actually did so less than the ones who did not report this. The study states, that these findings imply that there is a group of internet users who know they need to be sceptical of information found online and should thus verify it but do not make the effort to do so. This is an important result for our research. It indicates that giving the users information to help make a judgement about the credibility in a way that does not require them to actively find information, might be a step in the right direction.

3.2.2 Credibility perception on Twitter

For our research project, we have focused on Twitter as a social media service since the Twid-project is currently only being developed for Twitter. By focusing on this one service, we can go into more detail about which factors play a role in users' credibility perception. As opposed to general web pages, Twitter is limited in the capabilities of which the author of a message can customise certain things. In that way design will, for example, play a smaller role in the credibility perception since all Tweets look the same except for a few details, like profile picture or user name. Similarly, the way Twitter is being used is different from a web page. As such, it is social media, which means there are certain social behaviours such as sharing and following that we need to consider in this section. We will review how already existing literature describes the effects of different aspects of Twitter

on the user’s credibility perception and from there also draw conclusions to the design of our own research.

A study by Morris et al. (2012) has explored which aspects are relevant to a user’s credibility judgments on Twitter and which are not. In a pilot study, they first identified which features possibly influence a user’s credibility judgment. Examples, as mentioned by the participants, were the nature of the chosen profile picture and user names. They also discovered that some aspects, such as an author’s bio, were not investigated until prompted. Once they were explored, however, they were found to be useful. This is relevant to our study because our Twid labels could show some of these attributes at first sight. That way, a user would not need to investigate a bio to see more background information about the account holder. An additional advantage of Twid is that the information is also verified, while anyone can write anything in their bio.

Morris et al. (2012) continued with multiple follow-up studies, where they explored which factors are paid attention to and how some of the factors affect the credibility of a Tweet or the author of a Tweet. An important aspect that does play a role in the credibility of a Tweet is its content. An example that is found in this study, is that the use of non-standard grammar and spelling (e.g. abbreviations used in text messaging) decreases the credibility of a Tweet more than any other factor. Other examples of content-related aspects are the topic of a Tweet, where for example science topics are found to be more credible than those about politics. They further found that if the Tweet contained a URL leading to a high-quality site this would also increase its credibility. In a paper by Suh, Hong, Pirolli, and Chi (2010) they measured the likelihood of retweeting instead of credibility. They found that a Tweet containing a URL is more likely to be retweeted than a Tweet without one. This leads to the question of how sharing and credibility are related to each other, which we will further investigate in section *section 3.3 Sharing Behaviour and Credibility*.

Other aspects found to be relevant in Morris et al. (2012) study, are the account name and the profile picture. Here, they found that not replacing the default image of Twitter decreases the credibility of an author. Furthermore, they find that a user name related to the topic of the Tweet, is perceived to be more credible than a traditional user name or a user name in internet style (e.g. “Pickles_92”, “25th_Hour” (Morris et al., 2012, p. 446)). Since personal users of Twitter often simply use their personal user name which is not topic related, this shows again, that Twid is specifically interesting for personal users. Additionally, while some users might add a title (e.g. M.D., Dr.) to their username, the Twid labels are also a way of verifying this addition.

In Morris et al. (2012) research, the reputation of an author is said to have a positive effect on credibility. Reputation is measured by whether a user knows the account and the blue verification badge (since it indicates the account is of public interest), which Twitter already provides (see *section 2.1 Social Media*). In two separate studies, one by Vaidya, Votipka, Mazurek, and Sherr (2019) and one by Edgerly and Vraga (2019) it was found, however, that while more than half the users remember with certainty whether or not they have seen a badge, it seems to have little to no effect on the perceived credibility of a Tweet. It is important to note here, that the verification badge of Twitter only stands for the authenticity of a source and is thus not meant to have a direct effect on credibility. However, as described in *section 2.5 Credibility*, the concepts of credibility and authenticity are closely related to each other, which makes this a relevant result for us to discuss. As Vaidya et al. (2019) also state in their paper, these results indicate that users have a better understanding of the difference between authenticity and credibility than expected. Due to these findings, it is important to research the effect of the signature labels that, besides authenticity and integrity, provide additional background information to the readers.

Additionally, Morris et al. (2012) mention influence as a feature to enhance a Tweet’s credibility. They measure influence by follower, retweet and mention counts. While Morris et al. (2012) suggest that higher counts lead to more influence and thus increased credibility perceptions, a study by Westerman, Spence, and Van Der Heide (2012) found more nuanced results. In their study, they begin with the hypothesis that the feature of followers and follows, is “system-generated”, so produced by a system or machine. This leads them to relate these cues to the machine heuristic introduced by Sundar (2008) (see *section 3.2.1 Source features*) which would mean that, since the cues are created by a machine, they are perceived to be objectively true and thus higher numbers of followers would lead to higher perceived credibility. However, the results of the study show that this does not hold, but that there is a curve-linear pattern between credibility and followers. This means, that if a user has too few followers, they are not perceived as credible, however, having too many followers also makes them seem less credible. A possible explanation provided by this paper is that these “follower collectors” are viewed as spending more time on collecting followers, rather than providing useful content. Besides only researching the number of followers a source has, this paper also looks into the ratio between followers and follows. It finds, that a narrow gap, so having a similar amount of followers to follows, increased the perceived credibility. The fact that this ratio exists, suggests, according to Westerman et al. (2012), that in social media, users also expect certain social behaviour. In that way,

having a lot of followers, but not following a lot of people yourself, is not perceived as social and thus leads to users having a lower credibility perception of a source. Due to these findings of followers and follows having an effect on credibility, it is important for our experiment to use the same numbers among the different Tweets we will be looking at. That way we can be certain that any effects we will see can be accounted to the signature labels instead of the number of followers or follows.

3.3 Sharing Behaviour and Credibility

Social media has made it possible that everyone can access all information at all times while at the same time, also everyone can publish anything. This has increased the reach of information incredibly and with it the reach of misinformation. A large study done by a research team from the MIT (Vosoughi et al., 2018)² has shown that on Twitter, misinformation and information do not spread at the same pace, but false information spreads significantly faster than true news. In this study, they analysed data from about 126.000 stories that have been tweeted by about 3 million people and about 4.5 million times. The conclusion from this study was, that not only does false news spread about six times as fast as true news, but also that it reaches far more people. Furthermore, they found that this was the case, even though users spreading misinformation have significantly fewer followers, were less active, less often verified and spent significantly less time on Twitter. This also did not change when taking bots out of the analysis. According to their study, bots do accelerate the spread of news significantly, however, this is equally the case for false and true news. The only reason this study could suggest why false news spread so much faster is the novelty of false news and the human behaviour to spread novel news. However, the study also states that further research into these behavioural explanations is needed.

Knowing that false information spreads much faster on social media than true news, there has been plenty of research into the reasons and motivations for users to spread misinformation. These motivations do not always seem to be ignorance, but it shows that users are (at least sometimes) aware of sharing misinformation. One research (Chen & Sin, 2013), for example, shows, that even though truthfulness was valued by all participants, more than two-thirds could indicate to have shared misinformation at some point. The research goes further into the different motivations certain types of personality traits seem to have to share misinformation. Here they find that, for example, extroverted personality types are more likely to spread misinformation to socialise. Results of other studies have even shown that people

²Disclaimer: This study has been funded by Twitter

also share misinformation with bad intentions to disrupt and polarize (Osmundsen, Bor, Vahlstrup, Bechmann, & Petersen, 2021). These results are not very promising for Twid and we will look into this more in *chapter 6 Discussion & Conclusions* to discuss what kind of impact this can have on the Twid-project.

While there has been a lot of research into the motivation of sharing misinformation, and the motivation of sharing information on social media in general, the actual connection between credibility and sharing behaviour remains unclear. Some researchers (Vaidya et al., 2019) use sharing likelihood as an implicit measure of credibility, however, this connection cannot be found back in some other research. This is why, in this thesis, we decided to not use sharing behaviour as an indicator of credibility, but use it as a separate measure to explore.

3.4 Absence of indicators

Our research also aims to answer the question of how to deal with a missing Twid signature. For this, it is important to look at how users perceive missing indicators. Research has shown, “that users do not generally tend to notice the *absence* of an indicator” (Sobey, Van Oorschot, & Patrick, 2009, p. 6). This has been particularly researched in the context of web security, where one study even showed that all 63 participants ignored the absence of the *https* prefix and thus failed to recognise spoofing (Shi, Xu, & Zhang, 2011). To avoid that from happening in the Twid-project, a red label has been designed by the UX designer to signal the absence of a signature. However, there is also a chance that a red label is too strong of a signal and makes every Tweet unbelievable, which would not be accurate. This is because the red label only indicates, that a user who uses Twid and usually signs Tweets, chose to not do so this time. This indicates that a reader cannot be sure whether it was the account holder that posted the Tweet or a hacker, or whether it had been modified after posting it. Even in the case that the account was not hacked, a red label might indicate that the account holder does not have expertise in the concerning area. It is important to note, that these indications are no different from a normal Tweet without any signature label. This means, the only added information a red label gives, is the fact that the account holder who posted the Tweet usually does use Twid. Our research studied the red label to better understand what is a useful approach and how it should be implemented.

Chapter 4

Method

The goal of our research was to see what kind of effect on perceived credibility we could find, by adding different signature labels to Tweets. This chapter will describe the design and the choices we made to achieve this accordingly. We will explain how we designed the experiment and elaborate on the reasons for those choices. After that, we will go into detail about the materials we used for the set-up of the experiment and the group of participants we acquired.

4.1 Design of the Experiment

We used a between-groups design for our experiment, meaning every participant would only see one of the conditions to which they were randomly assigned. Our independent variable was the attached label being either *one (control condition)*, *signed by medical professional* or *active indicator of missing signature* giving us three conditions. Our dependent variable was the perception of credibility and indicated sharing likelihood of participants. As explained in the earlier section about *section 2.5 Credibility*, we distinguish between message and account credibility in this research. This section will explain the design choices we made and possible alternatives we could have chosen.

4.1.1 Conditions

Our independent variable, the conditions, were the different signature labels we attached to the Tweet. To answer our research question, these are either *none (control condition)*, *signed by medical professional* or *active indicator of missing signature*. In the following, we will explain our choice and elaborate on the considerations we have made.

At first, we could choose between using the labels as a signature or using

the labels as a verification of the user. Both are options the Twid-project considers currently and they differ in the interpretation of the labels underneath a Tweet. A signature is something that is put individually beneath a specific Tweet. A label attributed to a user, on the other hand, would be a property a user account has and would be shown on every Tweet, similarly to the verification badge of Twitter. We chose to focus on the implementation Twid is also currently focusing on, hence, resembling a digital signature. This means we placed both our labels underneath the Tweet instead of at the top. Twid currently focuses on this implementation because a signature does not only prove account authenticity but also message integrity since in this way a user has to actively sign every Tweet. This would not be possible for a hacker that had taken over the account. In the case of an absent signature, it thus means that the user has not verified their identity and that a Tweet could have been altered. Would we use the labels as an attribute to a user, a present label would simply mean the account has been authenticated at some point in time, comparable to the already existing verification badge of Twitter. It would not exclude the possibility of the Tweet being altered, since it does not guarantee the integrity of a Tweet in this way of implementation. An absent label would mean the same as in the case of a signature; thus, the account has not been verified yet and also the Tweet could have been altered. The labels were designed by the UX designer of the Twid team in the way they would look in the ideal case of development (Vervoort, 2021). They can be seen in Figure 4.1



Figure 4.1: The two labels as designed by the UX designer of the Twid team

To answer our research question, the first condition needed was the control condition, which is the Tweet the way it is currently the standard. The control condition is used to see how credible the Tweet is perceived without any labels so that the effect of labels can be investigated.

As a second condition, we chose a signature label that verifies the status of the author as a medical professional. The main reason we chose the role over the name is to provide relevant information that helps judge a Tweet's credibility. Additionally, this way we make use of IRMA's functionality so that the user can stay as anonymous as they wish while providing verified and relevant information. Furthermore, adding two labels at the same time would make it difficult to determine which of the labels had which effect, if any, during the study. We chose the medical profession as a label because the recent COVID-19 pandemic has made it apparent how medical misinfor-

mation is a prevalent topic on Twitter. Due to this the medical professional label also is currently the most prominent Twid label in this project.

As a third condition, we chose to have an indicator of a missing signature. We chose this because in the current state of the project it is unclear how to deal with this case. The question asked is, how to distinguish between users not using the Twid plug-in and users who use the plugin but consciously chose to not sign the tweet. The idea is that these should not be the same. The consideration in this is that while a missing signature can be a bad sign (i.e. account hacked or no expertise), this can also be the case with a normal Tweet. However, if a user usually signs their Tweets but then there is one Tweet by the same account that is not signed, this can be valuable information for a reader of a Tweet. As seen in *section 3.4 Absence of indicators* users do not always notice the absence of an indicator which means having an active indication of a missing signature might be of advantage. With this condition of the experiment, we want to investigate whether the red label is too strong of a warning sign or not.

4.1.2 Introduction of the labels

One design choice in this experiment was the question of whether or not we would explain the labels before the participant would get to see them. We decided to introduce and explain the labels first. This is due to our intention to adopt a more realistic Twid implementation simulating the case that Twid gets implemented as a plugin. Thus users would have to consciously download it, knowing about its function. Therefore, we assume an informed user that knows what the signature labels mean.

If we had chosen the second option, where the labels do not get explained it would have simulated the case that Twid gets adopted by Twitter and becomes a universal feature for all users. While this would be the ideal case, since Twid would then have more users, it is also the more unlikely scenario. It would further mean, that we assume an uninformed user. Had we chosen this scenario, positive results could have been a stronger indication of the effectiveness of Twid. This is because it would mean that even an uninformed user understands the labels. Since this is a first research into the effectiveness of Twid, however, we chose to not do this and first investigate its effectiveness with informed users.

4.2 Apparatus and Material

In the following section, we will discuss how we employed the above-mentioned decisions in the design of our material and experiment.

4.2.1 The Tweet

The Tweet’s content was created to be accurate in order to avoid spreading misinformation. We wanted the Tweet to be a fact that not everyone knew, and be a bit vague for participants to question its truthfulness. Furthermore, we wanted the Tweet to relate to the signature label we had added to it, thus the medical profession. In an informal conversation, we found a topic that is related to the human body, explaining how much energy it costs the body to store and digest food. We verified an approximation of the fact through multiple sources of grey literature (*Digestion and Energy*, n.d.; Gillespie, 2018; Grant Tinsley, 2018; Helen Kollias, n.d.; Pūtaiao, 2011; Williams, n.d.).

To design the Tweet as realistic as possible, we used a Tweet generator online (*Tweet Generator*, n.d.). For this, we also had to choose a username, account name, profile picture, and the number of likes and retweets. We chose these properties to be the same as in ongoing research done at the Radboud University (Schraffenberger, 2021-2022), to allow for a better comparison in the future. We did, however, change the posting date of the Tweet, so that it is not too far away from when the study would take place. The Tweet, as well as the three conditions used in the experiment, are shown in Figure 4.2. Besides the label, every participant saw the same Tweets, which means the same number of retweets, likes and date and time of posting.

Important to note is, that none of the Tweets had the verification badge provided by Twitter. We chose to not include this, because this badge is only meant for accounts of public interest, while the purpose of Twid is to add authenticity, integrity and background information to any user, also the less known users.



Figure 4.2: The three Tweets as used in the experiment. Control condition (top left), Signed condition (top right), Unsigned condition (bottom)

4.2.2 General set-up and procedure

We decided to do an online experiment using LimeSurvey (hosted on Radboud University servers) for this thesis. That way the environment in which the participants could join was as close to reality as possible, that is, online and using their own devices. Every participant in the study first needed to read and agree to a consent letter explaining the conditions under which they are allowed to participate and what they can expect. After agreeing to this, there were a few questions about their demographic background, which we later used to understand how representative our group of participants was. Among the background questions were also questions about their media literacy, which we took from an article published in the journal of media literacy education (Vraga, Tully, Kotcher, Smithson, & Broeckelman-Post, 2015). This research had already been used and applied in earlier research ongoing at Radboud University (Schraffenberger, 2021-2022) to assess media literacy on a 7-point Likert scale ranging from “*Strongly disagree*” to “*Strongly agree*”. Because we used a between-groups design for our experiment, every participant only saw one of the three conditions to which they were randomly assigned.

We have chosen to let users see the Tweet only once during the experiment, before they answer questions about it, instead of letting them see the Tweet continuously while answering questions about it. We made this decision, assuming that a user also scrolls through Twitter rather fast and makes the credibility judgement within that short amount of time. As one study found, the average time spent on a Tweet is about 2.9 seconds (Counts

& Fisher, 2011). Even for an interesting Tweet, where a user would consider whether the Tweet is credible or not and maybe even share it, users would still only spend slightly more than 3 seconds looking at it. For this reason, we have chosen the first scenario, where participants see the Tweet only once, because it is more realistic for the way Twitter is used so any effects found are more likely to also occur in reality.

After being exposed to the Tweet the participant had to answer the questions that would later be our measures. Additionally, we added some manipulation checks with questions about the Tweet the user had just seen. This was done to ensure the participant was paying attention while filling in the questions. At the end of the experiment, we also asked participants whether they would use the Twid plug-in and why and we also gave the possibility to leave feedback about the study. Lastly, we gave the participants some information about the study and the content of the Tweet. Knowing that information, the participant had to consent the last time to allow us to use their data for this research. Below, in Figure 4.3 a flowchart of the questionnaire the participants went through can be seen. In *Appendix C Questionnaire* the full experiment can be seen.

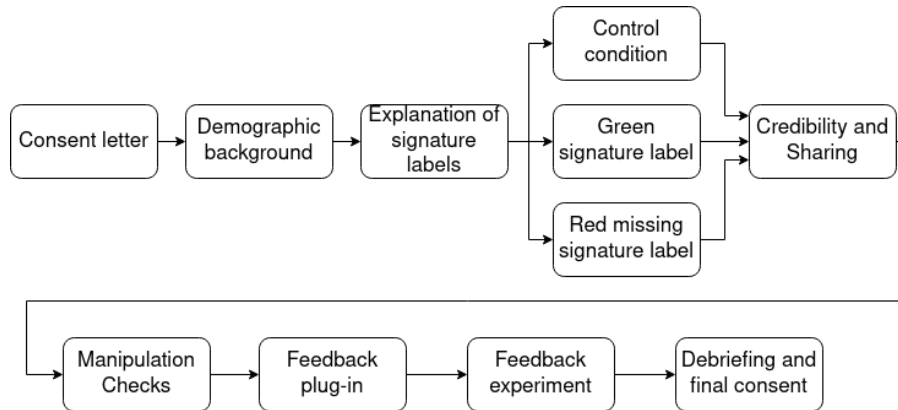


Figure 4.3: A flowchart of our experiment

4.2.3 Measurement scales

For message credibility, we used the scale of Appelman and Sundar (2016) where the user had to judge the content on a 7-point Likert scale ranging from “*Strongly disagree*” to “*Strongly agree*”. The participant was asked to judge the content on how *accurate*, *authentic* and *believable* they found the Tweet to be. Like in the original paper, we combined these measures into one compound score for message credibility to later analyse our results.

For account credibility, we use a 5-point Likert scale (ranging from “*Not at all*” to “*Extremely*”) as described in Metzger et al. (2020) and originally proposed by Flanagin and Metzger (2000). Here the participant judges how *biased*, *professional*, *trustworthy* and *credible* they perceive the account. As described by Metzger et al. (2020), we use the average, correcting for the reverse coded “*bias*”, of these aspects as a value in our analysis.

As we have shown in *section 3.3 Sharing Behaviour and Credibility*, the act of sharing is what spreads misinformation. For that reason, we added a question about how likely the participant is to share the information they had just seen in the Tweet. We use the scale from a study about the credibility effect of the blue checkmark of Twitter (Vaidya et al., 2019). In this research they ask participants to indicate on a five-point scale from “very likely” to “very unlikely”, how likely they are to share the Tweet they have just seen. For this thesis, we adopt that question as a measurement to see whether the Twid-project has an effect on users sharing behaviour. We do, however, add a qualitative question, of why the participant made that decision to see if we can judge whether their sharing behaviour is related to their credibility perception.

We have chosen these exact measurements because they are used in earlier ongoing research at Radboud University about the Twid-project (Schraffenberger, 2021-2022). That way, it will be possible in the future to compare results easier and thus learn more for the development of that project.

4.3 Participants

We ran a pilot study, to test the technical implementation, wording, design and timing of the online experiment. For this, we chose four participants who were also involved in the research of the Twid-project. We did this to receive good feedback on our experiment, but also knew that because they are so closely involved, their results might be skewed, which is why we did not use them in this research and only used the feedback to improve the experiment.

For the actual experiment, we acquired participants through multiple social media channels and by asking people to also share it in their social circles. This resulted in 115 complete responses on which we still had to apply the manipulation checks. These checked whether participants remembered the topic of the Tweet, the label of the Tweet or knew the account, even though it was a fully imagined account. We chose to do these manipulation checks from earlier ongoing research at Radboud University (Schraffen-

berger, 2021-2022). Participants answering either wrong or saying they did not know anymore were left out, which resulted in 100 complete responses we could use. It took the participants an average of approximately 656 seconds, so about 10.9 minutes, ($SD \approx 2000$, $\min = 144s$, $\max = 20181s$)¹ to fill in the complete experiment. One participant took thus a very long time, but since the manipulation checks still passed, we decided there is no reason to exclude the answers from that participant. That was the only outlier of that sort. The next replies were in the range of about half an hour², which lead to a median of 344 seconds, which is about 5.7 minutes.

The division between male and female participants was approximately equal ($Male=49\%$, $Female=50\%$, $Do\ not\ wish\ to\ say=1\%$). The majority of participants were in the age group 18-25 (78%) or 26-30 (14%) and had at least completed some secondary school education (42%) or a Bachelor's degree (38%). The big majority had Dutch as a first nationality (70%), with German being the second most common (17%) nationality. The majority of participants used Twitter less than every few weeks or never (62%) or every few weeks (15%). On a scale ranging from “*Strongly disagree*” about statements indicating their media literacy to “*Strongly agree*”, the mean of our participants self-reported media literacy got to 4.735 out of the 7-point Likert scale ($SD = .589$, $\min = 1.33$, $\max = 6.83$). See *Appendix A Demographics* for the complete descriptive statistics of the participants.

We had 100 participants randomly divided over three conditions with approximately equal frequencies as the second column of Table 4.1 shows. If we compare columns one and two in this table, we see that there is no significant difference in the number of people not passing the manipulation checks. For each of the participants, we had three measurements, message credibility, account credibility and sharing likelihood. These measurements were compounded in different ways as described in *subsection 4.2.3 Measurement scales*.

Condition	before manipulation test	after manipulation test
Control condition	38.6%	39%
Green signature label	33.1%	37%
Red missing signature	26.3%	24%

Table 4.1: Frequencies of conditions before and after manipulation tests

¹144s \approx 1.4min; 20181s \approx 5.6hrs

²The experiment was estimated to take a maximum of 10 minutes, however, as we later see English was not the first language for most of the participants, so translation could be an explanation for these longer than expected responding times.

Chapter 5

Results

In this chapter, we will look at the responses we got from the online experiment as described in previous sections. We will look at what kind of reactions we got and how we can use those to answer our research questions. We will furthermore also do some exploratory research with additional data we collected to see whether we can draw further conclusions from there.

5.1 The effect of signature labels on credibility

The first research question aimed to find out what effect the signature labels have on the perceived message and account credibility of a Tweet. In Table 5.1 the means for these measures are shown and in 5.1 the boxplots for the different conditions can be seen. Message credibility is measured along a scale from one to five, while account credibility lies on a scale from four to twenty (see *subsection 4.2.3 Measurement scales*).

Measurement	Minimum	Maximum	Mean	Std.Deviation
Message Credibility	4	19	10.96	3.848
Account Credibility	1	4.25	2.51	0.824

Table 5.1: Descriptive statistics of message and account credibility ($N = 100$)

Due to the fact that the residuals of one condition were not normally distributed, we chose to do a Kruskal-Wallis test to find the effect of the labels on a user's credibility perception¹. We found that for both, message credibility ($H(2) = 41.19, p < .001$) and account credibility ($H(2) =$

¹Talking to the SMAP desk (*SMAP*, n.d.) showed that an ANOVA test would have probably also still fine. However, it seems it leads to the same results, so here we continued with the test we had decided to do

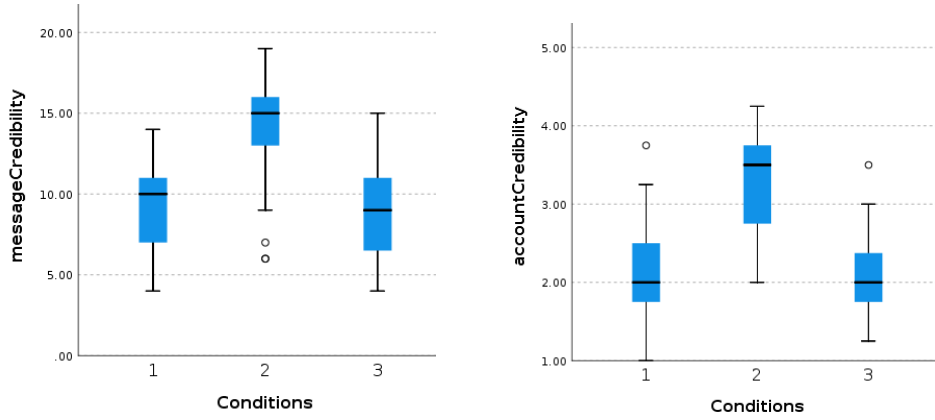


Figure 5.1: Boxplots of message and account credibility ($N = 100$)

38.619, $p < .001$), the difference was statistically significant. As can be seen in Table 5.2 the mean ranks already clearly indicate that the Tweet with the green signature label has a higher perceived account and message credibility than both the normal Tweet and the Tweet with the red “not signed”-label. While these mean ranks already show very clear results, we also carried out statistical tests to confirm between which groups the difference is significant. We performed a pairwise comparison using Dunn (1964) to confirm where the difference lies. The adjusted p -values using the Bonferroni correction for multiple comparisons are presented. We found that for message credibility the difference lies between the control condition compared to the green signature label ($p < .001$) and the green signature label compared to the red label ($p < .001$). The same holds for account credibility, where the control condition compared to the green signature label ($p < .001$) and the green signature label compared to the red label ($p < .001$) show a statistical difference.

Measurement	Condition	N	Mean Rank
Message Credibility	Control condition	39	36.82
	Green signature label	37	74.7
	Red label	24	35.42
Account Credibility	Control condition	39	36.97
	Green signature label	37	73.89
	Red label	24	36.42

Table 5.2: Mean Ranks as given by Kruskal-Wallis Test

5.1.1 Exploratory analysis: sharing behaviour

To address the problem of misinformation, Twid would ideally help prevent the spread of that misinformation. To see whether the labels in our experiment had an effect on users' sharing behaviour, we wanted to see whether there is a difference in the indicated sharing behaviour per condition. The median of our respondents was 1, which corresponds to the answer "Very unlikely". In Figure 5.2 the frequencies and the distribution of answers over the different conditions of our experiment can be seen.

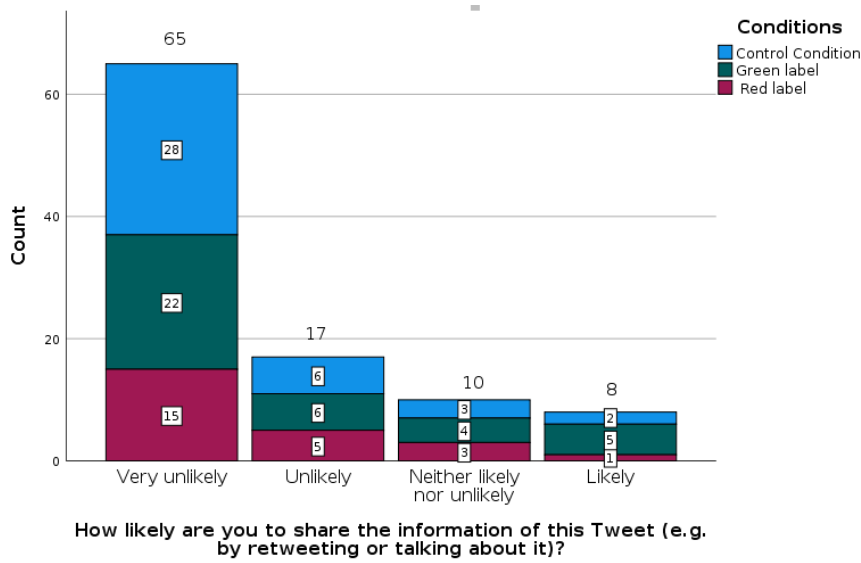


Figure 5.2: Distribution of answers about the sharing behaviour ($N = 100$)

We are again using a Kruskal-Wallis test to explore this since our sharing measurement is ordinal data. The results showed no significant difference ($H(2) = 1.715, p = .429$). This is not surprising, since only 8% of the participants replied with "likely", see Figure 5.2².

Because the Twid-project relies on the assumption that users are less likely to share a Tweet if they perceive it as less credible we investigate the users indicated sharing behaviour by checking for a correlation between sharing likelihood, message credibility and account credibility. We found the Spearman's correlation to be suitable, even though the assumption of a monotonic relationship was difficult to test, as can be seen in *Appendix B Assumption Spearman's correlation*. Since the assumption requires it to be

²Note that the category of "Very likely" is left out since none of our respondents indicated this as their answer

not non-monotonic, we concluded the assumption holds³. The results of this show that both message credibility ($r_s = .253, p = .011, N = 100$) and account credibility ($r_s = .291, p = .003, N = 100$) had a positive, statistically significant correlation. This means, that a higher measure for message and account credibility also resulted in a higher likelihood for a user to share the Tweet. It is important to note here, that the total number of users that actually did say they would want to share the Tweet (so indicated four or higher on the 5-point Likert scale) was only eight participants.

When looking at the qualitative data that we had collected, the reasons participants gave as to why they would not share the Tweet they had just seen, fell in most of the cases in one of the following categories:

- The participant did not find it interesting/relevant/useful or fun
- The participant does not use Twitter or does not actively post on Twitter
- The participant was not sure about the truth value of the message, often with the reason of missing source/proof or scientific research

Other reasons mentioned included the language and emoticons or the fact that participants did not know the account holder or were missing background information about them. It is important to note, that this last reason was only mentioned in the control condition and the red label condition.

Far fewer people indicated they would share the Tweet, but if they did so, the number one reason given was that the Tweet contained a fun or interesting fact. One participant also elaborated, saying there would not be “major consequences from sharing such a myth”.

5.1.2 Expected plug-in use

We further also asked participants to indicate whether they would use the Twid feature and why. From our participants 47% said “Yes”, 14% said “No”, 33% said “*I don't know*” and 6% did not give an answer. In Figure 5.3 we can see the answers in more detail and with different colours representing the different conditions. From these numbers, it is clear that most participants indicate they would use the feature and another big part of our participants do not know. Six participants chose not to answer this question. Since a lot of participants did not use Twitter, it is not surprising that a big part does not know whether they would use the feature.

³We confirmed our conclusion with the SMAP desk (SMAP, n.d.) again.

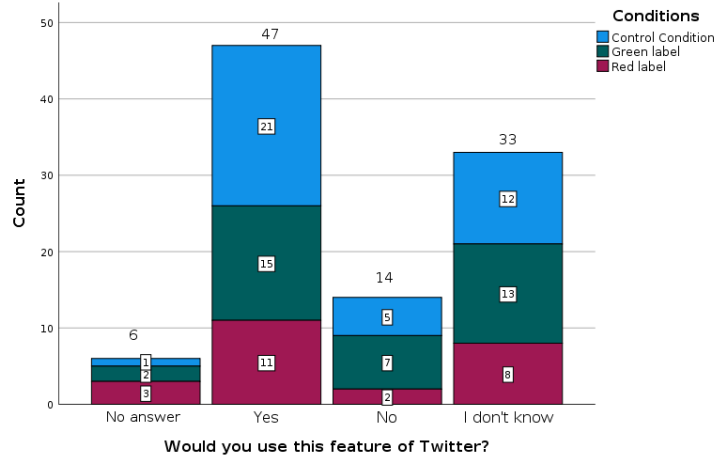


Figure 5.3: Distribution of answers about using the feature ($N = 100$)

Participants indicating they would use the feature often explained that it is a nice feature that makes it easier to judge credibility without a lot of effort.

The answers between the group of participants who indicated they would not use the feature or were unsure about it were comparable to each other. Besides not using Twitter, they argued that they would need to know where the signature gets the label from. This was often followed by the addition that they would not trust Twitter giving out these labels. Some also explained, that one expert can still be wrong and sources or peer reviews would be needed.

5.1.3 Feedback about experiment

Our last qualitative question left the room to leave any feedback about this experiment. The majority of answers were either empty or said everything was clear. Other feedback included some colours in the consent letter that had not been chosen ideally. The two most relevant categories of responses were that some people would have either liked a back button to see the Tweet again or at least a notification that this will not be possible. The other one indicated that an explanation of who authorises these labels would have been helpful for their judgement. Some feedback commented on the research and the measurements taken. Examples include the conditions shown and the adjectives chosen to measure message and account credibility or media literacy. Since these choices are explained and referenced in this thesis, we will not discuss this feedback in detail.

Chapter 6

Discussion & Conclusions

After presenting our results, we will now interpret what these results mean for our research and how they answer our research questions. We will also put these results into the context of earlier research and show how the labels contributed to the goal of helping users judge a Tweet’s credibility. We will further discuss the limitations of our research and make recommendations for future research to progress in this field. Additionally, we will look into some recent developments in the social media landscape that may have an effect on the development of Twid. Lastly, we will summarise some conclusions to take away from this research.

6.1 Discussion

In our results, we have found that a Tweet with a topic about the human body, signed by a medical professional, is perceived to have a higher message and account credibility than a standard Tweet or a Tweet with an indicator of an absent signature. This supports the hypothesis of the Twid-project that giving background information about an account holder helps users judge the credibility of a Tweet. Unlike expected, however, there was no statistically significant difference between an active indication of a signature and the normal Tweet or the signed Tweet. We had not expected this outcome since we believed participants would not be able to distinctly understand what it means if a signature is missing. As discussed in *section 3.4 Absence of indicators*, this indicator does not give more information than a normal Tweet, aside from the fact that the account holder does use Twid. This can mean that an account has been hacked or that the poster of the Tweet was not an expert in the respective topic, however, these are also possibilities for a normal Tweet. The only additional information the red label gives is that the account holder does usually use Twid. However, we expected users to see it as a strong warning and thus perceive it as less credible. So while not finding an effect was against our expectation, it indicates

that users have a clearer understanding of signatures than we expected them to have.

Our results did not show a significant connection between the indicated sharing behaviour and the labels. However, it did show a correlation between sharing behaviour and both message and account credibility. This is important for Twid since the goal of the project would of course be to prevent users from spreading misinformation. So if a decreased credibility perception indeed lowers the likelihood of sharing a Tweet, this would support the way Twid is intended to work.

Additionally, as seen in the qualitative data, one of the reasons for users to not share the Tweet, was that they did not know the person or were unsure about their background. Because this reason was only mentioned in the control condition and the condition with the red label, it seems that the signature helped users judge the account holder better. Interestingly, another reason to not share the Tweet, was that users did not know the truth-value of the message due to missing sources or scientific research. This reason was mentioned in all conditions. This indicates, that while the signature gives users knowledge about the account holder, they do not interpret it as a source and proof in itself.

6.1.1 Relation to Previous Research

As we had seen in *section 3.2.1 Visual features*, most users make a decision about credibility within seconds, simply based on superficial features. While we did not have the chance to design much, since Twitter is an already existing website, we did have the chance to design the labels attached to the Tweets. Because most people remembered correctly which label they had seen and it influenced their credibility, our findings indicate, that the design of these is visible and clear enough for a user to interpret them correctly.

Additionally, our results are in line with some of the heuristics as suggested by Sundar (2008) in the MAIN-model. The first one to look at is the authority heuristic, where a source that is viewed as official authority is perceived to be more credible. The second one is the identity heuristic, where the credibility judgment is influenced by whether or not a user is able to express their identity. The signature of an authorised medical professional plays a role in both heuristics. A medical professional is perceived as an official authority and since it is the account holder's qualification it also reflects part of their identity. Thus, it is not surprising that the Twid label in this research supports these heuristics of the MAIN-model.

The results in our research also indicate to be promising for the problem

Flanagin and Metzger (2007) research had found. The discrepancy found there indicated that while users seem to know they have to verify information found online, they do not do it. With Twid, users get background information without needing to look into it. As seen in the analysis of the qualitative data, the missing source or proof and not knowing the account holder were some of the main reasons not to share the post. This especially held for the conditions without the signature of the medical professional.

Our research did not show any effects of the labels on the sharing behaviour of users, however, future research is needed since we had very few respondents who were interested in sharing the shown Tweet. Our research did find a correlation between the perceived credibility and the indicated sharing behaviour of users, which supports the research that used sharing behaviour as an implicit way of measuring credibility (Vaidya et al., 2019). It also substantiates the assumption of Twid that giving users a better framework to judge the credibility of a Tweet might stop users from spreading misinformation by mistake.

6.1.2 Limitations

In the design of our study, we made the decision to explain the labels before using them in the Tweets our participants would see. This choice was made to simulate the introduction of Twid through a plugin the user would consciously choose to download. To further support the findings of this research, however, one would need to test this with uninformed users. The problem with introducing the labels is, that there is a chance we primed our participants into knowing what the research was about and thus paying special attention to some cues. In that way they would be responding to “demand characteristics” (Orne, 2017), answering in a way the participant thought was needed to prove the hypothesis.

Further, while our results seem promising, most of our participants indicated they used Twitter never or less than every few weeks. This is not surprising, since the majority of our participants were in a younger age group (18-25) than most of the Twitter users (25-34) (*Distribution of Twitter users worldwide as of April 2021, by age group*, 2022). Verifying these results with people, who use Twitter or are at least the age group of Twitter would be recommended.

As earlier research had shown, the average time a user looks at a Tweet is about three seconds (Counts & Fisher, 2011). Similarly, we had also seen that users make their first credibility judgement about websites in about 2-4 seconds (Robins & Holmes, 2008). This is why we decided to let participants see the Tweet only once. In our feedback, we did receive that as

a criticism, since participants were also not warned that they could not go back to the Tweet. While we had done this to simulate how a user typically goes through a Twitter feed, one could argue, that a user can also scroll back up in a feed. Future research can use this criticism to make it clearer for their participants.

Additionally, this research did not look at the time each user spent looking at the Tweet. This could have indicated whether users spent significantly longer than expected looking and thinking about the truth value of the Tweet. If future research would do this, it might be possible to conclude further, how the Twid labels can be applied in the reality of Twitter, where users make a quick credibility judgement.

Another feedback we received in our experiment was that we left out the explanation of where these signatures came from and who authorised them. The reason we had done this, was to make the explanation as simple and short as possible so that it is accessible to everyone and participants would not skip it. Furthermore, we did not want to go into detail to avoid priming participants. Future research can take this feedback into account and make sure participants know to trust the signatures. Additionally, this feedback is important for the development team of Twid. In their design, they should make clear to users where the signature label comes from and that it is a trusted system (i.e. IRMA) that issues these attributes. That way users know that the attributes in a signature are indeed verified and reliable, increasing their value to help judge a Tweet’s credibility.

An additional limitation of our research is the fact that we used a true fact, meaning we cannot distinguish, whether the green signature indeed helped the participants judge the credibility of the Tweet or whether it simply increased the credibility, no matter what the fact would have been. From this it also follows, that we need to research, whether experts are more likely to share true information, to know whether it is good to increase their credibility.

Similarly, we used a non-polarizing fact. Misinformation is especially critical in areas where facts polarise and due to the confirmation bias, it is especially hard to sway a user’s opinion in those areas. Since our Tweet contained a harmless fun fact, as also stated by participants in the qualitative question, the confirmation bias played much less of a role. It is important for future work to see whether the found effects still hold for polarising facts.

As the last limitation it is important to note that we made the choice for the content and the signature label to relate to each other since that is the way Twid would be intended to work. By providing verified and *relevant*

information an account holder can help a user judge credibility. However, earlier research has shown the existence of the lab coat effect (Khashabi & Samadzadeh, 2001; Shaw, 2013), where people wearing a lab coat are perceived as more authoritative, often even beyond their actual field of expertise. In Twid that would mean, for example, that a medical professional tweeting about climate change, would also be perceived as more credible than without a label. To understand how this affects the credibility of Tweets, future research is needed.

6.1.3 Future research

We have earlier seen which questions this research has answered and how this fits into previously done research. However, there are also some unanswered questions, that future research should investigate. For example, in Sundar (2008) MAIN-model the machine and bandwagon heuristic can also be related to Twid, but they have not been investigated in this research. The machine heuristic meant that a user perceives something as more credible if they believe something is shown because it has been chosen by a machine and not an editor. It is unclear, how in this research the user-perceived the Tweet to be chosen. The Tweet was constructed explicitly for this research, however, the participants were told to imagine they found it in their feed, which would be machine chosen. This also holds for the bandwagon heuristic, where Sundar explains that users perceive something to be more credible if other users chose that story. Here the same holds again, users were told to imagine they found the Tweet in their feed, where often Tweets of people one follows are shown. However, we do not know how the participants in our study perceived the Tweet to be chosen and whether or what kind of effect this had on their credibility perception.

As earlier research had found, the profile picture and name and with it the nature of an account (private, topic related etc.) play an important role in the credibility of a Tweet (Morris et al., 2012). In this research, we have only tested the Twid labels employing one account, with only one profile picture and name. In future research, it would be worth investigating how the Twid labels work for other kinds of accounts. This also holds for the number of followers and follows or retweets and likes of a Tweet.

Similarly, the same research (Morris et al., 2012) had shown that the content of the Tweet plays an important role in the credibility of Twitter. Aspects like a URL, grammar and abbreviations have not been investigated in this research but would be of value in future work. This also holds for the topic of the Tweet. We had seen that Tweets are perceived with different levels of credibility, depending on the topic they address with science gen-

erally having higher credibility than politics. Since this was a first research to see whether the labels work as expected, we have only used one topic, which was related to science. While science is a big part of misinformation on social media, the same holds for other topics, for example, politics. To know how Twid affects those areas, further research should look into the way the Twid labels change a user’s credibility perception about different topics.

As seen in *section 3.3 Sharing Behaviour and Credibility*, we know that ignorance is not the only reason for spreading misinformation. Research shows, that people also share misinformation to either socialise in different ways (Chen & Sin, 2013) or even due to bad intentions (Osmundsen et al., 2021). We do not know the effect of Twid labels on these motivations. We have found that one of the main reasons to share this Tweet in this research was that the user-perceived it as a “fun fact” to socialise with. Among the limited number of people who indicated they would share the Tweet, this argument accounted for about 50% of all given reasons.

Furthermore, as mentioned earlier, we had chosen a non-polarizing topic for the Tweet in this research. Besides the already mentioned confirmation bias that we do not take into account with this, we also have no information on what kind of influence Twid would have on users sharing misinformation with bad intentions. As speculation, we could hope, that Twid would also have a positive effect on this. There are two options for this, either someone is sharing misinformation with a label, where one could hope that the signature label leads to a higher feeling of personal responsibility, which would prevent users from posting misinformation. Or someone intentionally posts misinformation without a signature label which, according to our research, weakens a post’s credibility, and thus makes misinformation more harmless. However, further research is needed to see whether this is the case.

When considering a tool like Twid, one also needs to consider the option of an expert tweeting misinformation signed with a Twid label. This could either be due to bad intentions or incorrect/obsolete knowledge. In that case, this research would lead to assume that users still perceive the Tweet as more credible. Therefore, future research needs to investigate the assumption that experts are more likely to spread truthful information.

In *section 3.1 Current countermeasures against misinformation* we have discussed the confirmation bias, where people perceive information that suits their beliefs as more credible, as a major obstacle in fighting misinformation on social media. This search for confirmation can lead to a so-called echo chamber, where users are only exposed to their own beliefs and often increases the spread of misinformation Jost, van der Linden, Panagopoulos,

and Hardin (2018). By using the Twid labels, a Twitter user could choose to follow experts on certain topics instead of simply relying on their confirmation bias. Future research is needed to investigate the possible effects of Twid on the way users choose whom to follow.

As can be seen, there is a lot of room for future research, however, a first important step to help users judge credibility has been set. By giving users a way of judging the credibility of Tweets more accurately, a part of the problem is already addressed. Because misinformation is such an increasingly growing problem on social media, having only part of the solution addressed is already a start that might lead us in the right direction.

6.1.4 Recent developments around Twitter

In recent events, Twitter has been a popular topic of discussion in the news (Hawkins, 2022; Mike Isaac, 2022). Elon Musk, the currently richest man in the world, had bought 9.1% of Twitter's shares and immediately made announcements on changes to be made. As a response, the company offered him a seat on the board. This would allow him to own a maximum of 15% of Twitter. At first, Musk agreed to the seat but later declined, only to later announce he wants to buy Twitter. He also further declared that Twitter has to become a private company for the changes he wants to undertake. A private company means, that the shares will not be publicly traded on the stock market anymore, but it will only be owned by one person or a small group (*Oxford Learner's Dictionaries: private company*, n.d.). This usually results in them being bound to less strict regulations and requires less transparency to the public (*Differences Between a Public Company and a Private Company*, n.d.). The changes Elon Musk plans to make include making the algorithm open source, more room for freedom of speech and giving users more room to control what is shown in their feed. He also wants to clear all bots off the social network and *authenticate all users*.

While we do not know how this will affect Twitter and therefore the development of Twid, certain things can be said about it. For example, the fact that Musk wants to authenticate every user to eliminate bots, and give users a digital identity would be a part of what Twid tries to achieve. The possible authentication of every user, not only the accounts of public interest is one of the things Twid was developed for, in a privacy-friendly way. Secondly, Musk says he wants to give more room for freedom of speech on Twitter. While that is of course a noble endeavour, it is, as discussed, also the main reason that misinformation can spread so easily, because everyone can post anything. While Twid does not limit a user's freedom to post anything they want, it still gives the opportunity to support a Tweet's cred-

ibility with verified information about the poster. As researchers we want Twitter to be accessible and helpful to the public. Turning Twitter into a private company, where only one person or a small group has the authority about decisions and transparency to the public is decreased, could thus be a step to observe critically for an influential company like Twitter.

6.2 Conclusions

While the Twid-project might not solve all problems and a lot of future research is needed, this research has set an important first step to help users judge credibility on Twitter. We have seen that the usage of signatures underneath Tweets can help increase their credibility in a context where the signature provides relevant background information about the author. The observed connection between credibility and sharing behaviour indicates that Twid might be part of a solution to slow down the spread of misinformation on social media. Knowing these effects, which are essential for the functioning of Twid, is an important step in the development of Twid. We have against expectations not found any effect on the perceived credibility of a Tweet with a red warning sign of a missing signature. While this does not prove there is no effect, our observations on perceived credibility indicate that users do not misinterpret this label. From our observation, it seems that users can distinguish better between what a signature means and what a missing signature means compared to a normal Tweet than we expected them to be able to. These are important findings for the development of the Twid-project and will help us further in the fight against misinformation.

References

- About verified accounts.* (n.d.). Retrieved from <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts> (Accessed: 8 Oct 2021 and 29 May 2022)
- Alpár, G., & Jacobs, B. (2013). Credential design in attribute-based identity management.
- Alpár, G., van den Broek, F., Hampiholi, B., Jacobs, B., Lueks, W., & Ringers, S. (2017). Irma: practical, decentralized and privacy-friendly identity management using smartphones. *HotPETs 2017*.
- Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1), 59–79.
- Burkhardt, J. M. (2017). *Combating fake news in the digital age* (Vol. 53) (No. 8). American Library Association. Retrieved from <https://www.journals.ala.org/index.php/ltr/issue/viewFile/662/423>
- Cambridge dictionary entry: credibility.* (n.d.). Retrieved from <https://dictionary.cambridge.org/dictionary/english/credibility> (Accessed: 13 Apr 2022)
- Cambridge dictionary entry: disinformation.* (n.d.). Retrieved from <https://dictionary.cambridge.org/dictionary/english/disinformation> (Accessed: 8 Oct 2021)
- Cambridge dictionary entry: misinformation.* (n.d.). Retrieved from <https://dictionary.cambridge.org/dictionary/english/misinformation> (Accessed: 8 Oct 2021)
- Caramancion, K. M. (2020). An exploration of disinformation as a cybersecurity threat. In *2020 3rd international conference on information and computer technologies (icict)* (p. 440-444). Retrieved from <https://ieeexplore.ieee.org/document/9092330?denied=doi:10.1109/ICICT50521.2020.00076>
- Chen, X., & Sin, S.-C. J. (2013). ‘misinformation? what of it?’ motivations and individual differences in misinformation sharing on social media. *Proceedings of the American Society for Information Science and Technology*, 50(1), 1–4.
- Collins, D. (2017, Jan). ‘fake news’ inquiry launched. UK Parliament. Retrieved from <https://committees.parliament.uk/committee/378/>

- digital-culture-media-and-sport-committee/news/104981/fake-news-inquiry-launched/
- Counts, S., & Fisher, K. (2011). Taking it all in? visual attention in microblog consumption. In *Proceedings of the international aaai conference on web and social media* (Vol. 5).
- de Oliveira Albuquerque, R., Villalba, L. J. G., Orozco, A. L. S., de Sousa Júnior, R. T., & Kim, T.-H. (2016). Leveraging information security and computational trust for cybersecurity. *The Journal of Supercomputing*, 72(10), 3729–3763.
- Differences between a public company and a private company.* (n.d.). Retrieved from <https://legalvision.com.au/difference-between-public-and-private-company/> (Accessed: 19 May 2022)
- Digestion and energy.* (n.d.). Retrieved from <https://seven-health.com/2013/09/digestion-and-energy/> (Accessed: 11 May 2022)
- Distribution of traffic sources for fake news in the united states in 2017.* (2019, Nov). Amy Watson. Retrieved from <https://www.statista.com/statistics/672275/fake-news-traffic-source/>
- Distribution of twitter users worldwide as of april 2021, by age group.* (2022). Retrieved from <https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/> (Accessed: 24 May 2022)
- Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241–252.
- Edgerly, S., & Vraga, E. K. (2019). The blue check of credibility: Does account verification matter when evaluating news on twitter? *Cyberpsychology, behavior, and social networking*, 22(4), 283–287.
- Elisa Shearer, J. G. (2017, Sep). *News use across social media platforms 2017*. Retrieved from <https://www.pewresearch.org/journalism/2017/09/07/news-use-across-social-media-platforms-2017/>
- Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, 77(3), 515–540.
- Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society*, 9(2), 319–342.
- Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 conference on designing for user experiences* (pp. 1–15).
- Ghebreyesus, D. T. A. (2020, Feb). *Who director-general speech at munich security conference*. Retrieved from <https://www.who.int/director-general/speeches/detail/munich-security-conference>
- Gillespie, C. (2018). *Three ways the body uses energy*. Retrieved from <https://sciencing.com/three-ways-body-uses-energy-8706999>

- .html (Accessed: 11 May 2022)
- Grant Tinsley, P. (2018). *Do negative-calorie foods exist? facts vs fiction*. Retrieved from <https://www.healthline.com/nutrition/negative-calorie-foods#fact-vs-fiction> (Accessed: 11 May 2022)
- Graves, D. (2018). Understanding the promise and limits of automated fact-checking.
- Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. *Social media and democracy: the state of the field, prospects for reform*, 10–33.
- Gunther, R., Beck, P. A., & Nisbet, E. C. (2018). Fake news did have a significant impact on the vote in the 2016 election: Original full-length version with methodological appendix. *Unpublished manuscript, Ohio State University, Columbus, OH*. Retrieved from <https://cpb-us-west-2-juc1ugur1qwqqo4.stackpathdns.com/u.osu.edu/dist/d/12059/files/2015/03/Fake-News-Piece-for-The-Conversation-with-methodological-appendix-11d0ni9.pdf>
- Hawkins, A. J. (2022, Apr 29). *Elon musk buys twitter: all the news you need on one of the biggest tech deals of all time*. Retrieved from <https://www.theverge.com/23026874/elon-musk-twitter-buyout-news-updates> (Accessed: 18 May 2022)
- Helen Kollias, P. (n.d.). *Research review: A calorie isn't a calorie*. Retrieved from <https://www.precisionnutrition.com/digesting-whole-vs-processed-foods> (Accessed: 11 May 2022)
- Hilary, I. O., & Dumebi, O.-O. (2021). Social media as a tool for misinformation and disinformation management. *Linguistics and Culture Review*, 5(S1), 496–505.
- Hou, Y., Xiong, D., Jiang, T., Song, L., & Wang, Q. (2019). Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 13(1).
- How to retweet*. (n.d.). Retrieved from <https://help.twitter.com/en/using-twitter/how-to-retweet> (Accessed: 8 Oct 2021)
- How to tweet*. (n.d.). Retrieved from <https://help.twitter.com/en/using-twitter/how-to-tweet> (Accessed: 8 Oct 2021)
- Irma in detail*. (n.d.). Retrieved from <https://privacybydesign.foundation/irma-explanation/> (Accessed: 11 Oct 2021)
- Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A.-H. M., Hasan, S. M. M., Kabir, A., ... Seale, H. (2020). Covid-19-related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4), 1621 - 1629. Retrieved from <https://www.ajtmh.org/view/journals/tpmd/103/4/article-p1621.xml> doi: 10.4269/ajtmh.20-0812
- Issuance of mobile phone number attributes*. (n.d.). Retrieved from <https://privacybydesign.foundation/issuance-mobile/> (Accessed: 17

Oct 2021)

- Jacobs, B., Schraffenberger, H., van Gastel, B., Graßl, P., Botros, L., & Kleemans, M. (2021). *Twid: Fighting fake news on twitter*.
- Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current opinion in psychology*, 23, 77–83.
- Khashabi, J., & Samadzadeh, S. (2001). The role of lab coat on the reliance of the patients.
- Kim, J.-J., & Hong, S.-P. (2011). A method of risk assessment for multi-factor authentication. *Journal of Information Processing Systems*, 7(1), 187–198.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096. Retrieved from <https://www.science.org/doi/abs/10.1126/science.aao2998> doi: 10.1126/science.aao2998
- Metzger, M. J., Hartsell, E. H., & Flanagin, A. J. (2020). Cognitive dissonance or credibility? a comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research*, 47(1), 3–28.
- Michael Barthel, J. H., Amy Mitchell. (2016, December). Many americans believe fake news is sowing confusion. Retrieved from https://www.journalism.org/wp-content/uploads/sites/8/2016/12/PJ_2016.12.15_fake-news_FINAL.pdf
- Mike Isaac, L. H. (2022, Apr 25). *With deal for twitter, musk lands a prize and pledges fewer limits*. Retrieved from <https://www.nytimes.com/2022/04/25/technology/musk-twitter-sale.html> (Accessed: 19 May 2022)
- Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 441–450).
- Number of social network users worldwide from 2017 to 2025*. (2022). Retrieved from <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/> (Accessed: 8 Jun 2022)
- Obar, J. A., & Wildman, S. S. (2015). Social media definition and the governance challenge-an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy*, 39(9), 745–750.
- O’Keeffe, G. S., Clarke-Pearson, K., on Communications, C., & Media. (2011, 04). The Impact of Social Media on Children, Adolescents,

- and Families. *Pediatrics*, 127(4), 800-804. Retrieved from <https://doi.org/10.1542/peds.2011-0054> doi: 10.1542/peds.2011-0054
- Orne, M. T. (2017, jul). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. In *Sociological methods* (pp. 279–299). Routledge.
- Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3), 999–1015.
- Oxford learner's dictionaries: private company*. (n.d.). Retrieved from <https://www.oxfordlearnersdictionaries.com/definition/english/private-company?q=private+company> (Accessed: 19 May 2022)
- Pūtaiao, S. L. H. . A. (2011). *Energy requirements of the body*. Retrieved from <https://www.sciencelearn.org.nz/resources/1835-energy-requirements-of-the-body> (Accessed: 11 May 2022)
- Robins, D., & Holmes, J. (2008). Aesthetics and credibility in web site design. *Information Processing & Management*, 44(1), 386–399.
- Schraffenberger, H. (2021-2022). Personal Communication.
- Shaw, C. (2013). *Dress for success: The white lab coat effect and the subconscious experience*. Retrieved from <https://beyondphilosophy.com/dress-success-white-lab-coat-effect-subconscious-experience/> (Accessed: 17 May 2022)
- Shi, P., Xu, H., & Zhang, X. (2011). Informing security indicator design in web browsers. In *Proceedings of the 2011 iconference* (pp. 569–575).
- Smap. (n.d.). Personal Communication. Retrieved from <https://www.ru.nl/socialsciences/stip/facilities-support/facilities/smap/> (Accessed: 22 Mar 2022)
- Sobey, J., Van Oorschot, P., & Patrick, A. S. (2009). Browser interfaces and ev-ssl certificates: Confusion, inconsistencies and hci challenges. *Carleton University School of Computer Science, Canada, Technical Report TR-09-02*, 15.
- Stacks, D. W., & Salwen, M. B. (2014). *An integrated approach to communication theory and research*. Routledge. (Chapter 32 Credibility)
- Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 ieee second international conference on social computing* (pp. 177–184).
- Sundar, S. S. (2008). *The main model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA.
- Tandoc Jr, E. C., Ferrucci, P., & Duffy, M. (2015). Facebook use, envy, and depression among college students: Is facebooking depressing? *Computers in human behavior*, 43, 139–146.

- Tweet generator*. (n.d.). Retrieved from <https://www.tweetgen.com/create/tweet.html> (Accessed: 13 Dec 2021)
- Universal declaration of human rights*. (n.d.). Retrieved from <https://www.un.org/en/about-us/universal-declaration-of-human-rights> (Accessed: 2 Jun 2022)
- Vaidya, T., Votipka, D., Mazurek, M. L., & Sherr, M. (2019). Does being verified make you more credible? account verification's effect on tweet credibility. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–13).
- Valenzuela, S., Park, N., & Kee, K. F. (2009). Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. *Journal of computer-mediated communication*, 14(4), 875–901.
- Vervoort, L. (2021). Personal Communication.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
- Vraga, E. K., Tully, M., Kotcher, J. E., Smithson, A.-B., & Broeckelman-Post, M. (2015). A multi-dimensional approach to measuring news media literacy. *Journal of Media Literacy Education*, 7(3), 41–53.
- Westerman, D., Spence, P. R., & Van Der Heide, B. (2012). A social network as information: The effect of system generated reports of connectedness on credibility on twitter. *Computers in Human Behavior*, 28(1), 199–206.
- Williams, J. (n.d.). *How many calories does digestion use up?* Retrieved from <https://www.livestrong.com/article/320370-how-many-calories-does-digestion-use-up/> (Accessed: 11 May 2022)

Appendix A

Demographics

Below the complete descriptive statistics of the demographical background of our participants.

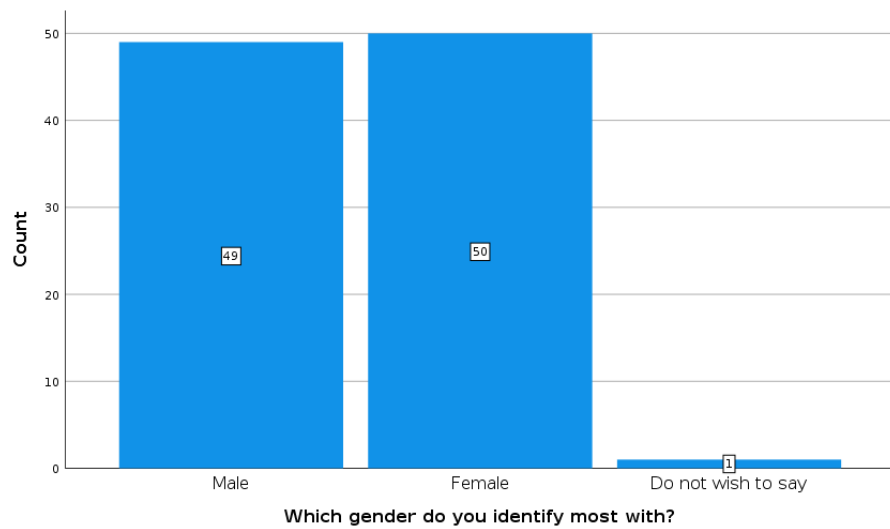


Figure A.1: Distribution of gender ($N = 100$)

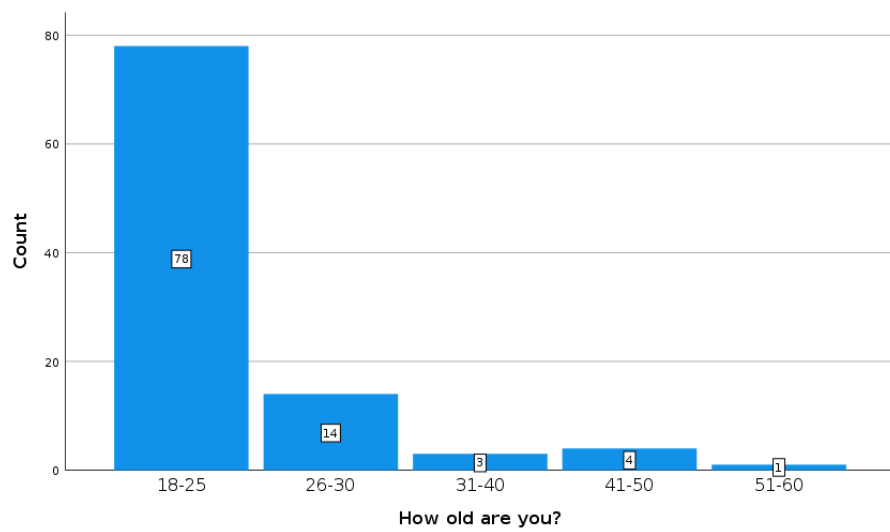


Figure A.2: Distribution of age groups ($N = 100$)

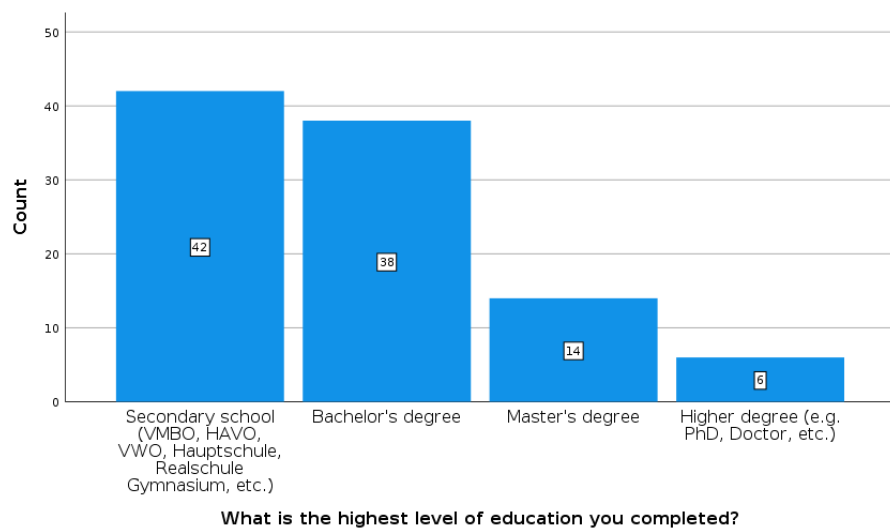


Figure A.3: Distribution of education ($N = 100$)

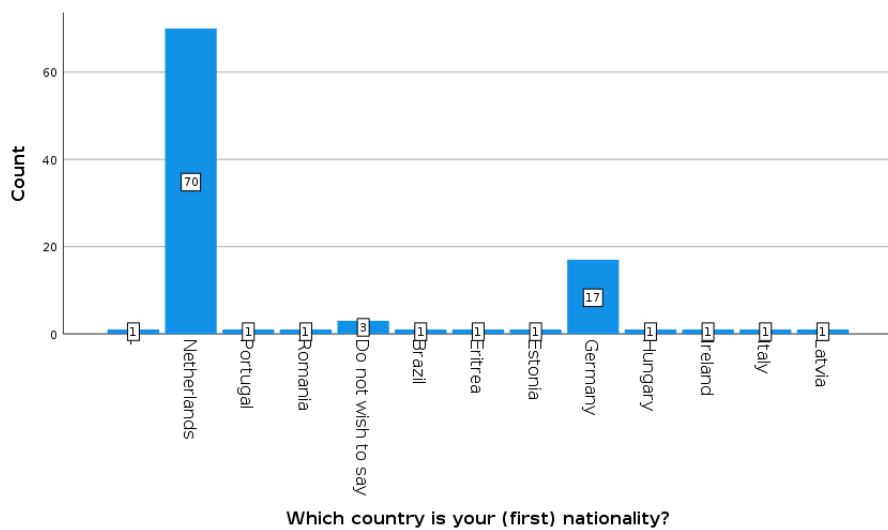


Figure A.4: Distribution of nationality ($N = 100$)

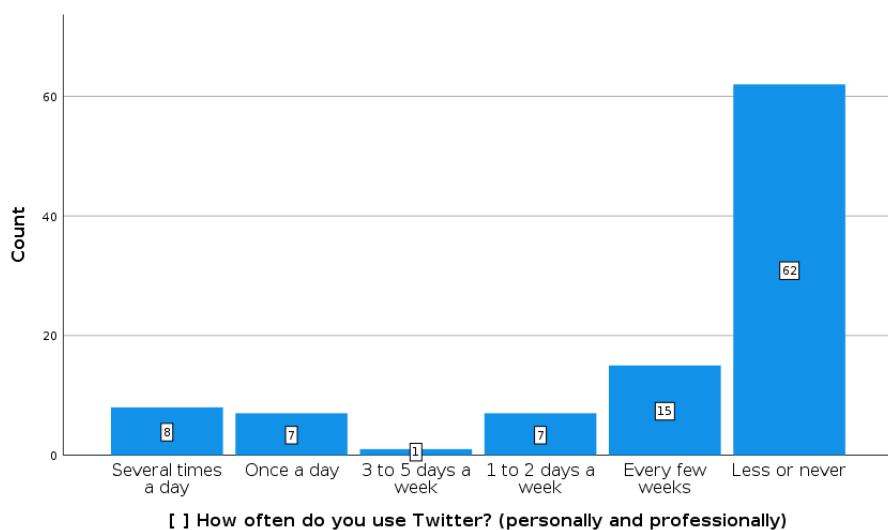


Figure A.5: Distribution of Twitter use ($N = 100$)

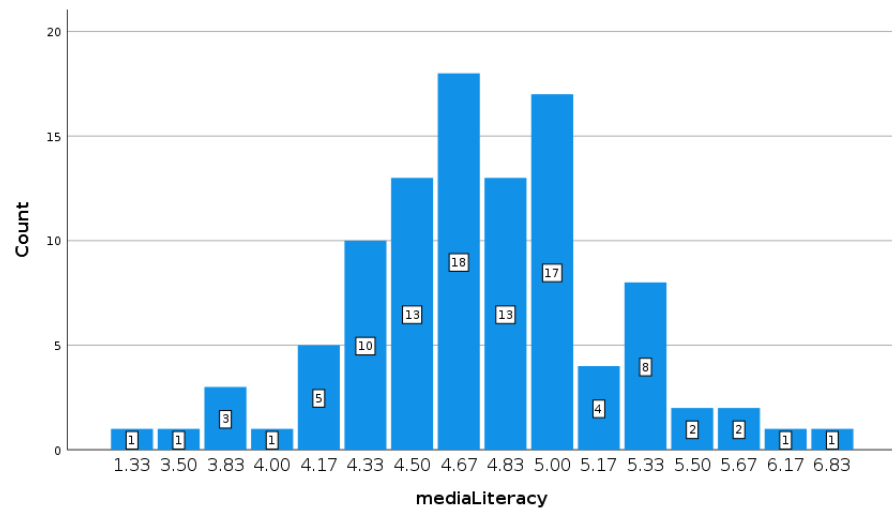


Figure A.6: Distribution of mediaLiteracy ($N = 100$)

Appendix B

Assumption Spearman's correlation

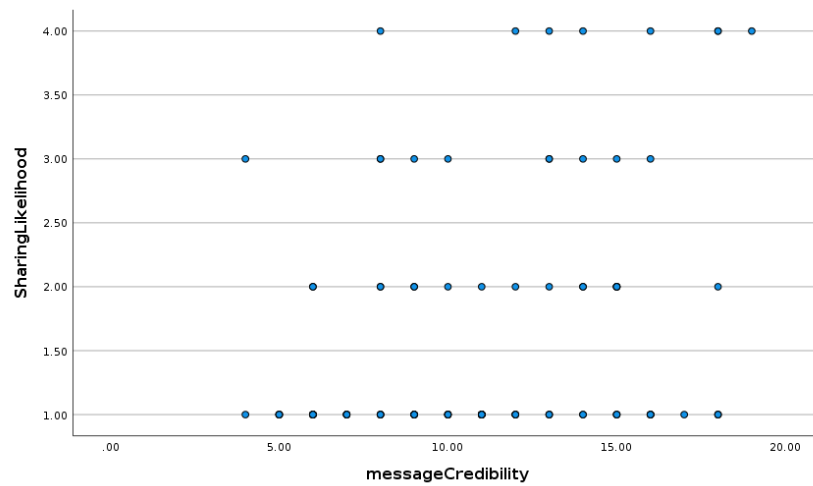


Figure B.1: Scatterplot between Message Credibility and Sharing Likelihood

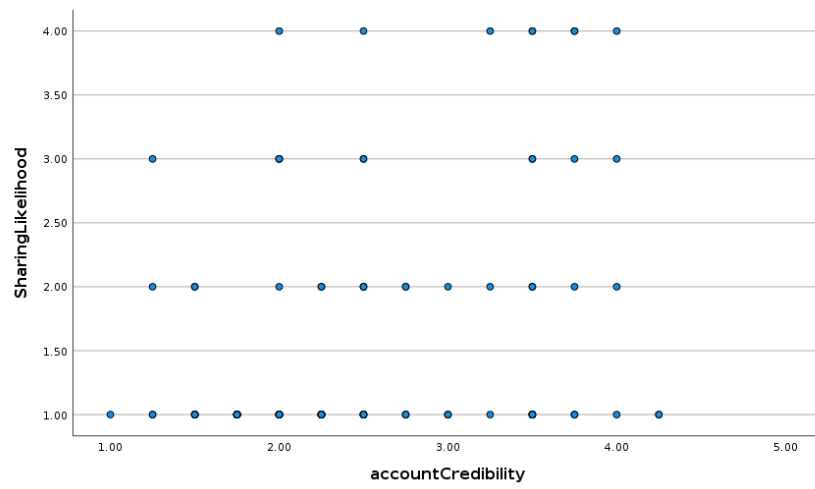


Figure B.2: Scatterplot between Account Credibility and Sharing Likelihood

Appendix C

Questionnaire

Here you can see the entire questionnaire participants as exported from lime survey.

Bachelor Thesis - Marie

Short summary:

This study is about the user experience of a new idea for an add-on for Twitter.

- You need to be 18 years or older to participate
- it will take 5-10 min
- you can quit anytime (but it won't save your results)
- data on your personal background is collected, but stored anonymized by the Radboud University. The data will be kept for the required term of 10 years, and erased afterwards.
- for more information, see these Radboud University guidelines (<https://www.ru.nl/english/vaste-onderdelen/privacy-statement-radboud-university/>) (<https://www.ru.nl/english/vaste-onderdelen/privacy-statement-radboud-university/>)).
- In case something is unclear or you need more information, please feel free to contact me (marie-sophie.simon@ru.nl (mailto:marie-sophie.simon@ru.nl)) or my supervisor (hanna.schraffenberger@ru.nl (mailto:hanna.schraffenberger@ru.nl)).

Welcome!

Welcome to my Bachelor thesis experiment and thank you for considering participating! First, a few things you need to know.

This study is about the user experience of a new idea for an add-on for Twitter. Participation is completely voluntary. If you decide to participate I will ask you to give consent on the next screen. Please read the following information carefully. In case something is unclear or you need more information, please feel free to contact me (marie-sophie.simon@ru.nl). To participate you have to be at least 18 years of age.

What you can expect

Participation in this research will take about 5 to 10 minutes. This survey will begin with a few questions about your background and your media behaviour. After that, you will see a screenshot of a possible Tweet. After looking at it, you will be asked to answer a few questions about it.

Risks and discomforts

There are no expected risks and discomforts from this research.

Voluntary participation

The participation in this research is completely voluntary. The study can be quit at any time during the survey by simply closing the page and without giving any reason.

What data is collected?

I will collect data on your personal background, such as age, gender, education, and nationality, as well as the answers to the survey. We will also collect timestamps (when a question group is answered) together with the answers.

What will happen to this data?

The data gathered during this study will be processed and stored anonymously. That way, the data we collect will not be traceable back to you. The anonymised research data will be kept for a period of at least ten years. Anonymised data will only be used for research purposes and might be shared with other researchers. All data are stored following the Radboud University guidelines.

More information?

If you have any questions about the research, feel free to contact me via marie-sophie.simon@ru.nl. 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 form

If you want to participate in this research, we ask you for your consent on the next page. By giving your consent, you declare that you have understood the information we have provided and consent to participate in this research study.

Please keep in mind:

It is not possible to save and continue later, so we would like to ask you to complete the survey in one sitting.

Kind regards and thank you for your time,

Marie-Sophie Simon

There are 23 questions in this survey.

Consent

Consent to the research

To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study.
- 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.

*

❶ Choose one of the following answers

Please choose **only one** of the following:

- ☐ I agree
- ☐ I do not want to participate

Demographical background

Which gender do you identify most with? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ Male
- ☐ Female
- ☐ Other
- ☐ Do not wish to say

How old are you? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ 18-25
- ☐ 26-30
- ☐ 31-40
- ☐ 41-50
- ☐ 51-60
- ☐ 61+

What is the highest level of education you completed? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ None
- ☐ Secondary school (VMBO, HAVO, VWO, Hauptschule, Realschule Gymnasium, etc.)
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ Higher degree (e.g. PhD, Doctor, etc.)
- ☐ Do not wish to say
- ☐ Other

Which country is your (first) nationality? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ Do not wish to say
- ☐ Afghanistan
- ☐ Albania
- ☐ Algeria
- ☐ Andorra
- ☐ Angola
- ☐ Antigua and Barbuda
- ☐ Argentina
- ☐ Armenia
- ☐ Australia
- ☐ Austria
- ☐ Azerbaijan
- ☐ The Bahamas
- ☐ Bahrain
- ☐ Bangladesh
- ☐ Barbados
- ☐ Belarus
- ☐ Belgium
- ☐ Belize
- ☐ Benin
- ☐ Bhutan
- ☐ Bolivia
- ☐ Bosnia and Herzegovina
- ☐ Botswana
- ☐ Brazil

- ☐ Brunei
- ☐ Bulgaria
- ☐ Burkina Faso
- ☐ Burundi
- ☐ Cabo Verde
- ☐ Cambodia
- ☐ Cameroon
- ☐ Canada
- ☐ Central African Republic
- ☐ Chad
- ☐ Chile
- ☐ China
- ☐ Colombia
- ☐ Comoros
- ☐ Congo, Democratic Republic of the
- ☐ Congo, Republic of the
- ☐ Costa Rica
- ☐ Côte d'Ivoire
- ☐ Croatia
- ☐ Cuba
- ☐ Cyprus
- ☐ Czech Republic
- ☐ Denmark
- ☐ Djibouti
- ☐ Dominica
- ☐ Dominican Republic
- ☐ East Timor (Timor-Leste)
- ☐ Ecuador
- ☐ Egypt
- ☐ El Salvador
- ☐ Equatorial Guinea
- ☐ Eritrea
- ☐ Estonia
- ☐ Eswatini
- ☐ Ethiopia
- ☐ Fiji

- ☐ Finland
- ☐ France
- ☐ Gabon
- ☐ The Gambia
- ☐ Georgia
- ☐ Germany
- ☐ Ghana
- ☐ Greece
- ☐ Grenada
- ☐ Guatemala
- ☐ Guinea
- ☐ Guinea-Bissau
- ☐ Guyana
- ☐ Haiti
- ☐ Honduras
- ☐ Hungary
- ☐ Iceland
- ☐ India
- ☐ Indonesia
- ☐ Iran
- ☐ Iraq
- ☐ Ireland
- ☐ Israel
- ☐ Italy
- ☐ Jamaica
- ☐ Japan
- ☐ Jordan
- ☐ Kazakhstan
- ☐ Kenya
- ☐ Kiribati
- ☐ Korea, North
- ☐ Korea, South
- ☐ Kosovo
- ☐ Kuwait
- ☐ Kyrgyzstan
- ☐ Laos

- ☐ Latvia
- ☐ Lebanon
- ☐ Lesotho
- ☐ Liberia
- ☐ Libya
- ☐ Liechtenstein
- ☐ Lithuania
- ☐ Luxembourg
- ☐ Madagascar
- ☐ Malawi
- ☐ Malaysia
- ☐ Maldives
- ☐ Mali
- ☐ Malta
- ☐ Marshall Islands
- ☐ Mauritania
- ☐ Mauritius
- ☐ Mexico
- ☐ Micronesia, Federated States of
- ☐ Moldova
- ☐ Monaco
- ☐ Mongolia
- ☐ Montenegro
- ☐ Morocco
- ☐ Mozambique
- ☐ Myanmar (Burma)
- ☐ Namibia
- ☐ Nauru
- ☐ Nepal
- ☐ Netherlands
- ☐ New Zealand
- ☐ Nicaragua
- ☐ Niger
- ☐ Nigeria
- ☐ Norway
- ☐ Oman

- ☐ Pakistan
- ☐ Palau
- ☐ Panama
- ☐ Papua New Guinea
- ☐ Paraguay
- ☐ Peru
- ☐ Philippines
- ☐ Poland
- ☐ Portugal
- ☐ Qatar
- ☐ Romania
- ☐ Russia
- ☐ Rwanda
- ☐ Saint Kitts and Nevis
- ☐ Saint Lucia
- ☐ Saint Vincent and the Grenadines
- ☐ Samoa
- ☐ San Marino
- ☐ Sao Tome and Principe
- ☐ Saudi Arabia
- ☐ Senegal
- ☐ Serbia
- ☐ Seychelles
- ☐ Sierra Leone
- ☐ Singapore
- ☐ Slovakia
- ☐ Slovenia
- ☐ Solomon Islands
- ☐ Somalia
- ☐ South Africa
- ☐ Spain
- ☐ Sri Lanka
- ☐ Sudan
- ☐ Sudan, South
- ☐ Suriname
- ☐ Sweden

- ☐ Switzerland
- ☐ Syria
- ☐ Taiwan
- ☐ Tajikistan
- ☐ Tanzania
- ☐ Thailand
- ☐ Togo
- ☐ Tonga
- ☐ Trinidad and Tobago
- ☐ Tunisia
- ☐ Turkey
- ☐ Turkmenistan
- ☐ Tuvalu
- ☐ Uganda
- ☐ Ukraine
- ☐ United Arab Emirates
- ☐ United Kingdom
- ☐ United States
- ☐ Uruguay
- ☐ Uzbekistan
- ☐ Vanuatu
- ☐ Vatican City
- ☐ Venezuela
- ☐ Vietnam
- ☐ Yemen
- ☐ Zambia
- ☐ Zimbabwe
- ☐ Other

Media Literacy

How often do you use Twitter? (personally and professionally) *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please choose the appropriate response for each item:

	Several times a day	Once a day	3 to 5 days a week	1 to 2 days a week	Every few weeks	Less or never
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please indicate for each statement how much it applies to you *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please choose the appropriate response for each item:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I have a good understanding of the concept of media literacy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have the skills to interpret news messages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I understand how the news is made	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am confident in my ability to judge the quality of news	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am not sure what people mean by media literacy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am often confused about the quality of news and information	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Randomisation

{if(is_empty(randnumber.NAOK),rand(1,3),randnumber.NAOK)}

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)


Explanation

The functionality of the new Twitter add-on:

In this survey you should assume the following. Twitter now has an add-on where a user can sign a Tweet. Below you see a few examples of how this could look and what that means.

No label

- This means the user does not use this plugin

Green label: 

- The user signed the Tweet. With this signature they prove that they are a medical professional at the ErasmusUMC.

Red label: 

- The user did not to sign that Tweet, even though they use the add-on and usually do sign their Tweets.

With this new add-on you now surf on Twitter (go to the next page)

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Conditions

**Caitlin G. Brown**

@CaitlinGBrown



Approximately 10-15% (!) of all energy we take in (so food) is used only to store and handle said food! 🤔😂

4:13 PM · Dec 3, 2021

101 Retweets 260 Likes



(/upload/surveys/974377/images/TweetResearchNewDate.png)

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.) and Answer was at question '8 [randnumber]'
(if(is_empty(randnumber.NAOK),rand(1,3),randnumber.NAOK))



Caitlin G. Brown
@CaitlinGBrown

...

Approximately 10-15% (!) of all energy we take in (so food) is used only to store and handle said food! 🤔😄

4:13 PM · Dec 3, 2021 ✓ Signed by medical professional at ErasmusUMC

101 Retweets 260 Likes



Only answer this question if the following conditions are met:
Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.) and Answer was at question '8 [randnumber]'
(if(is_empty(randnumber.NAOK),rand(1,3),randnumber.NAOK))



Caitlin G. Brown
@CaitlinGBrown

...

Approximately 10-15% (!) of all energy we take in (so food) is used only to store and handle said food! 🤔 😊

4:13 PM · Dec 3, 2021 Not signed

101 Retweets **260** Likes



Only answer this question if the following conditions are met:
Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.) and Answer was at question '8 [randnumber]'
(if(is_empty(randnumber.NAOK),rand(1,3),randnumber.NAOK))

Message Credibility

To what extent do you find the **content** of the Tweet... *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please choose the appropriate response for each item:

	Strongly disagree	Disagree	Disagree a little	Neither agree nor disagree	Agree a little	Agree	Strongly agree
accurate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
authentic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
believable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Account Credibility

To what extent do you find the Twitter account holder ... *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please choose the appropriate response for each item:

	Not at all	A little	A moderate amount	A lot	Extremely
biased	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
trustworthy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
professional	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
credible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Sharing likelihood

How likely are you to share the information of this Tweet (e.g. by retweeting or talking about it)? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please choose the appropriate response for each item:

	Very unlikely	Unlikely	Neither likely nor unlikely	Likely	Very likely
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Why would you (not) share the information of this Tweet?

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please write your answer here:

Manipulation check

What was the Tweet you read about? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ Climate change
- ☐ Concert
- ☐ Food intake
- ☐ Football
- ☐ I don't remember

Do you know the account? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ Yes
- ☐ No
- ☐ I don't know

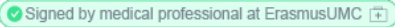

Which label did you see? *

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ A green label: 
- ☐ A red label: 
- ☐ No label
- ☐ I don't know

Preference and liking

Would you use this feature of Twitter?

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❗ Choose one of the following answers

Please choose **only one** of the following:

- ☐ Yes
- ☐ No
- ☐ I don't know

And why?

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please write your answer here:

Feedback

Was everything clear during this study or can something be improved?

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

Please write your answer here:

Final Consent

Please answer also this final question!

Study debrief:

The aim of this study is to investigate how user's credibility perception changes depending on the signature the Tweet has. Some participants saw a normal Tweet, others a Tweet with a green label, and others with a red label. The presented Tweet and account were fictional, the content, however, presents the truth.

If you would like to receive the final thesis about this research, please contact me via marie-sophie.simon@ru.nl. If you have any questions or concerns, please feel free to contact me via marie-sophie.simon@ru.nl or my supervisor Hanna Schraffenberger via hanna.schraffenberger@ru.nl

In case you changed your mind about this research, you can still have your data deleted. Please let us know below and we will delete your data as soon as possible:

*

Only answer this question if the following conditions are met:

Answer was 'I agree' at question '1 [Consent]' (To be able to participate in the study we ask you to give your consent. Please read the following statement 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 been given ample opportunity to think carefully about participating in the study. 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.)

❶ Choose one of the following answers

Please choose **only one** of the following:

- ☐ I do not want my anonymous data to be used for the study. Delete my data as soon as possible.
- ☐ I am fine with my anonymous data being used for the study.

Thank you for your time!

11.02.2022 – 16:01

Submit your survey.

Thank you for completing this survey.