UNIVERSITY
OF OULU

FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Eetu Huusko

CREDTWI: A RESEARCH TOOL FOR SOCIAL MEDIA CREDIBILITY ANALYSIS

Master's Thesis
Degree Programme in Computer Science and Engineering
August 2022

Huusko E. (2022) Credtwi: A Research Tool for Social Media Credibility Analysis. University of Oulu, Degree Programme in Computer Science and Engineering, 53 p.

ABSTRACT

In recent years, social media platforms have solidified their position as major information-sharing networks. People now also proactively look for information online and in social media for their everyday problems. There is a more sinister side to this development as well, as organizations and malevolent actors have taken the opportunity to spread false information and fake news online. For these reasons, it has become increasingly important to research the credibility of the shared information.

In this thesis, we designed and implemented a generalizable crowdsourcing research tool in the form of a browser plugin. Our tool, Credtwi, injects credibility questionnaires into the user's Twitter feed. These customizable questionnaires are attached to each tweet. Utilizing Credtwi, we carried out a week-long field study where participants assessed the credibility of tweets on certain topics. Analysing these assessments and the accompanying onboarding and post-experiment questionnaires, we identified which elements affect the participants' perceived credibility. These include factors such as the author's verification status, the linked information source, and the author's relevance to the topic. The participant's perception of Twitter as an information source, in general, had lowered statistically significantly after using Credtwi for a week.

Pulling together the results, the analysis, and the discussion at the end of this thesis, we contribute a timely piece of research to the domain of online content credibility. Further, we propose implications for crowdsourced credibility research with browser plugins including using a multi-dimensional credibility scale and adding cognitive load to the assessment process.

Keywords: Social Media, Credibility, Crowdsourcing, Browser Plugin

Huusko E. (2022) Credtwi: Tutkimus Apuväline Uskottavuuden Analyysointiin Sosiaalisessa Mediassa. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 53 s.

TIIVISTELMÄ

Viime vuosina sosiaalisen median alustat ovat vakiinnuttaneet asemiaan merkittävinä tiedon jakamis-verkostoina. Nykyään ihmiset myös ennakoivasti etsivät tietoa verkosta ja sosiaalisesta mediasta heidän jokapäiväisiin ongelmiinsa. Tähän kehitykseen liittyy myös pahaenteisempi puoli, kun organisaatiot ja pahantahtoiset toimijat ovat hyödyntäneet tämän mahdollisuuden levittääkseen valheellisia uutisia sekä tietoja verkossa. Näistä syistä on enenevässä määrin tärkeää tutkia jaetun tiedon uskottavuutta.

Tässä diplomityössä suunnittelimme ja toteutimme joukkoustamistutkimus yleistyökalun selain lisäosan muodossa. Työkalumme, Credtwi, lisää kyselyitä uskottavuudesta käyttäjän Twitter syötteeseen. Nämä muokattavat kyselyt on yhdistettynä jokaiseen tviittiin. Credtwiä käyttäen toteutimme viikon pituisen kenttätutkimuksen, jossa osallistujat arvioivat valittujen aiheiden tviittien uskottavuutta. Näistä arvioista sekä alku- ja loppukyselyistä tunnistettiin mitkä elementit vaikuttavat havaittuun uskottavuuteen esimerkiksi tviitin kirjoittajan verifikaatio tila, linkitetty tiedonlähde sekä tviitin kirjoittajan asiaankuuluvuus aiheeseen. Osallistujien havaitsema uskottavuus Twitteristä yleisenä tiedonlähteenä laski tilastollisesti merkittävästi heidän käytettyään Credtwiä viikon ajan.

Yhdistettynä tulokset, analyysit sekä loppukeskustelut edesautamme verkkosisällön uskottavuuden tutkimus-alaa ajankohtaisella tutkimuksella. Lisäksi ehdotamme seuraamuksia tulevaisuuden joukkoustamistutkimuksiin, jotka hyödyntävät selain lisäosia. Näitä seuraamuksia ovat esimerkiksi moni-ulotteisen uskottavuus-asteikon käyttö sekä kognitiivisen kuorman lisäys arviointiprosessiin.

Avainsanat: Ihmisen ja Tietokoneen Vuorovaikutus, Joukkoistaminen, Selain Lisäosa, Sosiaalinen Media, Uskottavuus

TABLE OF CONTENTS

FOREWORD

This thesis was done at the Center for Ubiquitous Computing (UBICOMP) in the Crowd Computing research group. I want to thank my supervisor Assoc. Professor Simo Hosio for all the support and guidance through the course of this thesis. Additionally, I want to thank all my colleagues in the Crowd Computing and CRITICAL research groups. Lastly, I am deeply grateful for my dear family and friends for their support along my study journey.

Oulu, August 31st, 2022

Eetu Huusko

LIST OF ABBREVIATIONS AND SYMBOLS

API	Application Programming Interface
UID	Unique Identifier
URL	Uniform Resource Locator
URI	Uniform Resource Indicator
JSON	JavaScript Object Notation

1. INTRODUCTION

Online social media platforms have solidified their position as major information sources people use to gather news and information around the world [1, 2]. Over half of U.S. Twitter users get their news from there regularly [3]. As online social media platforms' share as an information source increases, the spread of non-credible information is likely to increase. Low-credibility content spreads as effectively as high-credibility content on social media [4]. In response to this, the number of people expecting news content on Twitter to be accurate has decreased yearly. Additionally, the share of people feeling that news on social media has made them more confused about current events has increased [3]. Fortunately, online social media platform users are open to using tools to help them identify credible pieces of information from non-credible ones [5].

Credibility can be defined as "the quality of being trusted and believed in" [5]. As trust can be interpreted in other ways regarding human and computer interaction, "believability" is a more accurate definition [6]. A piece of information is credible if the reader believes it and an author is credible when the reader believes them. Credibility is a perceived quality, not a quality that someone or something inherently has. As such when discussing credibility, one is always discussing the perceived credibility of the perceiver [7].

The importance of the credibility of the content of a given platform becomes evident when considering the impact of the platform. For instance, Twitter has over 300 million active users worldwide and has a large role in the social discourse of news and topical events [8, 9]. Before the age of the internet and social media, there was an aspect of gatekeeping with information sharing. You had to go through specific channels to get your information to the wider public. In the case of written information, it was newspaper and book publishers. This liberation of information sharing brings along both good and bad aspects, and as such, it is important to research the online discourse from different angles and by different means. Twitter has been the platform of choice for a lot of credibility research. From rumors during crises [10, 11, 12] to fake news detection [13, 14], the credibility of discourse on Twitter has been under a microscope for years. Twitter users are willing to sacrifice some credibility in exchange for avoiding traditional media bias and getting information from people who share their views [15]. This preference for seeking like-minded content may have a risk of Twitter users forming "echo chambers" where misinformation and radical ideas can foster [16]. The spread of false information was evident in the recent COVID-19 epidemic and it sparked a lot of research into the field of social media information [12, 17].

This thesis introduces Credtwi, a novel and generalizable browser plugin. In this thesis, Credtwi was used for research on Twitter to enable the crowdsourcing of information and opinions related to tweets. The Credtwi plugin is used in this study to encourage users to reflect on the perceived credibility of tweets on their Twitter timelines. This is accomplished using a question form containing a scoring system and an open-ended question. The scoring system and the open-ended question are used to make the user assess the credibility of tweets.

As the Credtwi plugin integrates into the user's own Twitter page, they can perform credibility assessment utilizing multiple feature levels of a tweet [8]. The first of these feature levels is the content of the post, also called the post level, where the user is able to analyze the tweet's message characteristics, multimedia features, and sentiment features. The second feature level is the topic level, where the user is able to analyze the tweet in relation to the related topic of the tweet. The final feature level is the user level, where the user is able to analyze the different characteristics of the tweet's author's account e.g. the number of followers the author has or their ideological affiliation. The combination of these three feature levels is called the hybrid level. This feature level contains all the analysis levels utilized in credibility assessment. The thesis focuses on exploring what kind of aspects affect the perceived credibility of Twitter users when they judge the credibility of content on Twitter utilizing the hybrid credibility assessment level.

1.1. Motivation

A web browser plugin can be viably and successfully utilized as a research tool to study the credibility aspects of social media, as shown by previous studies [5, 9, 14, 18, 19, 20]. A research tool that can be easily shared across the web and be modified to suit the researcher's needs is a valuable asset when running studies concerned with the online discourse on social media platforms.

As Twitter encourages more concise content with its short character limit [21], users can be more susceptible to lazy thought processes, which in turn increases the likelihood of believing non-credible content [22]. As users are encouraged to think about the credibility of the information they are faced with, they are less inclined to share false information. Thus, additional cognitive load for the users in the form of credibility assessment might lead to increased accuracy of discernment of credible information [23].

These considerations led us to design and implement Credtwi and investigate social media credibility utilizing its crowdsourcing capabilities.

1.2. Objectives and Research Questions

We set out to design, implement, and test a customizable browser plugin that could be utilized in multiple crowdsourcing research studies. To test the utilization of the plugin, we planned to conduct a field study on Twitter interested in the credibility assessment of tweets. Thus, the thesis pursued two concrete objectives:

- O1: Design, implement, and test a Google Chrome plugin that allows crowdsourcing research on Twitter

- O2: Use the implemented plugin in a field study to gather data, focusing on credibility assessment

In the field study, we wanted to investigate if there is an effect on the general perception of social media credibility after using the implemented plugin, given how the plugin makes people reflect and study tweets carefully during the study duration. Additionally, we wanted to investigate the factors which affect the perceived credibility of the tweet's author and the tweet itself. Finally, we wanted to better understand using crowdsourcing on browser plugins for credibility research. Following these deliberations, three research questions were formulated:

- RQ1: Does reflecting tweet credibility with a plugin affect one's perception of social media credibility?
- RQ2a: What factors affect the author's perceived credibility?
- RQ2b: What factors affect the tweet's perceived credibility?
- RQ3: What implications for future crowdsourced credibility research with browser plugins can we derive from the study?

1.3. Structure

Following this introduction, Chapter 2 showcases the prior research and work related to the web as an information source, perception of credibility in general, credibility in online discourse, credibility in social media, and web browser plugins concerned with social media credibility. After the background work has been brought to the forefront, Chapter 3 describes the design and implementation of the Credtwi plugin, how preliminary tests were run for Credtwi, and the design of the field study. This is followed by Chapter 4 where the data provided by the field study is analyzed and the accompanying results are described. Following the study section, Chapter 5 discusses the results from the field study, various limitations of the conducted study, and possibilities for further research. Finally, Chapter 6 summarizes the whole thesis to bring it to a conclusion

2. RELATED WORK

This section provides the background information for the study. It is important to understand how information is perceived on the web, and what factors affect the perceived credibility of said information. As this study is focusing on social media platforms, especially Twitter, it is useful to understand how the perception of credibility has been studied in this context. Additionally, we highlight studies that have proposed or introduced web browser plugins designed to aid users in perceiving the credibility of information on social media.

2.1. Perception of Credibility

Before delving into specifics, one needs to understand the definition of credibility and how it is perceived in the human-computer interaction context. As the pioneers of this topic, Tseng and Fogg have found that a good synonym for credibility is "believability" [6]. These two terms are interchangeable in virtually every case. They propose four kinds of credibility types when working with computers: presumed credibility, reputed credibility, surface credibility, and experienced credibility. Presumed credibility is associated with the general assumptions in the perceiver's mind. Reputed credibility is associated with what other people have told the perceiver. Surface credibility is associated with the assessment of credibility by a simple inspection by the perceiver. Experienced credibility is associated with the perceiver's first-hand experience and how that applies to the object of credibility assessment. They also note that trust and credibility are not synonymous, even though these words have been used interchangeably by researchers. They suggest that a good interpretation of trust in the field of human-computer interaction might be "dependability".

Fogg and Tseng describe two credibility evaluation errors and three models of credibility evaluation [7]. The first credibility evaluation error is the "Gullibility Error", where the user perceives a computer product to be credible when it is not. For users to be less likely to make this error, they should be taught about information quality. The second error is the "Incredulity Error", where the user perceives a computer product not to be credible when it in fact is credible. This type of error is of more concern to the people who design and produce computer products. In addition to credibility evaluation errors, Fogg and Tseng introduce three models of credibility evaluation. These models are illustrated in Figure 1. On the vertical axis of the graphs is the user's acceptance of the computer product being evaluated. On the horizontal axis is the perceived credibility of the product. Binary Evaluation is the simplest of the three. The users either perceive the product as credible or not. This model of evaluation is usually adopted when the user is not familiar with the subject matter at hand or is not interested in it. The second evaluation model adds thresholds to the consideration and is aptly named the Threshold Evaluation. The users identify the product to be credible or not based on whether the perceived credibility passes or falls below the threshold. If the credibility is somewhere between the thresholds, the user perceives the product as somewhat credible. For there to be thresholds, the user has to be somewhat knowledgeable about the subject matter or be somewhat interested in it. The last and most complex of the evaluation models is Spectral Evaluation. This model has no fixed credible, non-credible categories. As such, no product is strictly credible or not credible, everything falls somewhere between those two extremes. In order for this evaluation method to be adopted, the user has to be very motivated and knowledgeable about the subject matter.

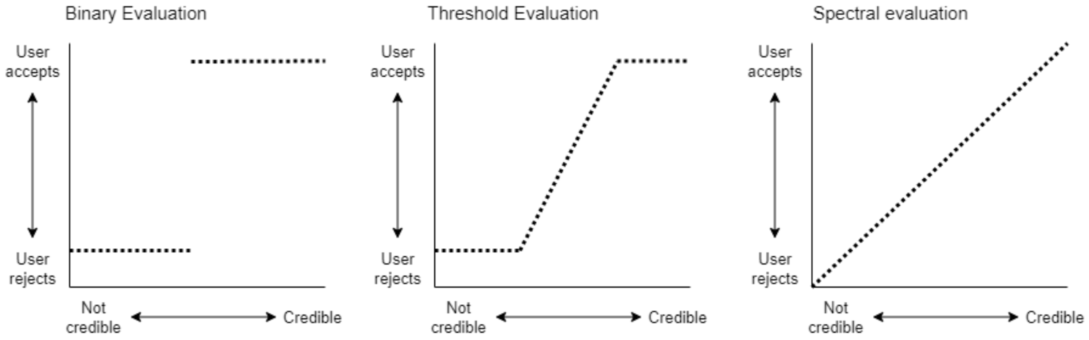


Figure 1. Three models of credibility evaluation. Adapted from Fogg B.J. & Tseng H. (1999) The elements of computer credibility. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 80–87.

2.2. The Web as an Information Source

As the base for online discourse, it is beneficial to understand how the Web is perceived as an information source. For instance, Sillence et al. showcased reasons why users trust and mistrust websites in their study which had 15 UK-based women search for health information online [24]. The large majority of factors (94%) contributing to website mistrust were design factors. The rest (6%) were content factors. Conversely, most factors (83%) contributing to website trust were content factors. The rest (17%) were design factors. These factors include:

- Design factors:
  - Mistrust:
    - Complex layout
    - Lack of navigation aids
    - Pop up adverts
  - Trust:
    - Clear layout
    - Good navigation aids
    - Interactive features
- Content factors:
  - Mistrust:
    - Irrelevant material
    - Inappropriate material

- Trust:
    - Informative content
    - Unbiased information
    - Clear, simple language

Del Vicario et al. analyzed how misinformation spreads online [16]. They found that the primary factor relating to the spread of misinformation is the user's tendency to expose themselves only to selected content which then generates homogeneous social clusters i.e. "echo chambers". As this clustering fosters user confirmation bias and polarization, an algorithmic approach to breaking these clusters does not seem to be the best option.

2.3. Credibility in Online Discourse

Continuing from the general perception of credibility, we now focus on the perception of credibility in online discourse. For instance, Fogg et al. conducted a survey of participants from Finland and the USA regarding their perceived credibility of websites [25]. From the gathered data, they created seven factors in which the questions could be loaded. These seven factors were:

- Real-World Feel
- Ease of Use
- Expertise
- Trustworthiness
- Tailoring
- Commercial Implications
- Amateurism

From these factors, Real-World Feel, Ease of Use, Expertise, Trustworthiness, and Tailoring affected positively the perceived credibility. Commercial Implications and Amateurism were impacting the perceived credibility negatively. In addition to these factors, they analyzed how different demographic groups differed in their perception of credibility. They found that even though there were significant differences in how different demographics assess websites' credibility, the differences were typically minor. This suggests that different demographic groups assess website credibility fairly similarly.

Hilligoss et al. propose a framework for online credibility assessment which consists of three levels: construct, heuristics, and interaction [26]. The construct level is the most abstract and thus the highest level associated with how a person self-defines credibility. The level below is the heuristics level which is concerned with general guidelines for assessing credibility which apply to several situations. Last is the interaction level which is associated with specific information sources and is more specific than the general guidelines in the level above.

Vosoughi et al. conducted a thorough investigation of the diffusion of true and false news stories distributed on Twitter [13]. They concluded that false information diffused significantly farther, faster, deeper, and more broadly in all categories of information. This was the case even when robot or bot activity was taken into account. False news was also more novel than true news, suggesting that people are more likely to share novel information.

In their study of the ability to discern fake news in the political context, Pennycook and Rand found that analytic thinking plays a large role in accurately identifying fake news [22]. Intuition-based thinking was less accurate. They also found that partisanship bias was unrelated to the participant's ability to discern political fake news.

Metzger et al. describe different cognitive heuristics online users employ when assessing online credibility [27]. These heuristics are:

- The reputation heuristic is the use of name recognition or authority when assessing the credibility of online content. If the user recognizes the name of the content provider or if they recognize the content provider as having authority in the topic of the content, the user assesses the information to be more credible.

- The endorsement heuristic indicates that users are more inclined to trust content without much self-made analysis if other users trust in it. This heuristic is associated with the assumption that if many others evaluate a piece of information to be correct or good, the user is more likely to as well.

- The consistency heuristic is the strategy to check if the information is consistent between different information sources. If the information appears consistent to the user, they are more likely to view the information as credible.

- The self-confirmation heuristic is the proclivity to believe online information if it consolidates their pre-existing beliefs or not to believe when the information counters their beliefs. This heuristic is associated with the self-confirmation bias that affects the user's credibility assessments.

- The expectancy violation heuristic is employed if the website in question fails to meet the expectations of the user, and thus, the user evaluates the website as not credible.

- The persuasion intent heuristic kicks in when online users come across content that appears biased such as commercial information or advertisements. This bias is seen as an ulterior motive by the users and thus decreases the perceived credibility of said information.

In their meta-analysis regarding credibility evaluation of online health information, Ma and Atkin concluded that source credibility may or may not affect the user's credibility evaluation of user-generated health information [28]. This was dependent on the online platform the information was posted. The information posted on platforms coded as "Websites" in the study, had a higher credibility evaluation than platforms coded as "Blogs" or "Bulletin Boards".

Graefe et al. found that readers tend to rate computer-generated news higher in credibility than their human-generated counterparts [29]. They propose a possible explanation that news tends to follow general conventions, and thus the algorithms that generate computer-generated news are made to follow these conventions. Even though the perceived credibility is higher with computer-generated news, the readers receive more pleasure from reading human-generated news.

Jung et al. noted in their research pertaining to diet information websites that when readers do not have much prior knowledge, they depend more on source expertise cues for credibility [30]. When readers are knowledgeable of the subject at hand they are more interested in the accuracy of the information than the information source. Thus even if the reader considers the information source to not be an expert, they perceive the information as credible if the information is in their view accurate.

Hämäläinen et al. presented an intervention program designed for improving sixth graders' credibility evaluation of websites [31]. They had 342 students which were divided into an intervention group and a control group. The intervention group received additional teaching on searching information online, information credibility, and synthesizing information. They found that the intervention program improved the student's ability to justify their credibility evaluations.

Armstrong and McAdams conducted a study on how gender cues affected the perceived credibility of blogs [32]. They found that blogs written by males were perceived as more credible than blogs written by females. They reflect on the results by adding that gender might not be used as a credibility heuristic by younger adults as often as older adults. And that in online discourse these gender-related cues might not be as prevalent as in other media.

Sun et al. investigated the differences between male and female Web advertising evaluation [33]. They conducted a laboratory experiment using a simulated website with banner advertisements. The participants were asked about their perceptions and attitudes towards pop-up advertisements. The results from their experiment indicate that males have a more positive attitude towards informativeness than females, and females have a more positive attitude towards entertainment than males. They found also that the interaction effect between informativeness and entertainment was significant for females.

Yin et al. investigated gender differences in microblog information credibility judgments [34]. They examined four cognitive heuristics platform credibility, source credibility, social endorsement, and vividness. They found that the effects of cognitive heuristics on microblog information credibility differ across genders. They found that microblog platform credibility was more important for men than women. Additionally, they found that source credibility was significant for men but insignificant for women.

2.4. Credibility in Social Media

Credibility research on social media is the most related area to this thesis, and a number of scholars have investigated the topic with an increasing volume during the past years. For instance, in their study analyzing tweets related to the Chilean earthquake of 2010, Mendoza et al. found that rumors are questioned a lot more by the Twitter community than true news items (around 50% vs. 0.3%) [10]. This finding leads to the possibility of detecting rumor tweets by using aggregate analysis.

By analyzing the distribution of fake images during Hurricane Sandy in 2012, Gupta et al. concluded that a minority of users contributed most to the distribution of fake images [11]. They analyzed over 16 000 tweets containing fake and real images from the incident. To identify if these images were real or fake, they used certain online resources e.g. a list of real and fake images made public by the Guardian news media company. They concluded that the top 30 (0.3%) of users resulted in 90% of the retweets or distribution of fake images.

Shariff et al. analyzed tweet features in relation to perceived credibility [35]. Through a credibility assessment task on CrowdFlower, they identify that the most important features affecting the perceived credibility are display name, link in a tweet, and user belief in the tweet topic.

Kouzy et al. searched for and analyzed 673 tweets from Twitter using keywords related to the COVID-19 pandemic [12]. Of the gathered tweets, 24.8% of tweets contained misinformation, and 17.4% of tweets contained unverifiable information related to the COVID-19 pandemic. This result showcases that misinformation and unverifiable information is a part of a large portion of social discourse happening on Twitter.

Morris et al. show in their study that Twitter users are poor at judging the truthfulness of a tweet based on the content of the tweet alone [36]. The users rely on other heuristics in combination with the content such as usernames and profile pictures to make their credibility assessments. In their pilot survey, they asked their participants to search for a local candidate in the U.s. Senate election and "think aloud" when viewing their tweets. The experimenter took notes of how the participants described the tweets and asked questions to better understand what factors users were looking at when analyzing tweets. These factors were then combined into a collection of 26 factors which then were rated by other survey participants on how much they paid attention to these factors and how much they affected their perceived credibility. These factors include things like verified account, author's follower count, and "author bio suggests topic expertise".

Shao et al. showcase how low-credibility content is propagated by social bots on Twitter [4]. The content is marked as low-credibility if it links or sources a site that has been identified as low-credibility by third-party news and fact-checking organizations or experts. There is a difference in the number of tweets shared in the replies between low-credibility and fact-checking sources, where the low-credibility-sourced tweets are not shared in the replies as often. Low-credibility-sourced tweets are mainly shared by the original tweets or retweets. They also found that when low-credibility content became viral, the spread of that content was concentrated on a few accounts and their behavior indicated an automatic method was used. Using a bot classifier, they were able to find out that these "super-spreaders" were likely to be bot accounts.

Mitra and Gilbert gathered a large social media corpus called CREDBANK, which has associated credibility assessments within [37]. The credibility of the corpus items is assessed utilizing crowdsourcing through Amazon Mechanical Turk. The assessment of this corpus suggests that a whopping 23.46% of the global tweet stream is not credible.

In their 2014 paper, Westerman et al. studied the relationship between social media platform content update recency, cognitive elaboration, and credibility [38]. Using a mock Twitter page for gauging credibility, they found a positive correlation between cognitive elaboration and credibility. This means that the more thought the participants put into analyzing the mock Twitter page, the more credible they found the information to be.

Pennycook et al. studied how people do not think about how accurate the information about COVID-19 was before they share it [23]. In the first study, the participants were asked one of two conditions: how likely they were to share a specific story related to COVID-19 online or whether the headline in the story is accurate. They found that the participants were bad at judging whether the information was true or false if they were not asked to assess the accuracy of the information. They also found that people were more accurate in their assessment when they utilized greater cognitive reflection and science knowledge. In the second study, they introduced an intervention method where a part of the participants rated the accuracy of a single headline before doing the same task as in the first study. They found that adding the accuracy task increased the truthfulness of the information sharing almost threefold.

With their Weibo study, Gao et al. explored different factors affecting Chinese users' perceived credibility of health and safety information [39]. They found that objective claims with low extremity increased the readers perceived credibility when they had prior knowledge of the subject. The claim extremity was defined as how much the information deviated from the central tendency. In addition, they found that negative comments had a decreasing effect on perceived credibility, whereas positive comments had no significant effect. Reposts of information that was perceived as credible increased the credibility, whereas, with less credible information, large numbers of reposts induced skepticism and lessened the perceived credibility of the post.

In their vast survey on credibility research on online social networks, Alrubaian et al. investigated 92 social media analysis papers showcasing their different methodologies in analyzing and assessing credibility [8]. The studies were divided into three main categories based on their credibility assessment approaches: automation-based, human-based, and hybrid. Both automation-based and human-based categories had 47 papers and there were 18 hybrid papers. The automation-based and human-based approaches were further broken into subcategories. There were 9 papers related to crowdsourcing, which was a subcategory of the cognition and perception approach subcategory. They noted that the dependence on the wisdom of the crowd is not ideal, as the participants or users might not have related prior knowledge to accurately assess the credibility of the information. They also noted that crowdsourcing is inherently inconsistent. The survey noted that information credibility in Twitter is typically analyzed at three different levels: post-level, topic-level, and user-level. The Post-level contains the analysis of a single post and assessing its credibility. This analysis can take into account the post's message characteristics, multimedia features, and sentiment features. This level of analysis is useful for automated credibility assessment tools as it can be done in real-time. The Topic-level analysis is done by analyzing the relevant topics by utilizing hashtags and relating URLs. The Topic-level analysis is utilized e.g. in research concerned with the information flow and propagation during crisis events and disasters. The third level of analysis is the user level in which the demographic information, the characteristics, and the social network of the user is brought under a microscope. The final level of analysis is the hybrid level which combines these three analysis levels.

In their crowdsourced study, Johnson and Kaye explored reasons why people rely on social media for their information even though they do not perceive it to be as credible as other sources [15]. They found that while users judged the social media sites as less credible than e.g. newspapers or broadcast television, they were willing to trade that credibility if other needs are met. For Twitter, the main motivation was to avoid traditional media bias. Another motivation for the use of Twitter was the access to information and opinions from people who share their views.

Shariff et al. explored the relationships between Twitter reader tweet features with the reader's perceived credibility of the tweet [40]. They showed in this study that readers use a combination of different features to assess the credibility of a single tweet. These different features are arranged into five different categories: author, transmission, auxiliary, topic, and style. The author category includes features like the Twitter ID or the profile picture of the person who published the tweet. The transmission category includes features like user mentions, retweets, and hashtags in the tweet. The auxiliary category includes links to outside sources and attached media in the tweet. The topic category includes alert phrases and topic keywords in the tweet. The style category includes features like the language of the tweet and if the tweet is the author's opinion or fact.

2.5. Web Browser Plugins for Social Media Credibility

As this thesis is concerned with using web browser plugins to research social media credibility, it is beneficial to understand what kinds of plugins have already been used in related research. As the first example of a successful research plugin, Gupta et al. presented a browser plugin usable on Twitter that allows its users to view the credibility of tweets in their timeline [5]. This credibility ranking system is semi-supervised, as the system needs a training set to be able to rate the credibility of each tweet. In this paper, they defined credibility as “the quality of being trusted and believed in." This set was collected using the Twitter API and the selected tweets concerned 6 crisis events in 2013. From this training set, 500 tweets were selected and their credibility was evaluated using a crowdsourcing platform called CrowdFlower. The CrowdFlower participants were first asked to rate the relevancy of the tweet to the event. Secondly, the participants were asked to rate the credibility of the tweet. This crowdsourced data was used to train the system to automatically rate the credibility of tweets using an SVM-rank learning scheme. The plugin fetches the tweet data, generates features, and computes the credibility score for each tweet. This score is then viewed by the user using a 7-point scale. The user can give feedback if the computed score is correct in their opinion, and provide a different score if they disagree.

Paschalides et al. present a fake news detection system called Check-It which is implemented as a web browser plugin and used on social media platforms [14]. At the time of writing the article, the plugin only supported Twitter as a platform. The plugin uses a deep neural network linguistic model for the evaluation of the social media article's text content. This linguistic model is trained using an open-source Fake News Corpus. In addition to text analysis, the plugin generates a user blacklist generated via social network user analysis which flags user accounts with high falsify scores. These two tools are accompanied by a list of known fact-checked articles and a list of known fake news domains. These four pieces of the detection system generate a fake news probability score for a certain social media article and then give a warning to the user if the plugin is highly confident the article in question contains fake news.

Dongo et al. propose a model to calculate the credibility of social media posts using three measures: text, account, and social impact [18]. To test this model, they created a proof of concept plugin for Twitter called World White Web. This plugin could be used to check a particular text piece for credibility by inserting it into the plugin or the credibility of the tweets on the current page by pushing a button in the plugin popup. The plugin would give a percentage score for the computed credibility which takes into account the three measures. The weights of these measures are user-definable so that users who value text credibility more than social impact, can do so.

Hartwig and Reuter designed and made a browser plugin called TrustyTweet, which helps Twitter users to detect possible fake news utilizing helpful indicators [19]. TrustyTweet was built to aid in increasing the user's social media literacy. Unlike some other credibility checking tools, TrustyTweet doesn't just show a numerical credibility score, it shows a warning indicator next to the tweet when a tweet might be considered potentially fake news. The warning indicator can be hovered over with the mouse to see the reason why the plugin considers the particular tweet to be fake news. These fake news indicators include, for example, account verification, capitalization, and excessive punctuation. These can be toggled on and off by the user to suit their personal needs.

Bhuiyan et al. introduced FeedReflect, a browser plugin designed to nudge users to assess the credibility of news content on Twitter [9]. The plugin has two nudging elements: highlighting mainstream news outlets' content that contains a question and dimming non-mainstream news content. The dimming is performed by reducing the opacity of the content in the user's Twitter feed. The dimming is accompanied by a tooltip informing the user why this content has been dimmed. The participants were made to reflect on the credibility of the information with an accompanying news credibility survey. They compared the nudging treatment to a control group and found that the treatment affected the perceived credibility of the news outlets. The mainstream news outlets were perceived as more credible and the non-mainstream as less credible when users were affected by the nudging treatment. They conducted interviews on the effects of the nudge treatment on the treatment group. They identified three types of effect themes from these interviews: make people reread and rethink the news, make people use external resources for credibility assessment, and make people actively participate in content credibility assessment.

Continuing on the theme of nudging, Bhuiyan et al. studied supporting news credibility assessment on social media utilizing informational content nudges [20]. They created a browser plugin named NudgeCred which added three nudges on news content on Twitter: Reliable, Questionable, and Unreliable. These nudges were in the form of tooltips. The plugin determines if the news tweet is from a mainstream news account or if the tweet has a link that directs to a mainstream news outlet. If that is the case, the news tweet is deemed either Reliable or Questionable. If the news tweet has replies which contain questions, it is deemed Questionable. If not, it is deemed to be Reliable. If the news tweet is not from a mainstream news account or does not contain a link directing to a mainstream news outlet, the tweet is deemed Unreliable. Utilizing these nudges in a simulated Twitter feed, they found that users in the intervention group assessed the Unreliable news tweets lower than the users in the control group. Additionally, the intervention group rated the Reliable news tweets higher than the control group. These findings indicate that the added nudges can help NudgeCred users to digest the Unreliable news tweets as less credible.

3. CREDTWI: A BROWSER PLUGIN FOR INJECTING QUESTIONNAIRES INTO TWEETS

The key technical output of this thesis is a browser plugin, Credtwi. The plugin was designed to be a generalizable research tool that could be used to run crowdsourced studies on social media. Twitter was selected to be the first social media platform of choice as it is an important platform for credibility studies due to its popularity and impact [5, 9, 14, 18, 19].

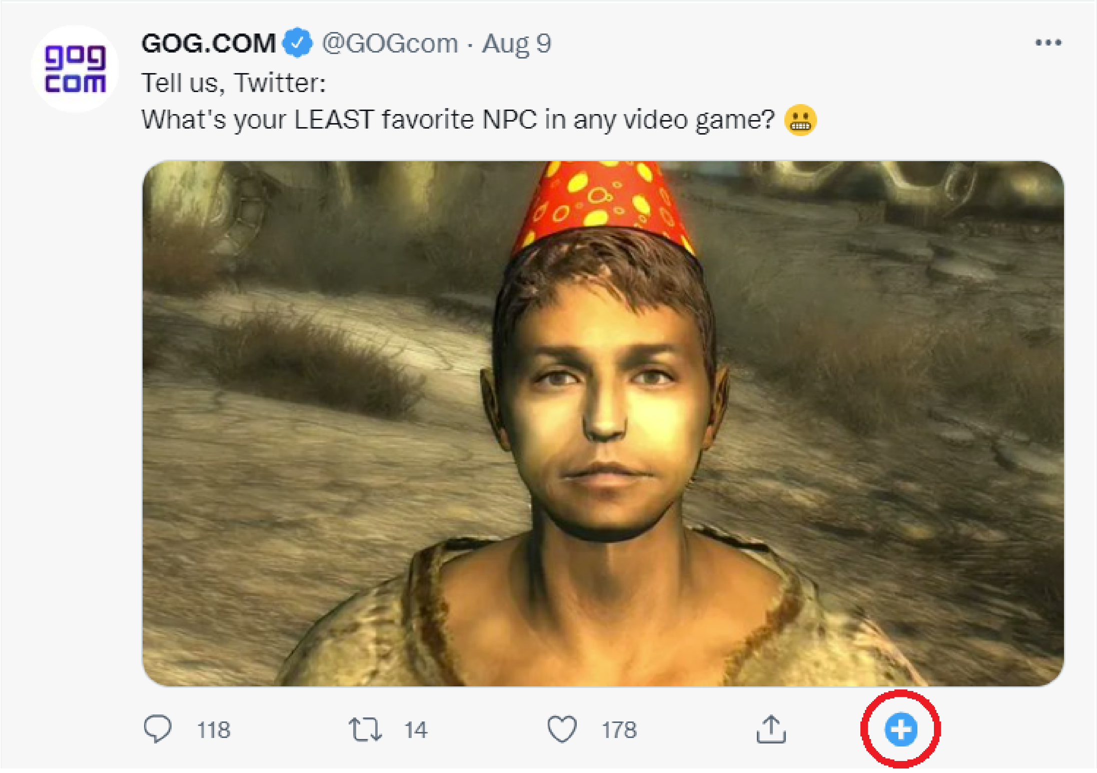
3.1. Plugin Design

The questionnaire injected into tweets by the plugin is designed to be customizable to serve different research needs. For this study, the participants were tasked to assess the credibility of tweets that they come across through various topics provided by the plugin. The questionnaire asks the user to rate the credibility of the tweet in question on a scale from 1 to 7: 1 being not at all credible and 7 being extremely credible. This rating scale lands between the Threshold and Spectral Evaluation models [7], as there are definite points in the scale where the tweet is assessed as not credible (1) or credible (7) for which the Threshold Evaluation model is known. But as there is a linear numeric scale between those two extremes, it also resembles the Spectral Evaluation model. Underneath the rating scale, there is an open-ended answer box, which prompts the user to elaborate on their selected credibility score. This open-ended answer box allowed the participants to elaborate on their numeric assessments.

The plugin was designed to be used in two ways: through its browser plugin popup and through the plugin button located in each tweet. The browser plugin popup contains the buttons for the onboarding questionnaire, the post-experiment questionnaire when applicable, and the three daily trends. The plugin button which is located in each tweet on the user's Twitter opens the questionnaire for the user. Details on these are provided in the following subsections.

The plugin database was designed to contain the following information for each of the assessments:

- Unique identifier (UID)
- Tweet's web address or Uniform Resource Locator (URL)
- Given open-ended answer
- Given credibility score on a scale from 1 to 7
- Time it took for them to evaluate and type out the questionnaire answer in milliseconds



GOG.COM @GOGcom · Aug 9

Tell us, Twitter:

What's your LEAST favorite NPC in any video game? 😬

Figure 2. The added button in the interaction bar of a tweet is marked with a red circle.

3.2. Plugin Implementation

The plugin was implemented utilizing Javascript, its library jQuery [1] , HTML, and CSS. The database was implemented using the cloud platform Heroku[2], and the open-source relational database management system PostgreSQL[3]. Heroku provided an easy way to host an online database and it had a ready-made add-on for using the Heroku PostgreSQL database service. A plugin itself can not interact with the database, so an API was needed to handle the conversation between the two. This was implemented using the open-source Node.js[4] which is a back-end runtime environment. As Node.js enables the usage of Javascript, it was the environment of choice.

The plugin added a button to the interaction bar included in each tweet. This button was in the shape of a blue circle with a plus sign in the middle. The button was designed to fit Twitter's color scheme. This added button can be seen in Figure 2 marked with a red circle. The button was added each time a new tweet appears in the user's timeline.

Dynamic loading of content performed by Twitter poses a unique challenge for code injection. Our solution listens for code changes and injects code into all newly loaded tweets. The appearance of new tweets was monitored with a MutationObserver[5] interface which interacts with the Document Object Model (DOM) tree. The DOM tree represents the web page document with a logical tree containing nodes. These nodes are HTML elements of the web page. The MutationObserver monitors the DOM tree

[1] https://jquery.com/

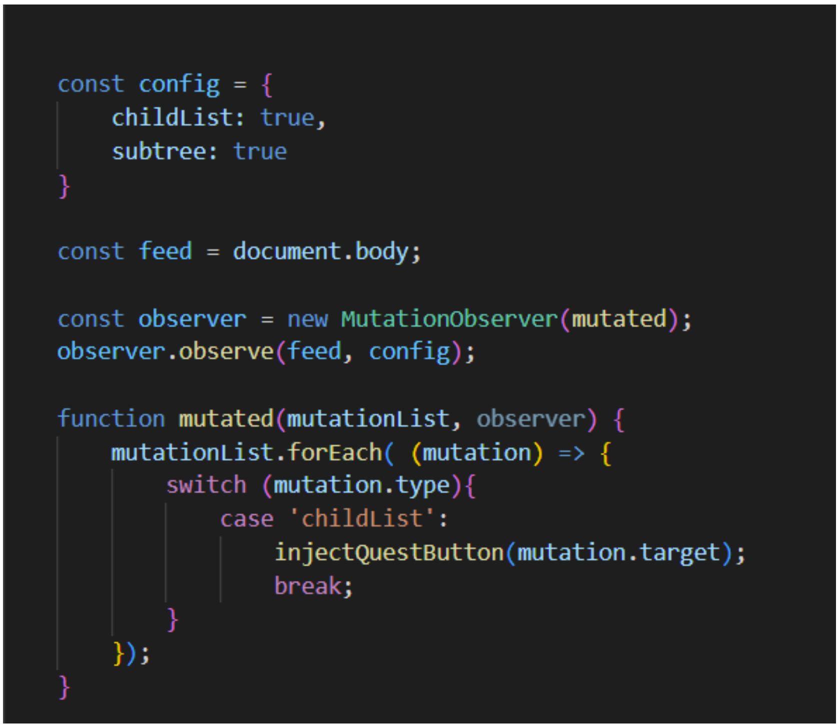
[2] https://heroku.com

[3] https://www.postgresql.org/

[4] https://nodejs.org/en/

[5] https://developer.mozilla.org/en-US/docs/Web/API/MutationObserver

for selected changes. The plugin creates a new MutationObserver which was given the Twitter document's body node as an observation target. The MutationObserver then monitors the target node and all of its child nodes if a new child node was added. If the MutationObserver notices a new child node in the target node tree, it will add the plugin buttons to each of the visible tweets if they do not have one. The configuration and the usage of the MutationObserver can be seen in Figure 3. When the MutationObserver notices a new child node in the DOM tree, the mutated function will be called. The code goes through each of the nodes where the mutation occurred and attempts to add the plugin button into the target if it is a tweet through the injectQuestButton function call.

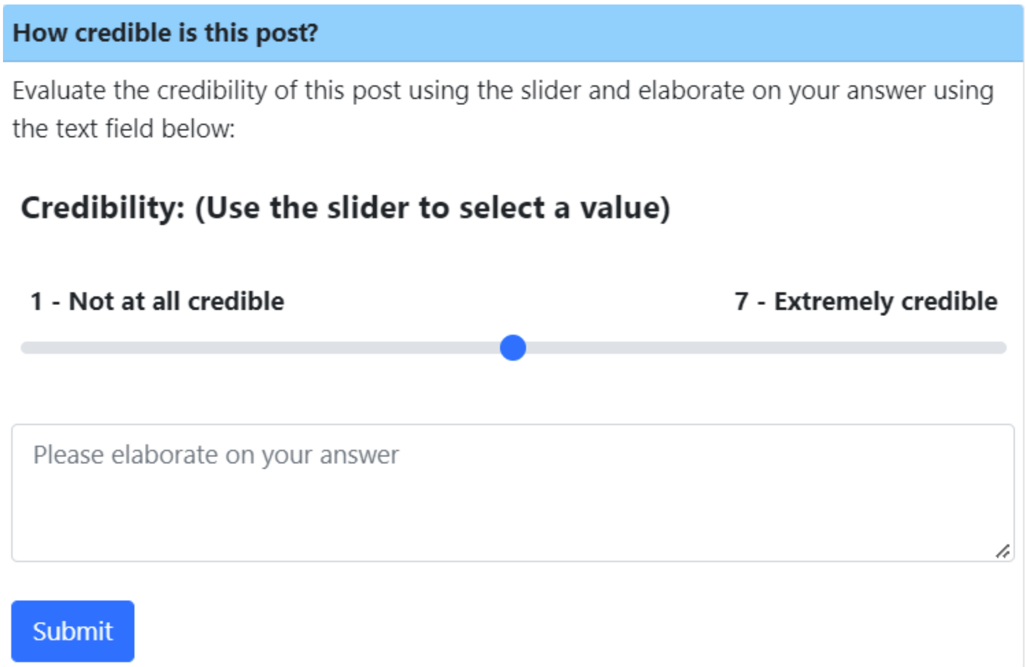


```
const config = {
    childList: true,
    subtree: true
}

const feed = document.body;

const observer = new MutationObserver(mutated);
observer.observe(feed, config);

function mutated(mutationList, observer) {
    mutationList.forEach( (mutation) => {
        switch (mutation.type){
            case 'childList':
                injectQuestButton(mutation.target);
                break;
        }
    });
}
```

Figure 3. The source code snippet for using MutationObserver to handle injecting content to dynamically appearing tweets.

When pushing the added plugin button, the plugin adds the questionnaire HTML element underneath the tweet in question. The questionnaire element can be seen in Figure 4. At the top of the questionnaire, there is the question being asked of the user. In this iteration, it was "How credible is this post?" Following that is a sentence to explain the use of this questionnaire to the user. The tweet's credibility is given by using a range slider from 1 to 7. The selected credibility score is shown on top of the slider. Before anything is selected the score section is empty and nudges the user to use the slider. Below the slider is the open-ended answer section which makes the user justify or elaborate on their answer. The assessment is sent to the database by pushing the Submit button.

The plugin generates a random 256 bits long token which is used as the unique identifier (UID) for each user. This UID is then stored in the browser's plugin memory



How credible is this post?

Evaluate the credibility of this post using the slider and elaborate on your answer using the text field below:

Credibility: (Use the slider to select a value)

Figure 4. The implemented plugin questionnaire.

using chrome.storage API [6]. The UID is used to link the assessments to the participant so that if there is any malevolent usage of the plugin, the participant can be identified.

The plugin popup receives the trends and the completion status of the study from the database. These trends were added to the plugin popup trend buttons as links. These daily trends lead the user to a new Twitter page with the corresponding trend searched. When the study had been completed, the post-experiment questionnaire was made available through the plugin popup by changing the completion status in the database. This completion status is stored as a Boolean value meaning it can be one of two possible values: true or false. At the start of the study, the status value will be false, which means that the "Final questionnaire" button is not visible or enabled in the plugin popup. When the study has run its course, the status can be changed manually in the database to true and the "Final questionnaire" button will appear on the plugin popup.

The information flow between the plugin, the API, and the database is visualized in Figure 5. Each of the requests is marked as a solid arrow showing the direction of the request. The requester is at the base of the arrow and the requestee is the entity where the arrow is pointing towards. Each request from the plugin to the API is followed by another request from the API to the database. The dotted arrows denote the response to the request above it in the diagram. When the user opens the plugin popup, both of the top two requests are sent to the API and from there to the database. The requests are sent via the Fetch API[7]. The requests are done with the GET request method which is directed to the relevant Uniform Resource Identifiers or URIs. In the case of the study status, the URI was the database URL where the Node.js API was also located with the path "/studystatus". The API then sends a query to the database and receives the Boolean value in the "study_status_table". The API then sends the received value

[6] https://developer.chrome.com/docs/extensions/reference/storage/

[7] https://developer.mozilla.org/en-US/docs/Web/API/Fetch_API

as Javascript Object Notation or JSON to the plugin. The trends are requested by the plugin with the GET request method from the API. The API then sends a query to the database for all of the trends from the "trends_table". The "trends_table" contained three different URLs manually changed daily by the study administrator. These trends are received as URL strings in JSON by the plugin. The plugin then places these URLs into the relating trend buttons on the plugin popup. Each time a new assessment is made, the plugin calls the POST method to "/ratings" URI. The API then inserts the new assessment and its data into the database "tweet_ratings_table". The plugin receives a status code 201 when this is done.

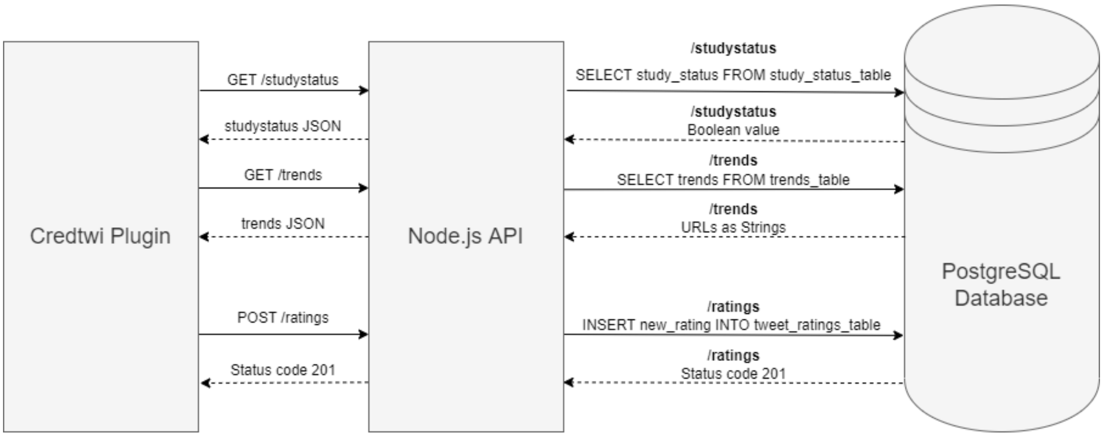
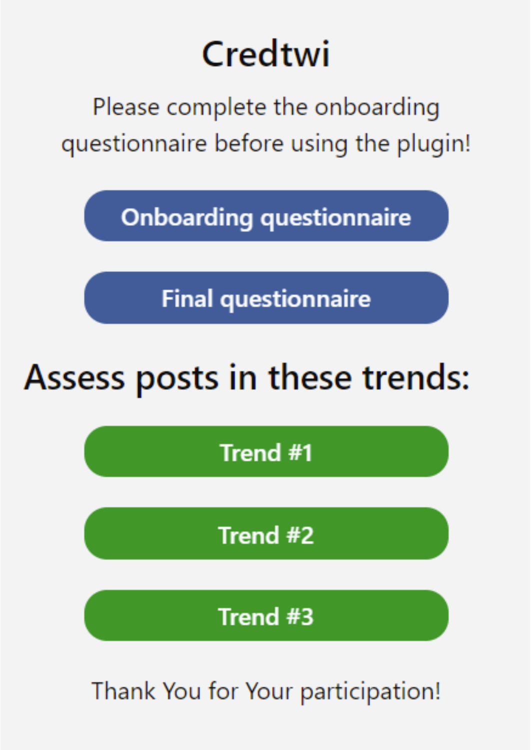


Figure 5. Diagram of information flow through the implemented plugin, API, and database.

The plugin popup can be seen in Figure 6. This is opened when the user clicks the plugin icon in their browser's plugin bar. At the top of the popup, there is the name of the plugin and brief instructions on how to use the plugin. Below them are the two questionnaire buttons. If the study is still ongoing and the study status queried from the database is negative, the "Final questionnaire" button will not be visible. Below these buttons are the trend buttons. Each of the numbered "Trend" buttons contains an URL to a Twitter search conducted using trending hashtags or relevant search words. These trends are used to give the participants something to assess. At the bottom of the plugin window, there is a message thanking the participant for their participation in the field study.

The plugin was designed to be generalizable for future studies on Twitter. This generalizability comes in the form of customizable question forms (shown in Figure 4) which are attached to tweets and additionally, guiding the participants to the content you want them to see through customizable links in the plugin popup's buttons (shown in Figure 6. For these aspects, this plugin can be used in varied studies which are concerned with content on Twitter. The plugin and the API source codes are publicly available on GitHub (https://github.com/EetuHuusko/Credtwi).



Credtwi

Please complete the onboarding questionnaire before using the plugin!

Assess posts in these trends:

Thank You for Your participation!

Figure 6. The implemented plugin popup.

3.3. Preliminary Testing

The plugin was submitted for review to be published on Chrome Web Store[8]. This was done to ease the distribution of the plugin during the testing and study phase. The plugin was reviewed by the Chrome Web Store reviewers in order to make sure the plugin complies with the Chrome Web Store developer program policies[9] and does not contain malware. As the plugin did not violate any policies, it was approved for publishing.

After publishing the plugin in the Chrome Store, we encouraged colleagues at the Crowd Computing research group in the Center for Ubiquitous Computing at the University of Oulu to test the plugin. After tinkering with the plugin on their Twitter pages, one major problem was discovered almost immediately: when opening and closing a tweeted picture on one's Timeline, the Timeline jumps to the beginning and doesn't anymore load more than five tweets. This problem meant that the plugin broke Twitter Timeline's infinite scrolling capabilities. This problem was surprisingly due to a redundant CSS script loading when opening a picture, causing parts of the Twitter website layout style to reset. This resulted in the Twitter Timeline collapsing on itself. This was fixed by removing the CSS script in question from the plugin package.

A pilot study was run with two participants who used the plugin for a day to test out the new version of the plugin. The participants were paid £3.40 each for their participation. The pilot study indicated that the core components of the plugin and field study setup were working correctly. However, there emerged a new problem:

[8] https://chrome.google.com/webstore/category/extensions

[9] https://developer.chrome.com/docs/webstore/program_policies/

the UID was not transferred over to the onboarding questionnaire if the user was too fast in opening the plugin popup and clicking the "Onboarding questionnaire" button. This was due to the nature of the code's asynchronous UID generation, which was not finished when the onboarding button was clicked for the first time. This problem was fixed by disabling the button as default and enabling it when the UID was generated and/or loaded from the Chrome storage.

After these two preliminary testing phases, a new fixed version of the plugin was uploaded and published to Chrome Web Store.

3.4. Study Design

The study followed a mixed-methods design, where we collected both quantitative and qualitative data from participants, i.e. the plugin users. The open answer field was mandatory for each of the assessment instances. From these open answers, we wanted to better understand what factors affect users' perception of credibility.

The study was designed to be run in a 7-day period where participants would rate the credibility of tweets on their Twitter timeline. During the 7-day study period, the plugin had daily changing topics that the participants could access. These daily topics contained at least one topic from the five topics used in previous studies by our research group and derived from a Finnish book concerned with common inaccurate online health claims [41]. These topics were modified to work as search terms on Twitter. The five modified topics from this book were "red processed meat", "vitamin D cancer", "aluminum vaccines", "fish oil depression", and "healthy foods". The remaining 16 topics were top weekly and monthly topics from the United States and the United Kingdom gathered from GetDayTrends [10]. We manually selected topics that were not related directly to music, TV shows, movie releases, or sports events. The topics were checked for ensuring that they directed to English content on Twitter. These daily trends are listed in Table 1. The trends which contain a hashtag symbol at the beginning are the ones that have been selected from GetDayTrends.

The participants of the study were recruited from the crowdsourcing platform Prolific. Prolific is a major online participant pool that can supply researchers with thousands of trusted participants on demand. Prolific is utilized by research centers such as Harvard and Yale in their crowdsourcing studies. The participants were a standard sample from all available countries offered by Prolific. They were prescreened using Prolific's filters for being fluent in English, using Twitter, and tweeting at least 4 times a year. The fluency prescreening filter was selected for the participants to be able to understand what the tweets are meaning, as the daily trends were from English-speaking countries. It was necessary also that the participants were able to convey their justifications for the assessments as clearly as possible. The Twitter prescreening filters were used for the participants to be familiar with using Twitter. These participant sample configurations can be seen in Figure 7.

At the start of the study, the participants were directed via Prolific to an online document that had the instructions for the entire study. First, the participants were informed about what the field study was about and what they would need to do in

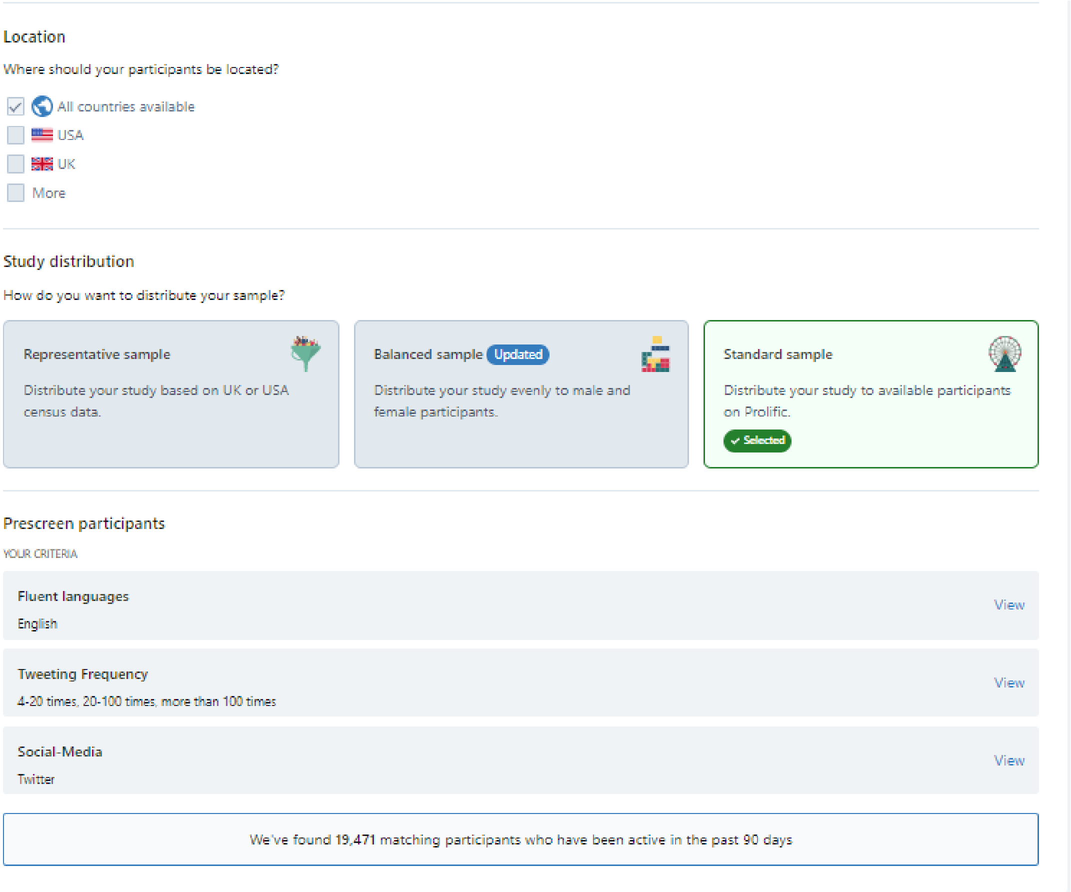
[10] https://getdaytrends.com/

Table 1. The used daily trends

Day	Trend or Topic
1	red processed meat #Eurovision #BansOffOurBodies
2	vitamin D cancer #crypto #EarthDay
3	aluminum vaccines #WeareallIdrissa #RoevWade
4	fish oil depression #MothersDay #FreePalestine
5	healthy foods #monkeypox #JohnnyDeppVsAmberHeardTrial
6	#HillaryForPrison #auspol #StopTheTreaty
7	#WorldBeeDay #AmberHeardlsApsychopath #BillGatesBioTerrorist

order to finish the field study. Secondly, the participants were instructed to download the plugin from Chrome Web Store, where the plugin was publicly available. If the participants did not have Google Chrome, a link to download it was provided. After downloading the plugin, the participants needed to fill out an onboarding questionnaire, which stored their Prolific IDs, plugin UIDs, and demographic information. After completing the onboarding questionnaire, the participants were informed how to use the plugin via instructions and instructive photos. The participants were instructed to assess at least 10 tweets per day. As we understood that not every one of the participants could be available every day for the duration of the field study, we were lenient on this aspect of the study. The participants were instructed not to assess advertisement tweets. After instructions on the usage of the plugin, the participants were informed about the post-experiment questionnaire and when would it be available. Lastly, contact information was provided if the participants had any questions or problems regarding the field study.

After a week of participating in the study, the post-experiment questionnaire was made available for the participants through the plugin popup. As there were no notification capabilities with the plugin, the participants were reminded of the necessity of completing the post-experiment questionnaire through the Prolific messaging system.



Location

Where should your participants be located?

☑ All countries available
□ USA
□ UK
□ More

Study distribution

How do you want to distribute your sample?

Representative sample
Distribute your study based on UK or USA census data.

Balanced sample Updated
Distribute your study evenly to male and female participants.

Standard sample
Distribute your study to available participants on Prolific.
✓ Selected

Prescreen participants

YOUR CRITERIA

Fluent languages
English
View

Tweeting Frequency
4-20 times, 20-100 times, more than 100 times
View

Social-Media
Twitter
View

We've found 19,471 matching participants who have been active in the past 90 days

Figure 7. The participant sample configuration on Prolific.

3.4.1. Questionnaire Design

Onboarding questionnaire

At the start of the study, the participants were instructed to complete an onboarding questionnaire. In the onboarding questionnaire, the participants were asked about their demographic information, what social media platforms they used, and how they perceived the credibility of Twitter as an information source. The demographic part of the questionnaire is based on Pew Research's Demographics Questionnaire [42]. As there is no established international standard for ethnicity classification [43], the suggested classification from the National Content test was used in the questionnaire [44, 45].

The onboarding questionnaire asked the participants:

- Prolific ID
- Credtwi UID (Provided by the plugin)
- Gender
- Nationality
- Marital status

- The highest academic qualification received
- Race, origin or ethnicity
- Employment status
- Annual income level
- Used social media platforms
- "How do you rate your self in terms of assessing credibility of social media posts, i.e. your critical social media reading skills?"
- "In your own opinion, how credible do you think Twitter, in general, is as an information source?"

The questions "How do you rate your self in terms of assessing credibility of social media posts, i.e. your critical social media reading skills?" and "In your own opinion, how credible do you think Twitter, in general, is as an information source?" were assessed with a 1 to 7 Likert scale. For "How do you rate your self in terms of assessing credibility of social media posts, i.e. your critical social media reading skills?" 1 was "not at all good" and 7 was "extremely good". For the question "In your own opinion, how credible do you think Twitter, in general, is as an information source?" the scale ranged from "Not at all credible" to "Extremely credible". The question "In your own opinion, how credible do you think Twitter, in general, is as an information source?" was in both of the questionnaires, onboarding, and post-experiment.

Post-experiment questionnaire

In the post-experiment questionnaire, the participants were asked to again assess the general credibility of Twitter as an information source. We were interested in if and how the perceived credibility of Twitter changes after using the assessment plugin for a week. After that, there was a Likert scale rating assignment for how much author and tweet features affect the credibility of the tweet. The scale of how much the features affected the credibility went from 1 or "not at all" to 7 or "extremely".

The author's features were derived from a previous study by Morris et al. and modified to fit the question we were asking "Please rate these factors by how much they affect the credibility of a Tweet. (In this context, the Author refers to the user who has sent the tweet)" [36]. The author features the participants were asked to assess were:

- Author's bio
- Author's verified status
- Author's profile picture
- Author's gender
- Author's ethnicity
- Author's nationality

- Author's follower count
- Author's tweet count

The tweet features were derived from a previous study by Shariff et al. and modified to fit the question we were asking "Please rate these factors by how much they affect the credibility of a Tweet." [40]. The tweet features the participants were asked to assess were:

- Tweet is well-written
- Tweet uses hashtags
- Tweet is a retweet
- Tweet has links to outside sources
- Tweet has pictures or other media attached

After each of these feature rating questions was an open-ended question box for the participants to elaborate or justify their ratings. These elaborations were followed by the question "What thoughts, if anything, concerning critical reading skills, do you have as a result of using Credtwi for a week?" This question was meant to gauge the feelings and thoughts the week-long field study had risen in the participants. After that question, the participants were asked "Would you recommend Credtwi as a tool to help people self-reflect on their critical reading skills?" The options were "Yes", "No", and "Maybe". From this question, we could know how did the participants feel about Credtwi. This question was followed by an open-ended question box for the participants to elaborate on their single-word answers. Lastly, the participants were asked, "How could a similar plugin such as Credtwi be used to improve people's critical reading skills?" From this question, we wanted to get new ideas on how to further develop Credtwi.

Plugin question form

The plugin contained a small question form which was added to each tweet. This question form can be seen in Figure 4. In this field study, we wanted the participants to assess the credibility of the tweets they come across on different topics during the week. For this field study, the question form asked the participant "How credible is this post?". The question form was connected to the tweet in such a way that the participant knew what tweet was being referenced by the question form. Below the question, there are instructions for the participant in order for them to understand what this question form is used for. The question form instructed the participants to assess the credibility of the tweet using a slider ranging from 1 to 7. This range slider was used as the Likert scale for the assessments. The slider had descriptions for each of the assessment extremes: the lowest was "Not at all credible" and the highest was "Extremely credible". Underneath the assessment slider, there was an open-ended answer box for the participants to elaborate on their given assessment. This open-ended answer box was used to gather qualitative data from the participants on their justifications for their assessments.

3.4.2. Qualitative Analysis Methodology

The participants were asked to rate Author and tweet-specific factors on how much they affect the credibility of a tweet. The participants were urged to elaborate on their ratings in an open-ended answer below. The open-ended answers from the post-experiment questionnaire are analyzed using conventional content analysis described by Hsieh and Shannon [46]. The answers were read to derive codes that highlight the key concepts. These codes were then sorted into categories for each of the open-ended questions. The definitions of these categories were then crafted. The same methodology was applied to a random sample of open-ended answers from the tweet assessments. The random sample was a collection of 183 assessments after removing one-worded answers from the selection.

4. DATA ANALYSIS

4.1. Data Cleansing

The data that was provided by participants for the onboarding questionnaire, but who did not provide the data for the post-experiment questionnaire, was removed. There were 5 participants who only provided the onboarding questionnaire data. Additionally, some participants did the onboarding questionnaire multiple times on multiple days. In these cases, the first submission was selected and the remaining submissions were removed. One participant answered their sex as "other" and elaborated the answer with an open answer that indicated that they misunderstood the question. The participant's demographic information was available through Prolific and was thus corrected.

4.2. Participants

A total of 30 participants started the study and 25 participants finished the study by submitting the post-experiment questionnaire. The participants were paid £11.00 each for their week-long participation and £4.00 bonuses were paid to participants who provided either high-quality answers or more answers than they were expected to. With the estimated time it took to participate daily, the average participant was paid 10.20£ per hour. The average age of the participants was 27.5, the youngest participant being 19 years of age and the oldest participant being 39 years of age. Of the participants, 16 reported themselves as male, and 9 reported themselves as female. The significantly largest portion of participants, the number being 16, was from South Africa. The second-largest portion was from Italy at 4 participants, then Portugal at 3 participants. Germany and Poland had 1 participant each. 15 participants reported their ethnicity to be black or African American, 9 reported to be white, and 1 reported to be white with a Hispanic, Latino, or Spanish origin. The large majority of the participants, at 21, reported never being married. The rest 4 of the participants were married. Out of the participants, 13 had a Bachelor's degree as their highest academic qualification, 7 had a Master's degree, 3 had a High school diploma, 1 had a professional degree, and 1 had a diploma. Most of the participants, numbering 15, were employed for wages, 5 were self-employed, 4 were students, and 1 was out of work and looking for work. The annual income level of 14 of the participants was reported to be below \$20,000, 7 participants reported their annual income level to be between \$20,000 and \$40,000, 3 participants reported their income level to be between \$40,000 and \$60,000, and 1 participant reported their income level to be between \$80,000 and \$100,000. The average self-assessment the participants gave themselves for their critical reading skills on social media was 6.08 on a scale from 1 to 7, 1 being not at all good, and 7 being extremely good. This demographic information of the participants is displayed in Table 2.

Before the participants started to use the plugin, they were asked about their social media usage in the onboarding questionnaire. Almost all of the participants, 24 out of 25, used Twitter. The second most used platforms were Instagram and Whatsapp with 23 participants using them. The third most used platform was Youtube with

22 participants using it. Other social media platforms used by the participants were Facebook with 16 participants, TikTok with 12 participants, Facebook Messenger with 10 participants, Telegram with 10 participants, Pinterest with 8 participants, Reddit with 8 participants, and Snapchat with 3 participants. QQ, Quora, and WeChat had 1 participant out of the 25. The social media usage can be seen in Table 3.

Table 2. Demographic information from the participants

Respondents	25
Average age	27.5
Minimum age	19
Maximum age	39
Gender:	
Male	16 (64%)
Female	9 (36%)
Nationality:	
South Africa	16 (64%)
Italy	4 (16%)
Portugal	3 (12%)
Germany	1 (4%)
Poland	1 (4%)
Ethnicity:	
Black or African American	15 (60%)
White	9 (36%)
White: Hispanic, Latino, or Spanish Origin	1 (4%)
Marital status:	
Never married	21 (84%)
Married	4 (16%)
Highest academic qualification:	
Bachelor’s degree	13 (52%)
Master’s degree	7 (28%)
High school diploma	3 (12%)
Professional degree	1 (4%)
Diploma	1 (4%)
Employment status:	
Employed for wages	15 (60%)
Self-employed	5 (20%)
A student	4 (16%)
Out of work and looking for work	1 (4%)
Annual income level:	
<$20,000	14 (56%)
$20,001 - $40,000	7 (28%)
$40,001 – $60,000	3 (12%)
$80,001 – $100,000	1 (4%)
Self-assessment of critical social media reading skills	
Average score (SD)	6.08 (0.81)

Table 3. Information related to the use of social media from the participants

	Number of participants (Percentage of participants)
Used social media platforms:	
Twitter	24 (96%)
Instagram	23 (92%)
Whatsapp	23 (92%)
YouTube	22 (88%)
Facebook	16 (64%)
TikTok	12 (48%)
Facebook Messenger	10 (40%)
Telegram	10 (40%)
Pinterest	8 (32%)
Reddit	8 (32%)
Snapchat	3 (12%)
QQ	1 (4%)
Quora	1 (4%)
WeChat	1 (4%)

4.3. Quantitative Analysis

After a week of using the plugin, a total number of 1830 assessments were gathered from the participants. The average credibility score given for tweets was 4.55 (SD = 2.18). The average time spent on providing the assessments was 54.2 seconds (SD = 83.0), the fastest time spent being 2.5 seconds, and the slowest time spent being 1437.3 seconds. The average length of the open-ended answer given as a justification for the number score was 66.5 characters (SD = 53.1), with 4 characters being the shortest justification, and 455 characters being the longest. This data can be seen in Table 4.

As indicated by previous studies [33, 34], there are differences in how males and females perceive information online. To investigate these differences, we grouped the data between the two genders present in the data and calculated the differences between the assessment characteristics. There were on average more assessments per participant per gender by females (102.4 per participant) than by males (65.8 per participant). A density plot of the given assessment scores is shown in Figure 8. The density plot shows the distribution of the assessment scores. The vertical dotted lines in the density plot show the mean scores for each of the genders. The mean assessment score given by males (M = 4.33, SD = 2.16) was lower than the scores given by females (M = 4.96, SD = 2.12). As these two groups are independent, the significance of this difference was analyzed with the Mann-Whitney U-test. The Mann-Whitney U-Test showed that there was a statistically significant difference (W = 402039, $p < 0.001$) between the two groups given assessment scores. The results can be seen in Table 5.

We were also interested in the difference in length of the open-ended answers for the assessments between the two groups. The open-ended answer lengths were measured in the number of written characters. The mean answer length for an assessment was longer for females (M = 89.81, SD = 59.67) than for males (M = 46.30, SD = 33.37). A density plot of the assessment justification answer lengths is shown in Figure 9. The

Table 4. Descriptive data from the assessments

Assessments	1830
Average score (SD)	4.55 (2.18)
Score:	
1	279 (15,2%)
2	178 (9,7%)
3	166 (9,1%)
4	97 (5,3%)
5	318 (17,4%)
6	325 (17,8%)
7	467 (25,5%)
Time:	
Average (SD)	54.2 (83.0)
Minimum	2.5
Maximum	1437.3
Open answer length:	
Average (SD)	66.5 (53.1)
Minimum	4
Maximum	455

density plot shows the distribution of the assessment justification answer lengths. The vertical dotted lines in the density plot show the mean answer lengths for each of the genders. The Mann-Whitney U-Test showed that there was a statistically significant difference (W = 257635, p < 0.001) between the two groups' answer lengths given for assessments. The results can be seen in Table 5.

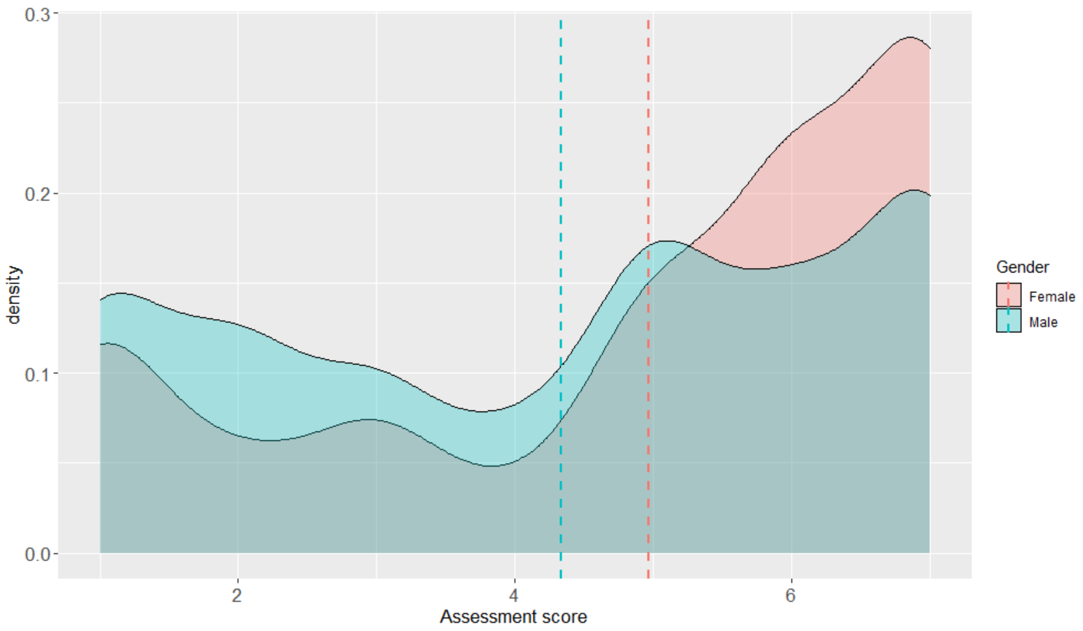


Figure 8. The density plot of given assessment scores for the genders.

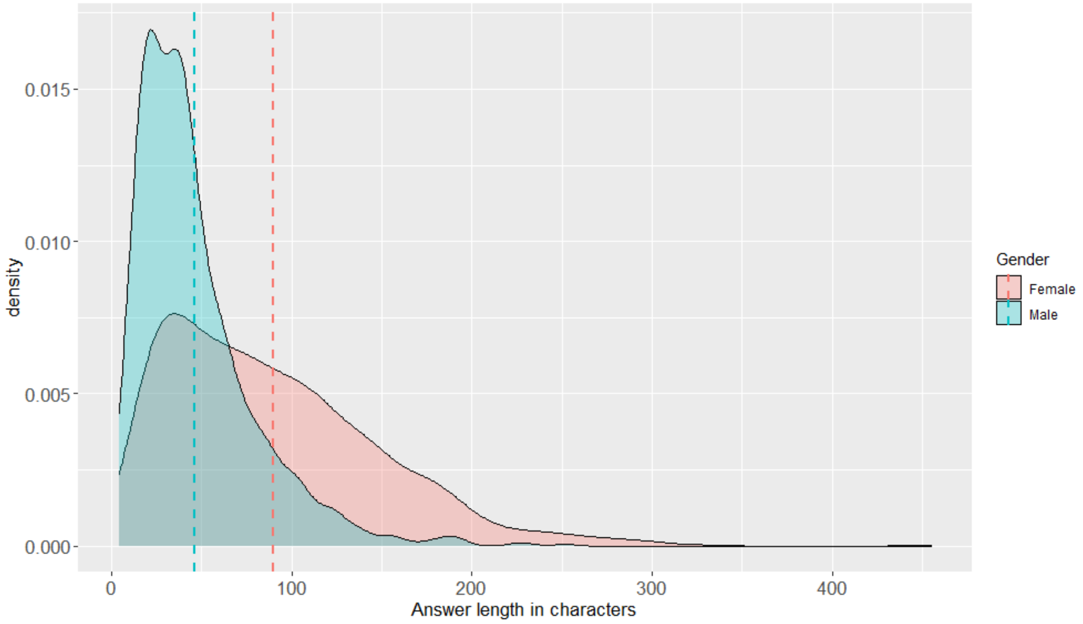


Figure 9. The density plot of assessment justification answers lengths for the genders.

A linear regression model was fitted to investigate whether the participant's perception of the credibility of Twitter as an information source predicted their assessment scores given through the plugin. The fitted regression model was: Assessment score = 2.98 + 0.34*(participant's given credibility score of Twitter). The overall regression was statistically significant (R^2 = .040, F(1,1973) = 82.63, p < 0.001). It was found that the participant's given credibility scores for Twitter as an information source predicted their given assessment scores through the plugin (β = 0.34, p < 0.001). The results from this analysis can be seen in Table 6.

The participants were asked before and after using the plugin for a week "In your own opinion, how credible do you think Twitter, in general, is as an information source?". The selected scale for this question was a Likert scale from 1 to 7, 1 being "Not at all credible" and 7 being "Extremely credible". The Likert scores and percentages of the answers are visualized in Figure 10. Before using the plugin, 12% of the participants rated Twitter 1 through 3 on the scale, 40% of the participants rated it as a neutral 4, and 48% of the participants rated it 5 through 7 on the scale. After a week of using the plugin, 32% of the participants rated Twitter 1 through 3 on the scale, 32% of the participants rated it as a neutral 4, and 36% of the participants rated it 5 through 7 on the scale.

Table 5. Differences in assessment scores and time spent between genders using the Mann-Whitney U-test (Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)

Assessments (n=1830)	Male	Female	*p*-value
Score: Mean (SD)	4.33 (2.16)	4.96 (2.12)	< 0.001***
Answer length: Mean (SD)	46.30 (33.37)	89.81 (59.67)	< 0.001***

Table 6. Linear regression model between the assessment scores and the perception of the credibility of Twitter as an information source (Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)

Model:	R^2	Adjusted R^2	Std. error of est.	
1.	.040	.040	0.037	
Coefficients:	Estimate	Std. error	t value	*p*-value
1. (Constant)	2.98	0.19	15.90	< 0.001***
Credibility	0.34	0.037	9.09	< 0.001***

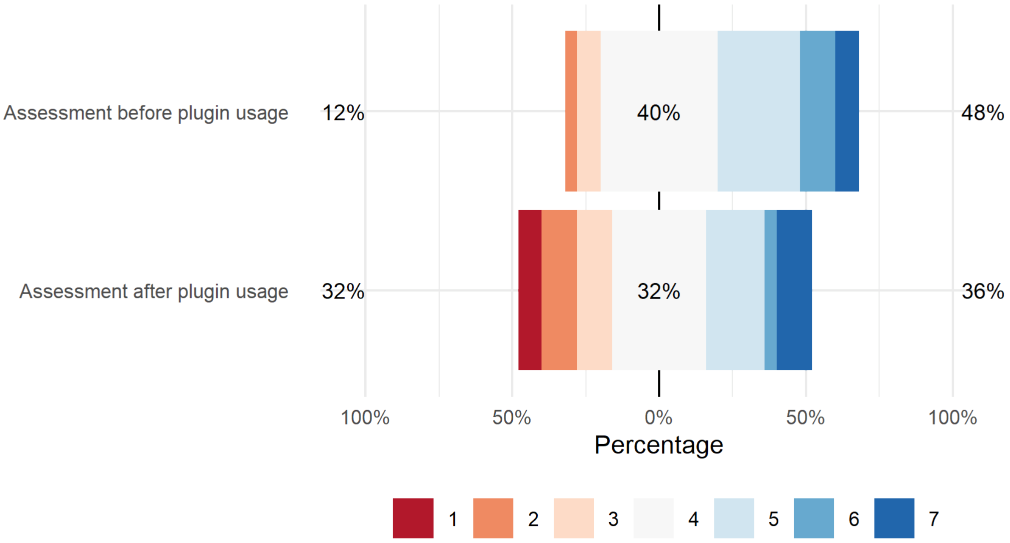


Figure 10. The perceived credibility before using the plugin and after using the plugin for a week.

The mean value of the perceived credibility of Twitter as an information source before using the plugin was 4.60 (SD = 1.19). After using the plugin for a week, the mean of perceived credibility was 4.04 (SD = 1.70). To assess the statistical significance of the differences between the two non-independent variables, the Wilcoxon signed-rank test was used [47]. Table 7 shows the results of the Wilcoxon signed-rank test. The difference was statistically significant (p < 0.05). The effect size of the difference was 0.184, which means that the effect is relatively small [48].

Table 7. Effect on perceived credibility of Twitter (Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)

Participants (n=25)	Before Mean (SD)	After Mean (SD)	Effect size	*p*-value
Credibility	4.60 (1.19)	4.04 (1.70)	0.184	0.034*

The participants were asked to rate how much different factors relating to the author of the tweet affect the credibility of the Tweet. There were 9 different factors that were rated on a scale from 1 to 7, 1 being "Not at all", 4 being "Moderately", and 7 being "Extremely". The 9 different factors were the author's bio, verification status, profile picture, gender, ethnicity, nationality, follower count, and tweet count. The Likert scores and percentages are visualized in Figure 11. The participants rated the author's verified status as the most effective factor with 76% of the participants rating it from 5 to 7. The second most effective factor was the author’s bio. The author’s follower count, profile picture, and tweet count were rated as moderately effective factors by the participants. The less effective factors were the author’s nationality, ethnicity, and gender. The author’s gender was rated as the least effective factor with 92% of the participants rating it from 1 to 3 on the scale. There were significant differences between the factors as tested with a Kruskal Wallis test ($p < 0.01$).

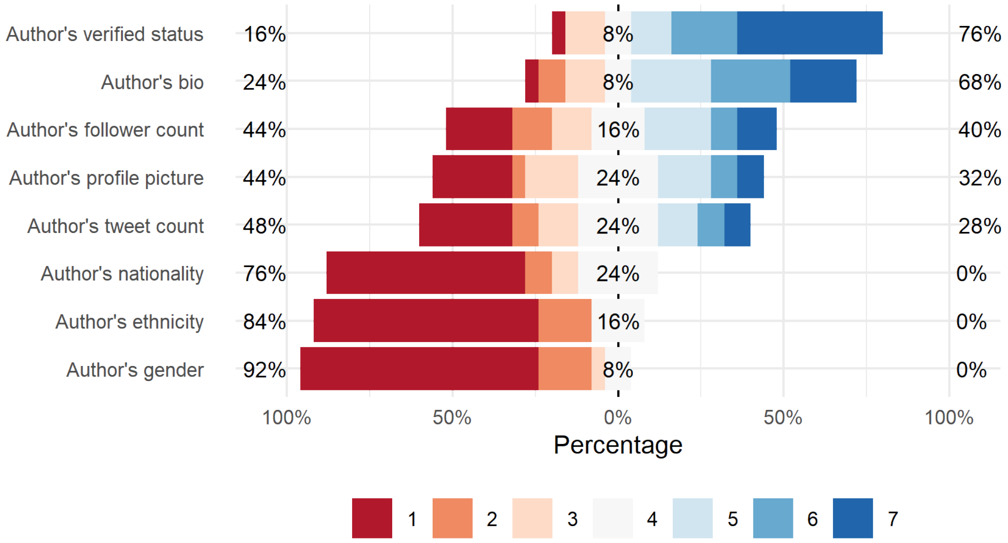


Figure 11. Likert visualization on how much author factors affects the perceived credibility of a Tweet.

The participants were asked to rate how much different factors relating to the content of the tweet affect the credibility of the Tweet. There were 5 different factors that were rated on a scale from 1 to 7, 1 being "Not at all", 4 being "Moderately", and 7 being "Extremely". The 5 different factors were "Tweet is well written", "Tweet uses hashtags", "Tweet is a retweet", "Tweet has links to outside sources", and "Tweet has pictures or other media attached". The Likert scores and percentages are visualized in Figure 12. The participants rated the factor "Tweet has links to outside sources" as the most effective factor with 84% of the participants rating it from 5 to 7 on the scale. The factors "Tweet is well written" and "Tweet has pictures or other media attached" were rated as more than moderately effectual. The factor "Tweet is a retweet" and "Tweet uses hashtags" were rated less than moderately effectual with the factor "Tweet uses hashtags" being the least effectual factor. There were significant differences between the factors as tested with a Kruskal Wallis test ($p < 0.01$).

The participants were asked after a week of using the Credtwi plugin "Would you recommend Credtwi as a tool to help people self-reflect on their critical reading skills?" 19 (76%) participants answered "Yes", 5 (20%) participants answered "Maybe", and 1 (1%) participant answered "No". The distribution of these answers can be seen in Figure 13.

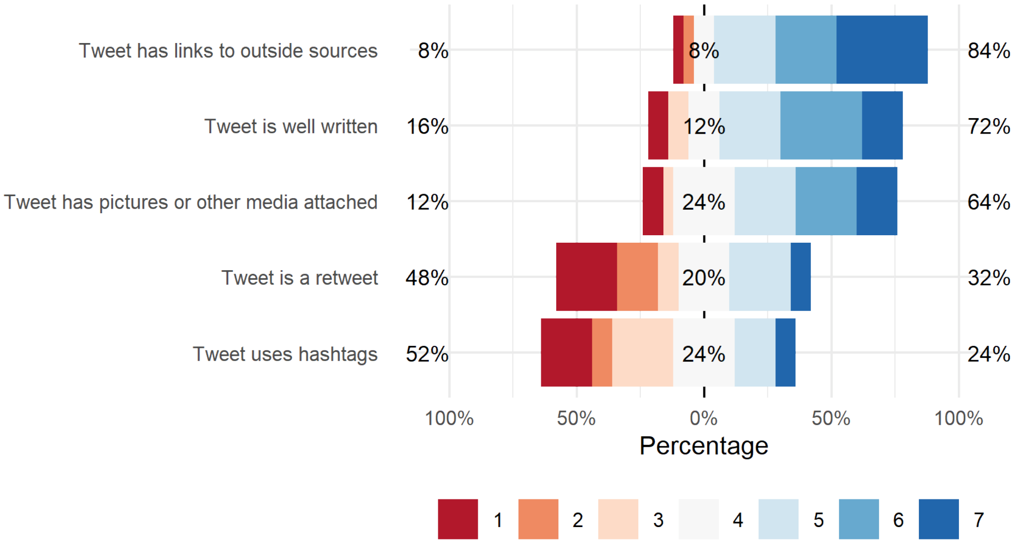


Figure 12. Likert visualization on how much tweet factors affects the perceived credibility of a Tweet.

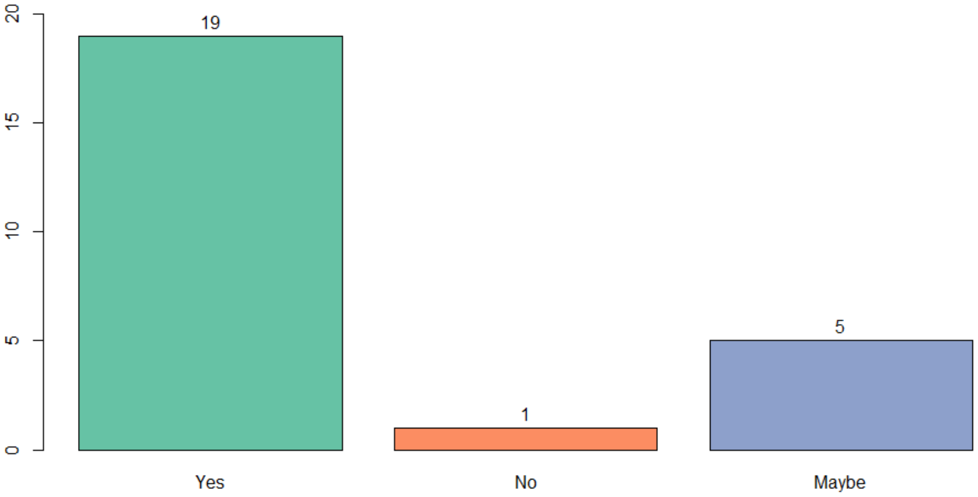


Figure 13. Answers to the question "Would you recommend Credtwi as a tool to help people self-reflect on their critical reading skills?"

4.4. Qualitative Analysis

Each of the open-ended question answers was separately coded and analyzed. These codes were then separated into different encompassing categories. In this section, the most prominent categories from each of the open-ended questions are highlighted.

4.4.1. Factors Affecting Author Credibility

The following categories were derived from the open-ended answers given to the credibility factors relating to the author:

Author's relevancy. The author's bio, profile picture, and account name give contextual information to the reader to determine if the author is credible. Authors that have experience, interests, or a career in a field that is relevant to the tweet in question are perceived as more credible. The closeness of the geographical location of the author in relation to the content of the tweet is also noted as being a possible credibility factor.

> *To help determinate[sic] if a tweet is truthful, I tend to check their bio to see how they would know or be aware about the information, if they're somehow close to what's happening...*
>
> – Female, Italy, 26

Verification status. The verification status of the author is an important factor when discerning if the author is credible. The participants voiced their concerns about fake accounts and impersonators, and a verified author alleviates those concerns. Verified accounts are perceived as more credible and have something to lose if they spread untruthful information.

> *If the person has a verified account I take it that the person has done their research before bringing information to the platform.*
>
> – Female, South Africa, 28

Following. The Twitter activity of the author and their number of followers impact their perceived credibility. If the author has a larger following, they are perceived to be more credible. The larger following also means that the author's tweets can be checked and scrutinized by public opinion.

> *While the success of the person who tweets is not a guarantee, if they have lots of followers, their content is checked and criticized by more people.*
>
> – Male, Italy, 27

4.4.2. Factors Affecting a Tweet's Credibility

Concerning the different tweet content factors affecting the credibility of the tweet, the following categories were observed from the answers:

Information source. When the information in the tweet is accompanied by the source where it came from, the tweet is perceived to be more credible. The participants also mentioned that if the information source is provided via a link and can be checked by the reader, it increases their perceived credibility.

> *When a tweet has more sources linked to it, anyone is able to verify the information written and see if it is reliable or not.*
>
> – Male, South Africa, 19

Supportive media. The credibility of the tweet is increased if there are relevant media, e.g. pictures, graphs, or video, accompanying the text content. With Twitter's character limit, the amount of information one tweet can hold is limited. As the saying goes: "A picture is worth a thousand words."

> *With media, is the photo/video related to information, if [sic] makes me believe them because not only captures attention but sometimes the issue is explained more easily there, unless the pictures are just random.*
>
> – Female, Italy, 26

Language. The structure and the language used in the tweet affect its perceived credibility. If the tweet is well-written, it is perceived as more credible. If the tweet contains curse words, it is perceived as less credible. Better structure and language also help the reader to better understand the message being conveyed.

> *The language is very important, because it allows us to understand better the message, (...) like someone took time to actually put well written data to let us know about something.*
>
> – Female, Italy, 26

Wisdom of crowds. As the tweet has been already liked and shared by other users, the tweet is seen to contain some valuable information. If the tweet survives through public scrutiny with users willing to spread the content of the tweet along, the tweet is perceived as more credible.

> *If the tweet has a large number of engagement that proves that they may have alot[sic] of people agreeing with what they have just tweeted.*
>
> – Female, South Africa, 28

4.4.3. Credtwi Improvement Suggestions

In addition to inquiring the participants about their assessment of the effects of different factors on credibility, they were also asked what thoughts did this week-long study bring up. Also, they were asked what kinds of things they might want to see added to a plugin like Credtwi. Following are the main categories organized from these questions.

What thoughts, if anything, concerning critical reading skills, do you have as a result of using Credtwi for a week?

Improved critical reading skills. Participants mention that after a week of using the Credtwi plugin, they are more careful when reading posts on social media. They also more frequently look for clues that might affect the credibility of posts. The method of justifying the reason for the credibility scores also helped the users to reflect on what they perceived as credible. The fact that justifying or reflecting on your given credibility score is mandatory forced the participants to really think about why they perceive credibility the way they do. The participants found it to be a good learning exercise to realize what factors they value when assessing the credibility of information on Twitter.

> *It surely makes you think twice about what you read on Twitter, or any social media actually.*
>
> – Male, Italy, 26

> *Over the course of the week I think Credtwi allowed me to be more analytical in my evaluation of reliable/credible sources because it forced me to quantify and think through various posts.*
>
> – Female, Portugal, 22

Information variety. Being exposed to different trends along the week and having to assess their credibility made the participants notice that there is a lot of false information on Twitter. The participants also noted that there exists a lot of opinion-based content along with true and false information, These personal opinions on various topics are hard to assess on a single dimension credibility scale.

> *Well mostly that everything you read is not necessarily true. A lot of information shared consists of personal opinions and is not evidence-based.*
>
> – Male, South Africa, 33

How could a similar plugin such as Credtwi be used to improve people's critical reading skills?

Guidance. The participants felt providing more guidance or information on how to detect fake news and non-credible information would be of help. This additional information in addition to the reflection gained by doing the assessments could allow the users to better detect fake news and non-credible information.

> *Having a guide to help spot usual fake news language.*
>
> – Male, Italy, 25

Affirmation. Having the ability to see how others have assessed the credibility of a certain tweet could be used to affirm the perceived credibility of the user. When something you perceive as credible would be labeled as non-credible or wise versa, it would make the user think more about what they are missing in the context of the credibility or if others have missed something they have perceived.

> *By providing stats around how other users are responding. By seeing how others rate the tweets would help get an indication of the validity of the tweet.*
>
> – Male, South Africa, 33

4.4.4. Tweet Assessments

A random sample was selected from the 1830 tweet assessments given by the participants. This random sample contained 183 assessments. These are the main categories observed from the open-ended answers.

Reputable source. The importance of providing sources for the news, statements, and information in the tweets was overwhelmingly the largest factor brought up by the participants. But just having the source in the tweet was not enough to warrant a credible assessment, the source also needed to be perceived as credible by the participants.

> *This is tweeting an article from a credible source.*
>
> – Male, South Africa, 22

Author. The relevancy, competency, occupation, and general credibility of the author were brought up regularly in the answers. If the participant felt the author had relevant knowledge in the field that the tweet was addressing, they perceived it to be more credible. The famousness of the author was perceived as a hindrance for them to be able to share false information. If the author of the tweet was verified, they were perceived to be more credible, and they were perceived as a real person rather than a fake account.

> *This tweet is from a public health scientist, this account is verified, so it is credible.*
>
> – Female, South Africa, 30

Information validity. The participants assessed the information given in the tweets as false, correct, or somewhere in between. This perception of the factual nature of the information was emphasized if the information was researched or not. Some participants would assess personal opinions and experiences as credible, and some would assess them to be non-credible and deem them to be not evidence-based. The participant's own prior knowledge of the information was also brought up as a reference point. Also if the participant agreed with the information or opinion given in the tweet played a role in how they perceived the credibility of said tweet.

> *The information contained in this tweet seems to be logical and I would agree with the sentiments being shared.*
>
> – Male, South Africa, 33

Supportive media. The pictures and videos included in the tweets were an important part of the perception of credibility for the participants. The supportive media gave context to the information, provided a visual proponent to the accompanying text, and helped the participants to better understand tweets that were based on data.

> *It looks credible, there's a video of the activist speaking and defeding[sic] their posture.*
>
> – Female, Italy, 26

5. DISCUSSION

The two objectives of this study were introduced earlier:

- O1: Design, implement, and test a Google Chrome plugin that allows crowdsourcing research on Twitter
- O2: Use the implemented plugin in a field study to gather data

Both of these objectives were carried out during the course of this study. O1 was fulfilled with the designing, implementation, and testing of the Credtwi plugin. The implementation of Credtwi followed the initial plugin design without large modifications needing to take place. The testing of Credtwi was done in two parts: the first was done by the Crowd Computing research group in the Center for Ubiquitous Computing at the University of Oulu, and the second was done by two recruits from Prolific.

After completing O1, the Credtwi plugin was used as a research tool to gather field study data from participants on Prolific. The field study was designed to be a credibility assessment study. In this study, the participants used the Credtwi plugin for a week assessing around 10 tweets per day. At the start of the study, the participants completed an onboarding questionnaire concerned with their demographic information. After completing the study, the participants filled out the post-experiment questionnaire that inquired about their opinions about how different author and tweet factors affect the perceived credibility of a tweet. This post-experiment questionnaire also asked the participants their thoughts about Credtwi and future suggestions for the development of Credtwi. The main data gathered was the credibility assessments provided by the participants through Credtwi. After a week of usage, a total number of 1830 assessments were gathered. This successful field study completed O2.

5.1. Answering Research Questions

This study had three research questions:

- RQ1: Does reflecting tweet credibility with a plugin affect one's perception of social media credibility?
- RQ2a: What factors affect the author's perceived credibility?
- RQ2b: What factors affect the tweet's perceived credibility?
- RQ3: What implications for future crowdsourced credibility research with browser plugins can we derive from the study?

RQ1 was answered by gathering and analyzing the participants' perceptions of social media credibility before and after the use of Credtwi. The analyzed data shows that there is a statistically significant effect on the perception of one's social media credibility after using a plugin that urges the user to reflect on tweet credibility. After using Credtwi for a week, the participants' mean perception of social media credibility was lower than before, but the effect size was small. Due to not having a control group, the effect of the usage of the plugin itself can not be made certain.

The two parts of RQ2 were answered by analyzing the data gathered from the post-experiment questionnaire. According to the participants, the most important factors affecting the author's perceived credibility were the author's verified status and the author bio. The least important factors were the author's ethnicity and the author's gender. According to the participants, the most important factors affecting the tweet's perceived credibility were that the tweet has links to outside sources and that the tweet is well written. The least important factors were that the tweet is a retweet and that the tweet uses hashtags. To understand the factors affecting the credibility assessments, a random sample from the participants' assessments was content analyzed. This was done to see what factor categories were most prominent and if there were any novel factors affecting the participant's perceived credibility. The most prominent categories from the credibility assessment justifications were the reputable source, the author, the information validity, and the supportive media. These factors from the qualitative analysis of the final questionnaire and credibility assessments were mentioned as prominent factors for credibility in previous studies [27, 30, 35, 36, 40].

As per RQ3, we wanted to better understand crowdsourced credibility research with browser plugins. The field study we carried out highlighted some key points to take into account when designing future research. If the content for which the participants are assessing their perceived credibility is varied, a multidimensional credibility scale might be in order. A credibility scale regarding only credibility might not translate well from tweets containing factual information into tweets containing jokes or personal experiences. A single-dimensional credibility assessment scale would work well in studies that concern themselves with false information such as fake news. As noted by previous studies [23, 38], additional cognitive load increases the user's accuracy of sharing correct information. This additional cognitive load in the form of thinking and writing an answer for justifying the credibility assessment was felt as a positive exercise by the participants. They felt the need for justifying the given assessment score and had them reflect on what aspects of the tweet they regard when assessing the credibility. Thus, adding cognitive load into future research designs is a good way of nudging the user's to think about what factors they value and deem credible. The usefulness of nudging users when assessing social media credibility has been found in previous browser plugin studies [9, 20]. Additionally, as the use of Credtwi lowered the participants' perceived credibility after a week of use, this might have to be taken into account when screening participant samples for future research.

The assessment data was grouped by the genders present in the data. There were significant differences between the male and female participants' assessments. The male participants gave lower credibility scores on average than the female participants. Prior research has shown that males have a more positive attitude towards informativeness and source credibility was more significant for males [33, 34]. These two factors might explain the more critical assessments from males. The male participants gave shorter answers than females as their justifications for their assessments. Additionally, the female participants contributed much more assessments on average per participant during the field study than their male counterparts.

5.2. Limitations

We wanted to investigate would an assessment intervention affect the participant's perceived credibility of the social media platform where the intervention took place. In this thesis, that platform was Twitter. One limitation our investigation had was the absence of a control group who would see the same topic-related tweets but would not assess them with Credtwi. This addition of a control group would have solidified the effect the assessment intervention had on the perceived general credibility. But as we wanted to gather assessment data from all participants, this was not done. From this study, we can not say if the effect of the general credibility of Twitter lowering was due to the usage of Credtwi or if through trending topics the participants saw content that lowered their perceived credibility.

Investigating gender differences with as small a sample size as in this field study was a notable limiting factor. The demographic sample in the field study was not balanced by gender as 64% of participants were male. For these reasons, it is hard to declare the gender differences found in quantitative analysis to be definite. Further research would be in order to investigate these differences deeper and with a larger sample size.

The credibility scale being a single-dimensional Likert scale brought difficulties when assessing the credibility of information and tweets which did not convey false or correct information or news. This issue was raised by a few of the participants as they felt that a single-dimensional assessment scale was not sufficient for assessing for example personal experiences or opinions shared by users on Twitter. Some participants assessed these as being credible and some assessed them as being not credible. The same could be said about assessing satire, humor, jokes, and memes. These kinds of posts are hard to be assessed if they are credible or not.

There were limitations in the data collection of the plugin. The plugin stored only the URL of the tweets, the perceived credibility score, and the user's justification for this score. As the tweet's content was not collected when the assessment was done, if the tweet was deleted, the context for the assessment was lost. This includes other contextual metadata like the number of likes and shares of the tweet. The tweets were later collected using screenshots, but the context information could have changed between the time of assessment and the screenshot.

There is a limitation in emphasizing the benefit of adding cognitive load for participants on the credibility assessment as the additional cognitive load itself was not measured. This additive cognitive load is derived from the fact that participants are forced to justify the assessment in a written form which adds complexity to having just a score slider.

5.3. Future Work

In order to make Credtwi a really versatile research tool, it would need easier modification options to its core components. The question form the researchers want to ask the participants was hard coded into the plugin source code as HTML. Making the procedure of changing this question form easier would broaden the applicability of the plugin. At the moment of this study, you would need to be somewhat code-savvy in order to make the necessary changes for the plugin to work in another context.

Further studies are in order to investigate the change in the perceived credibility of Twitter after using the plugin for a week. This might be due to the nature of the credibility assessment where the question of "How credible is this post?" makes the user suspect the credibility of the information at hand. Or, this decrease in perceived credibility might be caused by exposure to divisive trending topics.

There were several ideas for future additions to the plugin. One of the most requested additions the participants mentioned was the inclusion of some kind of gauge that would indicate how other users have assessed the tweet. This addition could influence the perceived credibility of the posts as the users are given an additional heuristic. Another addition mentioned by the users was a metric that would show how much the user has contributed to their assessments. This kind of gamification of assessment could motivate the users to contribute more data. On the other hand, it might lead to users giving lower quality data with an incentive that rewards the number of assessments given.

The plugin and the accompanying database should be broadened to store important metadata from the tweets at the time of assessment. This metadata could include the number of likes and shares, the text content of the tweet, and a screenshot of the tweet in question. This way the assessment can be examined in the same context as the users and if the tweet is deleted or removed, the assessment context is not lost.

A notification system for the plugin would be beneficial in the terms of study usage so that you can remind your participants to partake in daily assessments during a multi-day study. During the field study, the participants were reminded manually through the Prolific's messaging channel, which was time-consuming for the research conductor. The participants mentioned that they appreciated these reminders. If Credtwi would be made into a general usage credibility assessment tool, notifications could be used to entice the user to assess more tweets or notify when there were new tweets for the user to evaluate.

The plugin could be extended with natural language processing techniques to give real-time sentimental analysis of the tweet's text content. In this way, the plugin could provide sentimental analysis straight to the database when the assessment is made. This could be useful for data analysis for the researchers utilizing Credtwi.

5.4. Author's Reflections on the Progress

This thesis has taught me a lot about browser plugin development and the factors that come into play when dealing with multiple parts of a web software package. The importance of testing and multiple different point-of-views was highlighted when trying to work out the kinks in the plugin as one is predisposed to be blinded to certain things when working intensely on it for a long time. If even the problem does not seem to make sense from where you are looking, you have to change your perspective or try to divide the problem into smaller pieces.

During the field study, the importance of clear instructions was realized. It seemed that the participants from Prolific were not used to doing crowdsourcing studies lasting several days. Thus, at the end of the first day's activities, many participants were confused about how to finish the study. Even though this point of the study lasting for a week was mentioned several times in the instructions, it seemed that it went unnoticed.

Generating infallible instructions is a never-ending task, but one can always improve from before.

6. CONCLUSION

In this thesis, we designed and implemented a research tool browser plugin designed to conduct crowdsourcing studies on Twitter. After successfully implementing the plugin named Credtwi, a week-long field study was run utilizing participants from the crowdsourcing platform Prolific. The participants were tasked to assess the credibility of tweets pertaining to specific changing topics. During this week-long field study, the participants provided a total number of 1830 assessments. The participants also completed two questionnaires during the study. Comparing the participant's perceived credibility of Twitter as an information source before and after a week of using Credtwi, there is a small but significant decrease in the perceived credibility. The participants provided open-ended answers with their assessments as justifications for their answers. These justifications were mainly concerned with tweets having reputable sources, the credibility of the author, the validity of the information, and supportive media in the tweet. These justifications were supported by the qualitative analysis of the final questionnaire factors affecting credibility. From the process of implementing Credtwi and running a field study on it, we derived a few implications for future crowdsourced credibility research with browser plugins. The credibility scale used in the assessment portion has to be thought out properly, the addition of cognitive load can be beneficial for users, and when screening the participant sample, the possibility of credibility plugin usage lowering the participant's perceived general credibility needs to be taken into account.

7. REFERENCES

[1] Napoli P.M. (2015) Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. Telecommunications Policy 39, pp. 751–760.

[2] Chou W.y.S., Prestin A., Lyons C. & Wen K.y. (2013) Web 2.0 for health promotion: reviewing the current evidence. American journal of public health 103, pp. e9–e18.

[3] Shearer E. & Mitchell A. (2021), News use across social media platforms in 2020. URL: `https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/`.

[4] Shao C., Ciampaglia G.L., Varol O., Yang K.C., Flammini A. & Menczer F. (2018) The spread of low-credibility content by social bots. Nature communications 9, pp. 1–9.

[5] Gupta A., Kumaraguru P., Castillo C. & Meier P. (2014) Tweetcred: Real-time credibility assessment of content on twitter. In: International conference on social informatics, Springer, pp. 228–243.

[6] Tseng S. & Fogg B. (1999) Credibility and computing technology. Communications of the ACM 42, pp. 39–44.

[7] Fogg B.J. & Tseng H. (1999) The elements of computer credibility. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 80–87.

[8] Alrubaian M., Al-Qurishi M., Alamri A., Al-Rakhami M., Hassan M.M. & Fortino G. (2018) Credibility in online social networks: A survey. IEEE Access 7, pp. 2828–2855.

[9] Bhuiyan M.M., Zhang K., Vick K., Horning M.A. & Mitra T. (2018) Feedreflect: A tool for nudging users to assess news credibility on twitter. In: Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 205–208.

[10] Mendoza M., Poblete B. & Castillo C. (2010) Twitter under crisis: Can we trust what we rt? In: Proceedings of the first workshop on social media analytics, pp. 71–79.

[11] Gupta A., Lamba H., Kumaraguru P. & Joshi A. (2013) Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web, pp. 729–736.

[12] Kouzy R., Abi Jaoude J., Kraitem A., El Alam M.B., Karam B., Adib E., Zarka J., Traboulsi C., Akl E.W. & Baddour K. (2020) Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. Cureus 12.

[13] Vosoughi S., Roy D. & Aral S. (2018) The spread of true and false news online. Science 359, pp. 1146–1151.

[14] Paschalides D., Christodoulou C., Andreou R., Pallis G., Dikaiakos M.D., Kornilakis A. & Markatos E. (2019) Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web. In: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, pp. 298–302.

[15] Johnson T.J. & Kaye B.K. (2015) Reasons to believe: Influence of credibility on motivations for using social networks. Computers in human behavior 50, pp. 544–555.

[16] Del Vicario M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G., Stanley H.E. & Quattrociocchi W. (2016) The spreading of misinformation online. Proceedings of the National Academy of Sciences 113, pp. 554–559.

[17] Tsao S.F., Chen H., Tisseverasinghe T., Yang Y., Li L. & Butt Z.A. (2021) What social media told us in the time of covid-19: a scoping review. The Lancet Digital Health 3, pp. e175–e194.

[18] Dongo I., Cardinale Y. & Aguilera A. (2019) Credibility analysis for available information sources on the web: A review and a contribution. In: 2019 4th International Conference on System Reliability and Safety (ICSRS), IEEE, pp. 116–125.

[19] Hartwig K. & Reuter C. (2019) Trustytweet: an indicator-based browser-plugin to assist users in dealing with fake news on twitter. Proceedings of the International Conference on Wirtschaftsinformatik (WI) , pp. 857–871.

[20] Bhuiyan M.M., Horning M., Lee S.W. & Mitra T. (2021) Nudgecred: Supporting news credibility assessment on social media through nudges. Proceedings of the ACM on Human-Computer Interaction 5, pp. 1–30.

[21] Counting characters, twitter developer platform. URL: `https://developer.twitter.com/en/docs/counting-characters`.

[22] Pennycook G. & Rand D.G. (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition 188, pp. 39–50.

[23] Pennycook G., McPhetres J., Zhang Y., Lu J.G. & Rand D.G. (2020) Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science 31, pp. 770–780.

[24] Sillence E., Briggs P., Harris P.R. & Fishwick L. (2007) How do patients evaluate and make use of online health information? Social science & medicine 64, pp. 1853–1862.

[25] Fogg B.J., Marshall J., Laraki O., Osipovich A., Varma C., Fang N., Paul J., Rangnekar A., Shon J., Swani P. et al. (2001) What makes web sites credible? a report on a large quantitative study. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 61–68.

[26] Hilligoss B. & Rieh S.Y. (2008) Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Information Processing & Management 44, pp. 1467–1484.

[27] Metzger M.J., Flanagin A.J. & Medders R.B. (2010) Social and heuristic approaches to credibility evaluation online. Journal of communication 60, pp. 413–439.

[28] Ma T.J. & Atkin D. (2017) User generated content and credibility evaluation of online health information: a meta analytic study. Telematics and Informatics 34, pp. 472–486.

[29] Graefe A., Haim M., Haarmann B. & Brosius H.B. (2018) Readers' perception of computer-generated news: Credibility, expertise, and readability. Journalism 19, pp. 595–610.

[30] Jung E.H., Walsh-Childers K. & Kim H.S. (2016) Factors influencing the perceived credibility of diet-nutrition information web sites. Computers in Human Behavior 58, pp. 37–47.

[31] Hämäläinen E.K., Kiili C., Marttunen M., Räikkönen E., González-Ibáñez R. & Leppänen P.H. (2020) Promoting sixth graders' credibility evaluation of web pages: an intervention study. Computers in Human Behavior 110, p. 106372.

[32] Armstrong C.L. & McAdams M.J. (2009) Blogs of information: How gender cues and individual motivations influence perceptions of credibility. Journal of Computer-Mediated Communication 14, pp. 435–456.

[33] Sun Y., Lim K.H., Jiang C., Peng J.Z. & Chen X. (2010) Do males and females think in the same way? an empirical investigation on the gender differences in web advertising evaluation. Computers in Human Behavior 26, pp. 1614–1624.

[34] Yin C., Sun Y., Fang Y. & Lim K. (2018) Exploring the dual-role of cognitive heuristics and the moderating effect of gender in microblog information credibility evaluation. Information Technology & People .

[35] Mohd Shariff S., Zhang X. & Sanderson M. (2014) User perception of information credibility of news on twitter. In: European conference on information retrieval, Springer, pp. 513–518.

[36] 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.

[37] Mitra T. & Gilbert E. (2015) Credbank: A large-scale social media corpus with associated credibility annotations. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, vol. 9, pp. 258–267.

[38] Westerman D., Spence P.R. & Van Der Heide B. (2014) Social media as information source: Recency of updates and credibility of information. Journal of computer-mediated communication 19, pp. 171–183.

[39] Gao Q., Tian Y. & Tu M. (2015) Exploring factors influencing chinese user's perceived credibility of health and safety information on weibo. Computers in human behavior 45, pp. 21–31.

[40] Shariff S.M., Zhang X. & Sanderson M. (2017) On the credibility perception of news on twitter: Readers, topics and features. Computers in Human Behavior 75, pp. 785–796.

[41] Knuuti J. (2020) Kauppatavarana terveys–selviydy terveysväitteiden viidakossa. Helsinki: Minerva Kustannus Oy .

[42] Center P.R. (2015), Pew research center demographic questions web or mail mode 12-29-2015. `https://assets.pewresearch.org/wp-content/uploads/sites/12/2015/03/Demographic-Questions-Web-and-Mail-English-3-20-2015.pdf`.

[43] Connelly R., Gayle V. & Lambert P.S. (2016) Ethnicity and ethnic group measures in social survey research. Methodological Innovations 9, p. 2059799116642885.

[44] Mathews K., Phelan J., Jones N.A., Konya S., Marks R., Pratt B.M., Coombs J. & Bentley M. (2015) National content test: Race and ethnicity analysis report. US Department of Commerce, Economics and Statistics Administration, US Census Bureau .

[45] Jones N.A. & Bentley M. (2017) Overview of the 2015 national content test analysis report on race & ethnicity. US Census Bureau, Suitland-Silver Hill, MD .

[46] Hsieh H.F. & Shannon S.E. (2005) Three approaches to qualitative content analysis. Qualitative health research 15, pp. 1277–1288.

[47] Bauer D.F. (1972) Constructing confidence sets using rank statistics. Journal of the American Statistical Association 67, pp. 687–690.

[48] Tomczak M. & Tomczak E. (2014) The need to report effect size estimates revisited. an overview of some recommended measures of effect size. Trends in sport sciences 1, pp. 19–25.