University of Zurich[UZH]

‘Trust Me, I Am A Doctor’: The Credibility Of Doctor Titles On Twitter

Chuqiao Yan
Zurich, Switzerland
Student ID: 20-740-691

Supervisor: Stefania Ionescu
Date of Submission: April 28, 2023

MASTER THESIS – Social Computing Group

University of Zurich
Department of Informatics (IFI)
Andreasstr. 15, CH-8050 Zürich, Switzerland

ifi

Master Thesis
Social Computing Group (SCG)
Department of Informatics (IFI)
University of Zurich
Andreasstr. 15, CH-8050 Zurich, Switzerland
URL: https://www.ifi.uzh.ch/en/scg.html/

Abstract

The measurement of credibility for Twitter content has gained significant attention due to the difficulty in verifying the accuracy of posts, particularly those made by users who identify themselves as experts by including titles such as Dr.' or M.D.' in their display name. This study aimed to investigate three research questions. First, we assessed the credibility of users who display qualified titles on Twitter. Next, we analyzed the types of viewers who are most susceptible to the influence of such users, and finally we proposed strategies that can be used by actual 'Dr.' titled users to enhance their credibility on the platform. To gather data, a between-subject experiment and a survey were designed and conducted. The results indicate that users with professional titles in their display names are perceived as more credible than those without such titles. Additionally, the study found that individuals who have never used Twitter before are the most impacted by Twitter content. Our study suggests that real 'Dr.' titled users can increase their credibility by including a relevant bio in their profile and by including paper links in their tweets. By doing so, these users can more effectively persuade the public of their expertise.

Acknowledgments

I would like to express my greatest appreciation to my supervisor Stefania Ionescu who provides guidance, support, suggestion and comments during my work of this Master Thesis. I would also like to express my gratitude to Dr. Aleksandra Urman, Prof. Dr. Aniko Hannak and the entire Social Computing Group for their fully support, as well as this opportunity for me to work on this interesting topic.

Contents

Chapter 1

Introduction

Over the past decade, social media has taken an integral part of our lives. According to a study from Statista, 147 minutes per day was the average amount of usage that people worldwide spent on social media in 2022 [37]. Today, social media is widely used for a variety of purposes. Individuals use it to share personal life, exchange opinions with others; organisations use it to build public figures, implement media marketing strategies and promote professional networks in academia, etc. [8]. Social media provides users with a quick access to information, and enables the sharing of information with the public [30].

Among the increased availability of social media, Twitter has become the most popular microblogging social media service, gathering millions of users to publish and exchange information in a short format [5]. On Twitter, information can be propagated in real-time, and users can interact reciprocally. Twitter offers an environment to quickly disseminate information, and it has become a direct information source for more than 166 million daily active users [17]. Of those users, 56% used it as source of news and to seek advice from knowledgeable individuals. For instance, Twitter is the most popular social media for healthcare communication since it reduces collaboration barriers by allowing medical professionals to reach a broad audience, and users can easily connect with similar disease communities through the use of hashtags [30].

While social media has many positive functions, it can also easily spreads unconfirmed or inaccurate information. Twitter has interactive features that increase the speed at which false information is displayed to other users, the action of clicking 'Like' or 'retweet' can lead to misinformation spreading exponentially [4]. During the COVID-19 lockdowns, the usage of social media platforms was increased by more than 50% [15], resulting in a flood of ungrounded misinformation about the pandemic. This led to a "Massive Infodemic", announced by World Health Organization, which intensified the uncertainty of the crisis [46]. Access to fake news ultimately leads to unfounded conclusions, and poses significant dangers ranging from the level of individuals to our society. According to other researches [40, 46], people may rely on false information to make decisions, their attitudes and reception of true news can be changed, making it difficult for them to distinguish more credible sources from less credible ones. At another level, the threat of misinformation leads to a devaluation of the entire news system. Thus, it is crucial for online users to

recognize true information from credible sources, in order to prevent harms for the entire information ecosystem.

Fortunately, credibility is essentially believability [14], and believability can be measured by the professionalism of information sources. On Twitter, one of the most direct ways to report professionalism is adding the relevant titles of qualifications in users' display name (e.g., 'Dr.', 'M.D.'). This virtual label brings reputed credibility of the users and enhances the credibility of the information they share [44]. However, it remains unclear whether the presence of such titles in the display name has an effect on the credibility of their tweets and which viewers are impacted the most.

Another area of concern is the gender inequalities of users on Twitter, as gender is a commonly accessed stereotype which plays a role in human judgement [6]. As per the data from January 2021 [17], 68.5% of Twitter users are male, while only 31.5% are female. Suggested by previous studies, the gender of users may play a role in assessing credibility on social media [2], thus gender of users with different titled display names is included in this investigation as well.

Moreover, we acknowledge that users could claim to have a title without this being the case, and verifying such a claim is difficult. For example, it is hard to know if a user with 'Dr.' in their display name truly owns a PhD degree. Therefore, we also investigate what could real 'Dr.' titled users do to convince the public of their qualifications and, thus, further increase their tweets' credibility.

In short, my thesis studies the following three research questions:

RQ1: Does the presence of 'Dr.' and 'M.D.' title in users' display name increase the credibility of their tweets?

RQ2: What types of viewers on Twitter are most influenced by the users with an inclusion of a 'Dr. ' and 'M.D.' in their display name?

RQ3: How can the users who really own a PhD degree do to increase their tweets' credibility?

To answer those research questions, I have developed six related hypotheses which will be investigated via an online experiment. More precisely, I perform an exploratory experiment to investigate the relationship between the titles in users' display name and their associated credibility. An online survey is designed and participants are invited to take part in. During the experiment, participants who join the survey are playing the role of viewers who will read the twitter content and images, while users are portrayed as people who are posting tweets and presented in the form of images.

The six hypotheses are set as below:

Hypothesis 1 Viewers tend to trust tweets more when they come from 'Dr.' and 'M.D' titled users than no titled users.

Hypothesis 2 Viewers tend to trust females more than male users when the users have the same title.

Hypothesis 3 The higher Twitter usage viewers are more impacted by the 'Dr.' and 'M.D.' titled users' tweets statements on Twitter.

Hypothesis 4 Viewers who have higher activity on Twitter are more impacted by the tweets of the users with 'Dr.' and 'M.D' in their display name.

Hypothesis 5 Including valid publication proof in the tweets of users with 'Dr.' in their display name increases the credibility of their tweets.

Hypothesis 6 Viewers tend to trust the 'Dr.' titled users more when those users provide more relevant info in their bio information.

In conclusion, my thesis aimed to measure the credibility of Twitter users who have qualified titles in their display name. I make the following key contributions to achieve this goal. First, I did literature review from prior works and chose methodologies for measuring credibility and designing experiment. Second, I implemented an online experiment with survey to collect data. Third, I did data analysis to find answer for my hypotheses. My work contributes to the existing research on Twitter credibility by focusing on a specific group of users with titles in their display names.

The remaining thesis presents the related work, the methodology for conducting the survey, the procedure for the data collection and analysis, and suggestions for follow-up experiments. The related work and literature review are summarised in Chapter 2. Chapter 3 explains the survey design and preparation process. Chapter 4 describes the methodology used for conducting the survey online and describes the statistical tools chosen to analyze the results. I reserve Chapter 5 to present results of the study, along with the design of a follow-up experiment. The last chapter concludes this study with suggestions for further work.

Chapter 2

Related work

2.1 Background

2.1.1 Information and Its Credibility

Wathen and Burkell [44] defined *information* as the acquisition of meaningful knowledge for individuals who are searching for knowledge that can alleviate uncertainty. It can affect people's attitudes, behaviours, and decision-making. With lots of information encountered on a daily basis, people selectively filter out what they perceive to be useful. In this process, the primary criteria is the credibility of information [44]. For users, the Internet is a widely used method to acquire information. But, with the growing amount of information being spread, it has become difficult to find high-quality information amidst lower-quality information. Consequently, evaluating the credibility of online information has been studied a lot due to its importance and complexity.

The investigation of information credibility has been divided by researchers into three categories: Medium, message, and source credibility [11]. The first refers to the credibility of the specific medium individuals used to receive the information [28, 41]. Message credibility refers to the communicated message itself, including its accuracy and quality [26]. Source credibility refers to the expertise or trustworthiness of the source which provides the information [13, 21]. In this study, we keep the medium fixed and focus on the source and message credibility. First, we vary the perceived credibility of the source by considering users who have or do not have titles in their display names. Second, we modify the tweet content to observe how it affects the message credibility.

Four types of information credibility were identified by Tseng and Fogg [42]. First, *presumed credibility* refers to the extent to which the perceivers' trustworthiness is based on stereotypes of an object or source, such as assumption that 'car salesmen are dishonest'. Second, *reputed credibility* refers to the influence of information reported by third parties. For example, the official titles of 'Doctor' or 'Professor' can often enhance the perceived credibility of individuals via a virtual label. Next, *surface credibility* refers to the receivers' judgment based on a simple inspection of facial characteristics. Finally, *experienced credibility* is based on the receiver's judgment of prior experiences. Disregarding its

type, credibility is usually measured in one of three ways: (a) by directly asking whether the information is believable, (b) via proxy measures of knowledge change by checking whether the information can be recalled, and (c) via proxy measures of attitude change by checking whether the information affects attitudes and behaviours [31]. According to Wathen and Burkell [44], the first direct method was not always used, and the second method was based on a weak assumption that only credible information will be recalled. Therefore, in the main part of this study, we adopt the concept of reputed credibility and the third proxy measurement method of attitude change.

2.1.2 Twitter

Twitter is a widely used social media platform that has an impact on users' behavior in real life. It creates a virtual interactive community for users to communicate and engage with each other. As an information technology-based web application, it has shifted the traditional media platforms to a new type of information dissemination in forms of many-to-many communication [16, 47]. With has both consumption and interactivity built into its structure, it creates an environment to exchange users' generated content, and also integrates social functions of information industries, such as news and advertising [16, 20]. Through Twitter, registered users can post their own tweets which express ideas through messages of 280-character-long maximum. Moreover, they can engage with existing content by directly replying to other tweets and sharing or promoting other tweets through functions "retweet" and "like" [16, 25]. Users can also customize their own profile page by adding a personal short introduction within 160-character-long and a profile photo. Also, every user can choose to "follow" people they are interested in, so that they can know what's happening with the people they are interested in [45]. With these functions, Twitter serves as a great tool to stay connected with people, exchange messages and spread information.

People frequently get news on Twitter during emergency situations and large social events. Twitter's immediacy makes it a perfect source for breaking news updates, such as epidemic tracking, wildfires, floods, earthquakes, etc. [32]. News updates on Twitter provide real-time information which is not yet available in the mainstream media. The "trending topics" section on Twitter also provides a snapshot of sharp increased topics in popularity, similar to headline news in traditional media [19]. However, one of the drawbacks is that misinformation and rumours can also be spread rapidly on Twitter, even leading to negative consequences[4]. For example, the distribution of inaccurate medical information on Twitter can lead people to inappropriate treatment or messed medical care [36]. Thus, it is necessary to study the information credibility on Twitter.

2.2 Credibility Evaluation

Previous studies have provided valuable insights into how credibility can be measured. Wathen and Burkell [44] proposed a three-staged model for users to assess online information credibility. They suggest users first evaluate surface credibility of the website,

includes aspects of design, interface and organization of the website. The second step is evaluating the message credibility itself, includes factors such as information source and content, accuracy, relevance, etc. Finally, they suggest assessing the interaction of the message with user's cognitive state. Ferrel and Castillo [12] adopted three features of the source to measure online physicians' credibility [24]: competence, trustworthiness, and goodwill. These refer respectively to the source's ability to know the truth, the source's motivation to be truthful or biased, and whether the source has the best interest for receivers. These studies offer a useful framework for identifying the aspects of credibility measurement that my work can focus on, among the many categories available. Other studies choose to quantify the measurement of credibility on a scale. Jahng and Littau [16] used a 7-point likert scale to assess credibility of online journalists based on a list of criteria: experience, skill, activity, qualification and competence. A similar measurement can be found in the study by Edgerly and Vraga [9], which quantified the credibility assessment into a 7-point scale to evaluate the content of tweet as complete/incomplete; accurate/inaccurate; unbaised/biased, trustworthy/untrusworthy; credible/not credible; tell the whole story/not tell the whole story. In addition, Morris and Counts et al. [27] assess the Twitter credibility by asking participants answering two direct questions on 7-point likert scales from strongly disagree to strongly agree: "the tweet contains credible information" and whether "the author is credible". These studies have shown that using a 7-point likert scale for measuring credibility is appropriate and can also be utilized in this study.

Most studies measured credibility by conducting an experiment, usually an online survey, where participants do tasks to read tweets or related content, and then give their credibility evaluation. Since there are many factors that could influence the credibility on Twitter, each study always focuses on a small number of perceptions. For instance, a between subjects experiment was conducted by Ferrel and Castillo [12] to compare how a formal and casual appearance on a profile picture influences physicians' credibility on twitter. The study found that participants who had a regular health care provider and were assigned to the condition of a formal profile picture rated higher credibility compared to those with a casual profile picture. Jahng and Littau [16] investigated the credibility of journalists on Twitter by testing how factors such as gender, amount of personal information on their profile page, and amount of interaction with followers affect their credibility. The study found that highly interactive journalists are more credible than less interactive journalists. Another paper conducted an exploratory experiment on examining journalists' credibility on Twitter based on their gender and number of followers [10]. The other paper questioned how the perception of the number of retweets affects an individual's identity on Twitter [22]. Morris and Counts, et al.[27] conducted two experiments. First, they identified some features on Twitter that can impact viewers' attention and credibility; second, they tested the factors of tweets topic, user name, user image and their impact on credibility. The findings of these papers help me to filter out the perceptions about credibility I need to test, give me valid foundation for designing my experiment. I would choose profile photos of Twitter users as formal as possible; keep the retweet number at an intermediate level and as a fixed for all generated tweets images; modify the tweet images with/without a link; modify the user profiles with relevant/irrelevant info.

Some studies also developed algorithms to investigate automatic classification methods to assess online information credibility. Castillo et al.[5] used features from users' posting,

text of the content and external link contained to train two classifiers. One classifier can automatically separates newsworthy topics from conversations, and another classifier is able automatically assesses the level of credibility of newsworthy topics. Alrubaian et al.[1] proposed a new algorithm that analyses and assess the credibility of tweets and users. This new method consists of four components that work together: a reputation-based component, a credibility classifier engine, a user experience component, and a feature-ranking algorithm. These studies have explored the new automatic way to assess online information credibility, which highlights the ongoing interest and importance to evaluate online information credibility, also provides me new ideas to continue this study in the future beyond the current scope.

Previous studies have shown that Twitter credibility assessment can be approached from different perspectives with different methods and focuses. This study will focus on the credibility of 'Dr.' title in users' display name, mainly targeted in health related fields. Currently, Twitter is the most popular social media platform used for health communication [35]. According to Pershad and Hangge et al.[30], it is interesting for healthcare information on Twitter due to the emphasis on transparency and open sharing information for doctors and patients. But the potential risk for viewers who are seeking health information on Twitter is the difficulty of identifying real doctors and distinguishing true content from false one. Additionally, the content of health tweets can be too brief to be accurate, which leads viewers to the wrong conclusions [30]. For all those arguments, it is key to assess how shared medical information credibility on Twitter is assessed by regular viewers. This is precisely why in the rest of the thesis, I choose to focus on medical tweets shared by users with different declared titles. The goal of this thesis is to measure online information credibility by generating medical tweets in the designed survey.

Chapter 3

Methods

This chapter details the methodology employed to conduct the study. It starts by clarifying the variables to be tested, an overview of the survey design is followed, the content of the survey is presented here as it will be presented to participants. The last section will explain the plan of data collection procedure and statistical models for analyzing data.

3.1 Variables

To test the hypotheses presented in the introduction, I used different independent variables for each of them. First, three levels of titles will be compared: 'Dr.'; 'M.D.' title and no any titles. 'Dr.' title is the primary title of interest in this study and refers to a person who owns a PhD degree. In addition, the generated tweets used will be medical tweets, it is reasonable to include 'M.D' title in the comparison as well, which refers to medical doctor. Two genders of users on Twitter are compared: female and male. Because gender of users with different titled display name can impact the credibility assessment based on the remaining of previous work [2]. The viewers' Twitter usage and posting frequency are measured on a standard and five categorized level: daily, weekly, monthly, yearly and never. The last two hypotheses independent variables are the tweets credibility with/without an URL; and users' profile info level is tweet relevant/irrelevant. These choices are inspired by previous studies summarized in Section 2.2 to test different perceptions of credibility on Twitter, and theses features have not been tested yet.

As for the dependent variable, it is all the same for all six hypotheses: the viewers' perceived credibility level change of a statement. This change is measured by comparing the credibility level before and after the viewers read the tweets. This measurement uses the method of measuring credibility via proxy measures of attitude change, as suggested by Cacioppo et al. [31]. The credibility level is quantified on a 7-point likert scale, same as other studies [9, 16, 27]. The variables in this study is summarized in the Table 5.11

Hypotheses	Independent variable	Dependent variable
Hypothesis 1	Title: 'Dr.'Vs 'M.D' Vs no title	Credibility changes before and after reading a tweet on a 7-point Liker scale
Hypothesis 2	Gender: Female Vs male	
Hypothesis 3	Viewer Twitter usage: daily, weekly, monthly, yearly and never	
Hypothesis 4	Viewer tweets frequency: daily, weekly, monthly, yearly and never	
Hypothesis 5	Tweet content: With URL Vs Without URL	
Hypothesis 6	Profile level: Tweet Relevant Vs Tweet Irrelevant	

Table 3.1: Independent variable and dependent variable in this study

3.2 Survey Design

This survey is composed of two parts, aiming to test the six hypotheses. The first part focuses on collecting participants' demographic data, such as their age, education level, living country, Twitter usage frequency, number of followers, and number of accounts they follow. This data will provide an overview of participants and their familiarity with Twitter. The second part of the survey is the main part which measures the credibility change. Participants will first evaluate one statement, and then read a tweet on the topic of the statement, lastly evaluate the same statement again. This measurement process will be repeated three times, with only one variable being altered every time, while the other variables remain fixed.

All designs of the survey will be finished on the platform *Qualtrics*[34][1]. Block is the format that is used on Qualtrics to represent the separation of three repetitions. There will be three blocks in total.

The first block is controlling gender(Female Vs Male) X Title (non Vs 'MD' Vs 'Dr'), with six conditions in total, to test first four hypotheses. The second block is controlling Publication link (Include Vs Exclude) X Title (non Vs 'MD' Vs 'Dr.'), 6 conditions in total, to answer hypothesis 5. The third block answers hypothesis 6, will only be analysed within the 'Dr.' titled group, by controlling bio info levels (irrelevant Vs relevant) X Title ('Dr.'), with a total of 2 conditions. Table 3.2 gives a summary with all the conditions and variables listed under each block. Under each block, randomization is applied to make sure that every condition is equally likely to be presented to the participants. This is a between-subject treatment, participants are only assigned to read one tweet among all other tweets under each block.

3.3 Participants

Prolific [2] is a public platform frequently used for academic surveys [33], a public platform frequently used for academic surveys [29]. In this study, participants are recruited from

[1] www.qualtrics.com
[2] https://www.prolific.co/

Prolific. I restrict the survey to participants who are fluent in English, located in all around the world. Every participant who completes the survey dedicated with passing the attention check question is compensated 0.75 Euro.

3.4 Manipulation

All tweets and users' profile pages in the survey are presented to participants in the form of images, which are generated and manipulated as to address the research questions. For user gender, three females and one male are picked. In the first block, one male and one female are included for comparison, while the other two blocks feature two females. This decision is based on the finding from a previous research paper [10]. The study found that there is an interaction effect between users' followers number and their gender for female journalists instead of male journalists. While this survey need to keep gender fixed in the second and third blocks, female gender is preferred. For the user names, I use two from a previous study [10]; additionally, I randomly generated two names using an online tool, *namegenerators*. [3]. Altogether, the four user names are Mary Smith (female), John Brown(male), Railey Parker(female), and Elma Hartley(female). To select profile photos for these users, I used an online tool that generates face images [4]. I choose the formal profile photos for all four users. Based on the finding from Ferrel and Castillo [12], viewers with a regular health provider have stronger attention to engage with physicians who have formal appearance profile photo. Applying it to this study, the formal appearance of health care users on Twitter can gain more interaction with participants which may be influential to our study results. Together with the needed resources mentioned before, all tweets images for users to read are generated by the tool *faketweetmaker*[5].

In terms of the content of tweets, three health-related topics are chosen based on real tweets that I found with valid publication verification: depression treatment [39], diabetes [43], and high blood pressure [23]. It is necessary to have valid paper publications for the tweets that participants will evaluate, aiming to ensure the accuracy of the statements, and to avoid misleading participants. Because the tweet sentence are fixed for all generated titled users in one block, the tone of the sentence is chosen to be casual to make the tweets read as real as possible. I asked ChatGPT [6] to write it and slightly adjusted by me. Images of tweets that for participants to read are presented in the next section 3.2.

As per the survey design, conditions are identified and distinguished by users' display name and tweet content from generated tweet images. For the first block, there will be six diabetes tweets in total from either Mary Smith(female) and John Brown(male) with 'Dr.'/'M.D'/no title. For the second block, there will be six depression treatment tweets with/without paper link from Railey Parker with 'Dr.'/'M.D'/no title. In the last block, participants will read one high blood pressure tweet from Dr. Elma Hartley plus one profile image from her, with tweet relevant/irrelevant bio info contained. Table 3.2 below also gives a summary of all the conditions.

[3]https://namegenerators.org/us-female-name-generator-rd/
[4]https://generated.photos/faces/female
[5]https://www.faketweetmaker.com/fake-tweet-generator
[6]https://openai.com/blog/chatgpt

Blocks	Hypotheses	Conditions	Tweet Topic
First Block	H1, H2, H3, H4	Dr.Mary Smith/Mary Smith MD/ Mary Smith (Female); Dr.John Brown/John Brown MD/John Brown (Male)	Diabetes
Second Block	H5	Dr.Railey Parker/Railey Parker MD/ Railey Parker with paper link; Dr.Railey/Railey Parker MD/Railey Parker paper link without link (Female)	Depression Treatment
Third Block	H6	Dr.Elma Hartley with irrelevant/relevant profile (Female)	High Blood Pressure

Table 3.2: Conditions for three blocks in the survey

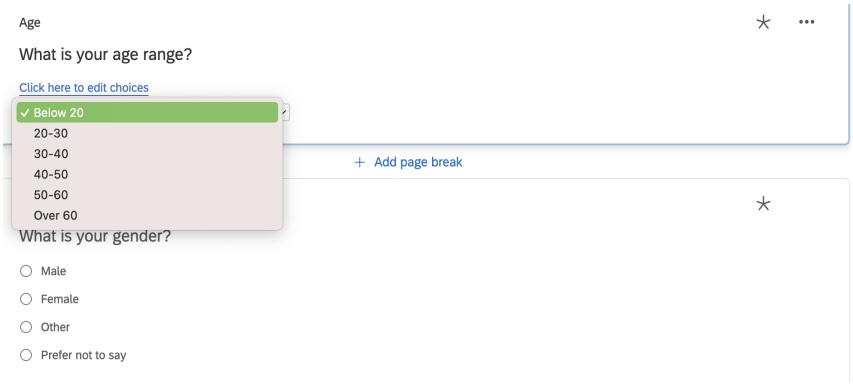
3.5 Survey Content

The survey starts with a brief introduction of what the survey is about, and what the participants will do in the next five minutes. The goal of this introduction is to give a welcome to participants and warm them up to answer the survey carefully. The introduction content is shown below.

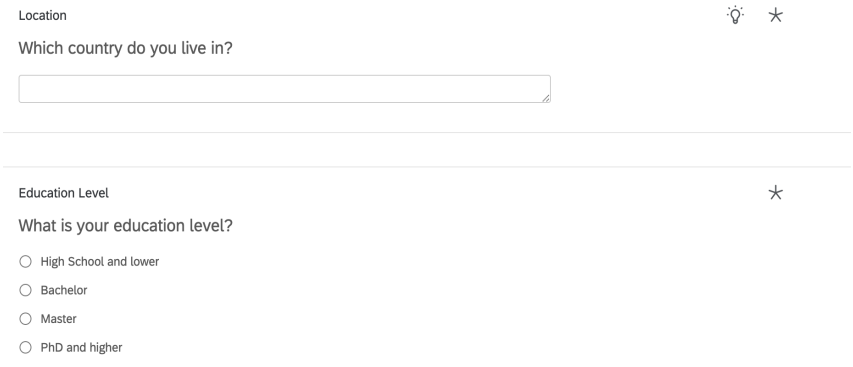
> **Motivation** The credibility of information matters. This is especially true for large social media platforms (e.g., Twitter), which became ideal spaces to disseminate information. In this study, we want to investigate the credibility of different types of users' titles. The ultimate goal is to understand better what influences credibility online and how real claims could be better promoted on Twitter.
>
> **Structure** This survey contains two parts. In the first part, you are required to answer some demographic questions. In the second part, you are asked to read three tweets and one user's profile page, after it you will have to respond to one follow-up question. The contents of tweets is health related. 'Dr.' stands for a person who has obtained a Dr degree. 'MD.' stands for Doctor of Medicine and indicate someone who has completed a medical school.
>
> This survey will take around 5 minutes. If you have any questions, please don't hesitate to contact us. We appreciate your time and honest response.
>
> Thank you in advance! Please click the blue button to start!

The participants are next asked to answer a series of demographic questions, which are all compulsory. More precisely, the survey asks them to report their age, gender, location, education level, familiarity with Twitter and activity on Twitter. The questions are mostly multiple-choice, as shown in Figure 3.1. This multiple-choice format of questions is straightforward and can be easily analyzed in later analyzing step, while also providing participants with clear and understandable options.

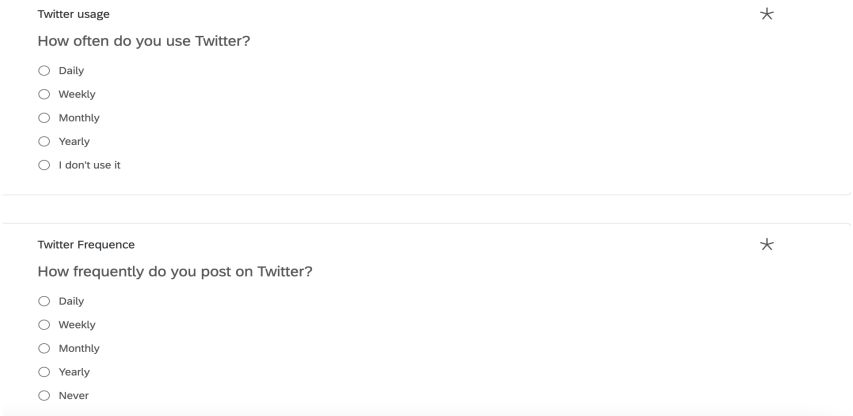
Figure 3.2 shows the three statements that participants will read and evaluate on a 7-point Likert scale. Additionally, the attention question here is to check whether the participants



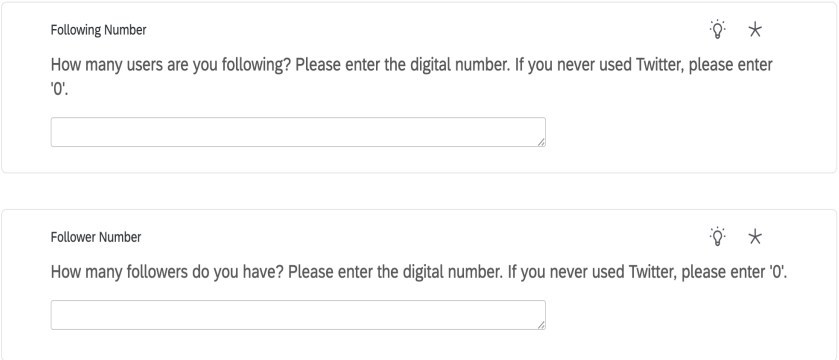
(a) Age and Gender



(b) Location and education level



(c) Twitter usage and post frequency



(d) Following number and followers

Figure 3.1: Demographic questions

are paying attention. If they answer this question wrongly, the participants' response will be rejected and the data will be recollected with other participants. At the end, I will only use the data from participants who answer the attention question correctly. This question used to measure the quality of participants' response, and to ensure I collect the high quality data.

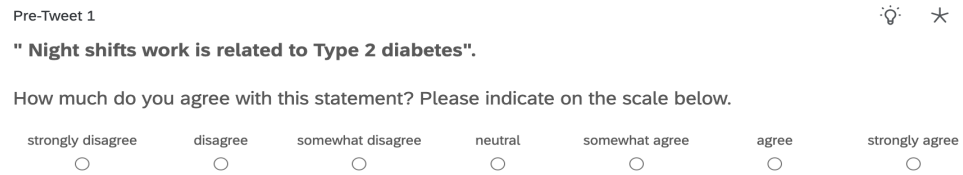
Figure 3.2 displays the three statements and attention question as participants read. Figure 3.3 displays six tweets with different conditions under the first block. The differences are users' gender and display names. Participants will read only one of those images by randomization in between of the two times evaluations for the first statement. Similarly, Figure 3.4 demonstrates all the six tweets under the second block, the differences are users' display name and paper link in the tweets. Still, participants will only randomly see one of the six tweet in between the evaluations for the second statement. Lastly, Figure 3.5 shows the two images participants will read in the third block. One is the tweet from Dr.Elma Hartley with a paper link in the tweet, and the second image is one of the profile page of her. In all tweet images, the index of "Views", "Retweets", "Quote tweets" and "Likes" are random odd numbers to make it look as real as possible, also for the tweet publishing time. These numbers were kept as not to be extremely high or low, due to the research findings from Lin and Spence [22], they suggest intermediate number of these perceptions would gain more credibility than extreme numbers.

At the end, it is the closing statement, shown as below. It provides the publication link and proofs that all the statements in the survey are verified. The paper links are given at the end has two purposes. First, it would not give participants any impressions and prior knowledge that all the statements are true. Second, it proves that all the statements are true and not mislead them to wrong conclusions. If the participants are interested and want to check the paper out, they can directly search for the paper with links.

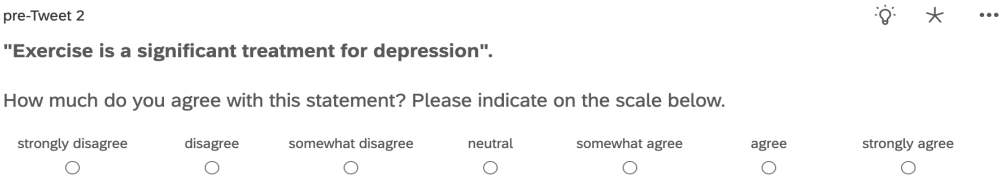
> We thank you for your time spent taking this survey. Your response has been recorded. The Twitter account and tweets information involved in this survey were generated and designed only for this study. The statement regarding to the healthcare topics are true. They are based on the evidence from scientific paper publication, if you are interested you can check them out with links below:
>
> First statement: https://pubmed.ncbi.nlm.nih.gov/26978184/
>
> Second statement: https://diabetesjournals.org/care/article/41/4/762/36957/Night-Shift-Work-Genetic-Risk-and-Type-2-Diabetes
>
> Third statement: https://www.ahajourna ls.org/doi/10.1161/HYPERTENSIONAHA.120.14695

3.6 Data collection procedures and Statistical Models

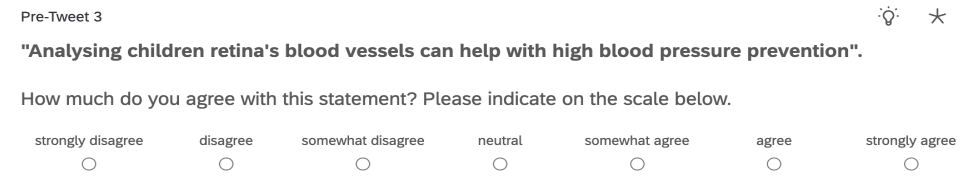
To ensure the survey questions are clear for participants and the tweets can effectively impact participants' attitude change, I will run a pre-test with 15 people before publishing



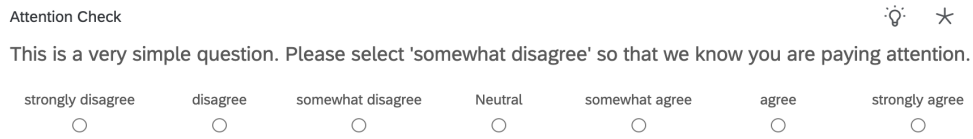
(a) Statement 1



(b) Statement 2



(c) Statement 3



(d) Attention check question

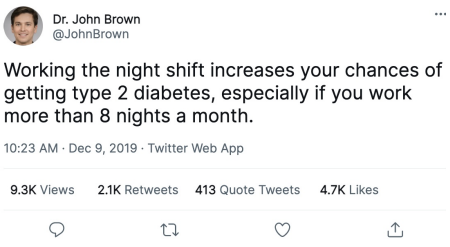
Figure 3.2: First evaluation of three statements

the official survey. The pre-test will focus on two aspects: the percentage of participants who pass the attention question, and the variation in participants' attitudes change towards the three statements. If more than 5 participants failed the attention questions or if the various of attitudes changes are not obvious. The survey design will be adjusted.

With the budget limit of 300 CHF, the official survey will be answered from 180 participants with 60 people per block at least. The questionnaire will take around 5 minutes for each participant. The data collection time will be maximum 1 week. Responses from participants who fail the attention check question will be rejected without compensation, and data collection on *Prolific* will continue until all 180 participants pass the attention check question. In addition, duplicated responses from the same user ID will be excluded, since some participants may attempt to fill out the survey twice to receive double compensation.

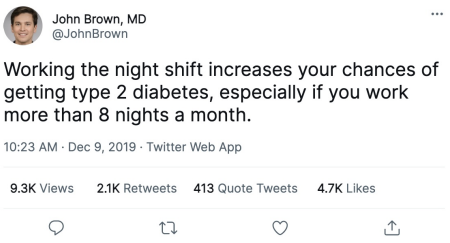
Since the purpose of this exploratory study is to learn whether the variables have effects on the credibility assessment, the primary model for analyzing the data is the ANOVA (Analysis of Variance), which has the NULL hypotheses that all group means of the variable are equivalent. Previous similar studies also applied ANOVA method to analyze the data. For example, Edgerly and Vraga [9] used ANONA to test the relationship of Twitter verification label and users' credibility. Ekanem [10] also used ANOVA to test

(a) Condition 1: Dr. female



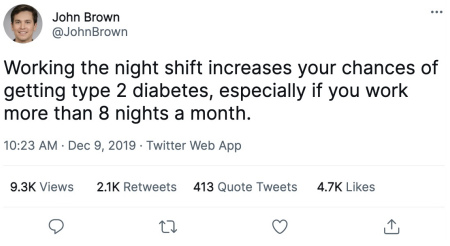
(b) Condition 2: Dr. male

(c) Condition 3: MD. female



(d) Condition 4: MD. male

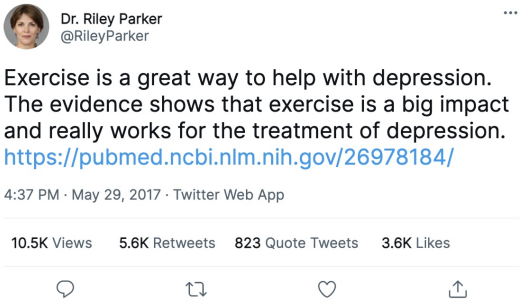
(e) Condition 5: no title female



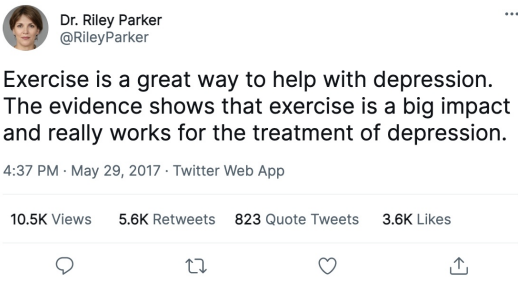
(f) Condition 5: no title male

Figure 3.3: Title x Gender conditions

journalists' credibility on Twitter regards to their gender and number of followers. Thus, ANOVA is a well performed method has been tested well in the previous studies. Student's T-Test will also be used, depends on the level of factors. If the independent variable has 2 levels, T-Test will be used; and if the independent varibale has more than 2 levels, ANOVA will be used. To assess how the factors and their interaction affect the dependent variable, one-way ANOVA and two-way ANOVA are both considered. The *aictab* method from the *AICcmodavg* library in R is selected to compare and choose the best model with the lowest AIC score. If the independent variable is found to have significant evidence on impacting credibility assessment, Tukey's Honestly Significant Difference (Tukey's HSD) will be used as following up the ANOVA model, to perform pairwise comparisons and find out the differences between each group from the factor of variable. P-value < 0.05 will be the threshold for all these tests, which means with 5% possibility the null hypothesis is true. If p-value > 0.05, the null hypothesis will be rejected.

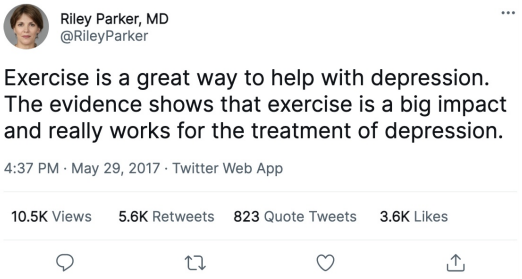


(a) Condition 1: Dr. with link

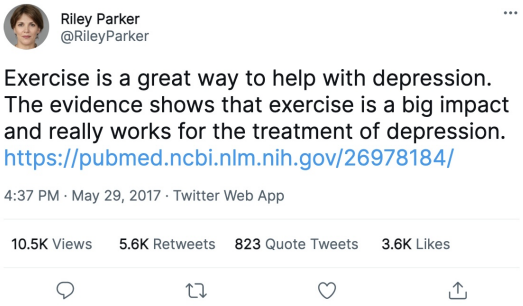


(b) Condition 2: Dr. without link

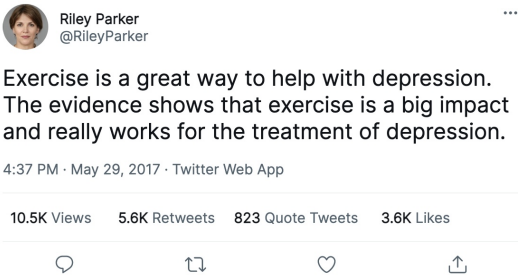
(c) Condition 3: MD. with link



(d) Condition 4: MD. without link



(e) Condition 5: no title with link



(f) Condition 6: no title without link

Figure 3.4: Title x Link conditions

(a) Condition 1: irrelevant profile

(b) Condition 2: relevant profile

(c) Third tweet

Figure 3.5: Title x Profile Level Conditions

Chapter 4

Implementation

4.1 Pre-test

All 15 people recruited for the pre-test answered the required questions, and passed the attention question check. These 15 people are excluded to join later official experiments to make sure no any of participant is familiar with the survey. There were 12 female and 3 male reported in the survey, most people were aged in the range of 20 to 30, and finished bachelor study. They reported a variety of attitude change towards the three statements, better than expected. For statement 1, the mean value of credibility difference is 0.5 (SD = 0.82). For the statement 2, the mean of credibility difference is 0.44 (SD = 1.03). Mean value of the statement 3 is 1.38 (SD = 0.62). The results showed that the survey questions were comprehensible for these participants, their attitudes change are various with quite high standard deviation. I am able to conduct further experiment with current version of survey.

4.2 Data pre-processing

4.2.1 Data Cleaning

The data was first processed by removing rows where users failed the attention check questions. There were five participants, representing 2.78% of the total number of participants. Second, there were three participants who are shown as time out and not finished the survey. These three data were excluded either. Lastly, there was one participant who answered the survey twice, both with attention question passed but different answers. This duplicated response is removed from the data set as well.

Afterwards, the data was divided into three subsets based on the three statements. Under each subset, labels were assigned to distinguish which condition the participant saw so that further comparisons could be made. For the first statement subset, a gender label and a title label were assigned. For the second statement subset, a link label and a title label

were assigned for each participant. For the third statement subset, a bio info level label was assigned. The following tables show examples of rows from each subset: Table4.1, Table4.2 and Table4.3.

ID	Pre Tweet	Dr.F	Dr.M	MD.F	MD.M	Non.F	Non.M	Gender Label	Title Label	Diff
5ddbea...	0	NA	1	NA	NA	NA	NA	Male	Doctor	1
6064b4...	-1	NA	NA	2	NA	NA	NA	Female	MD	3
5e89fb...	-2	NA	NA	NA	0	NA	NA	Male	MD	2

Table 4.1: Example of data cleaned for statement 1

ID	Pre Tweet	Dr Link	Dr No-link	MD Link	MD No-link	Non Link	Non No-link	Link Label	Gender Label	Diff
5ddbea...	1	NA	NA	1	NA	NA	NA	Nolink	MD	0
6064b4...	1	3	NA	NA	NA	NA	NA	Link	Doctor	2
5e89fb...	1	1	NA	NA	NA	NA	NA	Link	Doctor	0

Table 4.2: Example of data cleaned for statement 2

ID	Pre Tweet	Irrelevant Profile	Relevant Profile	Bio Label	Diff
5ddbea...	1	NA	2	relevant	1
6064b4...	1	2	NA	irrelevant	1
5e89fb...	0	1	NA	irrelevant	1

Table 4.3: Example of data cleaned for statement 3

4.2.2 Credibility Measurement

To measure credibility, the 7-likert scale assessment of each statement needs to be quantified as numeric value, so that the difference can be calculated. I used the standard to convert the scale according to Morris and Counts et al.[27]: Strongly disagree = -3, disagree = -2, somehow disagree = -1, neutral = 0, somehow agree = 1, agree = 2, strongly agree = 3.

4.2.3 Demographics

Of 180 respondents who completed the survey, 50% were female, 48% were male and the other 2% did not report their gender. I also include 'Prefer not to say' as an option in the survey, in fact non participants chose this option. The percentage of genders is demonstrated in Figure 4.1. The majority of participants, approximately 69%, were in the age range 20-30; followed by the group aged 30-40 (17%) and aged 40-50 (6%). Fewer

people reported in the other age ranges. 89 people reported owning a bachelor's degree and only 3 people had a PhD or higher education level. One worth-mentioning fact reflected from the collected data is that 82 people reported using Twitter daily, the highest number among other categories; however, only 15 people reported posting tweets daily, which is the least category. On average, these participants are following 236 people and they have 674 followers.

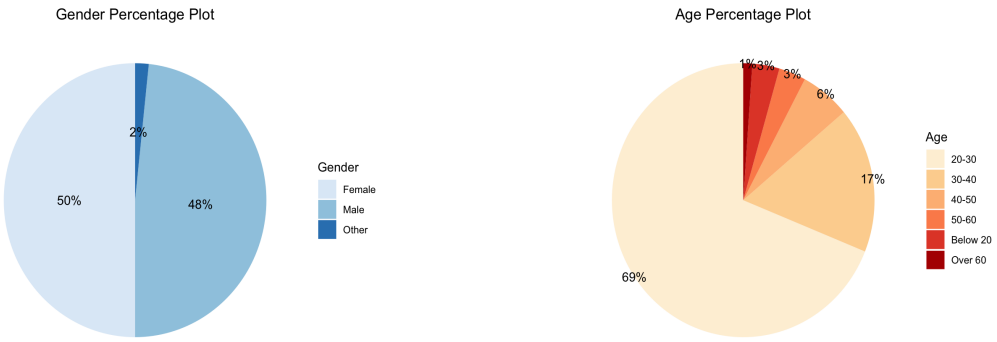


Figure 4.1: Demographics of participants age and gender

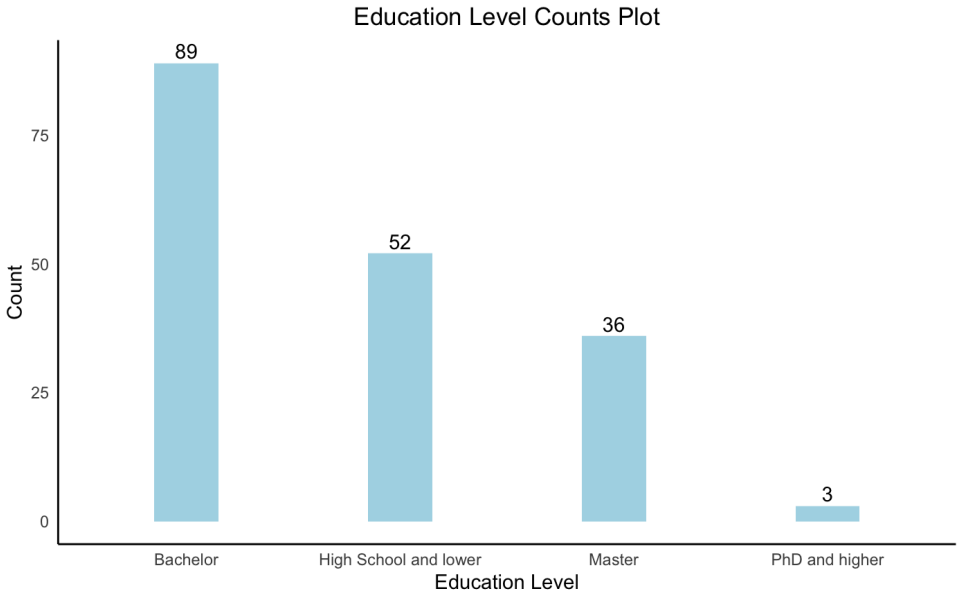


Figure 4.2: Education Level Distribution of Participants

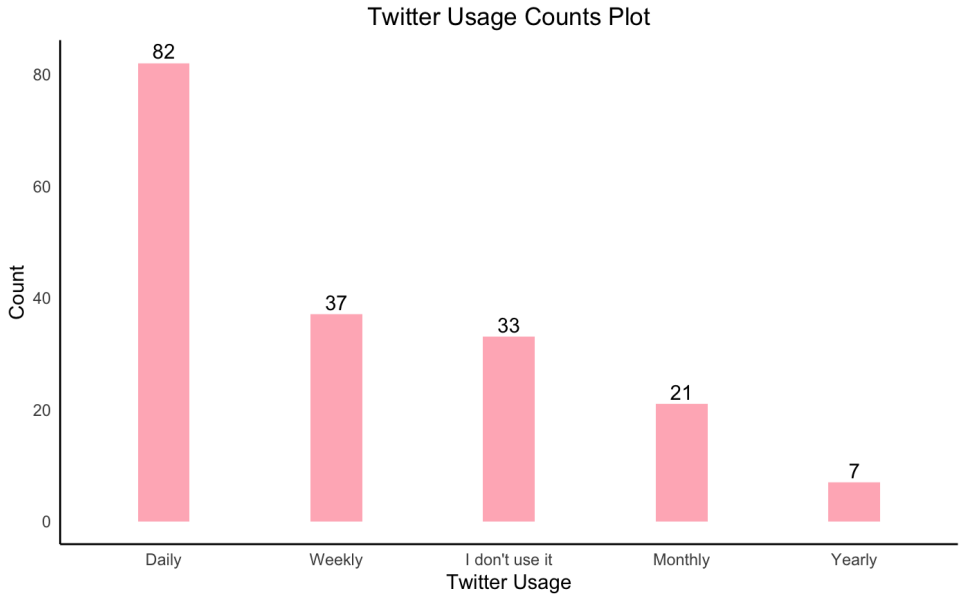


Figure 4.3: Twitter Usage Distribution of Participants

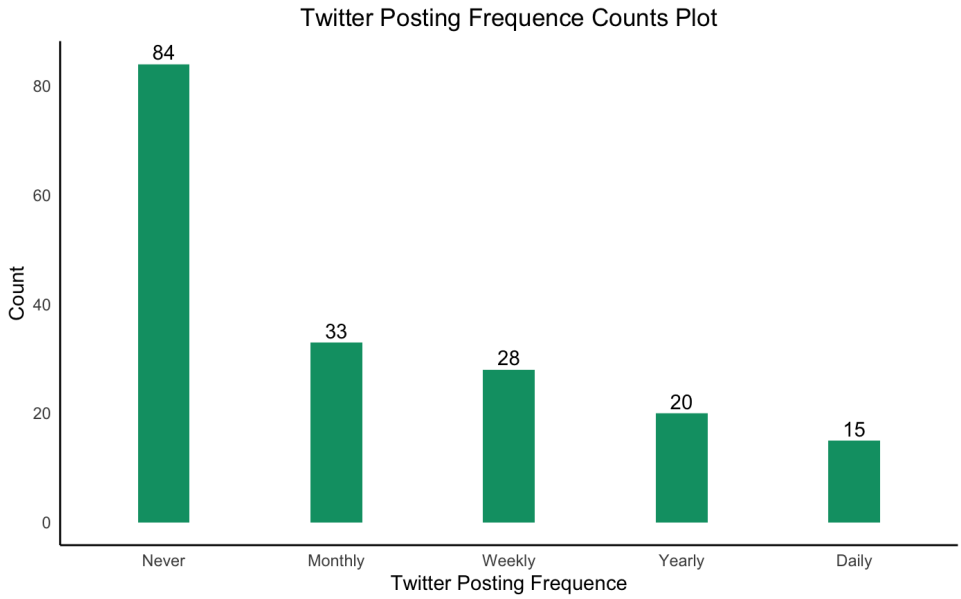


Figure 4.4: Twitter Posting Frequency Distribution of Participants

4.3 Model Comparison and Selection

For each hypothesis, one statistical model is selected to test. Model comparisons and selections are shown below.

Hypothesis 1 Only one model, no model comparison:

1. $title.anova = Difference \sim TitleLabel$

Hypothesis 2 Only one model, no model comparison:

1. $gender.t = Difference \sim GenderLabel$

Hypothesis 3 Three models are compared:

1. $usage.anova = Difference \sim Twitter.usage$
2. $usage\_title.anova = Difference \sim Twitter.usage + TitleLabel$
3. $interaction2 = Difference \sim Twitter.usage * TitleLabe$

The interaction model is the best according to Table4.4.

Model	AIC score
$interaction2$	521.51
$usage\_title.anova$	524.64
$usage.anova$	528.20

Table 4.4: Model Comparison for Hypothesis 3

Hypothesis 4 Three models are compare:

1. $frequence.anova = Difference \sim Twitter.Frequence$
2. $frequence\_title.anova = Difference \sim Twitter.Frequence + TitleLabe$
3. $interaction3 = Difference \sim Twitter.Frequence * TitleLabel$

$frequence\_title.anova$ is the best according to Table4.5. .

Model	AIC score
$frequence\_title.anova$	528.97
$frequence.anova$	532.91
$interaction3$	535.44

Table 4.5: Model Comparison for Hypothesis 4

Hypothesis 5 Only one model, no model comparison:

1. $link.t = Difference \sim LinkLabel$

Hypothesis 6 Only one model, no model comparison:

1. $bio.t = Difference \sim bioLabel$

Chapter 5

Evaluation

5.1 Results

Before fitting to statistic models, the distributions of changed credibility are investigated. Figure 5.1, Figure 5.2 and Figure 5.3 demonstrate the different levels of participants' credibility changed for the three statements. The most frequently observed change levels are 0 and 1, and few people showed extreme changes in levels -2, 3, 4, and 5. It means the majority of participants are positively impacted by the tweets and gained more trust from reading the tweets, with only a few being impacted negatively. The mean of statement 1 credibility change is 0.6111 (SD = 1.13). Statement 2 credibility change has mean value 0.17 (SD = 1.09), and statement 3 has 0.87 credibility changes on average (SD = 0.97). It suggests that there is a stronger impact on the statement 3 compared to other statements.

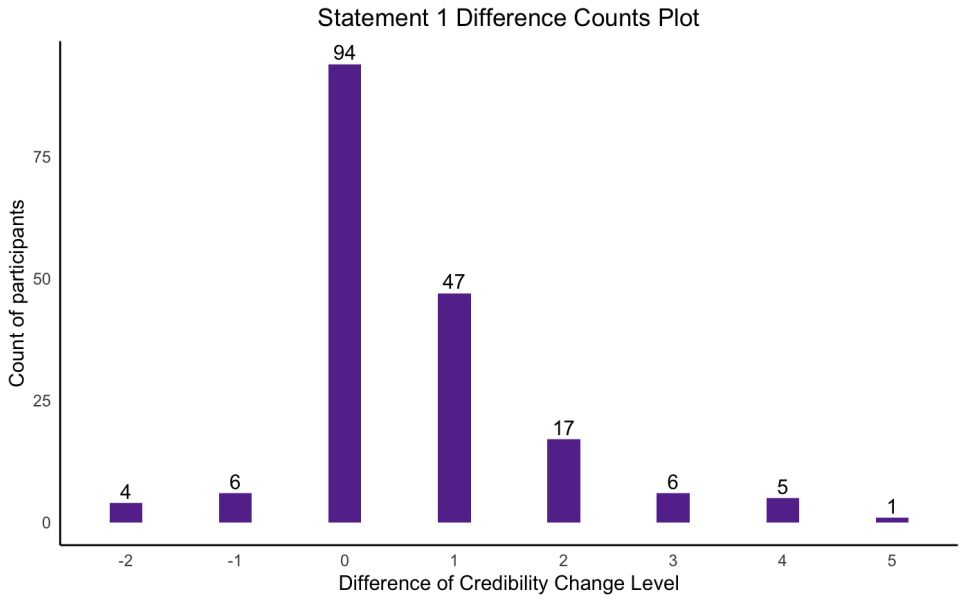


Figure 5.1: Counting number of participants on different levels of credibility change for statement 1

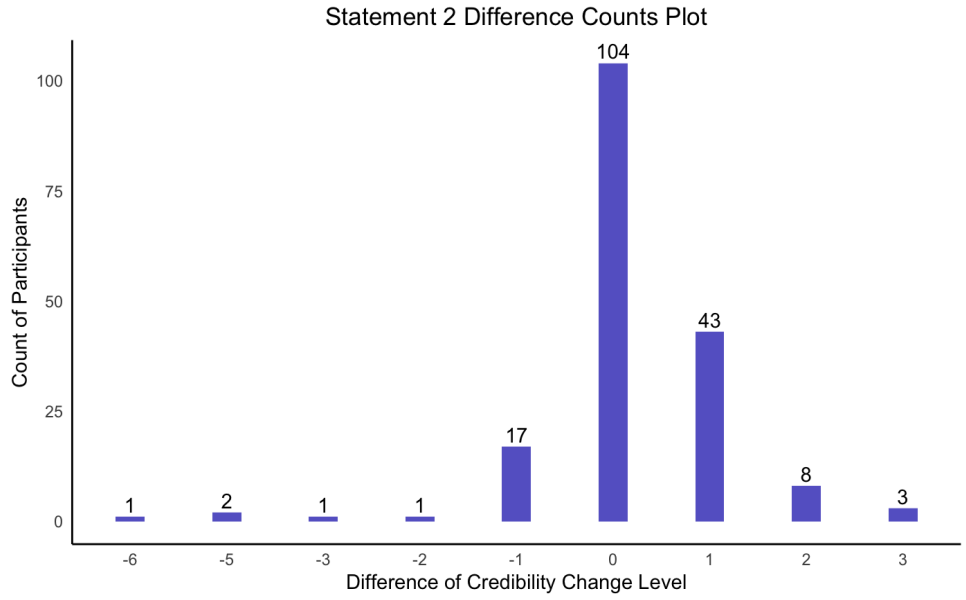


Figure 5.2: Counting number of participants on different levels of credibility change for statement 2

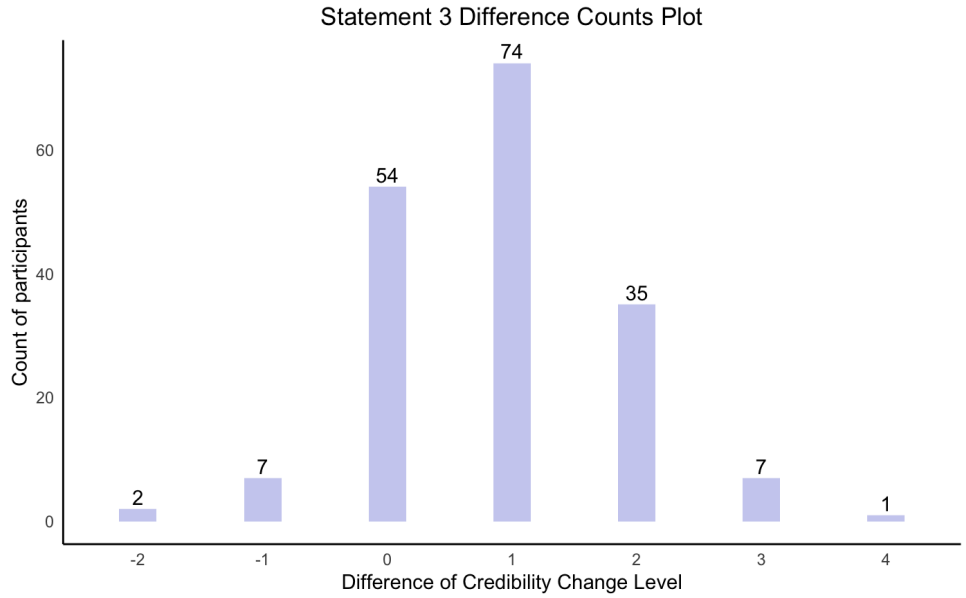


Figure 5.3: Counting number of participants on different levels of credibility change for statement 3

Based on the results from the ANOVA model used to test hypothesis 1, it finds that the title displayed to participants have a statistically significant impact on their credibility assessment, with a small p-value of 0.0202 as shown in Table 5.1. This small p-value is sufficient to reject the null hypothesis that there is no difference between all groups of the title factor. However, for hypothesis 2, the T-Test shows no significant evidence to support the assumption that users' gender has an impact on credibility assessment (p-value = 0.7623), although female users did appear to have a slightly greater impact on

viewers than male users(female mean = 0.7927; male mean = 0.7449), as shown in Table 5.2.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
TitleLabel	2	8.21	4.103	3.99	0.0202 *
Residuals	177	181.99	1.028		

Table 5.1: Anova Model for Hypothesis 1

Welch Two Sample t-test	
t	0.3030
df	154.23
p-value	0.7623
lwr	-0.2637
upr	0.3594
Female mean	0.7927
Male mean	0.7449

Table 5.2: Welch Two Sample t-test Hypothesis 2

Following up the ANOVA model, Tukey's Honestly Significant Difference (Tukey's HSD) is used to perform pairwise comparisons and to find out which title groups are statistically different from other title groups. The result is displayed below in Figure 5.4 with the difference of each comparison and its 95% confidence interval. Upon close inspection, Table 5.3 reveals that the difference between groups of 'Non title' and 'M.D.' is statistically significant (p-value < 0.05). Those who read tweets from users with the 'M.D.' titles have higher credibility changes compared to those who read no titled users' tweets, with 0.51 difference change on average. There is no evidence to support significant differences between other two group comparisons.

Tukey multiple comparisons of means				
	diff	lwr	upr	p adj
MD-Doctor	0.3191693	-0.1030446	0.74138311	0.1769766
Non title-Doctor	-0.1949153	-0.6556153	0.26578476	0.5778316
Non title-MD	-0.5140845	-0.9565654	-0.07160364	0.0181795

Table 5.3: Tukey Comparison for groups of Title

The results for Hypothesis 3 and Hypothesis 4 show that the participants' Twitter usage and their posting frequency have no main effect on credibility assessment. But the interaction of participants' Twitter usage and users' title has been found to have an effect on credibility assessment, with significant evidence that p-value is 0.0120. Therefore, a comparison of the means of all groups can be made based on the interaction factor of Twitter usage and TitleLabel. Figure 5.5 shows the group-wise comparison, where the group of participants who do not use Twitter and read 'M.D.' titled tweets has the highest

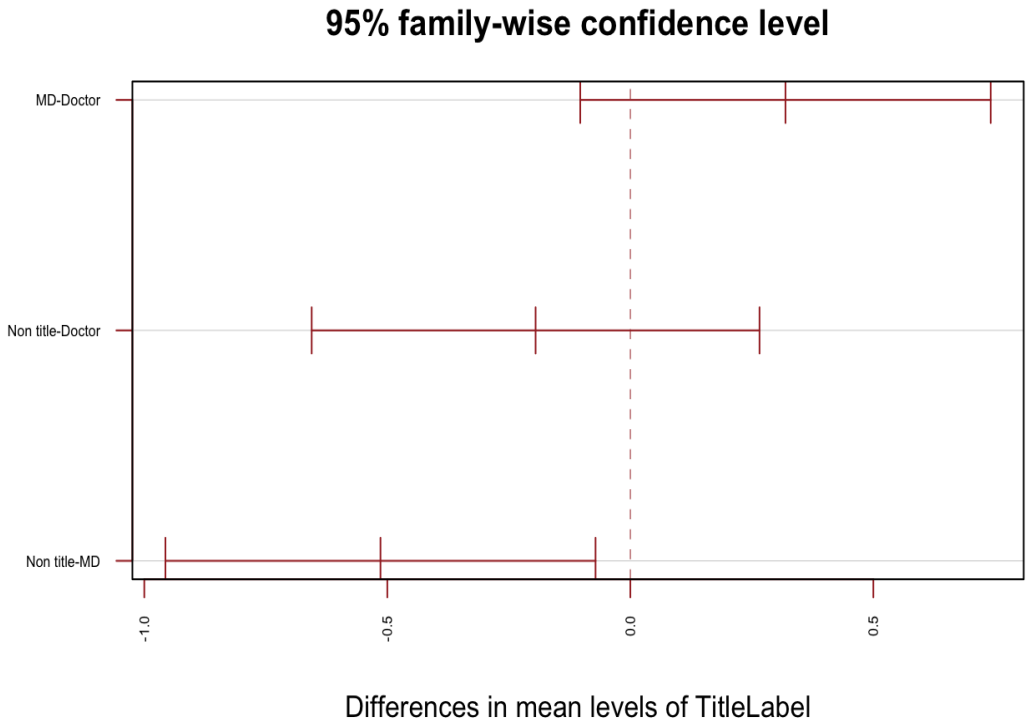


Figure 5.4: Tukey's HSD plot

mean value in credibility changes. The least affected group is monthly viewers who read tweets from users with 'M.D.' title.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Twitter.usage	4	5.24	1.309	1.366	0.2480
TitleLabel	2	7.96	3.978	4.151	0.0174 *
Twitter.usage:TitleLabel	7	17.93	2.561	2.672	0.0120 *
Residuals	166	159.08	0.958		

Table 5.4: Anova Model for Hypothesis 3

The results from the T-Test conducted for hypothesis 5 and hypothesis 6 do not yield significant evidence to reject the null hypothesis. The p-value of the factor publication link is 0.2942, and it is 0.4823 for the factor bio level from users, shown in Table5.6 and 5.7.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Twitter.Frequence	4	0.34	0.084	0.080	0.9883
TitleLabel	2	8.55	4.275	4.079	0.0186 *
Residuals	173	181.31	1.048		

Table 5.5: Anova Model for Hypothesis 4

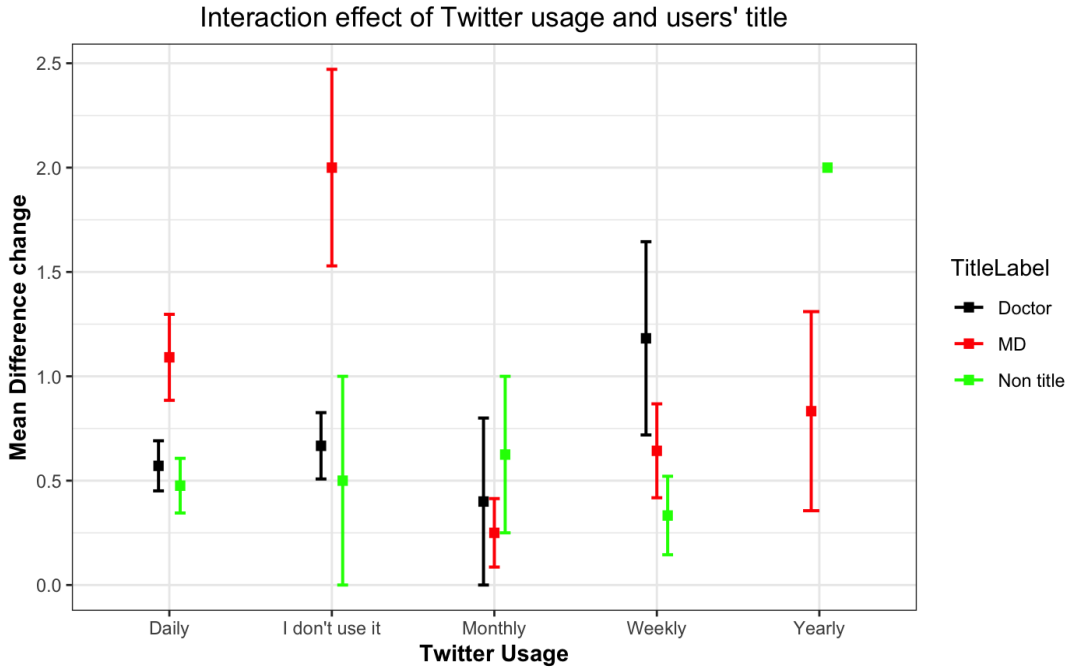


Figure 5.5: Group Comparisons from the interaction of twitter usage and users' title

Welch Two Sample t-test	
t	1.0535
df	124.03
p-value	0.2942
lwr	-0.1362
upr	0.4462
Link mean	0.6750
No Link mean	0.520

Table 5.6: Welch Two Sample t-test Hypothesis 5

5.2 Discussion

Through the data analysis, we have tested the six hypotheses and answered the three research questions. First, the data has explained that the titles in users' display names on Twitter can impact viewer's attitude changes in users' credibility. Participants showed

Welch Two Sample t-test	
t	-0.7041
df	177.8
p-value	0.4823
lwr	-0.3381
upr	0.1603
Irrelevant mean	0.9560
Relevant mean	1.0450

Table 5.7: Welch Two Sample t-test Hypothesis 6

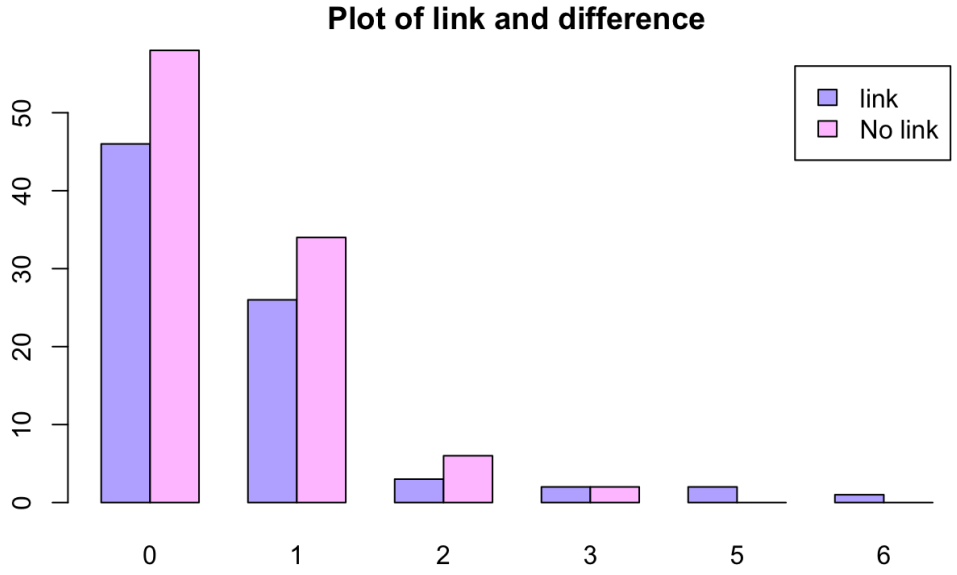


Figure 5.6: Including link or not and their effect on credibility difference

greater credibility when reading 'Dr.' titled and 'M.D.' titled users' tweets than reading no titled users' tweets. The most significant difference was observed between 'M.D.' titled users and no titled users. One reason for this could be that all our statement topics are health-related, and participants were informed at the beginning that 'M.D.' stands for 'Medical Doctor' and 'Dr.' refers to people who have a PhD degree. Thus, they tend to trust medical doctors more in this survey.

In terms of the second research question, there was no clear relationship between participants' Twitter usage and credibility judgments, nor was their tweet posting frequency found to be directly linked to their credibility judgments. However, the data shows an interaction effect between participants' tweet usage and users' title they were reading, which can cause changes in participants' credibility assessment. When participants read tweets from 'M.D.' titled users, people who never used Twitter are impacted the most; when participants read tweets from 'Dr.' titled users, people who use Twitter weekly are impacted the most. Participants who are monthly using Twitter are the least affected group by reading tweets from users with any type of titles.

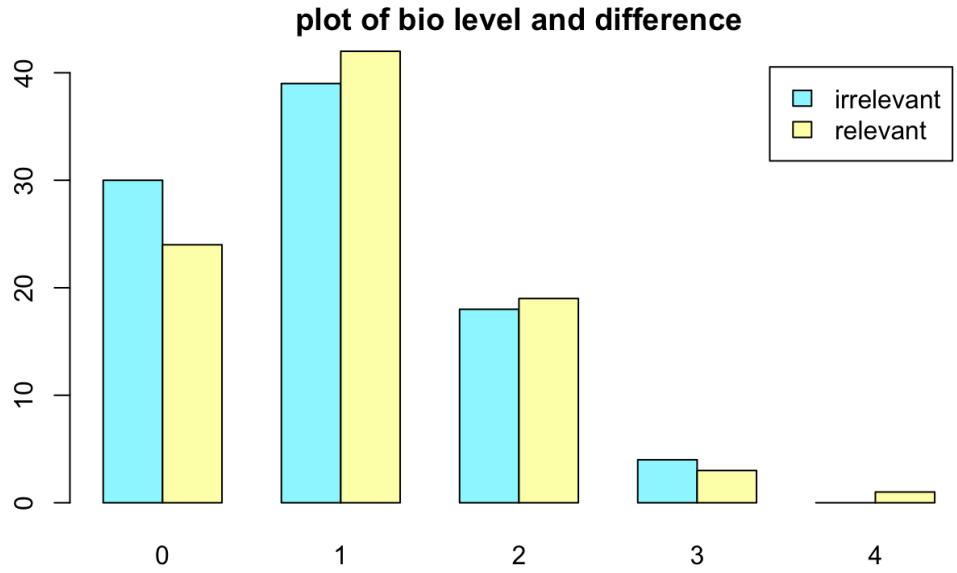


Figure 5.7: Users' bio level and their effect on credibility difference

For the last research question, there are no significant evidence to show that including a paper link and providing tweet related bio info in profile page can increase a Twitter user who have 'Dr.' title in the display name. The frequency of participants' credibility level changes regards to the factors of tweet link and bio level are shown in Figure 5.6 and Figure 5.7, respectively. These plots demonstrate that including a paper link or providing tweet-relevant bio info can lead to the most extreme credibility changes at levels 4, 5, or 6. However, only a small portion of participants are observed to have the extreme changes, the majority of participants show credibility changes in level 1 or no change at all. Comparing these two plots, it appears that including a paper link can be more effective to impact participants than providing tweet relevant bio info, with more extreme changes in credibility assessment. This suggests that it would be a better strategy for the real 'Dr.' titled users who want to increase the impact of their tweet.

5.3 Limitation

The present findings have some limitations regarding to its current scope. The data size are restricted at 180 participants, and survey question length is designed to be 5 min maximum, due to the budget limitation. Further research could benefit from a larger sample size, which could yield significant results.

Secondly, some of demographic data are self-reported by the participants, such as the variable Twitter usage, posting frequency, followers and following number on Twitter. Unfortunately, I cannot verify the validity of those numbers, and it may have the potential that extreme values affect the results of this study. If the study can hire more participants

and with more statistical power, it would gain better understanding of the extreme values and thus find an optimal solution to handle them.

Another challenge that presents in the survey design was the tone and style of the tweet, which relates to users' different titles in their display name. It has to find a balance between being academic and being casual since the tweets should read as real as the users' post, and it has the potential to impact participants' attitudes towards credibility. The solution in this study is using ChatGPT[1] to generate the tweet content and later rephrased by myself. This solution can be challenged for further study if more titles need to be tested.

There is also a constraint that the survey is conducted only on *Prolific*, and the participants are only selected who are fluent in English. This is not representative of the real world situation, where people may speak different languages on Twitter, and people who do not know the *Prolific* platform are not included at all.

5.4 Follow-up Survey

Given the budget limit and complexity of the credibility measurement, the present study left several areas for further exploration. This section combines the current findings and other perspectives which have not tested, provides two options for further follow-up investigation.

5.4.1 First Option

Since the previous experiment is exploratory, the follow-up experiment can be more precise. Based on current finding, the first option would be using power analysis to decide how many participants to recruit in the following for a desired significance level of results. The method used to do power analysis is *Fpower1* in R, and results are presented in Table 5.8. The value of Delta and Sigma used in the function *Fpower1* are from previous first statement results. 5% is set as the threshold because I want my study to have less than 5% chance of occurring that the null hypothesis is true. To obtain at least 0.90 with a family-wise significant level of 0.05, the follow-up experiment needs to hire at least 114 people per group, with a total of 684 people for 6 groups.

5.4.2 Second Option

The second option is still an exploratory experiment since there are many perceptions of Twitter credibility can be tested in the follow-up experiment. Keeping the research questions fixed, more hypotheses are going to be tested.

[1] https://openai.com/blog/chatgpt

alpha	nlev	nreps	Delta	sigma	power
0.05	6	112	0.611	1.131	0.895
0.05	6	113	0.611	1.131	0.898
0.05	6	114	0.611	1.131	0.901
0.05	6	115	0.611	1.131	0.904
0.05	6	116	0.611	1.131	0.907

Table 5.8: Power Analysis

The current result suggests that users with 'M.D.' title could have more impact on viewers' credibility judgement than 'Dr.' titled and no any titled users. It cannot be certain whether this result related to the medical content of the tweets. This followup study could consider to explore other tweet topics, as well as testing other titles. Morris and Counts et al. [27] used politics, science, and entertainment three topics to test and found that the topics do have influences on credibility assessment. The follow-up experiment will use science, business and politics, with three titles that can be representative in these topics. They are: Dr., CEO and traditional name which has no any titles.

As reported in an article [17], approximately two billion users are using social networks on internet. Aside from Twitter, other social medias are also popular among users, such as Facebook, Instagram, YouTube etc. [18]. Users on Twitter also has activities to share contents, networking [17]. These statistics provide inspirations to explore more about what types of viewers are more easily being affected on social media. It has been found that monthly viewers' attitudes can be affected the most by tweets in the current study. However, I want to know more about these viewers. What do they use Twitter for? Are they on social media for a long time? Do they like to interact with other users? Thus, the factors chosen to explore in the follow-up experiment are: viewers' interaction frequency with other users, their motivation on Twitter, years joined Twitter and other social media usage.

The third research question is going to explore how the real 'Dr.' titled users can do to tweet credible information on Twitter. Morris and Counts et al. [27] collected 26 features on Twitter that could impact viewers with an evaluation of credibility impact. There are four features that relate to this study: contains hashtag (credibility rating 3.48/5.00); contains URL (credibility rating 3.50/5.00); posted recently (credibility rating 3.59/5.00); author bio suggests topic expertise (credibility rating 3.66/5.00). Besides, the previous study did not reach my expectations by testing the users' bio info. Therefore, in the follow-up experiment, three factors will be tested. Users' profile info level: irrelevant/relevant; users' tweet containing URL/hashtag; users' tweet frequency: frequently/occasionally.

New hypotheses are formulated below for each research question.

5.4.2.1 New hypothesis to test

RQ 1 1. H1: Viewers tend to trust users who have 'Dr.' in their display name more than other titled users.

RQ 2

1. H1: Viewers who are on Twitter for getting news as motivation are the most impacted type of viewers on Twitter.
2. H2: Viewers who joined Twitter longer year show more trust towards tweet information.
3. H3: Viewers who are also active on other social media show more trust towards tweet information.
4. H4: The viewers who are more interactive with other users show more trust towards tweet information.

RQ 3

1. H1: Viewers tend to trust the 'Dr.' titled users more when those users provide tweets' relevant bio info in their profile page.
2. H2: 'Dr.' titled users' tweets have more impact on viewers if the tweet contains URLs other than hashtags.
3. H3: 'Dr.' titled users who post their opinion tweets frequently have more impact on viewers.

5.4.2.2 Survey and Experiment Design

To continue testing the three research questions, the follow up experiments can be designed as two rounds in order to keep simplicity of each survey. Similar as before. this is still a between-subject experiment, each participant will be assigned randomly to one condition under each section of the survey.

The first run will answer the first and second research questions. Each participant will be asked to answer some questions regarding the participants themselves in the first part of the survey, examples are shown in Figure 5.8. In the second part, they will assess three different topics of tweet with different users' titles in their display name. They need to answer the question "How much do you agree this tweet contains credible information?", and give an evaluation based on a 7-point likert scale, from strongly disagree to strongly agree. There are three types of titles and three topics in total, participants will be randomly assigned to one of three tweets under each topic group, thus evaluate three tweets in total. The nine different tweets are listed in Table 5.9.

The second run focuses on the third research question, participants will be asked to show their trustworthiness attitudes towards 'Dr.' titled users based on different features. First, participants will answer the question "How much do you agree this tweet contains credible information? " by reading one tweet containing either a paper link or a hashtag. Again, each participant will read one from each topic group, three tweets in total. The list of all tweets are shown in Table 5.10. In the second section, four tweets will be generated with two factors, bio level (irrelevant/relevant) X post frequency (occasionally/frequently). Participants will be randomly assigned to read one tweet among the four, and evaluate users' credibility on a 7-point likert scale by answering "How much do you agree this user is credible? " The four conditions are presented in Figure 5.9.

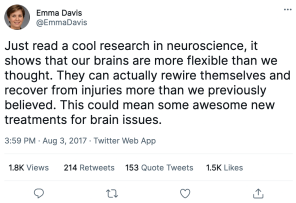
	Dr. group	CEO group	No title group
Economic Topic			
Science Topic			
Politics Topic			

Table 5.9: Second section of the first round in follow-up survey: tweets evaluation of different titles

The content of the tweets are based on three true studies [3, 7, 38]. To generate tweet screenshots, the same tools can be used as before for generating profile pictures [2], user names [3] and tweets content [4], and fake tweets [5].

For a clear clarification, Table **??** gives a summary of the follow-up experiment design.

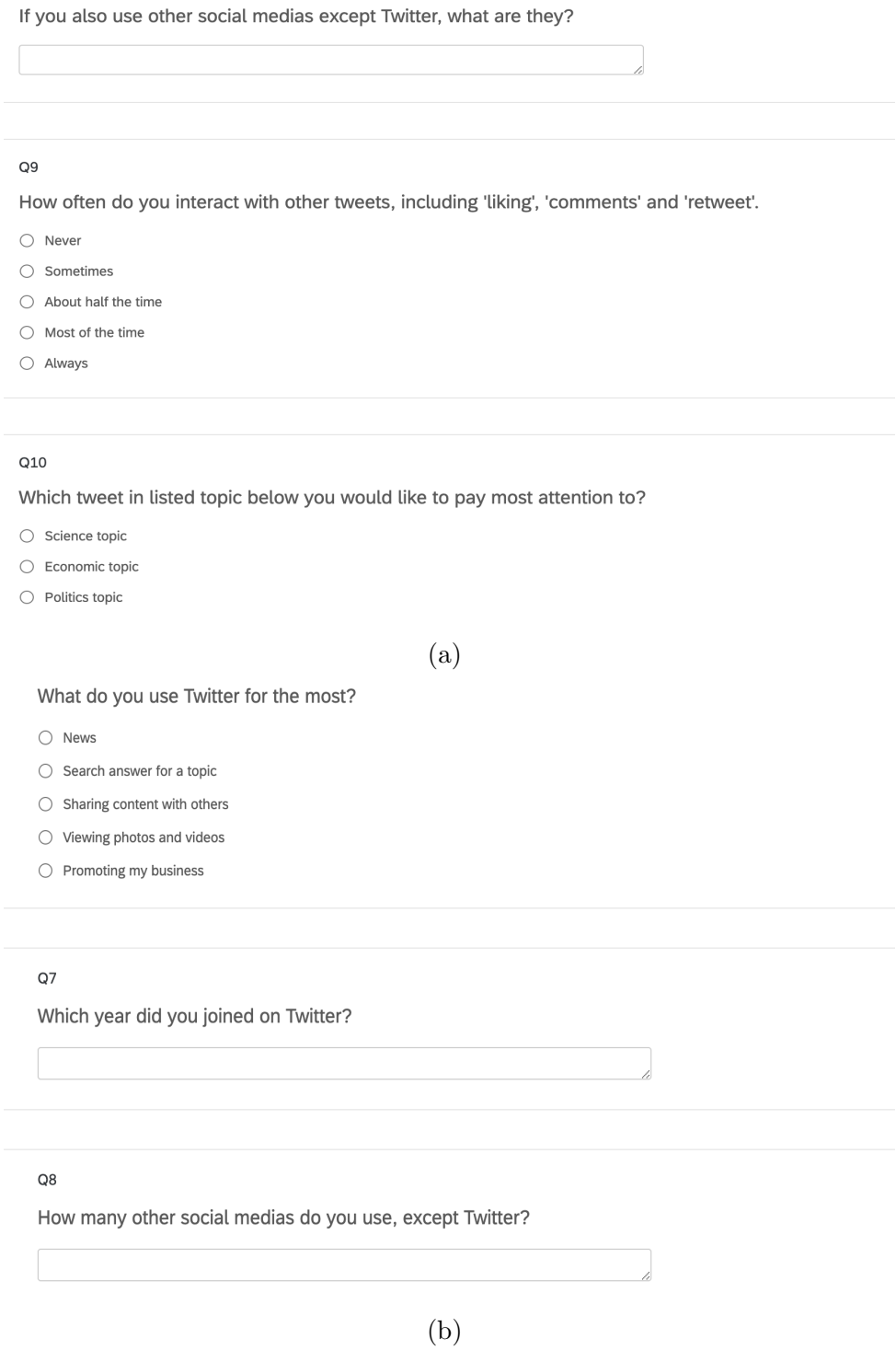
5.4.2.3 Budget and sample size

Since this is an exploratory experiment, we do not have enough information to do power analysis. The sample size can be chosen based on the budget. Table 5.12 lists some sampling plan for further considerations.

[2]https://generated.photos/faces/female
[3]https://namegenerators.org/us-female-name-generator-rd/
[4]https://openai.com/blog/chatgpt
[5]https://www.faketweetmaker.com/fake-tweet-generator

Economic Topic		
Science Topic		
Politics Topic		

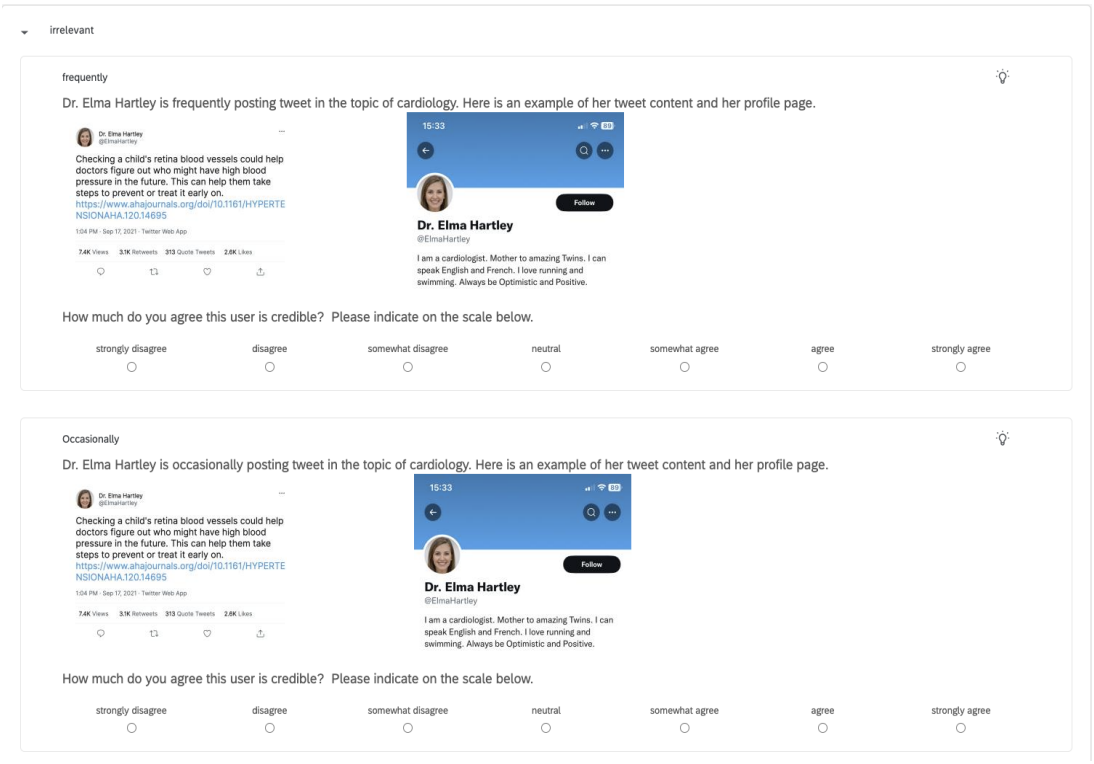
Table 5.10: First section of the second round in follow-up survey: tweets evaluation of tweets containing either URL/hashtag



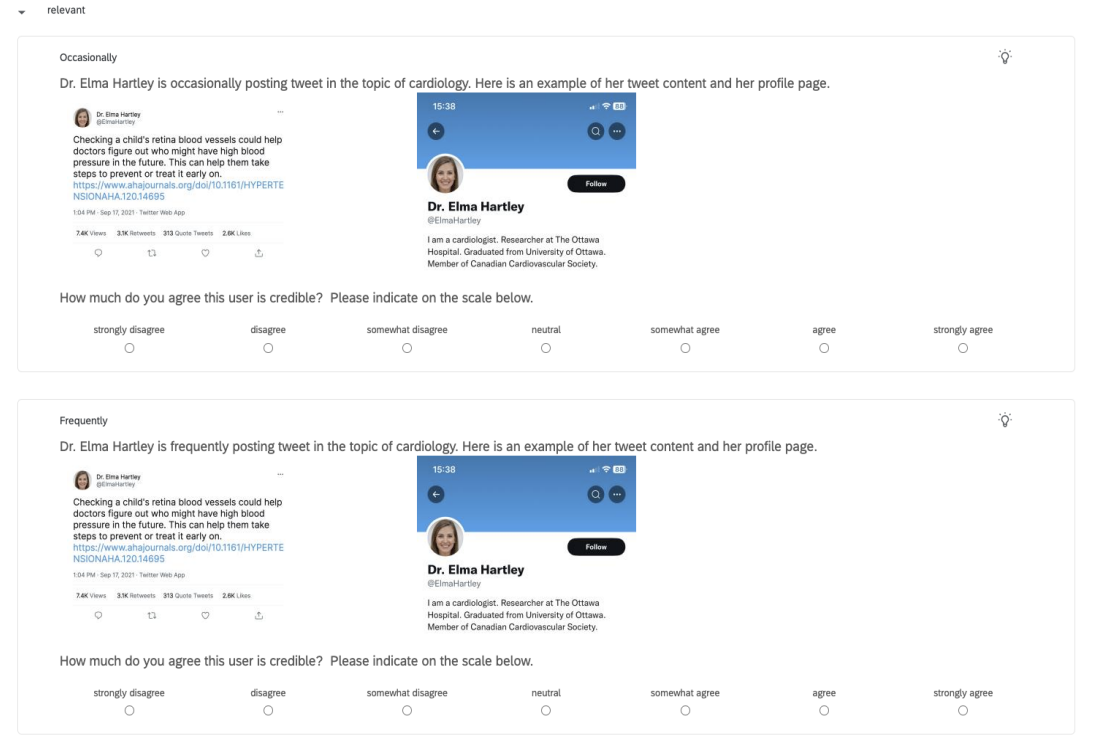
(a)

(b)

Figure 5.8: First section of the first round in follow-up survey: questions regarding to participants themselves



(a) Irrelevant group



(b) Relevant group

Figure 5.9: Second section of the second round in follow-up survey: users evaluation

Rounds	RQ	Dependent variable	Independent variable	Survey section
First round	R1, R2	Credibility evaluation on 7-point Liker scale	Participants' motivation (News/search answer for a topic/sharing content with others/viewing photos and videos/promoting Business); participants years joined twitter (numerical); and Participants' interactivity on Twitter (never/sometimes/about half the time/most of time/always); Participants' other social media usage (numerical)	Part one: about type of participants
			Users' title in display name ('Dr.'/CEO/no title)	Part two: participants' evaluation by reading tweets
Second round	R3	Credibility evaluation on 7-point Liker scale	Users' tweets containing URL/hashtag	Part one: participants' evaluation by reading tweets
			Users' posting frequency (occasionally/frequently) X users' bio info level (relevant/irrelevant)	Part two: participants' evaluation by reading tweets

Table 5.11: Summary of follow-up survey plan

Sample Size	First Round (5min)	Second Round (3min)
180	180£	108£
300	300£	180£
500	500£	300£
1000	1000£	600£

Table 5.12: Sample size and budget plan options for follow-up experiment

Chapter 6

Final Considerations

For many people, Twitter is a great source for getting news and searching for some information and others' opinions, thus the trustworthiness of the information source and content matters for these people, especially regarding healthcare related information which may impact viewers' diagnosis and treatments. On the other hand, Twitter has the disadvantage of easily spreading unverified misinformation and vague content due to tweets' posting length limitation. Therefore, my thesis is aiming to investigate the credibility of Twitter users with qualified titles in their display name.

This study explores Twitter users who have 'Dr.' and 'M.D.' titles in their display name and their credibility. I tested three research questions from three perspectives. The direct relationship between these titles in users' display name and their impact on viewers' credibility assessment; the type of viewers who are impacted the most; and how users with 'Dr.' title can impact on viewers. I set six hypotheses to test and finished the study in three phrases. At the first phase, I chose method of measuring credibility and measurement perceptions of Twitter based on previous studies. The second phase was about designing the experiment and survey, which was carefully considered and organised within budget constraint. I discussed the final version of the survey with experiment experts within the research group to collect opinions, and pretested it with 15 participants. The last phase was data analysis, ANOVA test and T-Test were used for different hypotheses. Based on the current findings, I provided further experiment design options as consideration.

This study finds that including 'Dr.' title and 'M.D.' title in users' display name can have significant impact on viewers than not any titles included. Also, users who have 'M.D.' title in their display name could gain more credibility than other titled users in healthcare topic tweets. When viewers reading the same 'M.D.' titled users' tweets, viewers who never used Twitter are impacted the most, and monthly viewers are impacted the least. This study did not find significant evidence to suggest how the users who own a real PhD degree can do to increase their credibility on Twitter. However, the results indicated that it is a good strategy for these users to include a paper link in their tweet to support their statement, or provide tweet related bio info on their profile page.

There is still much work that can be explored and tested in the future. For instance, future work can investigate additional credibility perceptions and utilize measurement

methods with larger sample size. Also, it can involve trained classifiers or other automatic assessment techniques to measure credibility. I believe the presented results are of key importance for both information seekers and information providers to spread credible information online. Both parties can benefit from the implicated strategies by understanding how credibility is assessed online. I believe my findings in this study provide a foundation for further work in this area to explore more about credibility online.

Bibliography

[1] Majed Alrubaian et al. "A credibility analysis system for assessing information on twitter". In: *IEEE Transactions on Dependable and Secure Computing* 15.4 (2016), pp. 661–674.

[2] Cory L Armstrong and Melinda J McAdams. "Blogs of information: How gender cues and individual motivations influence perceptions of credibility". In: *Journal of Computer-Mediated Communication* 14.3 (2009), pp. 435–456.

[3] Barbara Biasi and Heather Sarsons. "Flexible wages, bargaining, and the gender gap". In: *The Quarterly Journal of Economics* 137.1 (2022), pp. 215–266.

[4] Tom Buchanan. "Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation". In: *Plos one* 15.10 (2020), e0239666.

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter". In: *Proceedings of the 20th international conference on World wide web*. 2011, pp. 675–684.

[6] Kay Deaux and Laurie L Lewis. "Structure of gender stereotypes: Interrelationships among components and gender label." In: *Journal of personality and Social Psychology* 46.5 (1984), p. 991.

[7] Gustavo Deco et al. "The dynamic brain: from spiking neurons to neural masses and cortical fields". In: *PLoS computational biology* 4.8 (2008), e1000092.

[8] Helen Donelan. "Social media for professional development and networking opportunities in academia". In: *Journal of further and higher education* 40.5 (2016), pp. 706–729.

[9] Stephanie Edgerly and Emily K Vraga. "The blue check of credibility: Does account verification matter when evaluating news on Twitter?" In: *Cyberpsychology, behavior, and social networking* 22.4 (2019), pp. 283–287.

[10] Briana D Ekanem. "Journalists on Twitter: Followers, Gender and Perceptions of Credibility". PhD thesis. Ohio University, 2020.

[11] Gunther Eysenbach. *Credibility of health information and digital media: New perspectives and implications for youth*. 2008.

[12] DaJuan Ferrell and Celeste Campos-Castillo. "Cue the Doctor: An Online Experiment to Understand the Factors Affecting Physician Credibility on Twitter When Sharing Health Information". In: ().

[13] Andrew J Flanagin and Miriam J Metzger. "The perceived credibility of personal Web page information as influenced by the sex of the source". In: *Computers in human behavior* 19.6 (2003), pp. 683–701.

[14] Brian J Fogg and Hsiang Tseng. "The elements of computer credibility". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 1999, pp. 80–87.

[15] Ryan Holmes. *Is covid-19 social media's levelling up moment?* 2020. URL: https://www.forbes.com/sites/ryanholmes/2020/04/24/is-covid-19-social-medias-levelling-up-moment/.

[16] Mi Rosie Jahng and Jeremy Littau. "Interacting is believing: Interactivity, social cue, and perceptions of journalistic credibility on Twitter". In: *Journalism & Mass Communication Quarterly* 93.1 (2016), pp. 38–58.

[17] Allan Jay. *Number of Twitter users 2022/2023: Demographics, breakdowns & predictions*. 2023. URL: https://financesonline.com/number-of-twitter-users/.

[18] Simon Kemp. *Digital 2021: Global Overview Report - DataReportal â Global Digital Insights*. 2021. URL: https://datareportal.com/reports/digital-2021-global-overview-report.

[19] Haewoon Kwak et al. "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 591–600.

[20] Rick Levine et al. "The cluetrain manifesto: The end of business as usual". PhD thesis. Univerza v Mariboru, Ekonomsko-poslovna fakulteta, 2010.

[21] Ruohan Li and Ayoung Suh. "Factors influencing information credibility on social media platforms: Evidence from Facebook pages". In: *Procedia computer science* 72 (2015), pp. 314–328.

[22] Xialing Lin and Patric R Spence. "Identity on social networks as a cue: Identity, retweets, and credibility". In: *Communication Studies* 69.5 (2018), pp. 461–482.

[23] Giulia Lona et al. "Retinal vessel diameters and blood pressure progression in children". In: *Hypertension* 76.2 (2020), pp. 450–457.

[24] James C McCroskey and Jason J Teven. "Goodwill: A reexamination of the construct and its measurement". In: *Communications Monographs* 66.1 (1999), pp. 90–103.

[25] Marcus Messner, Maureen Linke, and Asriel Eford. "Shoveling tweets: An analysis of the microblogging engagement of traditional news organizations". In: *The official research journal of the International Symposium on Online Journalism*. Vol. 2. 1. 2012, pp. 74–87.

[26] Miriam J Metzger et al. "Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment". In: *Annals of the International Communication Association* 27.1 (2003), pp. 293–335.

[27] Meredith Ringel Morris et al. "Tweeting is believing? Understanding microblog credibility perceptions". In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 2012, pp. 441–450.

[28] John Newhagen and Clifford Nass. "Differential criteria for evaluating credibility of newspapers and TV news". In: *Journalism quarterly* 66.2 (1989), pp. 277–284.

[29] Stefan Palan and Christian Schitter. "Prolific. acâA subject pool for online experiments". In: *Journal of Behavioral and Experimental Finance* 17 (2018), pp. 22–27.

[30] Yash Pershad et al. "Social medicine: Twitter in healthcare". In: *Journal of clinical medicine* 7.6 (2018), p. 121.

[31] Richard E Petty et al. "Affect and persuasion: A contemporary perspective". In: *American Behavioral Scientist* 31.3 (1988), pp. 355–371.

[32] Kevin Poulsen. "Firsthand reports from california wildfires pour through twitter". In: *Retrieved February* 15 (2007), p. 2009.

[33] *Prolific, quickly find research participants you can trust.* URL: https://www.prolific.co/.

[34] *Qualtrics XM - experience management software.* 2023. URL: https://www.qualtrics.com/uk/?rid=ip&prevsite=en&newsite=uk&geo=CH&geomatch=uk.

[35] Wullianallur Raghupathi and Viju Raghupathi. "Big data analytics in healthcare: promise and potential". In: *Health information science and systems* 2 (2014), pp. 1–10.

[36] Thomas N Robinson et al. "An evidence-based approach to interactive health communication: A challenge to medicine in the information age". In: *Jama* 280.14 (1998), pp. 1264–1269.

[37] Published by S. Dixon and Aug 22. *Global Daily Social Media Usage 2022.* 2022. URL: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.

[38] Ana Lucía Schmidt et al. "Anatomy of news consumption on Facebook". In: *Proceedings of the National Academy of Sciences* 114.12 (2017), pp. 3035–3039.

[39] Felipe B Schuch et al. "Exercise as a treatment for depression: a meta-analysis adjusting for publication bias". In: *Journal of psychiatric research* 77 (2016), pp. 42–51.

[40] Kai Shu, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection". In: *Proceedings of the twelfth ACM international conference on web search and data mining.* 2019, pp. 312–320.

[41] S Shyam Sundar and Clifford Nass. "Conceptualizing sources in online news". In: *Journal of communication* 51.1 (2001), pp. 52–72.

[42] Shawn Tseng and BJ Fogg. "Credibility and computing technology". In: *Communications of the ACM* 42.5 (1999), pp. 39–44.

[43] Céline Vetter et al. "Night shift work, genetic risk, and type 2 diabetes in the UK biobank". In: *Diabetes care* 41.4 (2018), pp. 762–769.

[44] C Nadine Wathen and Jacquelyn Burkell. "Believe it or not: Factors influencing credibility on the Web". In: *Journal of the American society for information science and technology* 53.2 (2002), pp. 134–144.

[45] *What is the Twitter follow limit and ratio? | twitter help.* URL: https://help.twitter.com/en/using-twitter/twitter-follow-limit.

[46] Jeongwon Yang and Yu Tian. "Others are more vulnerable to fake news than I Am: Third-person effect of COVID-19 fake news on social media users". In: *Computers in Human Behavior* 125 (2021), p. 106950.

[47] Benxiang Zeng and Rolf Gerritsen. "What do we know about social media in tourism? A review". In: *Tourism management perspectives* 10 (2014), pp. 27–36.

List of Figures

List of Tables