AN EVIDENCE-BASED DIGITAL NUDGING IN SUPPORT OF HEALTH MISINFORMATION ASSESSMENT ON SOCIAL MEDIA SITES

by

HAMAD ALSALEH

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2022

Approved by:

____
Dr. Lina Zhou

____
Dr. Dongsong Zhang

____
Dr. Chao Wang

____
Dr. Shi Chen

____
Dr. Allison McCulloch

ABSTRACT

HAMAD ALSALEH. An Evidence-Based Digital Nudging in Support of Health Misinformation Assessment on Social Media Sites. (Under the direction of DR. LINA ZHOU)

In recent years, social media have dramatically improved the dissemination speed of information, which also includes health misinformation. To date, most of the computational approaches to addressing this problem have focused on detecting and flagging misinformation content. However, the majority of these approaches have overlooked many important aspects of health misinformation, such as the behavior of evidence sources and the sharing decisions of social media users. To address the limitations, this dissertation research develops an evidence-based approach to detecting health misinformation and to intervening user sharing intention on social media sites. This work takes on a new perspective regarding health misinformation by understanding user stance (i.e., for, against, neutral) due to their motivation of influencing others. Moreover, this research investigates arguments that combine both stance and evidence for assessing the credibility of health information for the very first time. Our analysis of evidence distribution in health information tweets shows that 70% of tweets contain source-based evidence, which provides the foundation for proposing an evidence-based approach to misinformation detection. Based on these results, we built argument detection models to identify stance positions within arguments. Our results demonstrate the importance of evidence-based features in identifying the stance within arguments on social media sites. Drawing on the evidentiality theory, information credibility heuristics, and consistency heuristics, we propose a research model that seeks to explain health misinformation detection and sharing behavior with evidence-based interventions. To test the research model, we designed and developed eleven types of evidence-based digital nudges and used them to conduct user experiments. The empirical results demonstrate that our nudge design improves credibility assessment of health misinformation. This dissertation makes several research

contributions. First, it extends an evidentiality theory and credibility cognitive heuristics provided by health experts to analyze the types of evidence included in health-related user generated content Second, it presents an evidence-based schema for categorizing evidence in user generated content . Third, it uses evidentiality theory as the kernel theory to guide the design of digital nudges. In particular, it illustrates how evidence-based design artifacts can be used to support augmented intelligence for mitigating the spread of health-related misinformation on social media sites. Finally, it combines cognitive heuristics to the design of digital nudges. Specifically, it uses information credibility and consistency heuristics to analyze user generated content on social media sites. The outcomes of this research have significant implications for augmenting users' assessment of health information credibility and enabling timely intervention of misinformation on social media sites.

ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr.Lina Zhou, chair of my dissertation committee and my advisor for her patience, motivation, and vast knowledge during the last years of my Ph.D. research. Her guidance helped me throughout this entire process of writing and researching my dissertation. I have known Dr.Zhou since my masters' studies, and I am very grateful for working under her guidance all these years. I could not have imagined having a better advisor and mentor while studying for both my master's and Ph.D.

Besides my advisor, I would like to thank Dr.Dongsong Zhang for his guidance and encouragement. His insightful comments gave me the motivation to widen my research scope from different angles.

Also, I would like to thank other members of my dissertation committee, including Dr. Chao Wang, Dr. Shi Chen and, Dr. Allison McCulloch, for their kindness and time through serving on my dissertation committee. Finally, a special thank you to my peers at the KAIM lab for their support through the toughest time while working on this dissertation. I am thankful to Louai Mohammed for his support since my first day at UNC Charlotte. Also, I would like to thank Ali Almadan for his insightful comments throughout this process. I am also thankful to one of my dearest friends in Saudi Arabia Fahad Alamari for his constant emotional support during my studies.

DEDICATION

This Dissertation is dedicated to my mother and father who taught me to be hopeful during tough times. Without their support, I would not have been able to accomplish this work.

To my siblings Abdullah and Monerah for their endless love and support. To my brother Mohammed for his constant encouragement during our times in Maryland and his endless love and support during my studies in the United States.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Problem Statement

This section discusses the problem of health misinformation on social media sites and the motivation of this research.

1.1.1 Misinformation on Social Media Sites

With the advancement of technologies, especially with the Internet and web 2.0, content creation and dissemination have become easier and faster. The development of social media sites (SMS) facilitate the spread of user-generated content (UGC) worldwide. As a result, the world is now essentially all under one roof. Posting content on SMS incurs little or no financial cost to reach a large audience. Consequently, SMS motivate the manipulation of content to fulfill any type of agenda or intent.

An increasing number of people depend on SMS daily as their outlet for local and world news and events. Today, people debate whether these sites will replace or already have replaced television-based media [1]. Due to the large amount of information consumed and the freedom of expression offered by these sites, the risk of running any type of information into misinformation is not negligible. On the one hand, SMS are utilized to connect to the world, but, on the other hand, they provide the potential to deliver misinformation [2].

Many studies have shown that health information consumers on SMS tend to share misinformation without further checking its veracity [3], [4]. In addition, previous research indicates that social pressure and selective exposure were among the reasons for social media users to believe misinformation [5]. Many account holders on SMS take advantage of such factors and spread misinformation that targets different communities with news that influences people toward psychological biases and causes them to believe misinformation [6].

One factor for amplifying the problem of misinformation on SMS is that they offer users the ability to customize their news feeds, which creates the problem of filter bubbles; such a phenomenon makes users less dependent on the news feeds that contradict their views and beliefs [7]. Because of the complexities involved in the production and evaluation of misinformation on SMS, scholars weigh on the importance of collaborations from the different fields of computer science, psychology, and social science to mitigate the spread of misinformation on SMS [1].

1.1.2 Health Misinformation on Social Media Sites

Health information such as looking for treatment advice or knowledge about a disease has been spread widely in the past years. The Pew Research Center has reported that around (60%) of e-patients' decisions (a term used for calling people who look for online health information) seek online health information in treating their illness or health conditions [8]. It was reported that 30% of e-patients consume health information found in UGCs on SMS [8].

An abundance of studies investigated the propagation and sharing of health misinformation. Health misinformation is defined as a "health-related claim that is based on anecdotal evidence, false, or misleading owing to the lack of existing scientific knowledge" [9]. One study [10] was conducted on Twitter to explore the misunderstanding and the misuse of antibiotics for treating certain health conditions. They found many individuals have been using antibiotics to treat cold, as the authors found that a significant number of users mention the words *cold* and *antibiotics*. Another misuse of antibiotics was reported in the study, such as sharing antibiotics or using antibiotics for treating *flu* conditions [10].

A major type of health misinformation distributed on SMS is the spread of fake cures. One of these examples is the debate on SMS that cannabis cures cancer [11]. This topic's discussion generated 4 million engagements between users on these sites, including replying, retweeting, and liking these posts. On the other hand, news that combat such false claims generated a few thousand engagements, which shows that misinformation propagates very quickly compared to trustworthy information.

One study analyzed how users discuss breast cancer screening guidelines on SMS. They found out most of the sharing behavior was distributed from nonexpert individuals tweeting about natural alternatives to breast cancer treatments with no mentioned support or scientific evidence of these alternatives [12]. This shows that breast cancer misinformation has also attracted the attention of information consumers on SMS.

Health misinformation during health crises has prevailed on SMS in recent years[13]. One study conducted a survey and found that 85% of the surveyed U.S population believe in one or more of the conspiracies related to COVID-19[14]. Another example that spread in the past is the discussion of the Ebola outbreak on SMS in 2014 the circulating misinformation content about the disease [15]. Consequently, health misinformation advice by non-experts on SMS has led to deaths for some individuals[1] and poses other individuals' health into critical condition[2].

[1] https://www.nbcnews.com/health/health-news/man-dies-after-ingesting-chloroquine-attempt-prevent-coronavirus-n1167166

[2] https://www.healthline.com/health-news/poison-control-calls-going-up

1.2 General Research Motivations and Questions

Previous studies have shown the impact of health-related information compared with information of other types on the believability of users on SMS [16]. Information quality experts found that users tend to be impacted by health misinformation compared to trustworthy health-related information despite the absence of scientific evidence sources (information source) in the content [12]. Scholars indicate that a major characteristic of information quality is the evidence source that accompanies the posted information [17]. Specifically, for health-related context, an important aspect is to evaluate the scientific nature of the evidence source when discussing such topic type [18], [19].

Previous research in evaluating the quality of health information online suggested one solution to obtaining trustworthy health-related information is to focus on expert-based sources when seeking health recommendations [20]. However, given the unstructured type of content on SMS and the freedom of expression on these platforms, users with any health- and non-health-related background have a similar voice when discussing health-related topics[21]. In other words, the structure of SMS do not notify users of expert and non-expert evidence sources. In addition, users are not required to list their profession or occupation on these sites; therefore, relying on developing computation-based models on user profiles is insufficient. Chafe [22] proposed evidentiality theory which highlights the need to assess and categorize the evidence source that accompanied and support the expressed information. Hence, it is important to develop an approach that considers the source-based evidence of authoritative or scientific nature accompanied by such information in categorizing health information on SMS.

Wardle and Derakhshan [23] mentioned that an SMS message could be generated from an organization or an individual. One property of misinformation online is that it can be in any form[24]. One of the UGC’s textual forms on SMS is a stance generated by a news organization’s story headline or a stance generated from opinions (expressing a stance without incorporating an evidence source) of any user on SMS [23], [24],[25]. Current approaches to detect health misinformation rely on keyword analysis of posts[26],[27]. However, a post that includes a fake cure keyword does not imply that the author’s stance supports taking this cure. Similarly, other works that developed computational models to detect health misinformation tend to focus heavily on news headlines. Even though computational methods have been widely developed in detecting misinformation in some areas, these methods do not deal with other aspects of misinformation, such as favoring stance toward misinformation within UGCs [2],[28],[29]. For example, one of the prominent stance types on SMS is the favoring stance toward the negative impacts of MMR vaccination [30], [31].

Studies in the area of argumentation mining and linguistics have highlighted the importance of studying the structure of content to determine the stance toward certain topics[32]. There are emerging studies of stance detection within opinions and arguments [32]. Previous studies in other online media have demonstrated the efficacy of stance detection in evaluating the credibility of arguments online[33], [34], [35],[36]. Viviani and Pasi [24] highlight the significant role of health-related arguments in the dissemination of health-related misinformation on SMS [37]. Castillo et al. [38]showed that users tend to believe arguments containing a supporting stance if it is incorporated with any type of evidence. Thus, identifying arguments that show supporting stance positions toward health misinformation would expand the scope of computation models in detecting other aspects of misinformation on SMS [39],[40]. It can further enhance the credibility judgment of information on SMS [41]. However, misinformation that propagates from UGCs on SMS, such as arguments toward a disease or a treatment, is rarely addressed[42]. Such an argument may expose the life of information consumers to danger when they are taken for granted [3]. Based on the previous discussion, we present the first general question in paving the road for this dissertation research.

How can we incorporate evidence source categorization to evaluate the credibility of arguments that shows a supportive stance toward health misinformation on SMS?

Current research on misinformation detection in SMS can be classified into three types: 1) developing computation-based approaches to detect misinformation [43],[44]; 2) analyzing user behavior while consuming misinformation [3],[45] and 3) developing hybrid systems that involve human-machine interaction as a major element in the decision-making process of information credibility [46], [47],[48], [49],[50]. The hybrid-based systems emerged in recent years [46], [47],[48], [49],[50]. Such systems have shown effectiveness in uncovering useful characteristics of information when analyzing information credibility on SMS [51]. They can be further categorized into a) computer visual analytics tools that help users with information credibility assessment by providing visuals and insights on the content [52],[49] and network features [52] of social media accounts; and b) human-assisted computation systems which trigger users to look into certain aspects of the information on SMS [47], [52], [48], [49]. However, these tools mainly focused on examining misinformation from the account level and do not distinguish the content level of misinformation in SMS [47], [52], [49]. An account that generates misinformation information on SMS does not imply a malicious account [52], [53]. Such generalized judgment could produce false positives, which may affect users' decision-making when assessing information credibility. Although such tools contain helpful features and insights on analyzing information credibility, existing studies do not evaluate the efficacy of these features on users' decision-making on information credibility on SMS [48],[49].

[3] https://www.nbcnews.com/health/health-news/man-dies-after-ingesting-chloroquine-attempt-prevent-coronavirus-n1167166

Human-assisted computation tools are merely developed [54], [55] in the context of misinformation on SMS. Other related tools are augmented intelligent systems, including digital nudging [56]. Such systems incorporate machine automation and human intelligence into enhancing users' decision-making [57]. Previous work [58] showed the effectiveness of the nudging systems in avoiding regrettable information disclosure on SMS [59]. However, nudging systems studies remain scarce in the area of misinformation on SMS and are limited to enhancing the detection of fake news accounts but lacking in evaluating misinformation content that originates from evidence sources within UGCs. Studies the develop an evidence-based categorization for nudging users on types of evidence incorporated with UGCs on SMS have not been conducted yet. Scholars call for developing augmented intelligent systems to mitigate the spread of health misinformation on SMS [15]. Such development will shed light on the effectiveness of the augmented-based systems compared to automated-based systems to detecting health misinformation on SMS [1], [51]. Based on the previous discussion, this dissertation proposal presents the second general research question:

How can we design a digital nudging system that incorporates evidence source categorization to effectively enhances users’ ability to assess the credibility of health misinformation on SMS?

1.3 Contributions of the Dissertation

This dissertation makes the following contributions to the literature:

First, this dissertation attempts to apply an evidentiality theory [22], a heuristics provided by health experts to analyze evidence sources incorporated with health-related UGCs, including favoring stance toward health misinformation on SMS. Applying an evidence-based approach to detect health misinformation in SMS within UGCs has not been investigated in previous works [22]. This dissertation identifies the gaps and the necessity of the approach for evaluating online health information.

Second, this dissertation constructs a source-based evidence scheme to categorize evidence within UGC. This research proposes a categorization scheme based on evidence credibility metrics[5] within UGC online. Such an effort aims to generalize the detection of misinformation on SMS to different information contexts.

Third, this dissertation expands the foundations of stance detection on SMS by differentiating between stance and argument-based UGCs on SMS. Moreover, it creates a dataset of favoring stance toward health misinformation on SMS and arguments that contain a favoring stance toward health misinformation with supporting source-based evidence.

Fourth, the dissertation extends evidentiality theory to the design of digital nudging, a type of augmented intelligence. In particular, it shows how an evidence-based approach is used to guide the design of a digital nudging system to mitigate the spread of health-related misinformation on SMS. To the best of our knowledge, this dissertation is the first to integrate source-based evidence into the design of digital nudges on users' credibility assessment and sharing intentions of health-related information on SMS.

1.4 Dissertation Roadmap

Chapter two discusses the related literature in misinformation detection and stance classification. The chapter identifies the research gaps that need further examination in health misinformation. In addition, we present a brief overview of the approach (evidence-based) we will follow in the study.

Chapter three offers empirical findings for the lack of source-based evidence of authoritative or scientific types in UGCs when discussing health-related issues on SMS. Given the critical nature of health topics, a computation process is needed to determine the quality of evidence sources embedded within arguments on SMS. Therefore, we developed a customized scraper to extract tweets and categorize the types of evidence within tweets. In the end, we provide statistical comparisons of types of evidence embedded within tweets that discuss a misinformation-related topic and non-misinformation-related topics.

Based on the findings of chapter three, in chapter four, we aim to detect types of stance expressed within arguments that relate to health misinformation on SMS. Given the essence of scientific information on health-related issues, we combine an evidence-based approach in developing argument classification models to detect arguments that support misinformation on Twitter.

Chapter five assesses how the evaluation of evidence within UGCs would contribute to user decision-making when assessing SMS misinformation. We designed EvidencEval, a system that comprises two categories of evidence-based nudges, evidence credibility nudge and viewpoint consistency nudge. The chapter presents the design of 11 types of nudges and examines their effectiveness in enhancing users' credibility assessment of health misinformation.

Chapter six describes the conclusions, findings, and contributions to the misinformation detection domain, followed by practical implications of the dissertation, limitations, and future directions.

CHAPTER 2: LITERATURE REVIEW

In this section, we discuss previous scholarly work in online misinformation detection and mitigation. We begin with the extracted features, which cover various major types of features extracted and used in misinformation analysis and detection studies. Next, we discuss the computational approaches, followed by the behavioral-based studies for misinformation detection. Finally, we cover the hybrid approach, which involves both humans and machines when analyzing misinformation on SMS.

2.1 Computational Approaches to Analyzing and Detecting Misinformation

One of the important research topics in information credibility assessment of online content is fact-checking. Fact-checking approaches are used by multiple computational models, such as fake news and rumor detection models, as well as on question answering systems for information verification.[60]. Fact-checking can be categorized into two approaches: computational and crowdsourcing [61]. Computational fact-checking addresses two significant issues, that is, check-worthy claims identification and claim veracity determination. Computational fact-checking approaches rely on external sources such as open web sources and knowledge graphs for checking the trustworthiness and the veracity of the claims [61].

Crowdsourced approaches exploit the vote of the crowds on SMS for differentiating reliable and less reliable sources [62]. An example of a crowdsourced website is Fiskkit, on which users can provide ratings and tags for news articles [43]. Expert crowdsourcing websites rely on expert judgments to judge the veracity of online news content. Several expert-based checking websites have emerged in past years, such as Snopes,[4] PolitiFact,[5] FactCheck,[6]

[4] https://www.snopes.com
[5] https://www.politifact.com
[6] https://www.factcheck.org

TruthOrFiction,[7] Full Fact,[8] and the *Washington Post* Fact Checker.[9] These sites cover different topics, but mainly American politics.

We categorize previous computational techniques developed for misinformation detection on SMS into two main categories: classification-based and propagation-based approaches. Classification-based approaches leverage traditional machine learning or state-of-the-art deep learning techniques in classifying misinformation into binary or multiple classes. Typically, classification-based approaches apply NLP techniques and require an annotated data set in order for the computational model to predict an outcome. On the other hand, propagation-based approaches consider the network's structure and features, propagation behavior of retweets, and temporal characteristics. Table 1 presents an overview of computation-based studies.

[7] https://www.truthorfiction.com
[8] https://fullfact.org
[9] https://www.washingtonpost.com/news/fact-checker/

Table 1: Summary of computation-based studies.

Article	Year	Misinformation Topic	Misinformation Type	Approach
[38]	2011	Trending Topics, Multiple contexts	Textual: news and rumors	Classification Based, Binary, Traditional Machine Learning
[63]	2013	Trending Topics, One Context (Crisis, Natural Disasters)	Textual: news and rumors	Classification Based, Binary, Traditional Machine Learning
[28]	2017	Normal Topics, Multiple Contexts,	Textual: news and rumors	Classification Based Multiclassification, Traditional Machine Learning + Deep Learning
[64]	2013	Trending Topics, One Contexts (Crisis, Natural Disasters)	Visual: news and rumors	Classification Based, Binary, Traditional Machine Learning
[65]	2012	Trending Topics, Multiple contexts	Textual: news and rumors	Classification Based, Multiclassification, Traditional Machine Learning
[66]	2016	Normal Topics, Multiple Contexts	Textual : news	Propagation Based method (Credibility Propagation with Conflicting Viewpoints CPCV)
[67]	2012	Normal topics, Multiple Contexts	Textual: rumors	Classification Based , Binary, Traditional Machine Learning
[68]	2013	Normal Topics, Social context	Textual + visual : rumors	Classification Based , Binary, Traditional Machine Learning
[69]	2013	Normal Topics, Multiple Contexts	Textual : rumors	Classification Based , Binary, Traditional Machine Learning
[70]	2015	Trending Topics , Multiple Contexts	Textual : rumors	Classification Based , Binary, Traditional Machine Learning
[71]	2016	Normal Topics, Multiple Contexts	Textual : rumors	Classification Based , Binary, Traditional Machine Learning
[72]	2016	Controversial Topics, One context	Textual: rumors and stance	Classification Based , Binary and Multi Classification, Traditional Machine Learning
[73]	2017	Controversial Topics, One Context	Textual: rumors and stance	Classification Based , Binary and Multi Classification, Traditional Machine Learning
[74]	2019	Normal topics, Multiple Contexts	Textual: Satire, Hoax, Propaganda, Clickbait	Classification Based Multiclassification , Deep Learning
[75]	2017	Normal topics, Multiple Contexts	Textual: Rumors, Hoax, Propaganda, Clickbait	Classification Based Multiclassification , Deep Learning
[76]	2014	Trending topics, One Context	Textual: rumors, fake news	Propagation Based method (Multiple Credibility Assessment)
[77]	2016	Trending Topics, One Context (Political Context)	Textual : news	Propagation Based method (Similarity between tweets/users and other characteristics

2.1.1 Classification Based Methods

Several studies on misinformation detection on SMS offer binary classification (fake, true) or multi-class classification models for combating different types of misinformation. The multi-class classification models label data into true or half-true news, unverified news, entirely true rumor, half false, and undefined rumor.

For detecting both textual fake news and rumors on Twitter, Mendoza et al. [38] conducted early work to detect time-sensitive malicious tweets from 2,500 topics in multiple contexts(e.g., political-related, health-related, disasters-related, etc.). They extracted different types of features, including content-, user-, topic-, and propagation-based features. Seven annotators generated data annotation on Amazon Mechanical Turk. Decision tree classifier J48 obtained the best result with (86%) accuracy. The best features in determining the credibility of tweets were hashtags, URLs(i.e., evidence source) existence, and user mentions. The same authors extended the previous work to detect fake news and rumors regarding natural disasters. For the creation of topic events, they used two methods; one is based on the frequency of keywords, and the other is based on clustering methods. The method based on the frequency of keywords was robust in classifying newsworthy topics. The model yielded a good result with (81%) for classifying newsworthiness and (82%) for post credibility [63].

Gupta et al.[65] also focused on analyzing the credibility of fake news and rumors on Twitter's trending topics of multiple contexts in 2011. They used the Support Vector Machine (SVM) rank classifier that achieved reasonable accuracy when mixing content- and user-based features. However, the models achieved higher performance when adding more content-based with context-based features, such as n-grams. Wang et al. [28] used traditional machine learning methods and deep learning methods to detect fake news and rumors on SMS. They build their methodology on a LIAR data set containing 12,836 statements from professionals on PolitiFact.com. The statements are labeled with six different veracity labels. They proposed a hybrid architecture by integrating textual information on the CNN model and information about the author of the statement by the long short-term memory (LSTM) layer. The model outperformed other models that used only content features with 26% accuracy by integrating content features and message meta-based features.

Several studies focused on identifying rumors only on Twitter. For instance, Kwon et al.[69] worked on identifying rumors in multiple contexts. The authors collected tweets related to 130 rumor topics classified from several fact-checking sites such as Snopes and Urban Legends.[10] The categorization ended up with 70 rumor topics and 60 non-rumor topics. After data collection, they hired four annotators to classify the tweets based on their relevance to the rumor and non-rumor topics. Three feature types were extracted from the data: linguistic features, temporal features, and structural features. Using random forest and logistic models to identify the most important features in distinguishing rumor and non-rumor tweets, both models showed that integrating periodic external shock features enhanced the rumor-detection performance. Giasemnidis et al.[71] also worked on determining the veracity of rumors of multiple contexts on SMS. After collecting about 100 million tweets of 31 false rumors and 41 true rumors, they applied a manual labeling method and extracted multiple features, including content-, user-, and network-based features. A decision tree classifier achieved the best performance with 96% accuracy using only six features.

By identifying rumors about trending topics on Twitter, Zhao et al. [70] exploited inquiring tweets to generate rumor clusters. The authors collected 10,417 tweets labeled as "verification" or "correction." Their technique is based on grouping similar posts on the same cluster and ranking clusters that contained only disputed factual claims (i.e., by clustering only

[10] https://www.liveabout.com/urban-legends-4687955

tweets that contained inquiry signals). For verification, they collected another set of trending events related to the Boston Marathon bombing, and they hired two annotators to label a total of 639 clusters. Thirteen features were extracted, including message-based features, network-based features, and percentage of signal tweets (e.g., tweets with words such as “Really?” and “Is this real?”). A decision tree classifier fed with these features yielded 52% accuracy in identifying rumors on Twitter. On Sina Weibo, Yang et al.[67] worked on identifying rumors in multiple contexts and accessing the rumors’ ground truth, which is a special service offered by Sina Weibo. They extracted three types of feature categories: content-, user-, and propagation-based features. They trained an SVM classifier on the data set and tested the model on three independent features (i.e., content, user, propagation) and had accuracies of 72.5%, 72.6%, and 72.3%, respectively. Adding the location of the event and the type of device used by the users to post enhanced the results, enabling the model to reach an accuracy of 78.7%.

Another type of work that uses traditional machine learning is detecting opinion spam of controversial topics. Stance detection is the task of classifying whether content posted by users supports or denies a specific claim [78]. Addawood et al.[72] introduced a gold-standard data set of 3,000 tweets related to the FBI and Apple encryption debate. The authors collected 531,593 tweets and generated four types of content sub-features, including basic features and psychometric features, which refer to dictionary-based features (e.g., the LIWC), linguistic features (e.g., sentiment features and LIWC summary variables), length of messages or whether the content contains mentions, and user-based features. They constructed a classifier that determines whether the tweets contain an argument or not. However, their classifier do not classify the quality of evidence type within these argumentative tweets. The SVM classifier achieved the best performance with an F1 score of 82% [72]. On other work that related to controversial topics, the same authors classified the stance on the same encryption debate into two subtopics, individual privacy and national security, by relying on the same dataset of the previous work., The authors further classified opinion stance into favor, against, or neutral on the term of their opinion toward the topics. Combining argumentative and sentiment features into the SVM classifier significantly improved the result with an F1 score of 91%[73].

Ghanem et al.[74] introduced a deep learning network combined with emotion features to detect fake news on Twitter and news article sources. They considered several emotion-based resources, including Emotion Association Lexicon (EmoLex), SentiSense, LIWC, and Empath, which resulted in 17 emotions. They chose the LSTM model with the word “embedding” (content- and emotion-based features) as an input for predicting false information types. For classifying false news, they used a random forest classifier as it showed better performance than the other two classifiers. The emotion-infused model, which used only emotion features, achieved better performance than other baselines with an accuracy of 96% in detecting false information on both data sets (i.e., news article sources and Twitter). They discovered that the words and sentiment showed significant statistical differences in identifying false news from real on both data sets. For differentiating between false information types, they found that clickbait tended to attract the readers' emotions by including the “surprise” emotion. For propaganda news, they tended to create extremely positive and negative emotions with calmness, triggering and providing feelings of confidence. In contrast, satire news tended to result in disgust emotions for humor purposes. Their work showed the importance of NLP, particularly in using emotion features to detect false information across different information types.

Volkova et al. [75] used CNN and RNN to detect false information types, including rumors, hoaxes, and propaganda. The data was collected from several websites that identified propaganda and hoax accounts, such as PropOrNot.com and fakenewswatch.com. For verified accounts, they manually labeled 252 accounts and checked if these accounts had the verification badge on Twitter. Moreover, they extracted several linguistic features that correlate to bias cues, subjectivity cues, psycholinguistic cues, and moral foundation cues. They then fed these neural network models' features and compared them with the baseline logistic regression model. Both models (i.e., CNN and RNN) outperformed the result obtained by logistic regression by reaching an accuracy of 95%.

2.1.2 Propagation Based Methods

Jin et al.[76] developed a hierarchical propagation model to detect fake news and rumors related to a trending topic on Sina Weibo. They introduced a three-layer approach to capture deeper semantics on the event. Their approach enhances credibility assessment of the event by passing it through events, subevents, and message layers. The approach was applied to a data set of tweets related to the lost Malaysia Airlines Flight MH370. The data set has 110,822 tweets annotated by authoritative sources such as the Xinhua News Agency and Sina Weibo. They extracted three types of feature categories: content-, user-, and message-based features (e.g., post time and the number of comments). The proposed method achieved an accuracy of 85%, which was better than the baseline methods. On the other hand, propagation-based methods have been applied to detect fake news in multiple contexts.

Jin et al.[66] exploited conflicting viewpoints of users’ comments when responding to news of untrendy topics on Twitter. They hypothesized that tweets with the same viewpoints tended to increase the credibility of the tweets. On the other hand, contracted viewpoints tended to weaken the credibility of tweets. Thus, the supporting and contradicting relations can help in judging the credibility of the tweets by extracting sentiment information and providing informative sentiments. The conflicting viewpoint topics were extracted using an unsupervised topic model technique, and the problem was formulated as a graph optimization problem. The authors constructed a data set from Sina Weibo with 49,713 tweets. The model they built took advantage of conflicting viewpoints from users’ responses to find credible statements. When comparing their special propagation model with other classification baseline models by[38], they found their model outperformed these models, reaching an accuracy of 84% in verifying tweets on SMS. In addition, for early detection of fake news on SMS, their model resulted in a faster and more accurate response than the classification-based methods, which shows the strength of propagation for credibility identification of news.

Propagation-based techniques have also been applied by Hua et al.[77] to detect trending fake news that relate to the political context on Twitter. They proposed a topic-focused framework for assessing the trustworthiness of tweets about real events. Their framework considered the relationships among authorship, retweeting behavior, and replying behavior to determine the trustworthiness of the tweets by considering both the contextual and social aspects of the tweets and users. Three types of features were collected: content-, user-, and message-based features (temporal and spatial). They compared their method with classification-based methods and found that the results of their approach outperformed classification models based on manual labeling, with the highest F1 score reaching (87%) in detecting topic-focused real events in 7 out of 10 Latin American countries.

2.2 Behavior-Based Methods for Analyzing and Detecting Misinformation

Behavior-based studies test users’ responses and behavior when consuming misinformation. Users' involvement in SMS amplifies the problem of misinformation diffusion [23]. This is mainly due to the receiver's cognitive biases, thus leading to misinformation spread on SMS [79]. Thus, it is important to study user behavior in consuming information on SMS misinformation.

Several studies conducted behavioral experiments to examine users' perception and credibility judgments toward content and non-content features. Lucassen et al.[80] evaluated the relationship between topic familiarity and the source of information on credibility judgment of Wikipedia articles that relate to multiple topics contexts. Half of the articles collected were familiar to participants, and half were not. The source cues availability factor was between subjects, and the former factors were within-subjects. Participants are presented with articles in the Wikipedia layout and other articles in the Microsoft Word layout without Wikipedia as a source cue. They found that the trust of users who are familiar with the viewed information was negatively influenced when they recognized the source was Wikipedia. Another finding involved is the credibility of participants in distinguishing between high-quality articles and low-quality articles, regardless of whether the source is mentioned or not.

Setfanone et al.[81] investigated the relationship between credibility and sharing behavior of news in the political context on SMS. They collected news articles from allsides.com,[11] a website designed to identify the orientation and biases of articles that originated from well-known global news sources. They presented the information in a layout that users typically see when browsing SMS. They used linear regression analysis on the users' answers and found that perceiving information as credible leads to sharing the information with the user's network. The distraction level and screen size did not affect sharing behavior.

In contrast, the political interest and religious engagement factor directly affected sharing behavior and perceiving the information as credible. Focusing on Facebook, Flintham et al. [3] utilized a think-aloud methodology in assessing content and non-content features that influence users to believe multiple contexts of fake news. They first conducted a survey and found that individuals rely on source reputation, headlines and full text, and plausibility to

[11] https://www.allsides.com/unbiased-balanced-news

assess news's credibility on Facebook. Through an experiment, they found that participants rely on the source's authority in judging the credibility of the information. Humorous or satirical stories were not perceived seriously, and they assumed such articles were not truthful. The date and formality of the news article were also taken as measures for believing the news. Additionally, users might not believe a news story just because they do not have an interest in it, which affects real news stories that are not aligned with users' interests.

On the other side, Morris et al.[82] applied a think-aloud methodology on Twitter users and evaluated the impact of content and noncontent features on their perceptions of news in multiple contexts. They found that features such as images with an avatar and the number of followees negatively correlated with credibility perception. In contrast, the association of the topic to the user's interest and retweet of the images positively affected credibility perception. Also, on Twitter, Shariff et al.[83] investigated the incorporation of visual features to content and noncontent features when examining users' perceptions. They aimed to assess users' backgrounds and their relations to credibility judgment on SMS. They found that the prediction tool provided different ratings from those annotated by the users. Despite adding visual features in their evaluation, they found that only topic and style features correlated significantly with users' demographics and backgrounds. The study revealed that the more educated the users were, the more cautious they were when making credibility judgments.

Kang et al.[4] assessed content and non-content features on users' perceptions on Twitter and Reddit. Their research highlighted that visual features affect users' perception to rate the information on SMS as credible. Besides, the content and source of the message were other determinants of online information. They tested users by showing images of three Reddit microblogs and four Twitter microblogs. The results showed that in contrast to the conducted survey, message metadata features (e.g., number of retweets and user mentions) and visual features had a significant positive effect on users' judgments on both platforms. In contrast, textual features showed a relative impact on users' judgment on Reddit. Their study highlights the importance of visual features on perceived credibility judgment, which may motivate the use of images to spread misinformation.

Focusing on a specific context, Vaidya et al.[84] tested the association between credibility judgment and account verification when consuming only false health-related information on Twitter. Their primary aim was to investigate whether the existence of authenticity indicators (i.e., the blue verification badge vs. the "Verified account" textual label) affects users' perception and credibility judgment. The result showed that users differentiate between authenticity and credibility on SMS such as Twitter, and there is no correlation between the verification badge on an account on Twitter and the intent for the user to share the information from that account. The majority of users were influenced by the content of the tweets more than non-content features, such as the verification status of accounts. Li and Suh [85] evaluated users' perception of content and non-content features toward false political news only. Drawing on the elaboration likelihood model (ELM), they explored the effect of medium and message as dimensions of credibility on Facebook pages. They found that medium transparency, argument strength, and perceived interactivity positively correlated to information credibility on Facebook pages.

There were previous studies that assessed the impact of content features only on users' perceptions. Lin et al. [16] reevaluated the MAIN model[86] and the effect of identity and authority cues on users' perception of SMS. In their study, they hypothesized that health information related to drug-resistant gonorrhea would be trusted when it came from an expert more than when it came from a peer, and the same information would be trusted when it came from a peer more than when it came from a stranger. They found that users trust the information from the CDC more than when it came from strangers or peers. Also, an expert source of information was rated the highest in terms of the goodwill and trustworthiness measures. On the other hand, they did not find differences in users' credibility judgments when the information came from a stranger or peer. Their study highlighted the effect of authority cues over identity cues.

Employing an eye-tracking device in their study, Simko et al.[87] examined user perceptions toward content features of multiple news contexts on a self-developed website that exhibits similar characteristics to Facebook. The goal was to assess participants' veracity rating selection of true and fake articles and their behavioral traits in sharing, liking, disliking, and reporting these articles. In their findings, they noticed participants shared and liked truthful articles more than fake articles. Furthermore, participants relied on their interest when examining the news article to determine their veracity by examining the articles' content deeply when the articles aligned with their interests. Unsuccessful participants tended to look at the feed more than their successful counterparts, whereas successful participants tended to look more at the content than unsuccessful counterparts. Another discovery is that participants' interests are reflected from the strength of opinions answered on the questionnaire.

Pal et al.[88] examined the motivations for debunking misinformation. They used denials to debunk the spread of health-related rumors on Facebook and Twitter. They conducted two studies to examine the effect of salient beliefs in combating rumors. In their first conducted survey, they identified three top salient beliefs underpinned by the theory of planned behavior (TPB)—behavioral, normative, and control beliefs—that have the potential to influence users' intentions to share denials. Then, they examined the relationship between those beliefs and users' intentions to share denials. They created fictitious pages with four health-related false rumors retrieved from Snopes.com. They found that the most influential pages, incorporating all three belief categories and their subcategories, enhanced the intention to share denials in comparison to the placebo group. Their research showed significant factors that influence the structure of the cognitive process, which forms the intention to share denials online and specifically in the health context.

On the other hand, Shu et al.[89] examined users' capabilities in detecting misinformation on Twitter. They analyzed users' profiles and found that user age is a determinant of fake news detection; older people fail more in recognizing fake news. They also found that popularity affects trust, in which popular people tend to trust real news. In terms of user characteristics such as personality, they applied an unsupervised machine learning technique drawn on the popular five-factor model (the Big Five personality test). They found that people who tend to be extroverted were more likely to trust real news. Females tended to believe in fake news more than males. Their study highlighted the users' characteristics aspect to detect misinformation. Table 2.1 for a comparison of the discussed studies in terms of data source , numbers of participants and the target misinformation topic. Table 2.2 described the discussed behavior studies n terms of their analysis methodology , and the evaluated cues.

Table 2.1: Summary A of Behavior-Based studies on Misinformation Detection.

Article	Year	Data Source AND Ground Truth Source	Number and Type of Participants	Misinformation Topic
[80]	2013	***Data Source***: Wikipedia ***Ground Truth:*** Wikipedia	***Number of Participants***: 41 ***Type***: Students	Multiple Contexts
[4]	2015	***Data Source***: Twitter and Reddit ***Ground Truth***: Not Specified	***Number of Participant***s: 81 ***Type:*** Twitter and Reddit Users	Multiple Contexts
[16]	2016	***Data Source***: Twitter ***Ground Truth:*** Real information: CDC; False information: Authors	***Number of Participants***: 696 ***Type***: Students	Health Context
[81]	2019	***Data Source***: allsides.com ***Ground Truth:*** allsides.com	***Number of Participants***: 207 ***Type***: Students	Political Context

[3]	2018	***Data Source***: Fictitious Facebook Page By Authors ***Ground Truth:*** Manually Developed by Authors	***Number of Participants***: 303 ***Type:*** Students	Multiple Contexts
[83]	2017	***Data Source***: Twitter ***Ground Truth:*** Tweet Credibility Tool	***Number of Participants:*** 754 ***Type:*** Twitter Users	Multiple Contexts
[84]	2019	***Data Source:*** Twitter ***Ground Truth:*** Authors	***Number of Participants***: 2041 ***Type***: Twitter Users	Health Context
[82]	2012	***Data Source***: Twitter pages ***Ground Truth:*** Authors	***Number of Participants***: 266 ***Type:*** Organization's website (not mentioned)	Multiple Contexts
[88]	2019	***Data Source:*** Fictitious unspecified pages by authors ***Ground Truth***: Snopes.Com	***Number of Participants***: 276 ***Type:*** University students, Facebook and Twitter users	Health Context
[85]	2015	***Data Source:*** Facebook Pages ***Ground Truth:*** Not Specified	***Number of Participants***: 146 ***Type***: Not specified	Political Context
[87]	2019	***Data Source*** :Manually created pages , Not specified ***Ground Truth :*** Slovakian Website (False and real information)	***Number of Participants*** : 44 ***Type***: Students	Multiple Context
[89]	2018	***Data Source :*** Twitter ***Ground Truth :*** Buzzfeed and PolitiFact	Not Specified	Multiple Context

Table 2.2: Summary B of Behavior-Based studies on Misinformation Detection.

Article	**Methodology**	**Survey**	**User Experiment**	**Focus**	**Evaluated Credibility Factors OR Cues**
[80]	Quantitative Analysis	✓	✓	Credibility Perception Factors (Content Features and Non Content Features)	Source cues , Topic Familiarity , and Article Quality
[4]	Quantitative Analysis	✓	✓	Credibility Perception Factors (Content and Visual Features)	Metadata Cues, Visual Cues, Textual Cues

[16]	Quantitative Analysis	✓	✓	Credibility Perception Factors (Non Content Features)	Authority Cues, Identity Cues, Bandwagon Cues
[81]	Quantitative Analysis	✓	✓	Credibility Perception Factors (Content Features and Non Content Features)	Sharing Behavior , Screen Size, Distraction level , religious engagement , and political interest
[3]	Quantitative+ Qualitative Analysis	✓	✓	Credibility Perception Factors (Content and Non Content Features)	Source Reputation, Content, Headlines, Personal and Professional Interest
[83]	Quantitative Analysis	✓	✓	Credibility Perception Factors (Content , Non Content Features, and Visual Features)	Geo location, Educational Background, Author Style, Topic, URLs, Hashtags, retweets, Mentions, Media Features
[84]	Quantitative+ Qualitative Analysis	✓	✓	Credibility Perception Factors (Content , Non Content Features and Visual Features	Authenticity Indicators, Content Features , and Visual Features
[88]	Quantitative Analysis	✓	✓	Misinformation Debunking stimuli	Behavioral, Normative, and Control beliefs
[85]	Quantitative Analysis	✓	N/A	Credibility Perception Factors (Content and Non Content Features)	Medium Dependency ,Interactivity ,Medium Transparency ,Argument Strength ,Information Quality
[87]	Quantitative Analysis	✓	✓	Credibility Perception	Reading Behavior, interest, and sharing

				Factors (Content Features)	
[89]	N/A	N/A	N/A	Misinformation Detection Abilities	Network, Content Features , Users Characteristics : Gender and Personality

2.3 A Hybrid Approach of Human–Computational Aspects in Analyzing Misinformation

The majority of previous studies relied on analyzing sources and content of misinformation to detect misinformation. Users are also a substantial part of misinformation diffusion on SMS [5]. However, existing studies lack the effort to evaluate the misinformation that is caused by users' actions, such as sharing and liking misinformation on SMS. Users apply different cognitive heuristics when evaluating the credibility of online information [5]. Consequently, it is important to assess users in making rational decisions when they are assisted by visual analytics tools or by interactive tools for assessing the credibility and their intention to share information on SMS. Here, we synthesize the previous studies that focused on the hybrid approach into three types:

- Computer visual analytics tools.
- Human-assisted computation tools.
- Augmented intelligent tools.

Computer visual analytics assistants have received some attention in recent years as tools to assist users in detecting misinformation on SMS. Such tools offer a deeper analysis of posts on SMS, and aids in user decision-making. Karduni et al. [52] developed a customized visual analytics system for analyzing fake news accounts of multiple contexts on Twitter and studied user behaviors when using the system. They analyzed these accounts from the following different dimensions: account-level view (e.g., includes tweet timeline), social network view (i.e., mentions and retweets), entity view (i.e., most mentioned people and organization entities), map view (i.e., tweet locations), and tweet panel view (i.e., tweet filtration and pattern inspection). They found that confirmation bias had no effect on participants' decisions in identifying fake and real news accounts.

On the other hand, they found that users depend on opinionated, fear language, as well as visuals of the social network of accounts when they evaluate real or fake news content generated by these accounts. Close analysis of user behaviors in identifying easy accounts that distribute real and fake news showed that language features and social network features were the bases for their decisions. For difficult accounts that exhibit cues of contradicted directions, users are more likely to make accurate decisions when they leverage using both quantitative (social network features, linguistic features, and tweet timeline) and qualitative features (opinionated language, style of text). Extending this work, the authors added a new dimension of features, mainly visual representations, by comparing images produced by the fake and real news accounts and then computing the similarity between them. Furthermore, from the language perspective, the system allows word comparisons among real and fake accounts [53].

Shao et al. [47] proposed a platform for tracking and visualizing fake news and rumors of multiple contexts on Twitter. The Hoaxy tool was developed as an effort to track news articles that originated from news websites and were then posted on Twitter. The authors collected fake news stories from 71 domains and associated them with seven fact-checking websites. The platform provides the ability for users to visualize the content of a tweet and display the origin of its websites and related fact-checking site. In addition, the platform offers different credibility ratings to assist users in making credibility judgments. For evaluating the credibility of context-specific fake news and rumor tweets, Gupta et al.[50] proposed a real-time platform for credibility during crisis and trending events called TweetCred, which displays the credibility ranking of tweets from 1 (low credibility) to 7 (high credibility). The platform uses a semi-supervised model trained on data of six crisis events during 2013. They used a set of 45 features under different feature categories, including content-based features, user-based features, network-based features, meta-based features, and linguistics-based features. The platform is deployed as a Chrome extension, and credibility scores are calculated within just six seconds from the tweet's original time.

Finn et al.[48] proposed another web-based tool, called Twitter Trails, that investigates rumors of multiple contexts and their propagation on Twitter. The tool collects relevant tweets, answers questions regarding rumors, and allows users to explore a rumor's bursty activities, temporal propagation characteristics, and retweet patterns. Information about the rumor is displayed on a summary page that discloses information on when the rumor story got exposed, the duration in minutes of the first 100 tweets, the number of retweets and propagation level, and the main actors responsible for spreading the rumor. Besides these features, the system provides information about the level of skepticism given to a particular rumor, which can help in enhancing users' credibility judgments.

RumorLens is another semiautomatic system for investigating rumors in multiple contexts on Twitter. The system is composed of two components; the first one is a rumor detector which identifies suspicious rumors based on keywords that indicate dubious sentiment. Second, a visualized interactive interface shows statistics about the rumors, such as the number of users who have seen the rumored content and whether users were exposed to the correction of a particular rumor or not. In the case of those who spread the rumor, the system also shows whether they spread tweets to mitigate false rumors [49].

Vosoughi et al.[55] presented a human-machine collaborative system that aims to identify relevant information about an event in multiple contexts. The system consists of two major components: an assertion detector and hierarchical clustering. The authors trained a classifier on 7,000 labeled pieces of data of real-world events to detect assertions. The classifier relied on syntactic and semantic features to detect those assertions. Then the outcome of the assertion detector was fed into a hierarchical clustering module which output a collection of clusters that share similar assertions. They evaluated the system on a data set of a world event that contains 20 million tweets. The system discarded 10 million tweets, and the hierarchical model built around 100-1,000 clusters, which helps journalists identify relevant rumors of an event.

On the other hand, Human-assisted computation for misinformation analysis and detection tools is underdeveloped. Such tools combine human intelligence and computational abilities to detect misinformation on SMS. Narwal et al. [54] developed a tool called UnbiasedCrowd that aims to detect fake news on Twitter. Their methodology is based on benefiting from the voting of the crowd to detect visual bias. Automatic bots call users to participate in the inspection of the bias by selecting those who access the hashtags that relate to the news story. They ran pilot studies and interviews to evaluate the effectiveness of the tool in combating such bias generated by news stories about Mexico’s energy reform protests. They collected and evaluated around 50 responses under the chosen hashtags. Users posted their opinions on the visual bias related to the context of the story. Additionally, users tended to inform their friends on Twitter who were interested in this particular story, warning them about the news media's bias in telling a particular story.

Other related tools are augmented intelligent tools. These tools aim to augment human intellect and influence their behavior to make certain decisions [57]. While augmented intelligence tools have been evolving in recent years and used in the healthcare sector for making decisions on clinical diagnosis and treatments, such tools are merely used to detect health-related misinformation online [90]. In online settings, nudging is an example of such a tool. The study of nudging has been widely seen in the area of security and privacy in recent years [91]. The deployment of such tools rose from the nature of the problem as a privacy-related problem that may lead to security breaches and regrettable disclosures [92]. Moravec et al. [93]tested two types of intervention (i.e., system one and system two) on Facebook users' perception of fake political news. They developed a mock of Facebook posts and added the sign "Disputed by 3rd Party-Checkers" and developed another set of mocks and added the sign" Declared Fake by 3rd Party Fact-Checkers." Their study claims that using the two signs (two systems) significantly reduced fake news's believability on Facebook. Similarly, Bhuiyan et al. [46] proposed a nudging system that is centered around obscuring content that consists of news articles that are of questionable credibility by obtaining fact checked articles from fact checking sites. See Table 3 for a comparison of the hybrid systems discussed in this research.

Table 3: Hybrid Systems.

System Name OR Article Title	Year	Platform	Purpose	Tool Type	Misinformation Topic	Misinformation Type	Evaluation Criteria
Verifi [52]	2018	Twitter	Visual Analytics Tool: Fake Accounts Analysis tool	Fully developed tool	Normal Topics, Multiple Contexts	Fake news	User Experiment+ Survey
Verifi 2[53]	2019	Twitter	Visual Analytics Tool: Fake Accounts Analysis tool	Functional tool	Normal Topics, Multiple Contexts	Fake news	User Experiment+ Survey
Automated Assistants to Identify and Prompt Action on Visual News Bias[54]	2017	Twitter	Human-Assisted Computation: Fake News Detection Tool	Functional tool	Trending Topics, Multiple Contexts	Fake news	User Experiment+ Survey
Hoaxy[47]	2016	Twitter	Visual Analytics Tool: Fake Accounts Analysis Tool	Functional tool	Normal Topics, Multiple Contexts	Fake news, rumors	N/A
TwitterTrails[48]	2020	Twitter	Visual Analytics Tool: Rumor analysis Tool	Functional tool	Normal Topics and Trending Topics, Multiple Contexts	Rumors	N/A
TweetCred[50]	2014	Twitter	Visual Analytics Tool: Content Credibility Ranking Tool	Functional tool	Trending Topics, Crisis, Natural Disasters	Fake news, rumors	User Experiment+ Survey
RumorLens[49]	2014	Twitter	Visual Analytics Tool: Rumor Detection Tool	Functional tool	Normal Topics and Trending Topics, Multiple Contexts	Rumors	N/A
A Human-Machine Collaborative System [55]	2015	Twitter	Visual Analytics Tool: Rumor Analysis Tool	Functional tool	Normal Topics and Trending Topics, Multiple Contexts	Rumors	N/A

System 1 and System 2 Interventions for Fake News on Social Media [93]	2020	Facebook	Augmented intelligent	Mockup	Political topics	Fake news	Survey
FeedReflect[46]	2018	Twitter	Augmented intelligent	Functional tool	Political topics	Fake news	Survey

2.4 A Summary of Relevant Studies Limitations

We discuss the major limitation of relevant studies to our context in the following points. First, existing computation based studies aim only to detect misinformation from news headlines [28], [38], [67],[68], [94].These studies do not deal with other aspects of misinformation that occur from the user’s stance expressed toward misinformation[21]. Based on the evidentiality theory, we integrate evidence source categorization to distinguish between two UGCs, including users’ stance and arguments to assess the credibility of information on SMS.

Second, the majority of hybrid systems assume that misinformation follows a similar pattern across different information contexts[53], [47], [46]. However, previous studies have shown that health-related misinformation exhibit different characteristics than other types of misinformation[19][20]. Thus, misinformation analysis should adopt a dynamic approach based on the type of information in context.

Third, existing hybrid systems focus on providing analytical results from the account level and not from the content level of information[47], [48],[49]. In other words, an account on SMS that publish few misleading information does not imply it is a misinformation source. Therefore, we address the existing literature gaps by focusing on the content level of information more than the account level.

Fourth, existing nudging studies in the area of misinformation detection mainly rely on fact-checked-based nudges to help users recognize the credibility of information on SMS[93], [95]. However, such type of nudges is time- and labor-intensive. In other words, the analyzed information relies on experts by third-party fact-checkers. In our study, we rely mainly on evidence source assessment to mitigate the aforementioned negative effect of fact-checking approach.

Fifth, health-related nudging on SMS is scarcely studied, and previously developed nudges majorly target politically related topics. Previous scholars showed credibility assessment of health-related information is different from those of political information[20]. In our study, we develop an approach to nudge users on health-related misinformation on SMS.

CHAPTER 3: AN ANATOMY OF TWEETS AROUND COVID-19 CONSPIRACIES

3.1 Introduction

The internet is a new venue in the health ecosystems [96]. Information consumers have been utilizing medical forums and social media sites (SMS) to seek medical advice and to make treatment decisions. The web evolution has enhanced communication between patients and physicians by benefiting patients, reducing the time for visiting physicians, and getting emotional support and medical advice with low or no costs[97]. Despite the multitude benefits, SMS also bring several disadvantages, such as low information quality and uncertainty.

Using SMS by patients has generated several problematic issues. A major problem is the loose ties on publication and less restricted rules for posting health-related content online include selling health-related products without medical expertise[98]. Previous work showed that illegal birth control pills, dietary supplements, and height booster pills have been circulating on SMS extensively in the past five years. The diverse background of social media users (i.e., the receiver of information) may affect their perception of scientific information due to the differences in their beliefs and social pressure [81]. Moreover, when legitimate information on social media is formulated in the form of biased cultural stereotypes, the perception of information could be altered [99].

With the rise of SARS-CoV-2, famously called the novel coronavirus or COVID-19, conspiracy theorists have gained new traction[12]. This attention has forced medical authorities to declare the state of "infodemic" and develop online sites devoted to fighting misconceptions and conspiracies about the virus. Conspiracy theory is considered a major component of misinformation on SMS [37]. Douglas et al. [100] defined conspiracy theory as "explanatory

[12] https://www.forbes.com/sites/petersuciu/2020/09/11/conspiracy-theories-have-gained-traction-since-911-thanks-to-social-media/?sh=41477df43ddb

beliefs about an individual or group of people working in secret to reaching malicious goals." Sunstein and Vermeule[101] defined conspiracy as an effort to explain some event or practice by reference to the machinations of powerful people who attempt to conceal their role. In a survey study [102], the authors discovered that around 29% of respondents believe that COVID-19 existence has been exaggerated, and 31% believe that virus was purposefully created and spread. Unfortunately, with the massive use of SMS in the world, the challenge of delivering trustworthy information at the proper speed to the public has become more complicated due to the significant similarities in writing between untrustworthy news and trustworthy news [1].

One example of the negative impacts of SMS such as Twitter in fueling COVID-19 is the spread of the hashtag #FilmYourHospital, which encourages people to go to nearby hospitals, proving they are empty, in a way to prove that the pandemic is just a hoax; it was later discovered in a study that right-leaning supporters sustain this hashtag [37]. Such hashtags paved the way for a list of conspiracies to begin spread on Twitter and other SMS. In this study, we analyze conspiracies that relate to COVID-19 and evaluate associated tweets' content and the evidence sources' traits to assist health authorities in their task of fighting misconceptions and COVID-19 related conspiracies on SMS and highlighting such conspiracies for the public.

3.1.1 Social Media Impact and COVID-19 Conspiracies

In recent years, control over information has been weakened by the development of websites that allow for sharing any type of information with minimal restrictions or verification of the information sender[1]. SMS facilitates the spread of UGC to multiple audiences worldwide. As a result, the world is now essentially all under one roof. Users of any background and any location can create any content and spread it to anyone in the world with just a few clicks and without any strict rules applied by SMS before post creation.

Because of SMS's affordances, these sites have become an attractive shelter for different people with various intentions to diffuse multiple types of information. The benefit of posting content on SMS with no cost led to an increasing number of people using SMS daily as their outlet for local and world news and events.

Different SMS have different characteristics. It has been reported that users on Twitter are more opinionated than people on other platforms, such as Facebook and Reddit [103]. Additionally, users on Twitter get news feeds mainly from two evidence sources, people who they are following, and location-based trending topics. This concludes that Twitter is an event- and social-centric platform, making the platform an alluring medium for malicious bots or maliciously intended or unintended users worldwide to publish fake news [6].

From an information-seeking viewpoint, previous work has shown that many individuals use Twitter for asking health-related questions [104], making the platform the top in the list for patients and health professionals discussions [105]. As a result, increasing the believability of any information regardless of its veracity. Another issue with Twitter is the speed of conspiracy propagation and evolution, which leads to increasing information believability and sharing [106].

It has been reported that Twitter is a main factor in the spread of COVID-19 conspiracies[13]. Based on interviews, one study found a positive relationship between believing conspiracy theories about COVID-19 and using Twitter as a source of evidence of COVID-19

[13] https://www.fox10phoenix.com/news/study-suggests-people-who-believe-covid-19-conspiracy-theories-get-the-misinformation-from-social-media

information [107]. In addition, those participants who believed conspiracies were more reluctant to apply handwashing or social distancing measures in a related manner.

3.1.2 Research Objectives and Questions

Wardle and Derakhshan [23] stated that the content modality of information is important to improving our understanding of misinformation spread on SMS and to developing context-specific systems to mitigate misinformation diffusion. Therefore, identifying the types of evidence sources posted by users promoting health-related conspiracies will provide a more transparent image of who is to blame for the rise of conspiracies about COVID-19.

Ahmed et al. [111] found that COVID-19's #FilmYourHospital conspiracy theory was promoted by users supporting their claims with right-leaning evidence sources on Twitter. Motta et al. [112] examined right-leaning and left-leaning media websites and people who consumed these evidence sources. They found that conspiracies were promoted heavily by right-leaning media, and people who consumed them are more likely to express false claims about the conspiracies. Therefore, it is important to explore the usage of partisan news media source-based evidence on SMS like Twitter for COVID-19 conspiracy theories spread on the site. Evaluating news media evidence sources social media users use to support their claim about conspiracy theories will shed light on if partisan media are actually the ones to blame for the spread of COVID-19 health conspiracy theories on Twitter.

On a related note, Broniatowski et al. [6] showed that Russian bots promoted anti-vaccine-related content to encourage vaccine skepticism in the United States on Twitter. Ferrara [109] highlighted the role of bots in the spread of conspiracies and misinformation on SMS. In the context of COVID-19, Uyheng and Carley [113] showed that social bots amplified malicious content by posting content that encourages hate speech in the United States and the Philippines during the pandemic. Therefore, we needed to determine whether it is automated bots or human users who populate conspiracies on SMS. From the perspective of evidence sources diffusion on Twitter, we aim to answer the following research questions:

RQ1: What types of evidence sources are used to support health-related conspiracy tweets?

RQ1-1: What types of partisan news media sources are referenced by health-related conspiracy tweets?

RQ2: Who is more accountable for the spread of health-related conspiracies on Twitter, automated bots or regular users?

Studies have investigated COVID-19 misinformation on Twitter and other SMS from different perspectives. Some have focused on the propagation and engagement of misinformation posts. For instance, a study identifies the share ratio of false, misleading articles between various SMS [108]. Others have explored detecting automated bots and their role in spreading misinformation on Twitter [109]. The third group of studies focused on non-textual features such as the relationship between user demographics, political orientation, and geographical locations using conspiracy terms [110]. Recent works on SMS found that posts that discuss political conspiracies exhibit different sentiment and topic features from those discussing non-conspiracy-related topics [38]. However, the sentiment and topic characteristics of COVID-19 conspiracies on SMS have not been examined. To address the above-mentioned limitations, we aim to answer the following research questions:

RQ3:How does the topic coverage compare between COVID-19 conspiracy and non-conspiracy tweets?

RQ4: How do sentiment expressions compare between COVID-19 conspiracy and non-conspiracy tweets?

The rest of the chapter is structure as follows: section 3.2 discusses related work, section 3.3 presents our methodology for data collection, topic selections, and evidence source credibility measure section 3.4 shows our data analysis and results, section 3.5 is a discussion of the results.

3.2 Related Work

Several works studied misinformation in SMS posts from various angles. However, studies of conspiracies highlight limited aspects. Wood and Michael J[114] conducted a content analysis of Zika related conspiracies tweets by specifically focusing on textual features. Their work lacks the analysis of evidence sources that support the discussion in those tweets. Similarly, Atlani-Duault et al. [115]conducted a content analysis of H1N1 Facebook posts. Their work highlights the blame that users put on public figures. Studies that highlight a COVID-19 conspiracies tend to focus on partisan news media sources. Ahmed et al. [111] focused on “film your hospital” conspiracy to determine if a particular group is associated with the spread of discussions of this conspiracy on Twitter. Similarly, another study conducted by Gruzd and Mai [37] assessed how film your hospital was fueled by prominent conservative politicians. There is a lack of studies that are focused on analyzing different types of evidence sources embedded in the UGCs of multiple types of conspiracies, regardless of whether they are credible or misinformation.

On the other hand, studies that focused on hidden topics and sentiment features of COVID-19 conspiracies on SMS are scarce. The majority of studies highlighted topic and sentiments features of misinformation on other health-related topics. Previous works applied topic and sentiment analysis focused on user attitudes toward health-related products[26], [116].

Chen et al. [117] found an association between e-cigarette usage and the mentioning of health symptoms-related topics on Reddit. Other authors have conducted LDA analysis and then applied binomial regression to model the likes and recognize topic preferences among followers [118]. Che et al. [119] applied LIWC to capture sentiment within fake news during the United States 2016 presidential election on Twitter. Plotkina et al. [120] also used LIWC for fake review detection.

3.3 Method

Our emphasis is on analyzing tweets around COVID-19 conspiracy theories. To answer the proposed research questions, we first collect tweets related to COVID-19 conspiracy theories based on the mentioned hashtags and terms. The list of conspiracy hashtags and keywords is compiled based on our review of scientific/authoritative evidence sources, in addition to previous works in analyzing COVID-19 on SMS [113], [121], [122].

We preprocessed the tweets by automatically converting the URLs of evidence sources into their absolute form [14]. To this end, we used a third-party Twitter URL expander[15] to transform URLs of evidence sources from their relative form to their absolute form.

After extracting evidence sources from tweets, we randomly sampled evidence sources from each conspiracy theory topic for manual annotation. Next, we compared the linguistic features between conspiracy and non-conspiracy topics. We applied text mining techniques such as Latent Dirichlet Allocation (LDA)[123] to extract topics from tweets, which are qualitative data. From the quantitative aspect, we apply Linguistic Inquiry and Word Count (LIWC) to extract sentiment features [124]. In addition, we applied our automated method to determine whether a tweet belongs to a bot or a human user within each conspiracy topic. Lastly, we performed an independent sample t-test on the conspiracy and non-conspiracy-related tweets to answer their related questions.

[14] https://help.twitter.com/en/using-twitter/url-shortener
[15] https://urlex.org/

3.3.1 Data Collection

We collected tweets related to COVID-19 that occurred between March 15th, 2020, and November 15th, 2020, using a scraper that leverages the public Twitter API. We chose Twitter for three reasons:1) previous research showed that the polarity of false claims of health-related topics on Twitter is more than other platforms [110],[125], 2) false news on Twitter were diffused and had more engagements than other platforms, and 3) COVID-19 information is evolving daily on Twitter since the start of the pandemic. We selected March as the start month of data collection because the World Health Organization (WHO) declared the COVID-19 outbreak as a pandemic on March 11th [99]. We limited tweets to English. In addition, we focused on five conspiracies that have received a large number of engagements during the pandemic [109]. We handcrafted a list of hashtags and keywords for retrieving the related tweets based on a list of keywords provided by the World Health Organization (WHO) [16] and the center and Computational Analysis of Social and Organizational Systems Center (CASOS)[17]. We incorporated additional hashtags and keywords from related works on COVID-19 related topics.

We selected conspiracies based on two criteria: 1) the conspiracy should have a sufficient number of tweets, 2) previous studies did not conduct a similar analysis of the conspiracies in relevant studies. After reviewing the related literature[108],[111],[121] and experimenting with the Twitter API with the hashtags/keywords, we selected five conspiracies (see Table 4).

[16] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

[17] https://www.cmu.edu/ideas-social-cybersecurity/

We developed two main heuristics for collecting data. For the conspiracy-related tweets, we constructed two lists: a generic list and a specific(contextual) list. The two lists were constructed based on previous work that found tweets during a crisis include a generic keyword in addition to a specific keyword.[114].The generic list contains COVID-19 related hashtags such as #covid_19, #coronavirus, and #COVID19. Sample hashtags and keywords associated with each of the conspiracy topics are listed in Table 4. For the non-conspiracy-related tweets, we developed the following heuristics: 1) a tweet must contain a keyword or a hashtag from the generic list, and 2) it does not contain any customized hashtag from the conspiracy-related hashtags/keywords.

We cleaned the data by removing noise such as duplicated tweets. We also removed retweets because they do not imply an endorsement of conspiracy theories, as suggested by a previous study [126]. After applying these measures, we obtained 390,655 and 86,501 conspiracy and non-conspiracy tweets, respectively.

Table 4: Selected Conspiracy Theories.

Conspiracy Theory Statement	Related Hashtag OR Keywords	Source
The virus is man-made and created by China	**Generic list and {**#Chinesevirus OR #Chinavirus OR #AsianVirus OR #coronachina OR #china_is_territorist}	[127]
The virus is man-made and created by the US	**Generic list AND** {Fort Dietrich OR Army Bacteria OR military OR government}	[128]
The virus is created for population control.	**Generic list AND** {#populationcontrol OR #plandemic OR #governmentconspiracy }	[129]
Vaccination is purposefully created for malicious cause	**Generic list AND** {#QuantumDotTattoos OR # anti-vaxxers OR # vaccineinjury OR # novax OR #DeepstateVirus}	[130]
Coronavirus non-existence.	**Generic list AND** {#coronahoax OR #covid19hoax OR #emptyhospitalsOR #coronavirustruth}	[111], [131]

3.3.2 Evidence Source Quality Measurement

Previous research showed that most misinformation contents on SMS were generated from low credibility evidence sources [43]. It suggests that the credibility of such content is positively associated with the evidence sources publishing this content. For the aim of answering our research questions, we need to determine the type of evidence sources. To this end, we classified evidence sources into four types: credible source," "mainstream news source," "misinformation source," and "other." The construction of this categorization drew on the types of information sources identified by previous studies of online media literacy [122] and fake news studies on SMS [132], [133], [134],[137]. We selected literature that specifically focuses on source evaluation of UGCs on Twitter. In addition to the four commonly used categories, including credible, misinformation, mainstream news media, and other, we included an additional category called "mixed " because we noticed a number of tweets containing more than one evidence source with different evidence source categories. We describe each category of evidence source next.

- Credible evidence source: include government agencies, education institutions, and research institutions. These evidence sources are compiled from the CDC website of State and Territorial Health departments [18], Data.gov list of health agencies[19] , and Wikidata list of education[20] and research institutions[21] worldwide.
- Misinformation evidence sources: various fact-checking websites (e.g., Snopes, Poynter, Politifact, Media Bias Fact Check, Fact Check, etc.) are operated by journalists verifying news or pieces of content circulating around the web. We combined information from five fact-checking sources, which conducted an extensive

[18] https://www.cdc.gov/publichealthgateway/healthdirectories/healthdepartments.html
[19] https://catalog.data.gov/dataset/hospitals-dcdfc
[20] https://www.wikidata.org/wiki/Q2385804
[21] https://www.wikidata.org/wiki/Q31855

analysis of misinformation evidence sources. These facts checking sources are Media bias/Fact Check [135], NewsGuard COVID-19 misinformation sources [136], Zimdars [122], Shao et al. [133], Wikipedia list of Fake news sites[132]. A tweet mentioning any of the five evidence sources will be labeled “misinformation.” Media/Bias Fact check labeled news evidence sources from “very high factual reporting” to “very low factual reporting” our scraper automatically identifies those tweets labeled as low factual reporting low and very low factual reporting to the list of misinformation evidence source tweets.

- Mainstream news media: mainstream news evidence sources are those news publishers (e.g., CNN, BBC, etc..) which publish mixed factual reporting [135]. Despite they are trusted as mainstream news publishers and famously known around the world. Such news evidence sources sometimes tend to be politically motivated and biased toward one of the political parties’ agendas [6], [112].
- Other: which contains any evidence source that was not mentioned by the credible and misinformation datasets. Examples of such evidence sources are social media posts, medication info sites, blogs, e-commerce sites, health magazines.

One important aspect of a tweet is the fact that a tweet may include one or more evidence sources. Hence, a tweet may contain a credible evidence source and misinformation evidence source. We label those tweets under the “Mixed” category because they contain mixed evidence sources. Such a step is important to recognize which tweets are incorporated with credible evidence sources only, misinformation evidence sources only, or a mixture of evidence sources. Figure 1 provides an illustration of the evidence source classification process.

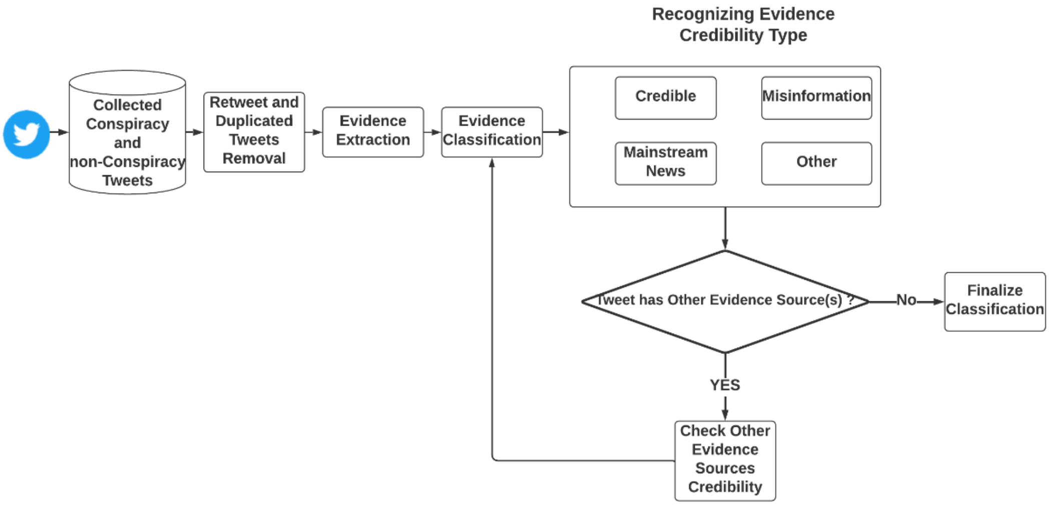

Recognizing Evidence Credibility Type
Collected Conspiracy and non-Conspiracy Tweets
Retweet and Duplicated Tweets Removal
Evidence Extraction
Evidence Classification
Credible
Misinformation
Mainstream News
Other
Tweet has Other Evidence Source(s) ?
No
Finalize Classification
YES
Check Other Evidence Sources Credibility

Figure 1: Tweet-Evidence Classification Process.

3.3.3 Bot Detection

Determining the possibility of an account being a bot is a challenging issue [137]. We explore if bots are responsible for spreading conspiracies by using a Botometer API (previously called BotOrNot) [138] which assigns bot scores to user accounts. Botometer is a machine learning framework trained to calculate the bot score of Twitter accounts. Botometer compares account features with a set of tens of thousands of annotated examples of bot accounts, yielding an accuracy of 95%. Botometer provides a likelihood score between (0 - 100%) with scores above 80% indicating the possibility of a bot account and between (20-80 %) to uncertain account , and below 20% to be likely human account[6].

Following the methods of [6], [133], [139], we randomly sampled each conspiracy topic by taking 10% of the accounts that posted original tweets and ran the sample through Botometer. For the non-conspiracy dataset, a random sample of 10% users was also selected for evaluation through Botometer. Such a process yielded a total of 32,046 unique users in the conspiracy and 4,650 unique users in the non-conspiracy dataset, respectively. Table 5 shows the number of users for each conspiracy topic and that for the non-conspiracy tweets.

Table 5: Conspiracy and non-conspiracy Dataset and Number of Evaluated User Accounts.

Conspiracy Statement	# of Users
The virus is man-made and created by China	10,544
The virus is man-made and created by the US	5,000
The virus is created for population control.	3,001
Vaccination is purposefully created for malicious cause	11,001
Coronavirus non-existence.	2500
Non-Conspiracy	4,650

3.4 Data Analysis and Results

The next section compares the conspiracy and non-conspiracy tweets in terms of the type of evidence sources incorporated into the two datasets. In the following section, we compare the two datasets from the perspective of textual content.

3.4.1 Comparisons of Evidence Source

The total numbers of COVID-19 related conspiracy and non-conspiracy tweets and tweets that contain at least one evidence source are reported in Table 6. The table shows that about 70% of tweets related to COVID-19 conspiracies included at least one evidence source. In addition, about 80% of tweets related to COVID-19 non-conspiracy topics included one or more evidence sources. The comparison shows that users were more likely to provide evidence in their tweets than not, and conspiracy-related tweets were less likely to contain evidence than non-conspiracy related tweets.

Table 6: Descriptive Statistics of Conspiracy vs Non-conspiracy Tweets.

Dataset Type	**# of Tweets**	**#of Tweets with at least 1 Evidence Source**
Conspiracy related	390,655	270,000
Non-Conspiracy related	86,501	67,000

3.4.1.1 Types of Evidence Source

We can observe from figures 2. A and 2. B that 25% of tweets that discuss conspiracy topics of COVID-19 mentioned evidence of misinformation sources. We can notice that credible evidence sources constitute around 5% of tweets discussing conspiracy COVID-19 topics and 22% of tweets that discuss non-conspiracy COVID-19 topics. This shows that people who discuss conspiracy topics lean toward including nonscientific or non-authoritative sources. The "mixed" evidence source constitutes the lowest percentage among all types of evidence sources. There were 40% conspiracy tweets and 6% non-conspiracy tweets that contained evidence sources not recognized by the scraper, respectively. In addition, tweets that mentioned "other" evidence sources dominate the majority of tweets for both datasets, accounting for around 30% and 40 % for conspiracy and non-conspiracy tweets, respectively.

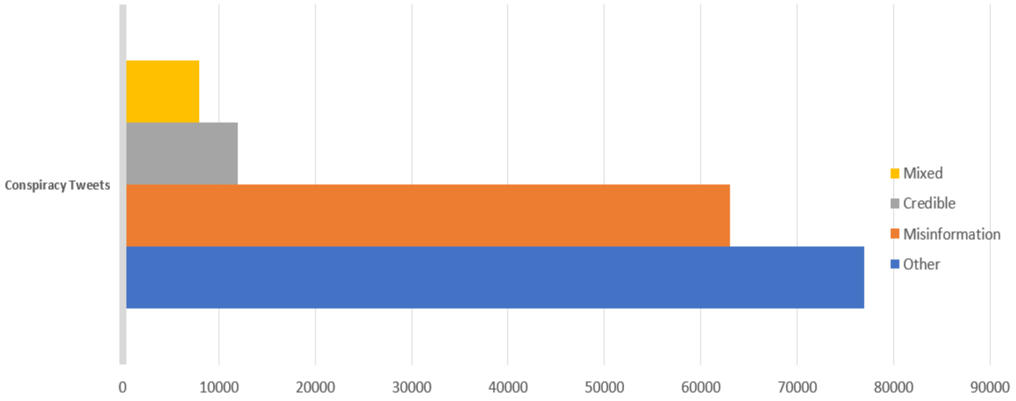


Figure 2.A: Frequency of Evidence Source Types of Conspiracy Tweets.

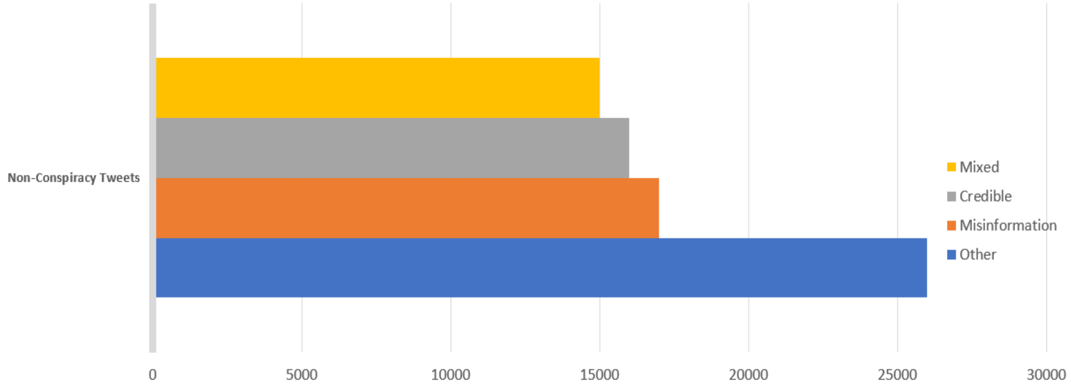


Figure 2.B: Frequency of Evidence Source Types of Non-Conspiracy Tweets.

3.4.1.2 Top News Publisher Evidence Source

Examining only news publishers' evidence sources referenced by tweets discussing conspiracy and non-conspiracy is important for determining the impact of news media on users' supporting evidence sources when discussing conspiracy-related topics [112],[122]. Figure 3 shows that around 20% of tweets that mentioned COVID-19 conspiracies include news publishers. On the other hand, 13% of tweets related to non-conspiracies mentioned news publisher evidence sources. This shows tweets that mention conspiracy-related topics lean toward referencing news media.

The correlation between political conspiracies and news media bias has been investigated in previous work [140] by identifying which news media group (right vs. left) highlights political conspiracies. To determine the affiliation of new media groups that highlight health conspiracies, we referred to Media Bias/Fact Check[22] , which employs experts in politics and journalism to assess local and global news media groups. We can see from Figures 4 and 5 tweets that mentioned news publisher evidence referenced left-leaning evidence sources (20%) more than right-leaning evidence sources (10 %). On the other hand,

[22] https://mediabiasfactcheck.com/

10 % of non-conspiracy tweets mentioned left-leaning evidence sources, and only 3% mentioned right-leaning evidence sources.

Figure 6 shows that left-leaning evidence sources such as the NYTimes ranked top in COVID-19 non-conspiracy topics. The same evidence source (Nytimes) is ranked second by those who discuss conspiracy tweets (40% of tweets referencing left-leaning evidence sources), as shown in Figure 7.

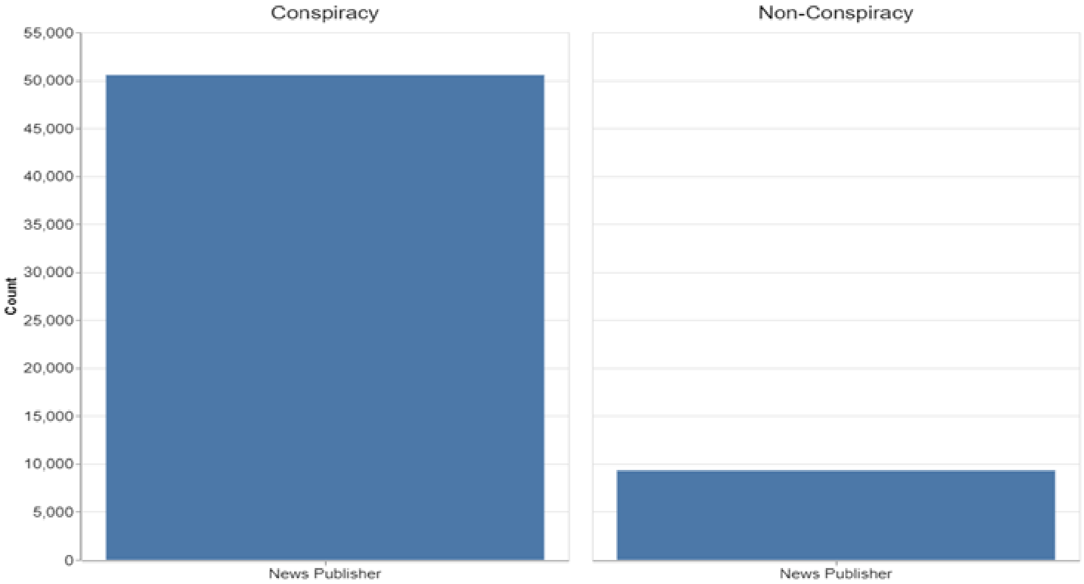


Figure 3: Frequency of Tweets Mentioning News Publisher Evidence Sources.

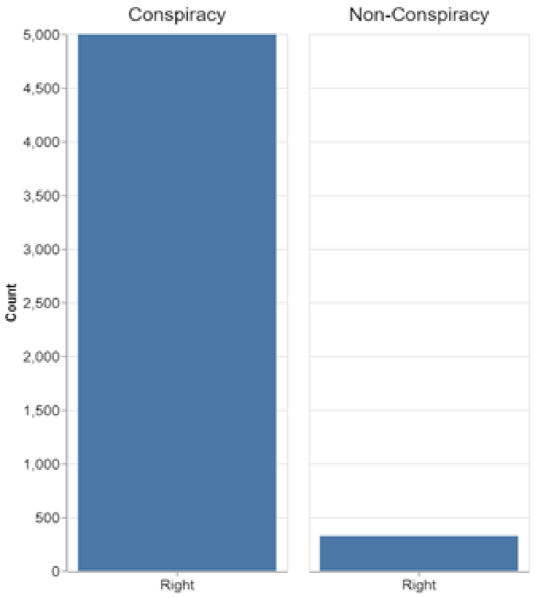


Figure 4: Right-Leaning Evidence Sources Referenced by Conspiracy versus Non-conspiracy Tweets.

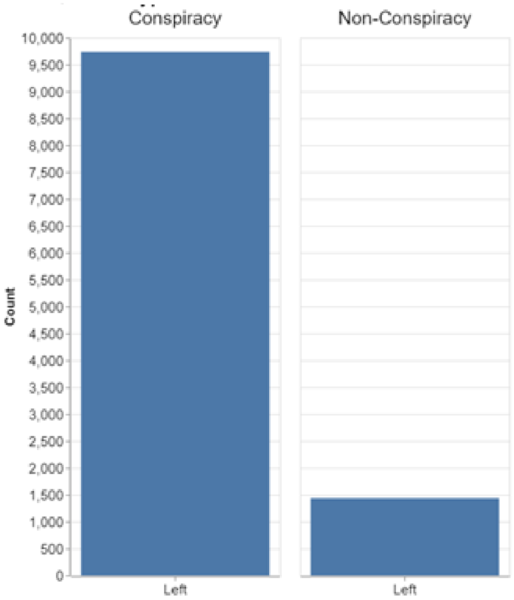


Figure 5: Left-Leaning Evidence Sources Referenced By Conspiracy versus Non-Conspiracy Tweets.

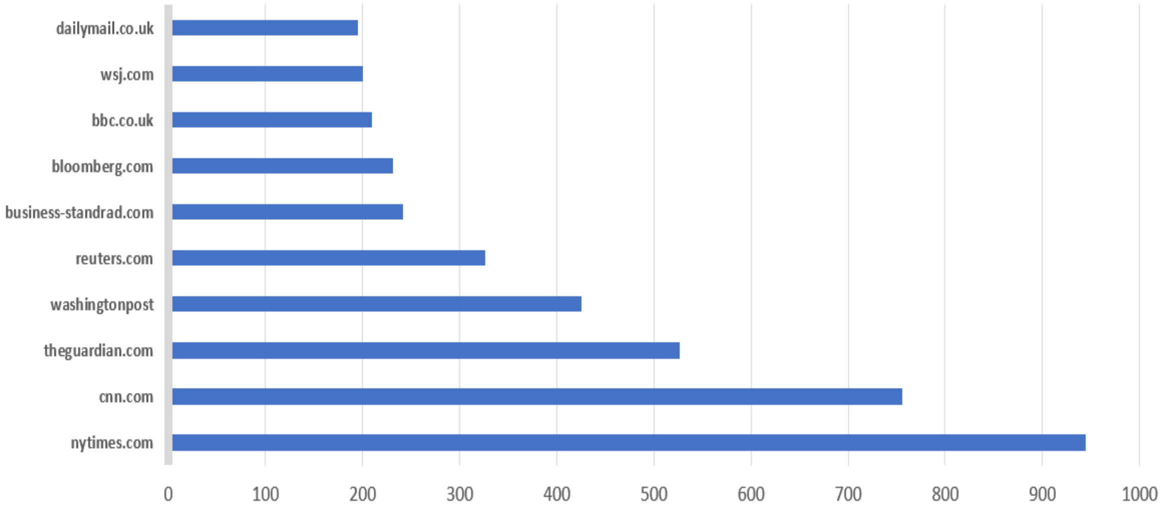


Figure 6: Top News Publisher Evidence Sources Mentioned by Non-Conspiracy related Tweets.

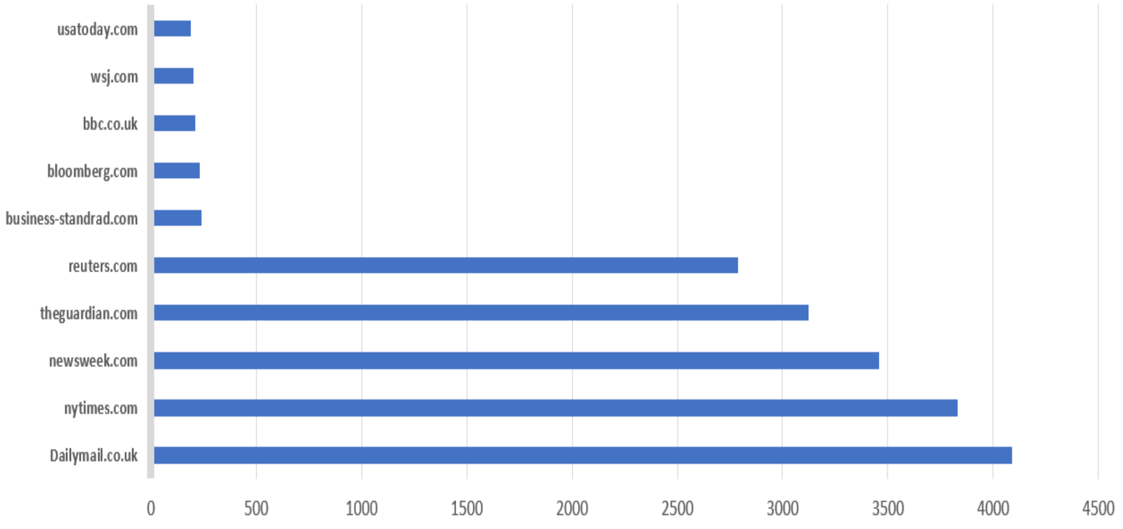


Figure 7: Top News Publisher Evidence Sources Mentioned by Conspiracy related Tweets.

3.4.2 Bot Analysis

We analyzed the distribution of bot accounts that posted conspiracy and non-conspiracy tweets. It is observed from Figure 8 that accounts that are likely to be bot accounts tend to post about conspiracies. Such accounts ranked top within conspiracy tweets (13%) in comparison to accounts that are likely to be human constitutes 1%. Accounts that are uncertain to be bot constitute are around 12%, and the rest of the accounts were not identified by the bot analysis tool. Accounts likely to be uncertain of their bot ranking tended to discuss non-conspiracy tweets (73%) and ranked first in this category. In contrast, bot accounts constituted only 14% and ranked second in this category.

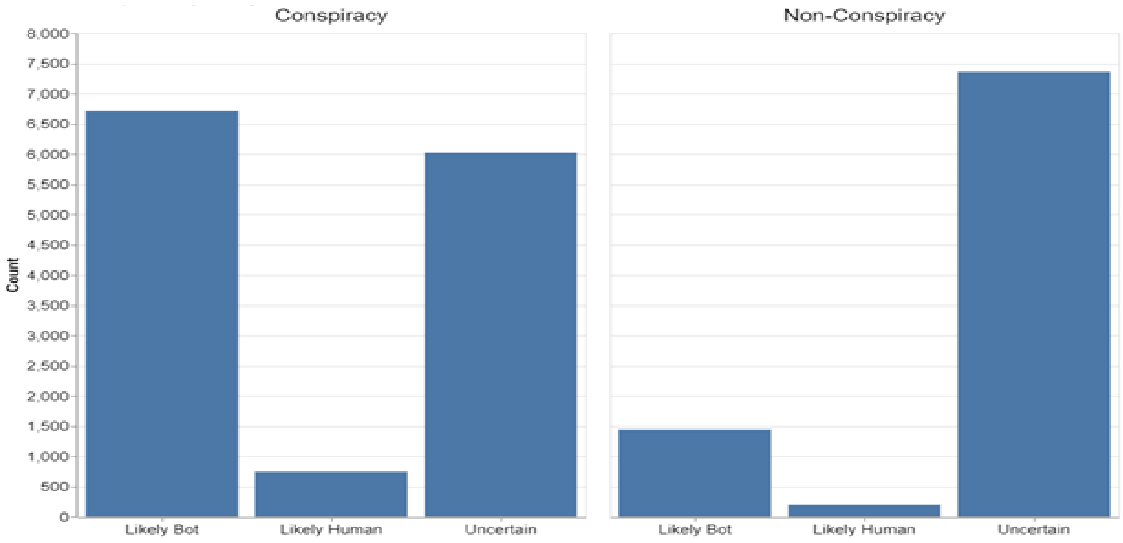


Figure 8: Frequency of Bot Accounts mentioning Conspiracy and Non-conspiracy Tweets.

Table 7 shows that the frequency of bot accounts that discussed COVID-19 conspiracy and mentioned news publishers is around 38% of all bot accounts that incorporate any type of evidence source (1740 accounts). Interestingly, news publisher is the most favored type of evidence sources for bots regardless whether it discusses conspiracy or non-conspiracy topics.

Table 7: Number of Bot Accounts That Mentioned News Publisher.

Dataset Type	**# of Tweets That Include New Publisher Evidence Source**
Conspiracy related	670
Non-Conspiracy related	150

We also compared the types of news publisher evidence sources spread by likely bot and human accounts. We can observe from Figure 9 that likely bot accounts tend to mention right-leaning news evidence sources such as Fox News and Daily Mail more when discussing COVID-19 conspiracy topics. Likely, bot accounts favor right bias new media when talking about conspiracy-related tweets and left bias media to come as the third and fourth used evidence sources among these accounts. Likely human accounts (Figure 10)tend to publish left bias evidence sources such as the New York Times, News Week, CNN, and BBC when discussing conspiracy-related topics. This type of news media was less used by likely bot accounts in comparison to right bias media.

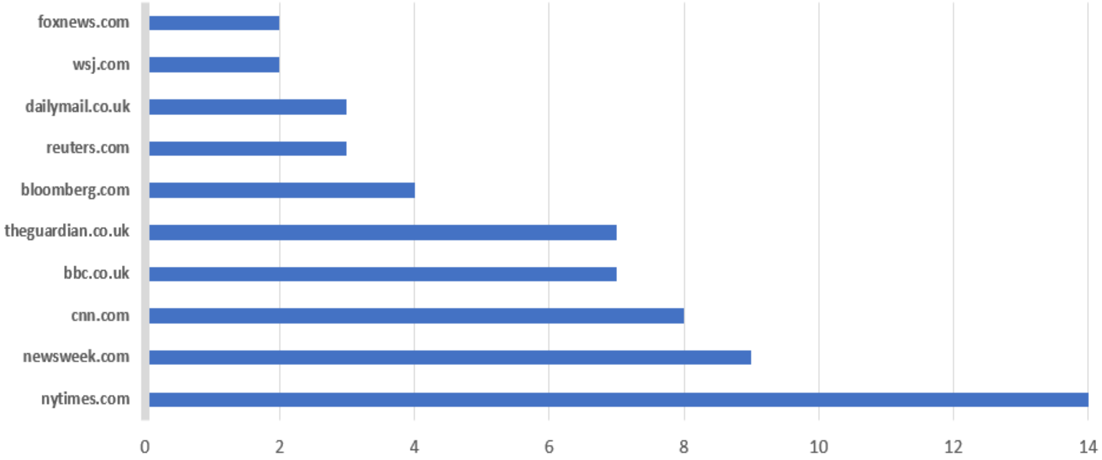


Figure 9: Top News Publisher Mentioned by Likely Human Accounts That Mentioned Non-Conspiracy Tweets.

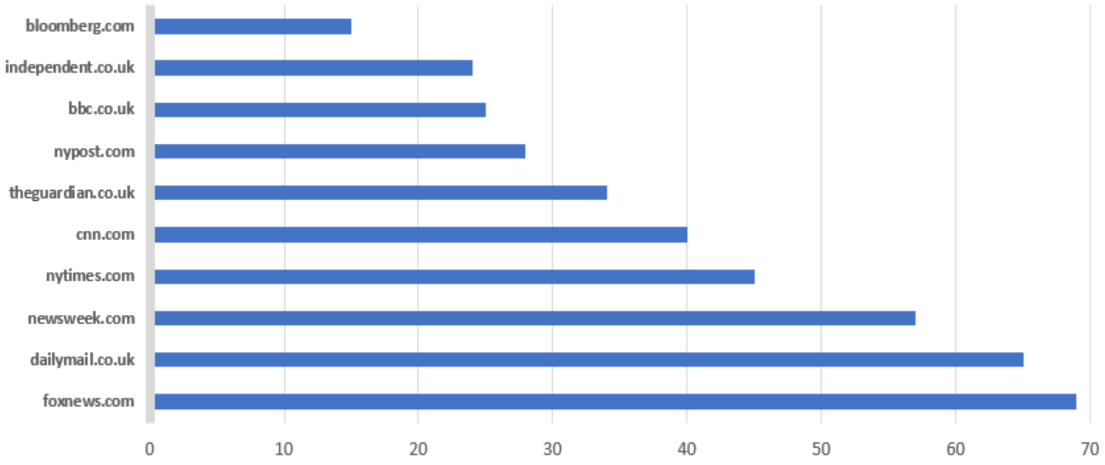


Figure 10: Top News Publisher Evidence Source Mentioned by Likely Bot Accounts That Mentioned Conspiracy Tweets.

3.4.3 Topic Modeling

Text analytics such as topic modeling is effective in detecting political misinformation on SMS [125]. Studies in related domains, including fake reviews [141], have demonstrated the efficacy of topic modeling in detecting misinformation posts. Thus, we employed LDA to extract hidden topics from both conspiracy and non-conspiracy datasets. Several preprocessing steps were employed, such as removing stop words. In addition, we removed terms that exist significantly across the two datasets, and these terms include COVID-19 related terms such as *covid19, covid-19, corona, virus*. In addition, URLs and emails were also removed since these features are not informative for this type of textual analysis[123]. We used the pyLDAvis [142] package to visualize conspiracy and non-conspiracy topic relationships and to tune the hyperparameters to reach an optimal number of cluster topics and a set number of words for each topic. Following the method used in previous work [143], we selected the number of topics by increasing the number of topics until there was a noticeable increase in the overlap between the words of the different topics. We selected the number of topics to be 3 for conspiracy-related tweets and 2 for non-conspiracy-related tweets, and the number of words per topic to be 10.

During our analysis of the conspiracy tweets, we initially increased the number of topics to 7; when we started to observe a significant number of terms existing in the two datasets(i.e., conspiracy and non-conspiracy). Words such as a *pandemic, people*, *US*, *China*, *scam*, *plan*, deaths, government, *health* were dominating conspiracy tweets. As we began decreasing the number of cluster topics, the terms became less overlapped between the clusters. Finally, we decided to keep the number of cluster topics to 3 when we reached the point of distinguishing the terms between them.

The results of topic modeling for conspiracy and non-conspiracy tweets related to COIVD-19 are reported in Table 8. We make the following observations from the table. First, conspiracy tweets consist of three major categories of topics: Wuhan, Telecommunication Technology, and Bill Gates. We label the first topic Wuhan because it contains several terms such as *source, animal,* and *lab,* which relate to the conspiracy of the origin of the virus. The second conspiracy topic relates to the 5G technology with terms such as *network* and *radiation;* thus, we label this topic as telecommunication technology. Lastly, we labeled the third topic as Public Figures because of the dominance in mentioning of public figures of this cluster topic. Words such as *Bill Gates, Fauci, Melinda*, and *foundation* exist under this topic. This shows the impact of false news reports that claim that Bill Gates and infectious disease expert Dr.Anthony Fauci could benefit from COVID-19 vaccine development[23]. We can observe from the analysis results that multiple terms are shared across all the three conspiracy-related topics, including *government and China*. Other terms such as are *weapon, and deaths* were mentioned in the Wuhan and the Telecommunication Technology topics.

[23] https://www.financialexpress.com/industry/bill-gates-making-200-billion-from-vaccines-microsoft-co-founder-explains-math-behind-returns/2092891/

Regarding non-conspiracy-related tweets, we labeled the first topic as cases because it contains terms such as the number of *cases*, *deaths, and positive*. The second topic contains terms such as *stay* and *lockdown;* thus, we labeled it as quarantine. For similarities in mentioned terms between non-conspiracy topics, we can notice the terms *health*, *protect,* and *stay.*

Table 8: Topics Generated by LDA for Conspiracy and Non-conspiracy related topics.

Dataset Type	**Topic**	**Top 10 Words**
Conspiracy	Wuhan	lab, cases, deaths, wuhan, animal, source, biological, Chinese, government, weapon
	Telecommunication Technology	Deaths, US, conspiracy, china pandemic, network, radiation , 5G , weapon, technology, government
	Public Figures	China, Bill. US, government, vaccine, gates, Fauci, foundation, money, Melinda
Non-Conspiracy	Cases	US , cases, government, protect, lockdown, positive, home, total, number, health
	Quarantine	Health, pandemic, lockdown, stay, home, support, mask government, safe, help

3.4.4 LIWC Analysis

Linguistic Inquiry and Word Count (LIWC) has been used to explore the sentiment differences in politically and non politically related conspiracies[124]. LIWC2015 is a dictionary-based tool that counts occurrence and the percentage of features representing different emotions and psychological states. We adopted LIWC 2015 to extract the sentiment expressed toward COVID-19 related conspiracy and non-conspiracy tweets. In view that positive and negative emotions have been used to analyze other health-related conspiracies on Twitter [144]. We focused on the overall positive and negative emotions and specific negative emotions such as anger, anxiety, sadness in our data analyses.

We compared the occurrences of positive and negative emotions between the conspiracy and non-conspiracy tweets(Figure 11). The figure shows that conspiracy tweets contain a higher level of negative emotions (4%) than non-conspiracy ones (2%). Specifically, there is a higher level of anger expressed in conspiracy tweets (1.7 %) than in non-conspiracy tweets(0.82%). Nonetheless, no notable difference in anxiety and sadness was detected between conspiracy and non-conspiracy tweets. Compared with the mean of a generic tweet on Twitter (2 %), [145], COVID-19 related express a higher level of negative emotion. In addition, people tend to express more positive emotions when discussing non-conspiracy tweets than when discussing conspiracy tweets related to COVID-19 (3 % vs. 2 %). Compared with the mean of a generic tweet (5%) [145], discussing COVID-19 yielded less expression of positively related sentiments.

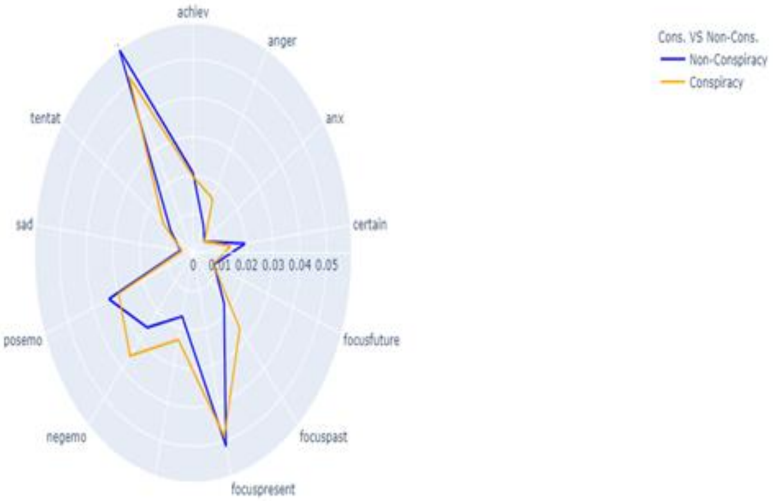


Figure 11: LIWC Analysis for Conspiracy Versus Non-Conspiracy Tweets.

3.5 Discussion

The current study examines health-related tweets during a health emergency. We analyzed the topic and sentiment features surrounding COVID-19 conspiracies and non-conspiracy tweets. We found that topics and sentiment variables differ between these datasets. In terms of topic differences, tweets discussing COVID-19 conspiracies tend to focus on entities such as individuals or cities.

To answer RQ1, we created a lexicon for collecting evidence sources from previous studies and fact-checking websites, and we classified tweets into different credibility categories. Overall, we found that around 70 % of the collected tweets mentioned evidence sources. People who discuss non-conspiracy topics tend to include evidence sources in their tweets more than people who discuss conspiracy-related topics (80%.vs 70%). Moreover, from the 390,655 collected conspiracy tweets, the proportion of tweets that include evidence sources is relatively high (270,000 tweets). Such a result is consistent with a previous study [130] that showed that people who discuss conspiracies of misleading information share evidence sources. We also found that 5% of tweets that discuss conspiracy topics mentioned only credible evidence sources that include authoritative evidence; on the other side, misinformation evidence sources transcended authoritative evidence sources in conspiracy tweets. For RQ1-1, we analyzed the proportion of new publisher evidence sources across the conspiracy and non-conspiracy tweets. We observed that conspiracy-related tweets mainly include news publisher sources to support their view of point, with around 20% of tweets. On the other hand, we found that only 13% of non -conspiracy tweets referenced new publishers. We explored these news publishers and analyzed which news publisher sources are left or right leanings sources by cross-checking these sources with fact-checking sites. Left-leaning partisan media evidence sources tend to dominate most of the conspiracy-related tweets (20%).

For RQ2, tweets published by automated bots have been the center of the focus of research on SMS. Such research highlighted the role of malicious bots in distributing fake news content benefiting a political party over the other [6],[133],[109]. However, previous research focused more on the political aspect than the health aspect of SMS information. In our health-related study, we found that 13% of conspiracy tweets were publisher account that is likely to be bot accounts. Likely human accounts constitute only 1% of the accounts. These bot accounts that discuss conspiracy topics (10 %) mentioned news media evidence sources.

For RQ3, we found that terms such as *biological*, *scam, Bill Gates, Wuhan, 5G,* and *Fauci* are discussed frequently in conspiracy-related tweets. In contrast, non-conspiracy tweets updates on the virus itself, such as updates on cases and quarantine-related topics. Terms such as *stay, protect, home,* and *lockdown* exist in non-conspiracy-related tweets.

In terms of sentimental features(RQ4), conspiracy tweets show more negative emotions, specifically sadness and less positive emotions than non-conspiracy tweets. The findings of this study call for the development of tools that consider topic and sentiment differences in detecting conspiracy-related tweets. In addition, this research calls the authorities to understand the reasons people express negative emotions when discussing conspiracy theories on Twitter.

CHAPTER 4: AN EVIDENCE-BASED APPROACH TO ASSESSING THE CREDIBILITY OF ARGUMENTS IN HEALTH-RELATED INFORMATION

4.1 Introduction

People turn to SMS nowadays as a source of information[146]. Such platforms have recently been known for discussing controversial topics and news on different topics such as political, economic, health, and social-related topics [1], [139]. Scientists, journalists, politicians, and many professionals have been reported to use such platforms to understand user viewpoints and perspectives on different topics. For instance, officials have been using SMS to understand users' arguments on vaccine usage [6], [30], which resulted in increasing efforts to raise user's awareness of vaccines and build computation models to monitor the public arguments on the vaccines [30].

An argument is defined as "giving a reason(an evidence) to support a stance that is questionable or open to doubt" [147]. In contrast, opinions are just people's stance without giving an explanation of how the stance was formulated [148]. Previous studies identified several arguments that populated health misinformation on SMS [36],[32]. These arguments negatively affect the quality of information; consequently, acquiring trustworthy information on health topics becomes challenging, especially during an emerging natural crisis [149]. Previous research in detecting fake news and rumors on SMS suggests that unverified information is exchanged heavily in the form of fake news and rumors [99] and acknowledges the scarcity of computation-based models for detecting misinformation in early propagation stages [43].

Although recent works applied sentiment analysis and opinion mining to understand users' feelings, expressions, and viewpoints on SMS platforms, they presented a limited solution to understanding users' stance and arguments toward misinformation[21], [150]. In addition, sentiments analysis and opinion mining does not identify argumentative structures or elements within a text. Argumentation mining was introduced [151] to verify evidence in other contexts, such as verifying the authority of articles in Wikipedia [33] and information in legal documents[32]. However, it has not been used in the context of misinformation on SMS.

Earlier works in stance detection on SMS have focused more on the political topics than health-related ones. Since health-related information is critical to an individual's life, it is necessary to develop computational models to detect users' stance in their arguments toward health-related misinformation on SMS. Guided by the evidentiality theory[22], this study aims to detect arguments about COVID-19 health-related misinformation and specifically vaccination.

4.1.1 Background

4.1.1.1 COVID-19 Misinformation on Social Media Sites

Numerous sources have raised concerns about COVID-19 related misinformation on SMS. Authoritative evidence sources such as the Center for Disease Control and Prevention (CDC)[24] issued a list of COVID-19 related misinformation. Their website claims that COVID-19 misinformation circulates heavily on SMS, such as Twitter, and specifically about causes[25] , cures[26], and preventive strategies[27] . The U.S. Food and Drug Administration (FDA) and the Federal Trade Commission (FTC) warned about companies selling fraudulent COVID-19 treatments that were claimed to cure or prevent COVID-19 [152]. Kouzy et al. [116] found an abundance of tweets with unreliable evidence in COVID-19 related datasets. Based on an analysis of 225 pieces of COVID-19 misinformation, Brennen et al. [108] found that misinformation in UGCs cover multiple topics including vaccination. By analyzing COVID-

[24] https://blogs.cdc.gov/publichealthmatters/2020/06/beware-scams/
[25] https://www.ncbi.nlm.nih.gov/research/coronavirus/docsum?filters=topics.Transmission
[26] https://www.ncbi.nlm.nih.gov/research/coronavirus/docsum?filters=topics.Treatment
[27] https://www.ncbi.nlm.nih.gov/research/coronavirus/docsum?filters=topics.Prevention

19 misinformation on SMS and Twitter, Singh et al. [121] found that misinformation majorly exists in COVID-19 preventive strategies and vaccination.

4.1.1.2 Fundamentals for Argumentation Mining

In the area of argumentation mining, an argument consists of three components: a)The ***premise*** (evidence) represents the reason for one's argument, b)a ***conclusion*** (which supports a stance), and c) the ***relation***(which shows how the premise led to the conclusion) [33], [147].

Ketcheam [153] defined *argumentation* as the "art of persuading others to think or act in a definite way." MacEwan defined argumentation as "the process of proving or disproving a proposition"[154]. Freeley and Steinberg [149] described argumentation as the reasoning or a justification of one's acts, beliefs, attitudes, and values. Despite the variety of definitions given by scholars, they all agree on the act of argumentation to persuade others.

Despite exiting argument annotated corpora, they focus on specific genres such as legal documents, newspapers and court cases, product reviews, and online debates [155]. There is a lack of argument annotated corpora of SMS content [156], specifically related to health-related topics [21]. In addition, the structure of social media text differs from other text types [32], [35], [36], [157], making the argument annotation of social media content that relate to health topics a new and challenging task.

Another challenge of analyzing SMS text is that such sites do not enforce users' following guidelines before posting content [157]. As a result, most content is written in an informal language or is unstructured, making its interpretation a complex task. Therefore, it is important to facilitate the task of consuming information of health-related topics. Assisting information consumers with understanding the argument position and the types of evidence that social media users include to support their arguments is crucial for them to construct strong and justifiable evidence for their arguments[33]. This in turn enhances individuals' credibility assessment of information on SMS [149], [157].

Existing argumentation mining systems are context-based[32], [157]. Such systems focus on one of the following issues: claim-premise mining, claim mining, boundary detection, dialogical argument mining, structure prediction for arguments, rhetorical category sentence classification, and evidence mining[36]. However, they have rarely investigated mining evidence from SMS. Differentiating arguments from non-arguments [33] would serve users in assessing the credibility of controversial health-related information on SMS by allowing users to focus on argumentative UGC rather than other types of UGC(i.e., stance, satire)

4.1.1.3 Argumentation Mining on Twitter

SMS like Twitter poses a challenge to authorities in addressing the polarity of stance and arguments about controversial topics that are health-related such as vaccination [30]. Previous studies have found an increasing amount of stance against vaccination on SMS [6], [30],[158], [159]. Such an upward trend has continued over the years, despite the platforms' algorithmic changes in an effort to fight misinformation after the 2016 presidential elections [160]. In addition, users were found to post content (e.g., arguments) that was not well justified [161]. One way to handle such an issue is to ignore such posts and focus on relevant posts that talk explicitly about the problem at hand [162]. However, a post is sometimes accompanied by weak evidence that weakens the overall arguments [33]. Therefore, identifying the type of evidence is crucial for determining the quality of the arguments on SMS, which in turn contributes to assessing the credibility of information [163].

4.1.2 Research Motivation and Questions

In this research, we followed a three elements schema to analyze the structure of an argument:

- **Topic**: is the short phrase that scopes the discussion[33].
- **Stance:** a statement that takes a favoring or opposing position[33].
- C**ontextual evidence**: a segment that supports the content in the claim[33].

Such schema has shown efficacy in categorizing evidence of Wikipedia pages in previous literature [33], [164]. The previous studies in stance detection on SMS did not differentiate argument-based text from the stance-based text[164]. This is in direct contrast to the linguistics literature [30], [33],[72], which highlights their differences as the following. An argument should include evidence explaining how the stance's position within an argument (into favoring, denying, or being neutral) has been reached [165]. Stance are just people's expressions, and they are not explanations of how those conclusions are formulated or made [148], [166]. [166]. Thus, we propose the following questions:

RQ1: How can we leverage argument-based tweets to build argument detection models for health-related misinformation?

RQ2: How can we categorize evidence incorporated into the arguments of health-related misinformation?

In section 4.2, we will discuss related work and the limitations of previous research. In section 4.3.2, we will discuss our data collection methodology; afterward, in section 4.3.3, we demonstrate the steps of data annotation, including evidence type. In section 4.4, we show our results from the data annotation approach we applied to analyze the selected arguments about vaccination. Finally, section 4.5 discusses our results and findings.

4.2 Related Work

Argumentation mining consists of two phases: argument structure annotation and argument content analysis[154], [155]. The majority of studies focus on the first phase [36], [157]. Rowe and Reed [34] developed a diagramming tool called Araucaria that supports the manual annotation of arguments. The tool supports convergent, linked arguments, missing premise(enthymemes), and refutations. They later released the AracuariaDB corpus, which was used heavily by researchers in argumentation mining. Focusing on Wikipedia talk pages, Schneider et al. [167] built their annotation guidelines based on Walton's schemes [168], and they reached an agreement (Cohen's k=0.48). However, a limitation of their work is the difficulty and intensiveness of applying their approach on a large dataset [167]. Habernal et al. [169] followed Toulmin's model to annotate a set of 990 documents of blogs, forums, and comments. In the first round of annotation, 524 instances were labeled as argumentative, the second round of annotation yielded 345 documents labeled as arguments with fine-grained annotations. Experimental results have not been reported for this corpus.

In regards to the second phase, Stab and Gurevych[155] prepared a dataset that classified arguments into different categories (none, claim, major claim, premise) by extending their previously annotated corpus and dataset, and their argument classification results achieved an F1 score of 0.72. However, they assume that arguments in documents should comply with a specific structure and documents should always contain arguments. Biran and Rambow [170] highlighted the need for a subjective claim on the blog threads. Nevertheless, they did not present detailed descriptions of the annotation guidelines [170]. Park and Cardi [171] developed a verification categorization framework for the premise within an argument. They classified propositions into unverifiable, verifiable non-experimental, or verifiable experimental. But their work exploits only legal documents and did not explore SMS content.

Studies that focus on analyzing evidence in an argument remain scarce. Rinott et al. [33] analyzed evidence in plain text such as Wikipedia pages based on a limited categorization. They are different from this research that aims to mine evidence from another type of argument like arguments within social media UGCs. A relevant work conducted on social media posts focuses on the domain of privacy [72]. Their works expand on the work done by Rinott et al. [33] by including more categories. However, they assumed stance and arguments to be the same text type. Such categorization is not aligned with the definition of argument from the linguistic literature[154].

In our work, we focus on differentiating arguments from non-argument tweets by analyzing the type of evidence sources around vaccination. In addition, we focus on health-related content because health information seekers on Twitter rely on evidence sources in tweets to verify the credibility of their UGCs [16], [38],[172].

4.3 Methods

This study aims to identify argumentative tweets toward COVID-19 vaccination. To answer the research questions, we present our overall method design in Figure 12. It consists of four main phases.

- Step 1: we filter the tweets related to COVID-19 vaccination and then identify the structure type, and if it is a stance, we classify evidence within the tweets.
- Step 2: we conduct a comparison between tweets of different stances and evidence types.
- Step 3: we clean and preprocess tweets.
- Step 4: using machine-learning techniques, we build classification models for arguments and evidence, respectively. The argument model classifies arguments into support-vaccination, against-vaccination, or non-argument categories, and the evidence

model classifies evidence into different types. To support the development of classification models, we identify important input features.

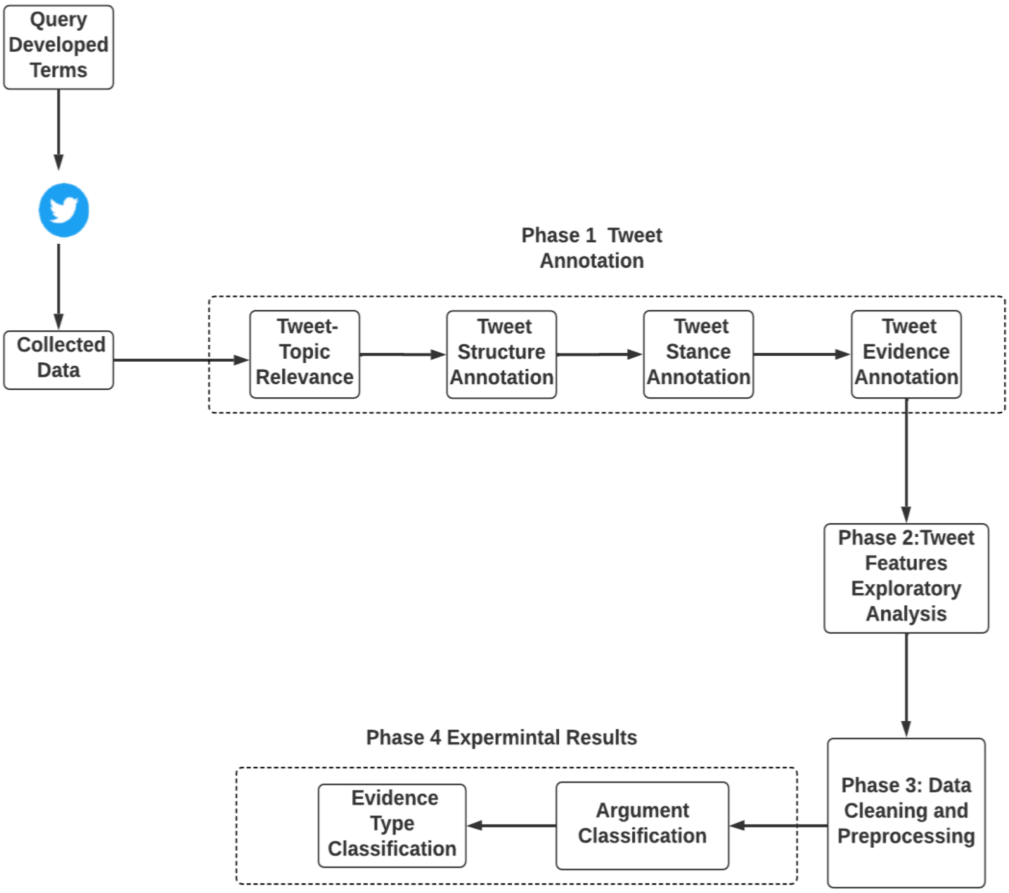

Query Developed Terms
Collected Data
Phase 1 Tweet Annotation
Tweet-Topic Relevance
Tweet Structure Annotation
Tweet Stance Annotation
Tweet Evidence Annotation
Phase 2:Tweet Features Exploratory Analysis
Phase 3: Data Cleaning and Preprocessing
Phase 4 Expermintal Results
Argument Classification
Evidence Type Classification

Figure 12: The Research Method.

To collect tweets relevant to the COVID-19 vaccination, we compiled their descriptions from the previous literature in COVID-19 [108], [121]. We examined both argument and non-argument-based tweets for the vaccination topic. The development of the coding scheme is introduced in Section 4.3.2.

4.3.1 Data Collection

This study develops a novel context-specific argument dataset that is focused on health-related UGCs on Twitter. We chose Twitter for two reasons: 1) previous research has shown that the polarity of arguments on health-related topics in Twitter is more than other platforms [26],[30]; and 2) users tend to include evidence sources to support their arguments on Twitter more than other SMS[173].

We developed a customized Twitter scraper using python 3.7 via the public Twitter application programming interface (API) and retrieved tweets related to the COVID-19 vaccine posted between March 15, 2020, and November 15, 2020. We chose March as the beginning month because the World Health Organization (WHO) declared on March 11th that the COVID-19 outbreak as a pandemic[99]. Additionally, misinformation started to propagate during March[112]. We chose November as the end month because we wanted to retrieve the maximum number of related posts without negatively affecting the performance of our scraper. We used a combination of python script and selenium. For the input, we used Symplur (Symplur LLC, Los Angeles, CA) [28] , a healthcare analytical social media tool for the list of top hashtags and terms about COVID-19 being used on Twitter. Our search is limited to extracting tweets in English.

[28] https://www.symplur.com/topic/coronavirus/

We created a set of seed search terms for COVID-19, such as #covid_19, #coronavirus, #COVID19, which led to a collection of 578,077 tweets. We then filtered those tweets using search terms related to vaccination, such as vaccination and vaccine, which yielded 108,169 tweets.

Next, we created a sample list of keywords (terms) and hashtags for a preliminary grouping of tweets into support- or against- vaccination categories (see Table 9 and Appendix A). First, we selected a list of hashtags related to vaccination from previous work[108],[121],[159]. Then, we used Symplur[29] which is a healthcare hashtag finder that specifically analyzes posts related to healthcare on SMS to determine the stance of the hashtags. Such as step further reduces the sample size to 11,801 tweets that contain stance position toward vaccination.

Table 9: Search Keywords for support and against vaccination.

Support	Against
VaccinesSaveLives VaccinesWork WorldImmunizationWeek VaxWithMe	LearnTheRisk VaccineInjury VaccineDeath VaccineDamage

4.3.1.1 Feature Extraction

A major challenge of conducting research on Twitter is the ability to extract useful features and metadata. Based on our review of the features used by related studies in detecting fake news[28], [38], [50],[65], we group those features into two levels: user-level and message-level features. In this study, we propose a novel type of features - evidence level features. See table 10 for the detailed list of features and their descriptions. Going beyond the existing tools [43],[89], we developed a customized scraper to extract the three types of features.

[29] https://www.symplur.com/topic/vaxxed/

- **User-level features:** characteristics of the author of the tweets (13 features)
- **Message-level features:** metadata features of a tweet content (11 features)
- **Evidence-level features:** Based on the information from the Media Bias/Fact Check[30] website, we extracted the following five types of features from each evidence source: 1) the type of evidence source; 2) bias ranking of the evidence source ; 3) factuality of information reported by the evidence source; 4) evidence source expertise ranking; 5) evidence source entity type.

Table 10: Extracted Features.

Feature Level	Feature	Description
USER	USER ID	Twitter Handle/ USER ID
	USER NAME	User name of the user who posts the tweet
	GENDER	Gender of Twitter user
	IS VERFIED	Verification status of Twitter user
	REGESTRATION	User’s registration date
	NUMBER OF FOLLOWINGS	Number of people followed by the user of the account
	NUMBER OF FOLLOWERS	Number of users following the use of the account
	BIO	Biography or profile of the user
	EVIDENCE IN BIO	If there is an evidence source in the user’s biography
	EVIDENCE SOURCE IN BIO	EVIDENCE source in the BIO
	NUMBER of EVIDENCE SOURCES IN BIO	Number of EVIDENCE sources in the user profile
	COUNTRY	The country where the user is tweeting
	GEO LOCATION	Longitude and latitude of the user
MESSAGE	THREAD ID	Thread ID or link of the tweet
	Link to post	URLs of the post
	TIME	Time of the tweet
	CONTENT	Tweet full content
	NUMBER OF COMMENTS	Number of comments to a tweet
	NUMBER OF LIKES	Number of likes to a tweet
	NUMBER OF EVIDENCE SOURCES	Number of evidence sources in a tweet
	EVIDENCE IN MESSAGE	The absolute form of evidence source in a tweet
	NUMBER OF RETWEETS	Number of retweets
	INFLUENCE SCORE	Influence level of the tweet
	IS RETWEET	Is the tweet original OR a retweet
EVIDENCE	BIO-EVIDENCE SOURCE TYPE	The type of evidence source found in bio
	BIO-EVIDENCE SOURCE BIAS RANKING	The bias ranking of evidence source found in bio
	BIO-EVIDENCE SOURCE FACTUALITY RANKING	Factual reporting ranking of evidence sources found in bio
	TWEET EVIDENCE TYPE	The type of evidence source found in a tweet

[30] https://mediabiasfactcheck.com/

	TWEET EVIDENCE BIAS	The bias ranking of evidence found in a tweet
	TWEET EVIDENCE EXPERTISE	Expertise of the evidence source in relation to the health related topics.
	TWEET EVIDENCE ENTITY	Evidence entity if it belongs to individual or organization

4.3.2 Data Annotation

Building a corpus of argument and non-argument-based tweets requires a selection of tweets related to the context of our research. A high-level description of the annotation strategy is shown in figure 13. Based on our applied automated filtration strategy for identifying support and against vaccination tweets, we select only tweets that are either supportive or against vaccination that are incorporated with evidence sources. Next, we verify manually if the selected tweets are related to vaccination. After that, those tweets related to vaccination are further manually checked if a stance toward vaccination is present in the selected tweets. Finally, the selected tweets are moved to be labeled in terms of the specific stance position and the evidence type. The outcome of this step is fed into the training classifiers to detect argument and non-argument tweets and classify tweets based on their evidence types.

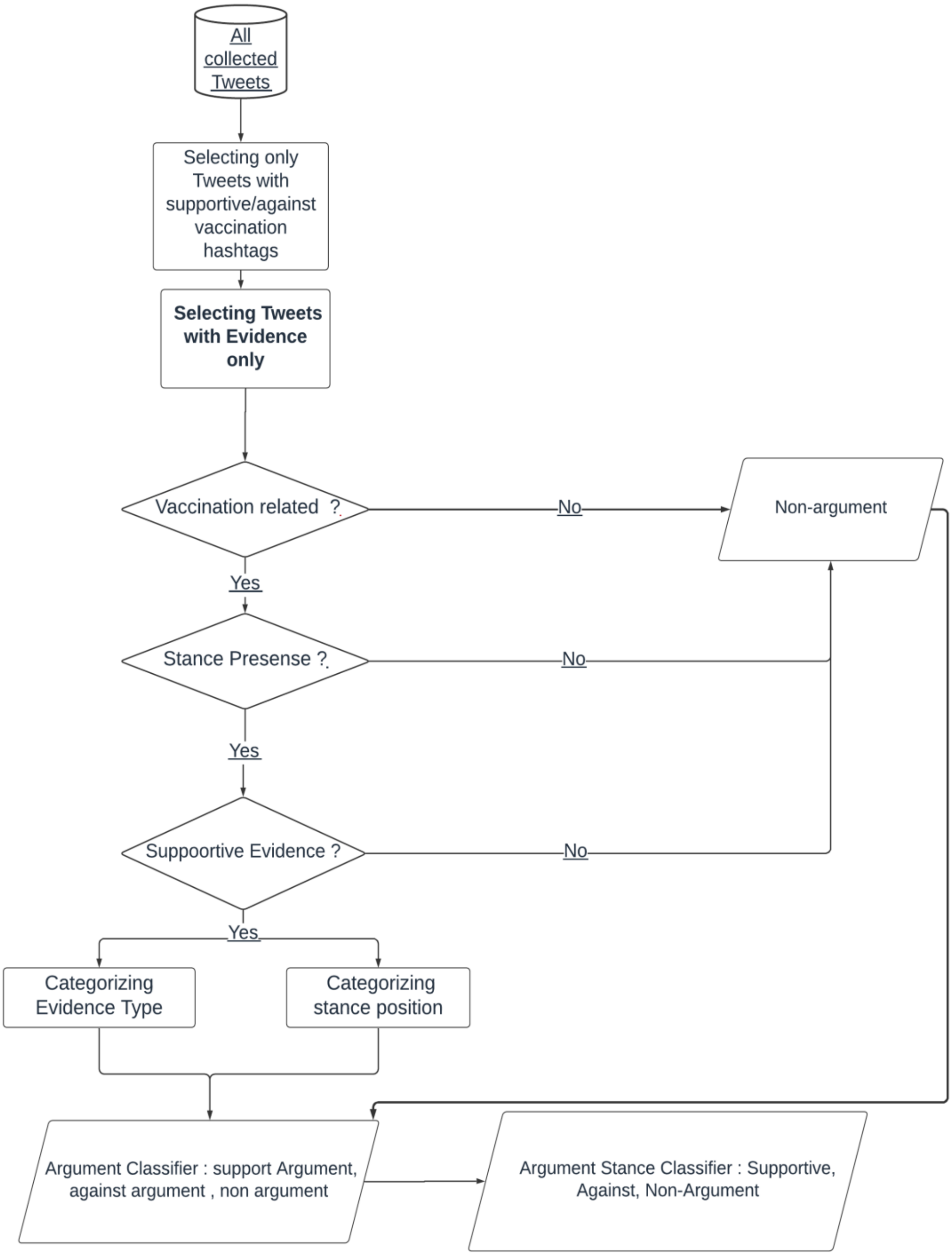

All collected Tweets
Selecting only Tweets with supportive/against vaccination hashtags
Selecting Tweets with Evidence only
Vaccination related ?
No
Non-argument
Yes
Stance Presense ?
No
Yes
Suppoortive Evidence ?
No
Yes
Categorizing Evidence Type
Categorizing stance position
Argument Classifier : support Argument, against argument , non argument
Argument Stance Classifier : Supportive, Against, Non-Argument

Figure 13: Data Annotation Workflow.

4.3.2.1 Topic Relevancy

To verify whether the selected tweets are truly relevant to the vaccination topic, we performed a content analysis on the filtered data. We selected structural coding and thematic analysis method [174], which went through multiple iterations (see appendix B). The final coding scheme for topic relevance consists of the following categories:

- **Vaccination:** A) a tweet contains at least one keyword or hashtag that is related to COVID-19 and B) a tweet mentions any vaccination news, availability, and anything about vaccination of COVID-19.
- **Not vaccination-related**: A tweet that is not related to vaccination such as mentioning any issues related to COVID-19 (e.g., social issues, economic issues, political issues, OR other health issues of COVID-19) or other topics not related to COVID-19.

4.3.2.2 Tweet-Structure and Stance Annotation

Detecting stance toward health misinformation in UGCs requires recognizing the structure type of tweet to recognize their eligibility for stance evaluation.

We employed a two-step annotation procedure to identify the structure within the argument and non-argument-based tweets. The first task is determining the structure of the tweets. We divided the tweet structure into two types: argument-based and non-argument-based: (see Table 11 for examples)

- **Argument-based:** mentioning a stance about COVID-19 vaccination while incorporating evidence such as external link(s) or tweet mentions.

- **Non-Argument-based:** mentioning a stance about vaccination but with unrelated evidence, expressing a stance about COVID-19 but not about vaccination, or containing non-sense jokes or questions.

Table 11: Sample tweets of different structure types.

Structure Type	Example
Argument	*“How can we explain to the Enlightened Ones who like things "natural" that #VaccinesSaveLives? Microsoft billionaire Bill Gates said Thursday that his foundation was funding the construction of factories for seven #coronavirus #vaccine candidates <u>https://www.wsj.com/articles/bill-gates-to-spend-billions-on-coronavirus-vaccine-development-11586124716 via @WSJ”*
Non-Argument	*“@AstraZeneca potential two billion corvid-19 vaccine doses A vaccine must be seen as a global public good - a people's vaccin You get first in line #UN globalists ???????????? #Globalists #BillGatesIsEvil #vaxxed #coronavirus #NewNormal @UnitedNations https://www.bbc.co.uk/news/business-52917118 https://pic.twitter.com/894NAz77mH*

The second strategy is determining the stance expressed toward the chosen topics. An overview of the instructions given to annotators is shown in table 11 (see Appendix B for detailed codebook). Based on our review of existing work [41], [161], [162] and a sample of tweets, we developed three types of stance positions toward the vaccination topic. The following stance positions are adapted to the scope of tweets and the scope of this research.

- One position is *support* when a user shows a favoring position toward vaccination. Concurrently, we cover the case when a user favors trustworthy- fact-checked information about vaccination.
- Another position is *against* when the user shows a denying position toward vaccination. As we did with the first position, we check if the stance covers the case where the user shows a denying position toward fact-checked vaccination information.

- A third is *neutral* position, where the user shows mixed positions toward the vaccination topic.

Table 12: Types and Descriptions of Stance.

Stance Position	**Definition**
Support	A tweet that its context shows positive sentiment, which is interrupted as a **favoring stance position** toward vaccination. **OR** A tweet that its context shows positive sentiment, which is interrupted as a **favoring stance position** of **trustworthy** and **fact-checked vaccination information**.
Against	A tweet that its context is showing negative sentiment, which is interrupted as **denying stance position** toward vaccination. **OR** A tweet that its context is showing negative sentiment, which is interrupted as a **denying stance position** of **trustworthy** and **fact-checked vaccination information**.
Neutral	A tweet that **exhibits** a **favoring stance position toward vaccination** **OR** A tweet that its context shows positive sentiment, which is interrupted as a **favoring stance position** of **trustworthy** and **fact-checked vaccination information**. **AND- AT THE SAME TIME** A tweet that its context is showing negative sentiment, which is interrupted as **denying stance position** toward vaccination. **OR** A tweet that its context is showing negative sentiment, which is interrupted as a **denying stance position** of **trustworthy** and **fact-checked vaccination information**.

4.3.2.3 Recognizing Evidence Types Supporting User Arguments

We applied a hybrid method for recognizing the type of evidence in the tweets(see figure 14). During our data collection process, we configured our scraper to label tweets that contain evidence that was not identified with the “not recognized” label. Such tweets are moved to the manual annotation process for manual evidence identification. The annotators labeled the evidence that was not recognized by following two sub-methods: The first method is choosing the type of the evidence. Based on previous works that analyzed the type of evidence sources shared on SMS [175],[176],[177], we identify the following eleven main types of evidence.

- **News publishers:** different types of news publishers, including mainstream, fake, hyperpartisan (biased), and satirical news sources.
- **Government agencies:** government or federal based sites, and US nonprofit foundations.
- **Education and research institutions: p**ublic and private educational and research institutions
- **Charity:** any type of entity that provides money or food to people.
- **Medical associations :** any professional organization developed to spread health high standards.
- **Magazines:** different kind of magazines, including health and non-health related magazines.
- **World blogs:** a regularly updated website or web page, typically one run by an individual or small group or an organization, that is written in an informal or conversational style.
- **Online health sites:** website related to drug information, health discussion communities.

- **Scientific publishers:** websites that publish scientific and papers peer-reviewed journals
- **Commercial sites:** E-commerce sites that offer services and products and second hand used items
- **Social media sites:** Social media posts (textual, imagery) from Twitter users or other social media platforms (e.g., Facebook, Instagram, Reddit, YouTube, etc.)

Then, the annotators were asked to annotate the bias ranking if it applies. Then they see they check if the evidence entity is a person or organization. Lastly, they check if the evidence source is an expert in health-related issues by checking the evidence source website page and Wikipedia page for more information about the evidence. Table 13 summarizes the number of evidence sources collected for each evidence type.

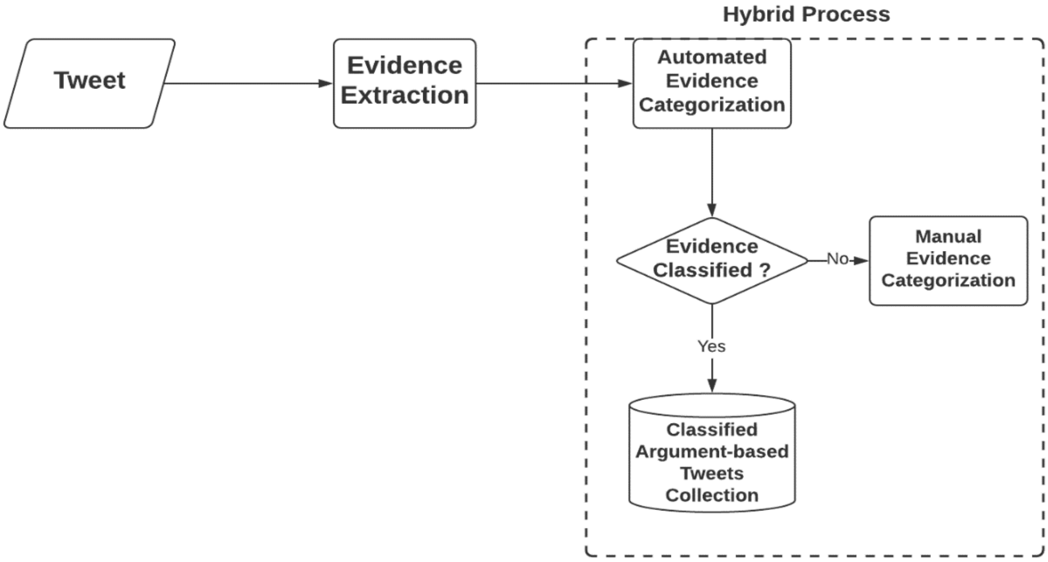

Hybrid Process
Tweet
Evidence Extraction
Automated Evidence Categorization
Evidence Classified ?
No
Manual Evidence Categorization
Yes
Classified Argument-based Tweets Collection

Figure 14: Hybrid Methods for Evidence Classification.

Table 13: Categories, subcategories, and counts of Evidence Type.

Evidence Type	**Sub-categories**	**Evidence Source Frequency in the Lexicon**	**Example**
News	Mainstream news	1352	cnn.com, bbc.com,etc.
	Hyperpartisan news	35	Newcenturytimes.com , usuncut.com.etc.
	Fake news sites	1355	ActivistPost.com, The Onion.com, etc.
US Government Federal agencies And Medical Associations	Government and state-level	91	cdc.gov/,usa.gov, cdph.ca.gov/ health.maryland.gov.
	Medical and Non- profit associations	149	www.texmed.org/, americanheadachesociety.org/
Education and Research Institutions	Public and private educational and research institutions	6089	Harvard, Mayoclinc, Peking University,etc.
Online Health Sites	Health discussion communities	226	Healthunlocked, patientslikeme,etc.
	Drugs information	149	MedlinePlus, WebMd, etc.
	Health magazines	27	self.com, alive.com,etc.
World Blogs	NONE	268	Wordpress.com, wix.com
US and World Scientific, academic publishers and Academic health specialized papers	General scientific publishers	15240	IEEE, ACM, thelancet,etc..
Commercial sites	E-commerce sites	36	Amazon.com, ebay.com, Craigslist.com
Social media sites	Worldwide social media sites	388	Twitter.com, Facebook.com

4.3.2.4 Challenges

We encountered several challenges during the data annotation process. First, our annotators disagreed on several tweets where the stance position tends to be more neutral toward vaccination than showing a stance position toward vaccination. As a result, a third annotator was involved in resolving the stance label. Second, users sometimes included more than one evidence source when showing a stance toward vaccination. We considered one type of evidence source during the annotation process. Third, some tweets retweeted other tweets that show a stance towards vaccination; however, we did not include such tweets as having a stance. Fourth, some users showed a stance toward vaccination, and they included an evidence that is irrelevant to the stance position. We decided to exclude such tweets because an argument should contain a stance with evidence that support the stance. Such a process was not direct because we had to access the evidence source page and see if the stance of the evidence source supported the stance of the tweet.

4.3.3 Data Pre-processing

To prepare training data for building classifiers, the data went through the following preprocessing steps. First, due to the character limit imposed by the website, users tend to use slang and abbreviations in their tweets. We reversed such words to their original word phrases (e.g., 4 u => for you) and lowercased capitalized words. After identifying the evidence type mentioned in the argument-based tweets, we removed the URLs, mentions, and replies from the content of the tweets.

4.4 Results

4.4.1 Annotation Results

Based on the annotation results and discussion, revisions were made to the original coding scheme. We analyzed the inter-annotator agreement on tweets that are argumentative or non-argumentative toward vaccination using Cohen's Kappa [178]. The value of Cohen's Kappa yielded around 86% agreement ($\kappa=0.858$, $p<0.001$) for the argument and non-argument tweets. To determine the stance of the arguments, we analyzed the agreement between the two annotators. This step yielded a Cohen's Kappa of around 96% (κ=0.957, p<0.001) for the argument classification task.

During the annotation process, we discovered that argument-based tweets with neutral stances constitute a very small portion of our dataset(less than 2%). As a result, we excluded such tweets from the total annotated tweets. Thus, the total count of argument and non-argument-based tweets in our dataset is 816, with 52% being argument-based tweets (see table 14). The table shows the frequency of argument-based tweets that are supportive or against vaccination.

Table 14: Distribution of Argumentative and non-argumentative Tweets in our dataset.

Argumentation Class	**Counts**
Supportive-Argument	127
Against-Argument	303
Non-Argument	386

4.4.2 Experimental Results

Table 15 shows the results of the binary classification between argumentative and non-argumentative tweets. First, we treated the problem as a supervised machine learning approach and conducted a binary classification task to distinguish tweets that have a stance and evidence in discussing COVID-19 vaccination(i.e., argument). Furthermore, we applied a multi-classification task to distinguish tweets that show a supporting, against stance with supporting evidence or tweets that do not show a stance toward COVID-19 vaccination. Although our focus in this research is argument-based tweets, and due to the scarcity of related works in detecting arguments in SMS, we conducted a review on related works that focus on stance detection in the context of SMS to choose proper classifiers. Based on our review, we chose three classifiers that have been frequently used by such literature: Support Vector Machine (SVM) [179], [180]; Decision Trees(DT) [181], [182], and Random Forests (RF) [183].

We trained and tested the models using four features sets(message level features, message and account level features, message and evidence level features, and all features). As a traditional supervised machine learning problem, 70% of the annotated data was used for training and 30% for testing(for models' setup, see appendix c). For evaluation purposes, we used the following five evaluation metrics[184].

- Accuracy: This is calculated by taking the amount of correctly classified examples (TP, TN) and dividing their sum by the number of all observations (TP, TN, FP, FN).
- Precision: which is the percentage of relevant instances obtained from the total number of instances and is calculated by taking the (TP) and dividing it by the total of (TP/FP).
- Recall: which is the percentage of relevant instances obtained from the total number of relevant instances and is calculated by taking the (TP) and dividing it by the (TP/FN).
- F1-score: which is the harmonic means(aka weighted means) of precision and recall and is calculated by the following formula : (=2*(Precision*Recall) /(Precision + Recall)).
- Macro averaged F1-score is computed by taking the arithmetic mean (aka unweighted mean) of all the per-class F1 scores.

We can see from table 15 that the best F1 score performance was around (90.28 %), which was achieved by both DT and RF using all features. The second-best performance was achieved by SVM (89.19%) using a combination of message and evidence level features, which was only one percent lower than the best performance.

Table 16 reports the performances of classifying stance positions used within argumentative-based tweets. We can notice that SVM scored the best performance (84.25%) when using evidence-based features in combination with message-level features. In contrast, using the three classifiers on the remaining three feature sets yielded noticeably lower performance than combining the message and evidence level features.

In addition, we used macro averaged F1 score to validate our results. Using macro-averaging, each class contributes equally to performance, avoiding the result dominating performance on a single class when the classes are imbalanced[185]. Again, we can note from table 17 that using all features yielded the best performance for both the DT and RF classifiers(90%). For the performance of argument stance classifiers (Table 18), the result shows that using message and evidence features has yielded the best performance using the SVM classifier(80%).

To evaluate our results with the results of relevant studies, we compared our top classifiers with three baselines. Since no study developed argument detection models on SMS (i.e., our first classification problem), we decided to compare our results in the second classification(i.e., argument stance position classification) with stance detection literature. Table 19 shows the comparison between our proposed method and the baselines. Research suggests that the unigram baseline is difficult to beat for some types of debate[186]. Thus, we selected the best performer that used message level feature as a baseline. We can see from the table that message and evidence feature combined (80.44%) outperformed all baseline methods. Specifically, the proposed model outperformed using message-level features alone by 23.44% in detecting vaccination stance within arguments. Also, we compared our method with a popular stance detection system[41].We can see that our method outperformed using only word and character n-grams, as well as those obtained using external resources, such as word-embedding features from additional unlabeled data, by around 10%. Additionally, our method outperformed other work[187] that added sentiment features to their feature set by 12.72%.

Table 15: Performance of Argument Classification.

Feature Set	SVM				DT				RF			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
Message Level Features :Unigram+Bigram	81.94	89.96	79.43	**84.24**	80.56	90.32	77.35	83.33	79.83	91.61	75.94	83.04
Message +Account-level Features	59.38	95.48	57.36	71.67	69.1	68.39	72.6	70.43	68.75	80.65	67.57	**73.53**
Message+Evidence-level Features	88.54	87.74	90.67	**89.18**	86.11	82.58	90.78	86.49	86.11	84.52	89.12	86.75
All Features	82.29	86.35	82.29	82.82	90.28	90.33	90.28	**90.27**	90.28	90.28	90.28	**90.28**

Table 16: Performance of Argument Stance Position Classification.

Feature Set	SVM				DT				RF			
	Acc.	Pre.	Rec.	F1.	Acc.	Pre.	Rec.	F1.	Acc.	Pre.	Rec.	F1.
Message Level Features :Unigram+Bigram	68.59	78.22	68.59	**71.59**	59.62	61.94	59.62	60.64	62.18	74.28	62.18	66.93
Message +User-level Features	67.31	91.16	67.31	**74.82**	71.79	76.22	71.79	73.37	64.74	82.08	64.74	71.00
Message+Evidence-level Features	82.05	89.74	82.05	**84.25**	80.13	81.4	80.13	80.41	80.05	86.57	80.77	82.37
All Features	69.87	93.06	69.87	76.5	76.28	77.22	76.28	**76.66**	69.87	82.03	69.87	74.00

Table 17: Macro Average F1-score for Argument Position Classification.

Features Set	**SVM**	**DT**	**RF**
Message Level Features :Unigram+Bigram	82.0	80.0	73.0
Message +User-level Features	50.0	69.0	69.0
Message+Evidence-level Features	89.0	86.0	83.0
All Features	81.0	**90.0**	**90.0**

Table 18: Macro Average F1-score for Argument Stance Position Classification.

Features Set	**SVM**	**DT**	**RF**
Message Level Features :Unigram+Bigram	57.06	36.0	62.0
Message +User-level Features	50.22	69.0	50.0
Message+Evidence-level Features	**80.44**	76.0	79.05
All Features	62.81	73.0	62.0

Table 19: Performance Comparison between Proposed Approach and Baselines.

Feature Set	**Macro Average F1-Score**
Proposed Model	**80.44**
Message Level Features :Unigram+Bigram	57.06
Word n-grams and character n-grams features [41]	70.03
Lexical and sentiment features [187]	67.72

4.4.3 Feature Analysis

In order to identify the most important features for argument classification, and argument-stance classification, we used RF built-in feature importance function in the scikit-learn package. Figure 15 lists the top-N most important features. The figure shows that the three credibility attributes are ranked the top. Similarly, we can see from figure 16 that classifying the stance position toward COVID-19 vaccination among argument-based tweets weighed heavily on the evidence-based features, particularly evidence bias.

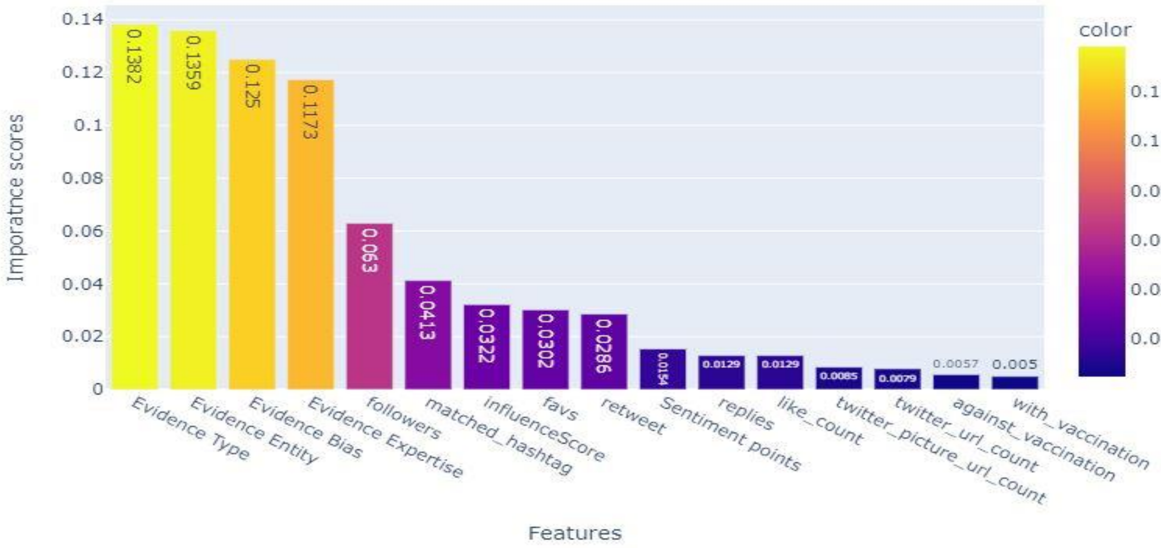


Figure 15: Feature Importance for Argument Classification.

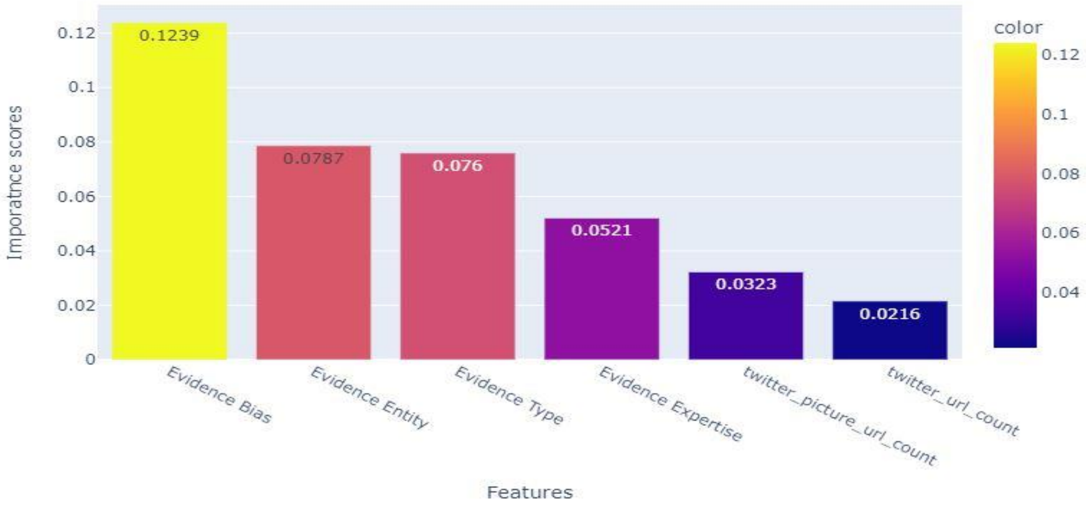


Figure 16: Feature Importance of Argument Stance Classification.

4.5 Discussion

This study develops a theoretically driven classification scheme based on the evidentiality theory to classify argument-based tweets (i.e., stance with evidence) on SMS. In addition, we developed a semi-automated approach for classifying the evidence within COVID-19 related UGCs. First, our approach automatically classifies the evidence from a lexicon of evidence sources we collected from multiple online sources. For those evidence sources that were not automatically classified, they were manually analyzed using a codebook we developed over an iterative process. The results demonstrate the importance of integrating credibility metrics of evidence characteristics (i.e., expertise, and bias) [5] when assessing argument-based tweets on SMS.

Our results show that a significant number of posts on SMS are non-argumentative in that they do not provide evidence toward COVID-19 vaccination. This is consistent with the findings of a previous study that claims users on Twitter tend to post health-related topics without providing evidence to support their claims[172]. We developed prediction models to distinguish arguments(tweets that contained supporting evidence) from non-argument tweets. The DT and RF classifiers achieved the best performance (around 91% in F1-score). Given the importance of differentiating types of stance expressed within arguments, we improved our coding scheme to detect arguments on SMS by integrating the types of stance expressed within these arguments. We developed detection models for various stances within arguments (supporting or against COVID-19 vaccination) that achieved a high F1 score (84%) by incorporating evidence features with the message features. The results of analyzing the feature importance in the classification models confirm our hypothesis about the importance of evidence-based features in detecting arguments in tweets. The results showed that evidence bias and expertise (from the credibility heuristics), evidence entity and type (from the authority heuristics) were the most influential features to detect argument types within tweets.

Our study makes a novel contribution to the literature of fake news detection and stance detection on SMS. First, we differentiate between stance and argument in the context of health-related topics and specifically on COVID-19 vaccination on SMS. Such differentiation is essential in identifying those arguments with evidence and those without evidence for information receivers on SMS. Second, we integrated evidentiality theory to the area of detecting stance on SMS. Such effort aims to not just detect the stance of a tweet but also evaluate the strength of the argument in the tweet by evaluating the incorporated evidence when discussing controversial topics, such as COVID-19 vaccination. Third, we built detection models that detect arguments of different stance positions. Fourth, we enhanced the detection of arguments by including evidence-based features derived from credibility metrics(bias and expertise) mentioned in the variety of information credibility literature. To the best of our knowledge, this is the first work that developed a codebook scheme to distinguish arguments from a stance and detect arguments on SMS.

CHAPTER 5: AN EVIDENCE-BASED DIGITAL NUDGE FOR HUMAN ASSESSMENT OF HEALTH INFORMATION CREDIBILITY ON SOCIAL MEDIA SITES

5.1 Introduction

With the emergence of SMS, the spread of misinformation has become inevitable. SMS are becoming a hub of content and a reference for user exchange of information. Current research highlights the challenges in detecting misinformation on SMS [44], which can be partly attributed to the evolving technologies and the sophisticated strategies that spreaders use in spreading misinformation. One of such strategies is mixing misinformation with trustworthy information to elevate the 'truthfulness' of their stories [140]. For instance, it has shown that in the months preceding the 2016 U.S election, foreign organizations with political agendas conducted organized efforts to spread mixed misinformation on Twitter and Facebook [188].

In recent months, misinformation about SARS-CoV-2, the virus causing the novel coronavirus (COVID-19) pandemic, has dominated SMS and other online media. This pandemic becomes a hotbed of multiple types of misinformation related to the virus [130].

Cognitive heuristics are information processing strategies or rules of thumb that may ignore information to make decisions with a few efforts than complex methods, helping reduce the cognitive load during information processing [5]. Cognitive heuristics play a major role in assessing the credibility of health-related information online [15]. Information overload, in turn, is one of the primary motivations for applying cognitive heuristics when assessing the credibility of information on SMS [5],[39], [189]. Literature from multiple disciplines, including psychology, medicine, and politics, suggests the impact of cognitive heuristics on the decision-making process [190], [191]. One example of cognitive heuristics is self-confirmation [192]. It has resulted from selective exposure, where people tend to engage with information that is consistent with their beliefs while ignoring information that contradicts their beliefs [7],[43]. However, such a heuristic might lead to the creation of echo chambers or filter bubbles [45].

Experts in the area of information credibility [20] emphasize the importance of the credibility heuristic for assessing the quality of health-related information [16]. The study finds that one major criterion for assigning the credibility of source (we call it evidence in this research) is whether evidence is an official authority or not. However, current SMS' affordances lack the functionality to differentiate authoritative evidence from non-authoritative ones [19]. Moreover, SMS do not require including an evidence even when discussing health-related topics [172]. Therefore, exposing users to authoritative evidence is critical for the users' assessment of information credibility in SMS.

Nudging was introduced in behavioral economics [193] as a paternalistic intervention to poke individuals toward certain behaviors [59]. Thaler and Sunstein[194] have extended the idea of nudging to help people overcome cognitive biases in decision-making. Previous studies showed the impact of digital nudging on enhancing users' disclosure decision on Facebook [59]. Acquisti et al. [92] stated that nudging addresses the problem of incomplete or asymmetric information within the user interface. Such kind of problem also presents itself on SMS [172], especially when users post information without an evidence or when users include unfamiliar or untrustworthy evidence [164]. Lazer et al. [1] suggested that nudges should be used as a solution to reducing the effect of cognitive biases when evaluating information credibility online. If the selection of information with varying levels of credibility on SMS is reconstructed as a choice and not taken for granted, the choice can be offered through a nudge.

This research proposes two novel digital nudges to help with information credibility assessment on SMS: evidence source credibility nudge and viewpoint consistency nudge. The design choice of these nudges stemmed from our motivations to extend the evidentiality theory when discussing health-related topics to UGC on SMS. In this research, we focus on UGC that relates to health information and specifically to COVID-19 on Twitter. We define UGC as any user-generated content in form of claims, arguments, or news about COVID-19 associated topics in SMS. We focus on Twitter for the following reasons: 1) Twitter users tend to tweet non-scientific information about health-related issues more than scientific information [160]. Another study [195] showed that users tend to discuss health-related issues without providing evidence within their tweets. 2) SMS do not differentiate scientific evidence from non-scientific evidence [106]. In addition, Twitter automatically embeds a relative form instead of an absolute form of a source within individual tweets, which anonymizes the authority aspect of sources [196]. As a result, it hinders the application of heuristics endorsed by medical experts for evaluating information credibility, such as authority heuristics.

Researchers have made progress in developing visual analytical and human-assisted computation tools to analyze misinformation on SMS [16],[48], [49], [183]. However, we observe several main limitations to their research. First, it lacks the evaluation of the effects of these tools on users' assessment of information credibility. Previous scholars highlight the necessity to evaluate the efficacy of system features for users' decision-making[51]. Second, the developed systems are focused on accounts that distribute news stories on SMS [47],[49], but have not considered the news stories within UGCs, such type of UGC was highlighted by previous studies [30],[160] that showed that UGC contains a significant number of misinformation evidence sources accompanied with the content from the user. Third, they mainly target political information on SMS[54], which is different from health information. Fourth, the design of evidence credibility nudge in the majority of previous studies is based on third-party fact checkers[198],[199], [200]. Such an approach is labor-intensive and time-consuming [201]. In addition, nudges that present related information are biased toward mentioning a specific evidence source. Several studies mentioned that users on SMS are affected mainly by their existing beliefs (identity cues)[16]. For example, nudging users with a single evidence source of one political party might influence the belief of a user that follows an opposed political party .

In this study, we develop EvidencEval that implements the proposed two types of nudges. We evaluate the effect of EvidencEval on user assessment of the credibility of information on SMS. Such evaluation will shed light on the effectiveness of proposed nudge designs on users' ability to detect misinformation. In addition, we will extend the scope of UGC to analyze news stories within UGC..

In the remainder of this chapter, section 5.2 introduces the background of this research. Subsequently, section 5.3 discusses related work. Section 5.4 presents the basis of our design for EvidencEval. Section 5.5 list of hypotheses about the effects of the proposed EvidencEval system. Section 5.6 presents our method for evaluation, and sections 5.7 and 5.8 show the results of our method of evaluation. Finally, in section 5.9, we provide our discussions for the results.

5.2 Background

5.2.1 Cognitive Bias and their Role in Misinformation Spread on SMS

Confirmation bias states that individuals tend to look to an explanation that confirms their beliefs and values without further verification of alternate explanations [190]. As a type of cognitive bias, confirmation bias may lead to systematic errors [202].

Selective exposure is a primary reason for amplifying the effect of cognitive biases on SMS [7]. Although exposure has been applied to facilitate information presentation to users on

SMS, but it may lead to worsen users' confirmation bias [140], [203]. Based on an analysis of around 10 million Facebook users, a study [25] concluded that a primary reason for believing misinformation is holding to content and sources that are aligned to the individual's choices and beliefs. Pennycook and Rand [204] found that right-leaning people tend to believe misinformation more than people with other political affiliations. Consequently, with the increasing use of SMS, selective exposure has become more apparent and led to more misinformation sharing [7], [205]. Lazer et al. [1] call for an approach that directs people's attention toward scientific information to build a solution to mitigating misinformation on SMS. Therefore, it is important to provide a computerized support that prompts users and influences their behavior toward looking for scientific evidence while consuming UGC on SMS.

5.2.2 The effectiveness of URLs in Assessing Content Credibility on SMS

Ulicny and Baclawski [206] studied user perception when analyzing information credibility on Twitter. They found that including hyperlinks was a determinant of the credibility of debatable and controversial topics. Suh et al. [207] found that users tend to retweet tweets that contain hashtags and links more than those without these elements. Several studies (e.g., Suh et al.[207] Fogg et al. [20]; Stewart and Zhang[208]) have confirmed that adding hyperlinks in the content conveys a positive signal of credibility.

Twitter offers multiple cues to determine the credibility nature of the content on the platform. Kinsella et al. [209] investigated content and metadata features for classifying the topic and the credibility of the tweets. They concluded that adding metadata features such as hyperlinks helped boost the performance of credibility classification. Castillo et al.[63], [38] studied both content and meta-features of news sources and Twitter arguments under normal and crisis (emergency)conditions. They concluded that including hyperlinks, usually used in health topics to support a claim (referred to as evidence in our context), affected users' perceptions of UGC credibility.

Several studies consider the presence of an URL as an input feature in building classification-based models for the credibility of news on SMS. Slimi et al. [177] built random forests models that predict the credibility of news using different features, with the models based on URL-based features performing best.

Eysenbach and Köhler[210] evaluated the credibility perception of users when evaluating health-related content online, and they found that users focus on the existence of URLs without checking the "about us" page of these sites. Their study suggested the importance of informing users of the authority of the websites when searching for health information on SMS. Park et al. [211] investigated the problem of sharing cancer-related information on Twitter; they found that multiple users shared information from non-health-related websites. They also identified a problem with URLs on the platform, which automatically shortened URLs when users included them in their tweets. Song et al. [212] compared between American and Korean cultures when seeking health-reported information online. They found that users from these countries tend to seek information from SMS sources rather than from professional or authoritative websites (66 % vs 94%). In a similar context, Pulido et al [160] found users tend to tweet more non-scientific evidence than scientific evidence. They highlight the need of making authoritative sources available to SMS users to facilitate their task of credibility assessment.

5.2.3 Augmented Intelligence

Augmented intelligence (AUI) is defined as a design that involves the collaboration between human and artificial intelligence (AI) [57]. It focuses on the assistive role that cognitive technology can be used to enhance but not replace human decision-making [213],[214]. In the healthcare domain, Bollier [215] anticipated the role of AUI in intelligence and improving the skills of physicians. IBM's Watson has targeted to improve the diagnostics and treatments in the healthcare domain [216]. However, AUI has received little attention in the assessment of misinformation online, especially in the deployment of solutions to assess misinformation on SMS. Scholars called for a collaboration between machines and humans to enhance users' decision-making when assessing information on SMS[1].

Digital nudging is a form of AUI that has started to receive attention in recent years [56]. Thaler and Sunstein defined nudges as *"to push mildly or poke gently in the ribs especially with the elbow."*Weinmann et al.[56] defined nudging as using computer-user interface design components to influence online users' behavior or choices. Unlike digital nudging, traditional nudging tends to influence individual actions in offline settings such as organ donations and political voting [217], [218]. In a digital setting, the idea of nudging is to influence users' behaviors or actions to improve their experience or decision-making choices. Nudging in the context of SMS involves issuing warnings to enhance users' decision-making when consuming information [219], among others.

5.3 Related Work

We group the related works in the design of tools (human-machine collaboration)for analyzing misinformation into three streams: computer visual analytics tools, human-assisted computation tools, and augmented intelligence or digital nudges.

Computer visual analytics assistants have started to receive some attention in recent years, specifically analyzing image and metadata features of UGCs in SMS. Shao et al. [47] proposed a system called Hoaxy for tracking and visualizing fake news and rumors of multiple contexts on Twitter. The tool was developed to track news articles that were originated from news websites and then posted on Twitter. The platform provides users with the ability to visualize the content of a tweet and display the origin of its websites and related fact-checking site. In addition, it offers credibility ratings to assist users in making credibility judgments. Gupta et al.[50] proposed a real-time tool for assessing information credibility during crisis and trending events called TweetCred, which displays the credibility ranking of tweets from 1 (low credibility) to 7 (high credibility). The platform is deployed as a chrome extension, and information credibility scores are calculated within just six seconds from the tweet's original time. Finn et al.[48] proposed another web-based tool, called Twitter Trails, that investigates rumors of multiple contexts and their propagation on Twitter. The tool collects relevant tweets, answers questions regarding rumors, and allows users to explore a rumor's bursty activities, temporal propagation characteristics, and retweet patterns. RumorLens[49] is another semi-automatic system for investigating rumors in multiple contexts on Twitter. The system is composed of two components. The first one is a rumor detector that identifies suspicious rumors based on keywords that indicate dubious sentiment. The second component is a visualized interactive interface that shows statistics about rumors. Vosoughi et al.[55] presented a human-machine collaboration system that aims to identify relevant information about an event in multiple contexts. The system consists of two major components: an assertion detector and hierarchical clustering. A hierarchical model was built around 100-1,000 clusters, helping journalists identify relevant rumors of an event. The majority of hybrid systems assume that misinformation follows a similar pattern across different information contexts[53], [47], [46]. However, previous studies have shown that health-related misinformation exhibit different characteristics than other types of misinformation[19][20]. In addition, the majority of developed computer visual analytics assistants focus on providing analytical results at the account level but not at the content level of information[47], [48],[49]. This generalization in the analysis may introduce false negatives when a suspicious account posts trustworthy information.

Human-assisted computation [220] involves outsourcing tasks within a computation process to humans [221]. Crowdsourcing is one example of such a technique [222]. Crowdsourcing has been heavily used in the area of misinformation in the past years, especially in the development of machine learning models for detecting misinformation on SMS [13],[41],[44],[63],[65],[67]. Such crowdsourcing activities have been employed during the ground truth generation for the collected data and to train machine learning models. For instance, Narwal et al. [54] developed a tool called UnbiasedCrowd that uses the voting of the crowd to detect bias in the news on Twitter. Automatic bots notify users of news they are interested in, and it invite them to evaluate the content bias of news. When a news feed exceeds a specific threshold, the tool warns the users of such a potential bias. However, the involvement of humans is scarce in areas beyond ground truth generation [44].

The third stream of related work is in augmented intelligent tools such as digital nudging [56]. It is most relevant to the current research, so we discuss it in detail in the next subsection.

5.3.1 Digital Nudges

In misinformation research, few studies have investigated the effect of digital nudges on assessing information credibility on SMS. Bhuiyan et al. [46] developed a tool called FeedReflect, which aims to assess and assist users in examining news evidence sources on SMS. Their design obscures content that consists of sources that are of questionable credibility.

The system analyzed political-related evidence sources only, and its scope was limited to a small number of mainstream and non-mainstream sources.

Another study [198] focused on examining the efficacy of fact-check alerts in sharing political misinformation on SMS. In a controlled experiment, the authors examined user sharing behavior when fact-checked alerts notify users on misinformation content from credible news evidence sources versus non-credible news evidence sources. They found a higher impact of fact-checked alerts on sharing misinformation from mainstream news sources than those from non-mainstream sources. In a similar study, Moravec et al. [93] tested two types of interventions (i.e., system one and system two) on Facebook users' perception of fake political news. They developed one set of mocks of Facebook posts that included the sign of “Disputed by 3rd Party-Checkers” and developed another set of mocks incorporating the sign of ” Declared Fake by 3rd Party Fact-Checkers.” Their study claims that using the two signs (two systems) significantly reduced fake news's believability on Facebook.

Acquisti et al. [92] presented a digital nudge for influencing individuals’ online revelations behavior and raising user awareness of privacy on SMS. In another online environment, Johnson et al. [223] reported a significant impact of digital nudging on user selection of the most effective healthcare plans. Acquisti et al. [92] discussed digital nudging via an online privacy lens and categorized the dimensions of nudging into information, presentation, and defaults. Information creates awareness of security risks; presentation concerns how information is presented within the privacy context; defaults relate to how default settings serve individuals' needs and their privacy expectations, incentives, reversibility, and timing.

The issue with existing AUI systems is the assumed generalizability of their designs. The majority of these tools assume that misinformation follows a similar pattern despite their different contexts. However, a previous study has shown that health-related misinformation exhibit different characteristics than other types of misinformation [24]. Thus, misinformation analysis should adopt a contextualized approach based on the type of misinformation being investigated [74]. Also, despite the intellectual grounds that these tools are built on, they do not address the problem that an individual on SMS would rely on an evidence source that aligns with his/her beliefs regardless of its veracity[1]. Therefore, it would be promising to develop systems that question the credibility of an evidence source or provide evidence from news media of different affiliations. Previous studies [224], [225] suggest that such systems can offer SMS users an opportunity to question the veracity of a news evidence source or receive a comprehensive consensus on a particular news story.

Another issue with existing human-machine collaboration systems is that the majority has focused on providing analytical results from the account level but not from the content level of SMS posts [47],[48] [49],[53]. An account that publishes a few pieces of misleading information does not necessarily imply it is a malicious account. Additionally, most of these tools focus on misinformation UGC in the form of fake news only. Analyzing fake news within UGC of other forms (i.e., opinions, arguments) is less commonly conducted [21]. Although these tools are developed with multiple analytical features for analyzing social media accounts, they do not evaluate users' decision-making process of information credibility assessment [47], [18],[176]. Thus, there is a lack of empirical evidence for the efficacy of AUI systems for users’ credibility assessment of health related news on SMS.

The majority of nudging systems in the misinformation detection studies developed fact-checked-based nudges to help users assess the credibility of information on SMS [198]. However, the development of such a type of nudges is time-consuming and labor-intensive. The analysis of information relies on experts and/or third-party fact-checkers. In addition, the developed nudges alert users of news-related UGC on SMS only [95] without considering claims or arguments type of UGC. Furthermore, health-related nudging on SMS is understudied. Previous nudges mainly target politically related topics [198],[226]. Studies have shown [20] that credibility assessment of health-related information is different from those of political information.

Lastly, analyzing the nudging impact on users' credibility assessment of health misinformation and when using a nudging system is less commonly investigated compared with other misinformation context types. Identifying which features trigger users' credibility of health misinformation would assist health officials in designing systems that aim to help users' decision-making of information on SMS[227].

5.4 The Design of Nudges

Our approach to designing nudges with heuristic cues draws on design principles from two perspectives: the nudge perspective and the heuristic perspective.

5.4.1 Nudge Perspective

Researchers suggest considering two primary aspects in designing nudges: the mode of thinking involved (automatic vs. reflective) and the degree of transparency (transparent vs. nontransparent)[228].

- **Mode of thinking:** In psychological research, cognitive psychologists developed a concept called dual processes, a theory that explains how humans make decisions. Based on the theory, there are two primary modes of thinking: automatic and reflective[229]. The automatic mode is characterized as fast and instinctive. It relies on knowledge of the past or past behavior and has a minimal cognitive capability. In contrast, reflective thinking critically examines the effects of options before the selection and thus uses more cognitive capacity to make a goal-oriented choice, which makes such thinking slow, effortful, and needs more concentration.

- **Transparency:** Researchers classified nudge designs into two categories (transparent and non-transparent) based on the level of epistemological transparency (i.e., if users would comprehend the purpose of the nudge)[228]. Thaler and Sunstein [230] emphasized the use of transparency in nudging design due to the concern that a designer will manipulate people to choose an option when the reason to display the option is displayed on the interface.

Using a combination of these two dimensions Hansen and Jespersen[228] classified existing nudges into four categories: reflective transparent, automatic transparent, reflective non-transparent, and automatic non-transparent. We chose to focus on the transparent dimension because it allows the user to understand the motives behind it easily. Combining the transparent design with the two modes of thinking leads to two types of design choices: transparent with an automatic mode of thinking, and transparent with a reflective mode of thinking. The reflective mode of thinking nudges users implicitly and needs users to engage in a cognitive effort to determine the credibility of information on SMS. In contrast, the automatic mode of thinking nudges users explicitly on the information credibility of an unfamiliar topic[231].

5.4.2 Heuristic Perspective

Models of effective heuristics have been proposed by cognitive psychologists[232]. One example of such heuristics is the endorsement heuristics where people are inclined to believe information when others do so[5]. Another example is the expectancy violation heuristics where people tend to believe an information from an evidence source only if the source satisfies their expectations [5]. Todd and Gigerenzer [233] identified the robustness of the decision model as a success criterion for heuristic design. They also stated that using a more limited set of information can result in a more robust computational strategy. We selected two heuristics (credibility heuristics and consistency heuristics)to build the design of the nudges based on heuristics for information consumers to determine the credibility of information or sources online [5].

- Credibility Heuristic: credibility is defined as the believability of the evidence source[176]. We compiled the credibility of news evidence sources on SMS based on our review of a diverse set of literature in journalism and online fake news [175], [176],[177]. Then, we classified those news sources based on three credibility metrics: trustworthiness, expertise, and homophily. We selected these credibility metrics because of the consensus of previous literature in information quality and credibility [20],[234], [235], [236], [237]. In addition, the availability of these metrics for the majority of evidence sources is available online. We operationalized these credibility metrics with the following three variables:

    1) **Profitability:** a sub-component of the trustworthiness metric is the profitability of the source [237]. In our context, we measure profitability if the online evidence source's main revenue is its news website.
    2) **Bias:** a subcomponent of the homophily metric is the bias of the evidence source[238]. Homophily is defined as the similarity of the evidence source to the receiver of the information. In our context, we measure bias if the evidence online source has a political affiliation. We classified the bias of evidence sources with right and left political affiliations based on the information from the allsides website.

3) **Expertise:** If the online evidence source's main expertise is health-related issues.

- Consistency Heuristic: consistency is defined as "validating information by checking to see if it comes from different evidence sources.[5] We limit our retrieval of news headlines to those containing information across different types of evidence sources [5]. For the consistency of information, we select supporting, opposing, and mixed news headlines related to the topic of COVID-19 vaccination, as we will explain the rationale for these designs in the hypothesis section.

5.4.3 System and Stimuli Design

We designed EvidencEval to apply the selected heuristics(i.e., credibility heuristics, consistency heuristics) in the development of the nudges to the context of SMS. The system is based on two types of socio-technical nudges: the evidence source credibility and the viewpoint consistency. The evidence source credibility contains low credibility and high credibility evidence source nudges. The low credibility evidence source setting is triggered based on the evidence source of the news headline. Thus, it relies on credibility heuristics. The credibility heuristic follows three credibility variables: profitability, bias, and expertise of the evidence source. These variables are binary (i.e., profitable vs. non-profitable, biased, non-biased, expert, nonexpert). After analyzing the types of evidence sources shared on SMS and by previous literature [172], [30], we utilize four main types of evidence source in the design of EvidencEval (see table 20 for category 1,2,3,4 and their properties) because they covered the majority of evidence sources on SMS. To operationalize transparency, EvidencEval will present three types of messages to users (see table 20 in the message type column for the three message types) right below each news headline. Figure 17 shows a snapshot example of our designed low evidence source credibility nudge. The nudge shows three types of messages below the evidence source mentioned in the UGC. The exclamation mark next to each message was adopted from the nudge currently employed by Twitter[31]. In addition, for presentation reasons, we highlighted each credibility metric.

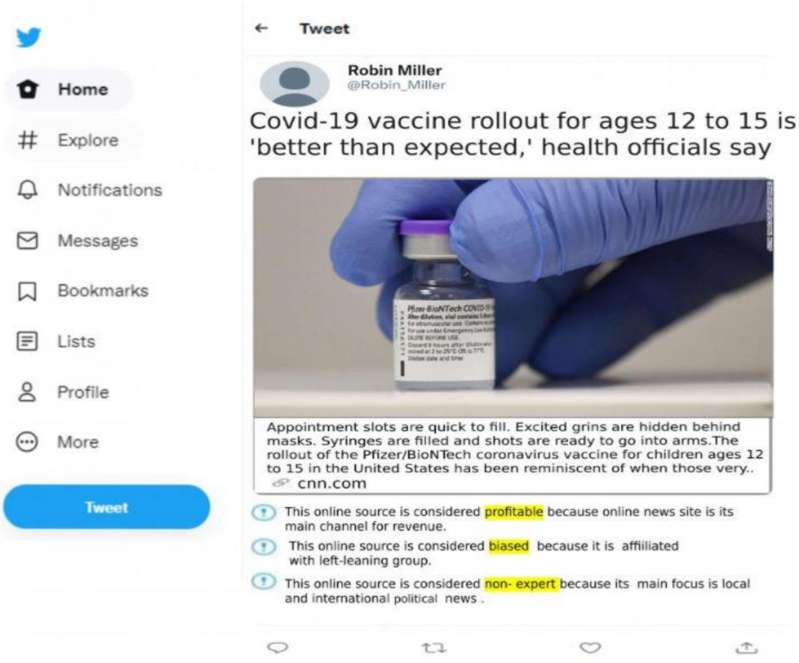
In the high credibility evidence source nudge, EvidencEval presents factual information and references to the CDC web page that are relevant to the topic of the tweets. For experiment purposes and by following the design of a previous work [239], we selected four references of factual information from the CDC to be displayed under the tweets. See figure 18 for a snapshot example of the designed high evidence source credibility nudge.

Table 20: Evidence Source Main Categories and their Properties in EvidencEval.-Low Evidence Source Credibility Nudge.

Evidence Source Category	**Profitable Category**	**Bias Category**	**Expertise Category**	**Message Type**	**Example of Evidence Source**
Category 1	Profitable	Non-Biased	Expert	Message 1: “This online source is considered **profitable** because online news is its main channel for revenue”. Message 2: “This online source is considered **non-biased** because it is not affiliated with any political group”. Message 3: “This online source is considered an **expert** because it is main focus is health-related issues.”	Health.com
Category 2	Profitable	Biased	Non-Expert	Message 1: This online source is considered **profitable** because online news site is its main channel for revenue. Message 2:“ “This online source is considered **biased** because it is affiliated with the left-leaning group”. Message 3:“This online source is considered a **non- expert** because its main focus is not health-related issues.”	boingboing.net
Category 3	Profitable	Biased	Non-Expert	Message 1: This online source is considered **profitable** because online news site is its main channel for revenue. Message 2: “This online source is considered **biased** because it is affiliated with the right-leaning group”. Message 3: “This online source is considered **non-expert** because its main focus is not health-related issues.”	InfoWars

[31] https://techcrunch.com/2020/05/11/twitter-to-add-labels-and-warning-messages-to-disputed-and-misleading-covid-19-info/

Category 4	Non Profitable	Non-Biased	Expert	Message 1: "This online source is considered Non-**profitable** because online news site is not its main channel for revenue". Message 2: "This online source is considered **non-biased** because it is not affiliated with any political group". Message 3: "This online source is considered an **expert** because it is main focus is health-related issues."	CDC.gov



Robin Miller
@Robin_Miller

Covid-19 vaccine rollout for ages 12 to 15 is 'better than expected,' health officials say

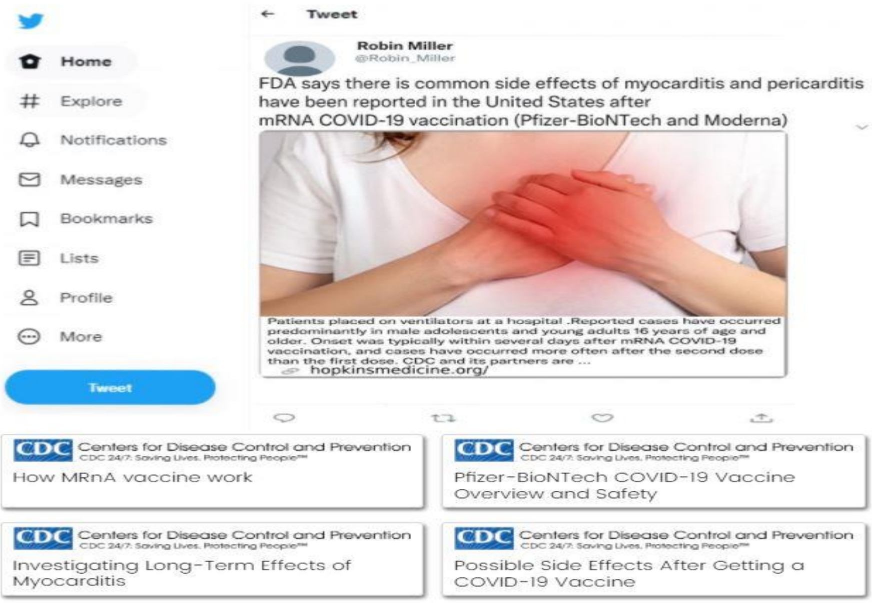
Appointment slots are quick to fill. Excited grins are hidden behind masks. Syringes are filled and shots are ready to go into arms.The rollout of the Pfizer/BioNTech coronavirus vaccine for children ages 12 to 15 in the United States has been reminiscent of when those very..
cnn.com

This online source is considered profitable because online news site is its main channel for revenue.

This online source is considered biased because it is affiliated with left-leaning group.

This online source is considered non- expert because its main focus is local and international political news .

Figure 17: Sample Low Credibility Evidence Source Nudge.



Robin Miller
@Robin_Miller

FDA says there is common side effects of myocarditis and pericarditis have been reported in the United States after mRNA COVID-19 vaccination (Pfizer-BioNTech and Moderna)

Patients placed on ventilators at a hospital .Reported cases have occurred predominantly in male adolescents and young adults 16 years of age and older. Onset was typically within several days after mRNA COVID-19 vaccination, and cases have occurred more often after the second dose than the first dose. CDC and its partners are ...
hopkinsmedicine.org/

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
How MRnA vaccine work

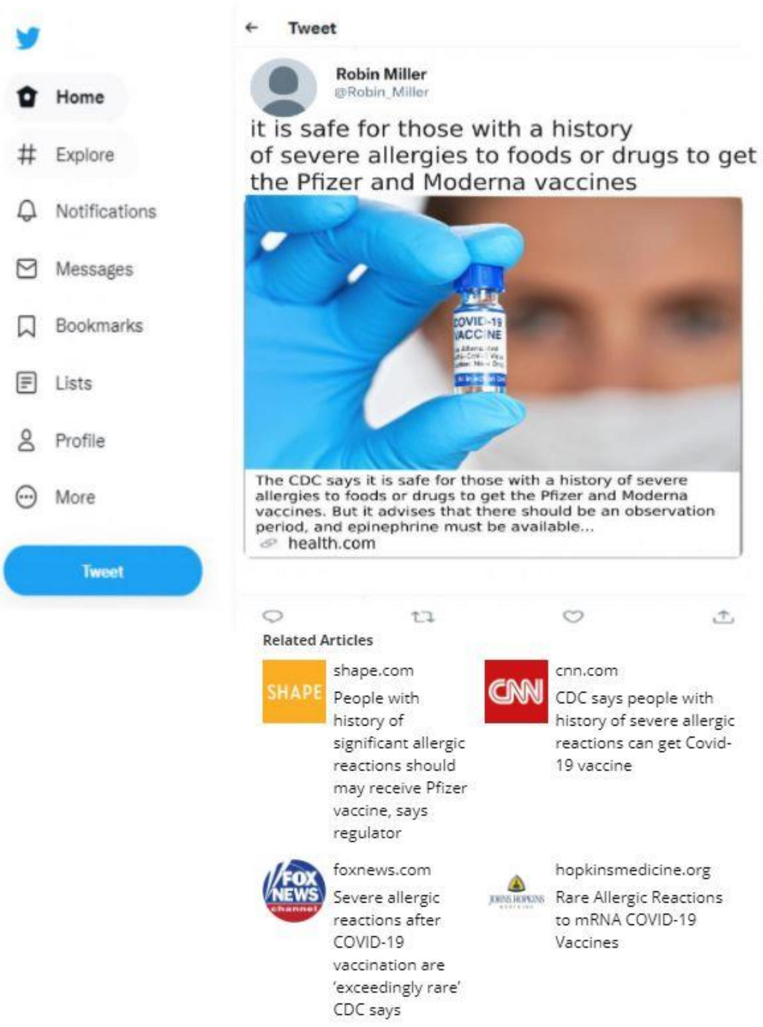
CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Pfizer-BioNTech COVID-19 Vaccine Overview and Safety

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Investigating Long-Term Effects of Myocarditis

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Possible Side Effects After Getting a COVID-19 Vaccine

Figure 18: Sample High Credibility Evidence Source Nudge.

For the viewpoint consistency setting, see figure 19 for a snapshot example of viewpoint consistency nudge. For the design of the nudge , we chose four headlines that vary in the profitability , bias, and expertise metrics. Such number is selected based on four categories of evidence sources we considered in the design of EvidencEval. The nudge shows four news headlines below the tweets. The goal of such design is to provide a viewpoint from different categories of evidence sources. For presentation purposes and to align with the tweet in the figure, we chose the headlines to be displayed along with the evidence sources' logos.



Robin Miller
@Robin_Miller

it is safe for those with a history
of severe allergies to foods or drugs to get
the Pfizer and Moderna vaccines

The CDC says it is safe for those with a history of severe allergies to foods or drugs to get the Pfizer and Moderna vaccines. But it advises that there should be an observation period, and epinephrine must be available...
health.com

Related Articles

shape.com
People with history of significant allergic reactions should may receive Pfizer vaccine, says regulator

cnn.com
CDC says people with history of severe allergic reactions can get Covid-19 vaccine

foxnews.com
Severe allergic reactions after COVID-19 vaccination are 'exceedingly rare' CDC says

hopkinsmedicine.org
Rare Allergic Reactions to mRNA COVID-19 Vaccines

Figure 19: Viewpoint Consistency Nudge Example

5.5 Hypothesis Development

Research interests have grown concerning the development of strategies to influence decision-making in an environment that is heavily affected by cognitive biases. Metzger et al. [240] highlighted that in an environment overloaded with information, people do not have the cognitive capacity or time but tend to use cognitive heuristics to assess the credibility of information, including the evidence credibility heuristics and consistency heuristics.

5.5.1 Evidence Source Credibility

The Elaboration Likelihood Model (ELM) is a framework for understanding the implicit success of persuasive communication that influences individual attitudes [240]. It suggests that factors such as evidence source credibility may influence judgments differently depending on how combinations of those factors affect an individual's motivation and capacity[241]. ELM defines two routes taken by a persuasive message for influencing an individual’s attitude including the central and the peripheral routes. Petty and Cacioppo[240] described the central route as the receiver of information thinks critically about issue-related arguments and scrutinizes the merits and relevance of those arguments before forming an attitude about the advertisement or product. In the central route, facts are used to persuade people of the validity of an argument [242]. On the other hand, the peripheral route uses peripheral cues to link positivity to the message in an indirect manner[243]. Such a route focuses on facts and a product's quality, uses associations with positive qualities like positive emotions and other endorsements [244] and multiple metrics can be used to assess the credibility of the evidence.

In this research, we operationalize the central route as high credibility evidence source which takes the authority heuristic (consider only official sources) as the basis for evidence credibility judgment[86]. The peripheral route is operationalized as low credibility evidence source that considers the three-credibility metrics (profitability, bias, expertise) [20] as the basis for evidence credibility judgment.

5.5.1.2 Low Credibility Evidence Source

Metzger and Flangin [5], [245] mentioned that evidence source credibility might be gauged by examining the source of information and the source's qualifications. Sundar [86]was the first to demonstrate that the assessment of the information credibility online is guided by authority heuristics. He argues that cues within a web-based environment and the presentation of features offered by the technology trigger user's decision-making when evaluating the credibility of the information. He proposed agency affordance which is defined as "if users are aware of the source of information" as a part of the MAIN model that guides the triggering of information credibility heuristics such as the authority heuristics.

Castillo et al. [38] found that individuals tend to perceive UGC to be more credible if it contains any type of evidence source regardless of its credibility. Vraga and Bode [19] also found that users tend to believe health information on SMS when any type of evidence source is embedded within UGC. Pulido et al. [160] claimed that UGCs with non-scientific evidence sources were tweeted more than UGCs with scientific sources on Twitter. However, individuals may put their lives in danger when they select information from non- scientific evidence sources in treating a medical condition. Because of SMS's lack of affordance cues in their design to trigger the evidence source credibility heuristics, it is necessary to provide such cues to facilitate users in assessing the credibility of information within UGCs.

The majority of previous studies in psychology and communication encompass two core dimensions of evidence source credibility: expertise, namely the extent to which the source of the evidence is perceived to be capable of making correct assertions, and trustworthiness, which is the willingness of source evidence to provide the assertions he or she considers most valid [224]. According to Kelman and Hovland [247], an evidence source that profits from persuading the evidence receiver is judged to be less trustworthy, thus having less influence on the attitude of evidence receivers. Thus, we chose profitability as an indicator of trustworthiness in this study. The third dimension of evidence source credibility is homophily, which is the similarity in beliefs and background between parties[234]. The dimension is understudied by the previous literature. One criterion to evaluate homophily is the fairness to different views relating to the discussed topic [235]. Thus, we selected evidence source partisan bias toward particulate view(s) as a measure of homophily of evidence source.

Researchers noted that the interface's affordances could produce positive outcomes [86]. In particular, these outcomes are observed when the design features of an affordance are implemented successfully in an interface. As a result, users are likely to react positively to such affordance for its ability to trigger their cognitive heuristics when assessing information credibility within UGC, which contributes to better decision-making in judging information veracity. Clayton et al. [248] showed in an experimental study that nudging users with "rated false" tags besides the original UGC enhances users' perceived accuracy of Facebook articles. Pennycook et al. [95] indicate the importance of explicitly nudging users by asking about UGCs' accuracy to lower sharing behavior and increase discerning misinformation within UGC on Twitter. Thus, digital nudging might help reduce the spread of misinformation in an SMS environment. Therefore, we propose the following hypotheses (refer to figure 20 to see the research model):

H1a.Using a low credibility evidence source nudge will lead to increased performance in credibility assessment of health *misinformation* in UGC compared with *not* using any nudge.

H2a. Using a low credibility evidence source nudge will lead to decreased intention of sharing health *misinformation* in UGC compared with *not* using any nudge.

5.5.1.3 High Credibility Evidence Source

Sunstein[193] proposed a list of possible nudges that spread information. One of them is the disclosure nudge, which adds supplementary information from authoritative viewpoints. For instance, health-related authorities have used disclosure nudges, like putting nutrition facts on cigarette packets to show the associated risks of smoking [193]. Also, disclosures have been used successfully in online privacy to enhance user choices [59].

One strategy endorsed by health authorities to mitigate health-related misinformation is using denials[32], which refer to messages or statements to debunk misinformation [249], [250]. One of the approaches to spreading denials is displaying high evidence from authoritative sources on a particular topic; such an approach aids in reducing confirmation bias that occurs from believing information aligned with our beliefs [192] and enhances individual information credibility assessment of UGC on SMS [251]. Therefore, we hypothesize the second set of hypotheses as the following.

H1b.Using a high credibility evidence source nudge will lead to increased performance in credibility assessment of health *misinformation* in UGC compared with *not* using any nudge.

[32] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

H2b. Using a high credibility evidence source nudge will lead to decreased intention of sharing health *misinformation* in UGC compared with *not* using any nudge.

5.5.2 Viewpoint Consistency

Studies show that people typically review a few sites to check whether online news is consistent [245]. Consistency heuristic is a common technique for judging news credibility by checking if the news is consistent across different sources[245]. Metzger and Flanagin [5] defined the consistency of news on SMS as the consensus of other sources that either oppose or support this news. However, due to the nature of SMS, polarized news is common in health-related controversies[30]. Researchers have demonstrated the effectiveness of offering diverse viewpoints on a given topic to counter cognitive biases, thus hinting the potential of controversial articles related to fake news[252]. Some researchers suggest using supportive articles to enhance users' decision making on articles related to the topic[239]. In contrast, some researchers have extensively discussed the concept of "considering the opposite" to overcome cognitive biases[253]. This method involves pointing people directly to the opposite side of a topic or question. Other studies have shown [254] that users may be able to detect fake news more easily when articles stating the truth are available. The "illusory truth effect" [226] suggests that articles aligned with a similar position of the consumed news may enable users to detect fake news or become hesitant in sharing the news even if they don't know the truth. Experimenting nudging by providing mixed viewpoints have been used in the area of online medical reviews by displaying user reviews alongside two different treatments and prompts users to carry out a comparative inquiry, preventing a focus on a single treatment [255]. Other studies explored the effect of providing mixed viewpoints on online pro-social behavior, and they found a subtle change in a video stimulus performed by social robots can influence human altruism [256]. Thus, we propose the third set of hypotheses.

H1c. Using a nudge that expresses a) mixed viewpoints of *COVID-19 information*, b) supporting viewpoints, and c) opposing viewpoints will lead to increased performance in credibility assessment of health *misinformation* in UGC compared with not using any nudge.

H2c. Using a nudge that expresses a) mixed viewpoints of COVID-19 *information*, b) supporting viewpoints, and c) opposing viewpoints will lead to decreased intention of sharing health *misinformation* in UGC compared with not using any nudge.

Previous scholars examined the role of emotion in opposing viewpoints while consuming news[257]. They found that partisan-motivated reasoning can be tamed by increased anxiety, which has implications for communication researchers, political psychologists, and others who have long sought to reduce confirmation biases and tribal identity protection [258]. A combination of correlational and causal evidence suggests that reliance on emotions increases belief in fake news: self-reported emotional use was positively associated with belief in fake news [259]. Fear nudges were introduced to evoke feelings of fear, loss, and uncertainty. Zaalberg and Midden [260] discovered that simulating floods by listening to heavy rainfall and watching a river rising gradually was able to motivate individuals to purchase flood insurance. In the context of negative news, regardless of whether it is fake or real, people tend to be attracted to alarming news because it is fear-inducing [261]. Salvi et al. [262] hypothesized that fearful news should predict a greater likelihood of sharing and discerning the credibility of such news. Thus, we propose the following hypotheses:

H1d.Using a nudge that expresses opposed viewpoints with a fearful sentiment will lead to increased performance in credibility assessment of health *misinformation* in UGC compared with using a nudge that spreads opposed viewpoints only.

H2d. Using a nudge that expresses opposed viewpoints with a fearful sentiment will lead to decreased intention of sharing health *misinformation* in UGC compared with using a nudge that spreads opposed viewpoints only.

The design of the two types of nudges: evidence source credibility and viewpoint consistency, are based on the two alternative routes of the ELM model, as discussed earlier. They are expected to play complementary roles in the credibility assessment of UGC. Therefore, combining both types of nudges is expected to have a stronger effect on assessing the credibility and intention to share fake news than using low evidence alone.

H1e. Combining a low credibility evidence source nudge with a viewpoint consistency nudge will lead to increased performance in credibility assessment of health *misinformation* in UGC compared with *using a low credibility evidence source nudge alone*.

H2e. Combining a low credibility evidence source nudge with viewpoints consistency nudge will lead to decreased intention of sharing health *misinformation* in UGC compared with *using a low credibility evidence source nudge alone*.

H1f. Combining a low credibility evidence source nudge with a viewpoint consistency nudge that expresses opposed viewpoints with fearful sentiment will lead to increased performance in credibility assessment of health *misinformation* in UGC compared with *using low credibility evidence-based source nudge alone*.

H2f. Combining a low credibility evidence source nudge with a viewpoint consistency nudge that expresses opposed viewpoints with fearful sentiment will lead to decreased intention of sharing health *misinformation* in UGC compared with *using* with *using a low credibility evidence source nudge alone*.

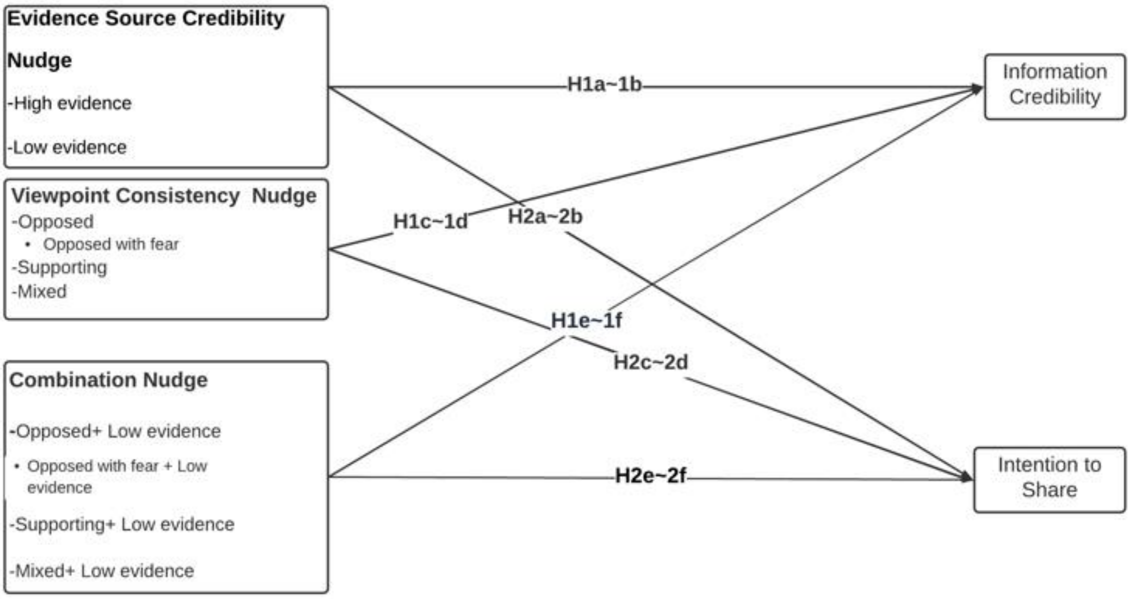


Figure 20 The Research Model

5.6 Method

5.6.1 Participants

To test our hypotheses, we recruited participants from Amazon Mechanical Turk. Information systems and marketing studies have found that online crowdsourcing markets are just as good as or even better than student samples in representing the U.S. population [263]. In total, we recruited 800 participants but we selected 600 adult participants because of the following factors: a) they are a reasonable sample for constructed conditions based on relevant literature [52],[87],[239] and b) a power analysis was conducted to determine an appropriate sample for analysis using an estimated effect size of f2=0.15 (which corresponds for medium effect) and power (1 - β) of 80% and sample size of N=600 will need 50 participants by each condition. We removed 50 participants' answers who did not pass the screening survey, 120 participants' answers who failed in the training session , and 30 participants' answers were not included because they did not finish the entire study. Each participant was paid a total of $2.1 USD. This study has been approved by our Institutional Review Board(see Appendix D).

Participants were required to have completed at least 50 tasks with an approval rate of 95% or higher, as suggested by the research[264]. Among the participants, about 59.1% were male, 68.5% had an undergraduate degree, 23.2% held a graduate degree, and 8.4% had never attended college. The participants were spread across different age groups: 11.1% were between 18-25 years old, 25.7% were 25-34 years, 25% were 35-44, 25.5% of them are 45-60 years old, and 12.7% of respondents are over 60 years old. The participants' political preferences were dominated by Democrats (54.7%), followed by Republicans (28.5%) and the remaining 10.8 % were independent and 6 % had no political preference.

5.6.2 Tasks and Procedure

The experiment follows a mixed design with the type of UGC content as a within-subjects factor (factual vs. misinformation) and nudge condition (with vs. without nudge) as a between-subjects factor. The experiment conditions consist of the following settings:

- Control(no-nudge).
- High credibility evidence source only
- Low credibility evidence source only
- Mixed viewpoints.
- Supporting viewpoints only
- Opposed viewpoints only
- Opposed-Fear viewpoints
- Combination of low credibility evidence source with supporting viewpoints
- Combination of low credibility evidence source with opposed viewpoints
- Combination of low credibility evidence source with opposed-fear viewpoints
- Combination of low credibility evidence source with mixed viewpoints

Each experimental setting consists of four identical tasks. (See figure 21 for Tasks and Procedure). Each participant was asked to read 16 UGCs consisting of news headlines(see Appendix E for the selected headlines) that covered two sub-topics about COVID-19 vaccination: precautions and side effects. These topics were among the most questioned topics on SMS during the period of conducting this research.[33]. We chose a number 16 news headlines because related studies with similar experiment designs also examined participants, on an average, 16 news headlines [16],[95],[265],[266], [267]. In addition, since we focused on two sub-topics of COVID-19 vaccinations, we divided the 16 news headlines to each of these topics making them eight headlines for each topic. The eight headlines per each topic consist of four true and four false news headlines. To avoid any headline-specific effects we retrieved news headlines from left-leaning, right-leaning, and scientific sources. We used a gender-neutral

[33] https://www.pewresearch.org/fact-tank/2021/08/24/about-four-in-ten-americans-say-social-media-is-an-important-way-of-following-covid-19-vaccine-news/

Twitter handle as the author of these news headlines. Moreover, the news headlines were randomized in presentation order and treatment (e.g., text and news evidence sources).

The study proceeded in the following steps (the full list of questions is in Appendix F). First, participants answered questions concerning the knowledge and concerns about the COVID-19 pandemic. Second, participants proceed to the training session by viewing the EvidencEval nudge interfaces and answer questions about the components of the interface. Third, participants viewed UGC from one of the two types of settings, 1) *control condition(no nudge)* that consists of UGCs but without any nudging type, 2) *treatment conditions* that consists of UGCs with evidence nudging (mentioned earlier). Finally, participants were asked to respond to a post-experiment survey that contained questions about demographics, social media usage, and their level of trust in news organizations.

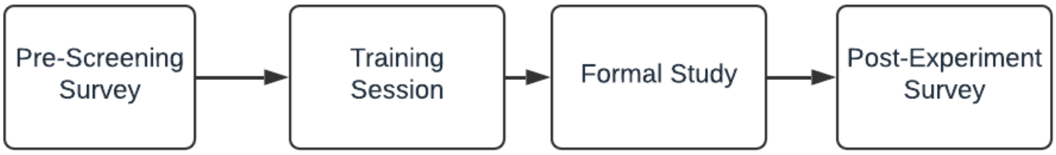


Figure 21: Tasks and Procedure.

Dependent Variables. For measuring information credibility variable, the participants were asked the following credibility assessment question for each viewed UGC, *"Based on the claim in the news headline posted by the author, How likely do you think the claim is accurate and truthful?"* we used an item of the message credibility to assess individuals' perceptions of the accuracy of each news headline [268]. Participants reported the extent to which the viewed news headlines are fake or real on a 7-point Likert-type scale (from 1 = Very unlikely 7 = Very likely). Also, to measure the sharing intention variable participants are asked the following intention to share questions *"If you were to see the above claim on Social Media, would you consider sharing it?"*An item of the message sharing scale was used to assess individuals' perceptions of their intention to share each news headline[198].Participants reported the extent to which they would share the viewed true and fake news headlines on an item of 7-point Likert scale (from 1 = Very unlikely 7 = Very likely).

Headline credibility assessment was categorized as (0=fake, 1=factual). Similarly, intention to share was coded as (0=not intend to share, 1=intend to share). Users with an average answer score of four or less (for either of the credibility assessment or sharing intention question) are categorized as not tending to believe/share fake news. In contrast, users with an average answer score of more than four tend to believe/share fake news.

Control Variables. We considered six control variables(see Appendix G for control variables measurement). The following question used a seven-point Likert scale(individual's frequency of news reading on SMS, individual's frequency of news sharing on SMS both questions used a seven-point Likert scale, individual's political leanings on seven-point Likert scale, trust in news organizations on seven-point Likert scale). Reading and sharing news frequency their dummy variables are measured using a nominal scale, which we used once a year for both of the items for comparison. For political leanings we used users who answered of not having political leanings as a reference in the group to measure their dummy variables. For COVID-19 concerns, and COVID-19 news checking we used a five-point Likert scale. The first four variables showed positive associations in terms of discerning fake news in the area of fake news detection [93], [198]. The latter two variables were selected following the suggestion of previous work[95].

5.6.3 Nudges' Headlines Evaluation

Since our viewpoint consistency nudge consists of four types of viewpoints (i.e., supporting, opposing, opposing-fearful, mixed), we evaluated the selected viewpoint nudges for each of the viewpoint setting with two annotators to verify that each headline is actually a representative of its own viewpoint nudge. For the supporting and opposed viewpoints settings each participant views 16 news headlines and each nudge settings shows four headlines incorporated under the main shown headline, the total number of headlines for both viewpoint settings (i.e., supporting and opposing) is 128 headlines. We examined annotators agreement total agreement using Cohen Kappa and this step yielded 100 % ($\kappa=1$, $p<0.001$) between the annotators. For the opposed fearful nudges, we used IBM Watson tone analyzer[34]. The demo provides an analysis of different types of sentiments (joy, fear, sadness, anger, etc.). We analyzed each of the four nudged headlines presented under the main news headline with the tool and each of these headlines shown a fear sentiment.

5.7 Analysis

We tested the research model using the logistic regression analysis. To gain a deeper understanding of the effect of the nudges we analyzed the results with and without the control variables.

[34] https://tone-analyzer-demo.ng.bluemix.net/

5.8 Results

We analyzed subjects' homogeneity to ensure comparability across the different conditions. Following the previous work[239], [93], we chose three variables (gender, education, and age) to test the homogeneity between the treatments(see table 21). We used chi-square tests for discrete variables such as gender and education and ANOVA for continuous variables such as age. Test statistics indicate that there are no significant differences in participants' age (*p* = 0.194), gender (*p* = 0.881), or education (*p* = 0.195) between treatments and between treatment and control groups. Therefore, the participants were comparable between different conditions.

Table 21: Homogeneity Analysis between groups using Chi-Square and ANOVA analysis of variable.

Variables	χ^2	**F**	**P**
Gender	5.882	-	0.881
Education	27.427	-	0.195
Age	-	1.394	0.194

5.8.1 Nudging Effect on Fake News Credibility Assessment and Sharing Intentions

Table 22 shows the descriptive statistics of news credibility assessment. We report the logistic regression hypothesis testing results of the nudges in table 23, where the no-nudge setting was used as the baseline.

Table 22: Descriptive Statistics for Credibility Assessment

Treatment	**Mean**	**Standard Deviation**
Control (no nudge)	4.085	1.287
High Credibility Evidence Source	4.220	1.282
Low Credibility Evidence Source	3.764	0.792
Mixed Viewpoints	3.744	1.101
Opposed Viewpoint	3.846	1.130
Opposed-Fear Viewpoints	3.506	0.723
Supporting Viewpoints	4.879	0.983
High Evidence + Low Evidence Source Credibility	4.635	1.167
Low Evidence Source Credibility + Opposed Viewpoints	3.800	0.939
Low Evidence Source Credibility + Opposed-Fear Viewpoints	5.229	1.380
Low Evidence Source Credibility + Supporting Viewpoints	4.883	1.145
Low Evidence Source Credibility + Mixed Viewpoints	3.826	0.827

We can see from table 23 that several nudges showed effects on credibility assessment of fake news headlines. The logistic regression model showed that low credibility evidence source nudge had a positive effect (β = -0.232**, *p* = 0.009) on credibility assessment of fake news. However, high credibility evidence source nudge did not yield significant scores on users' credibility assessment of fake news (*p* > .05). Thus, hypothesis H1a was supported, H1b was not supported. In contrast, for sharing intentions of fake news headlines H2a and H2b, the results showed that low credibility evidence source nudge yielded a negative effect on users' intention to share fake news headlines (β = -0.156*, *p* = 0.043). However, high credibility evidence source nudge did not affect user sharing intention of news headlines (β = 0.175, (*p* > 0.05). H2a was supported, but H2b was not supported.

The analysis results of viewpoints nudge showed that both mixed viewpoints nudge (β = -0.238**, *p* = 0.016) and opposed viewpoints nudge (β = -0.177*, *p* = 0.033) yielded a significant effect on users' credibility of fake news assessment. However, the supporting viewpoints nudge did not affect users' ability to detect fake news. Therefore, H1c was partially supported. In terms of sharing intentions H2c, the results showed no viewpoint nudge affected users' intention to share fake news. Therefore, H2c was not supported.

Table 23: Logistic Regression Test Results for Fake News Credibility Assessment and Sharing Intention

	Fake News Headlines Credibility Assessment	**Fake News Sharing Intention**
High Credibility Evidence Source	0.78	0.175
Low Credibility Evidence Source	-0.232**	-0.156*
Mixed Viewpoint	-0.238**	0.048
Opposed Viewpoint	-0.177*	0.098
Opposed-Fear Viewpoint	-0.263**	-0.040
Supporting Viewpoint	0.180	0.254
High Credibility Evidence Source + Low Credibility Evidence Source	0.130	0.207
Low Credibility Evidence Source + Opposed Viewpoint	-0.044	-0.024
Low Credibility Evidence Source + Opposed-Fear Viewpoint	0.295	-0.118
Low Credibility Evidence Source + Supporting Viewpoint	0.310	0.225
Low Credibility Evidence Source + Mixed Viewpoint	-0.248**	0.074
Control Variables		
Reading_Frequency > Once a Week	-0.080	0.170
Reading_Frequency > Once a Day	-0.141	0.435
COVID19_Concerns	0.002	0.002
COVID19_News_Check	0.004	0.031
Trust_In_News_Organization	0.026	0.027
Democratic	0.299	0.108
Republic	0.357	0.153

5.8.2 Comparison of the Effects of Different Types of Nudges on Fake News Credibility Assessment and Sharing Intentions

We conducted several Wald tests [269] to determine which nudges yielded the strongest effect in comparison with other nudges. Wald Test is a widely used method of behavior studies in relevant fake news literature.[204], [93], [270]. In addition, the Wald test has the benefit of comparing two treatments in hypothesis testing [271].H1d and H2d argued that opposed-fearful viewpoints would have a stronger effect on credibility assessment and sharing intentions of fake news than showing only opposed viewpoints. The results show(see table 24) that opposed viewpoints with fearful sentiments is more influential ($p= 0.022$) than using only opposed viewpoints in helping users detect fake news. Thus, H1d was supported. However, in terms of the effect of the opposed viewpoints with fearful sentiment nudge in sharing fake news with opposed nudge, the results showed no significant difference between the two nudges (p>0.05).H2d was not supported. Surprisingly, combining high and low evidence source credibility nudge has affected users' credibility assessment of fake news more than using only a low evidence-based source credibility nudge.

Although it was mentioned in the previous section that mixed viewpoints nudge has shown a significant effect on users' ability to detect fake news on SMS, combining it with low evidence-based nudge did not yield significant results in comparison with using low evidence-based nudge alone(p>0.05). Based on the previous discussion, we conclude that H1e and H1f were not supported. Previous results showed that despite the influence of the opposed or low evidence-based nudge in helping users discern fake news, combining low evidence-based source credibility nudge and opposed nudges did not provide a notable effect on users ability to detect fake news than in comparison with using low evidence source credibility nudge alone($p>0.05$). Surprisingly, using low evidence source credibility nudge alone has shown a significant effect ($p=0.001$) than using opposed-fear viewpoints and combining it with the low evidence source credibility nudge. Similarly, low evidence source credibility nudge has shown a significant effect ($p=0.001$) on users' ability to discern fake news in comparison with using combining low evidence source credibility nudge with supporting viewpoints.

For the low evidence source credibility nudge vs combination of low credibility evidence source nudge and the viewpoints consistency: a-opposed, b-opposed-fear, c- mixed, the results showed no significant effect of such combination over low credibility evidence source nudge ($p>0.05$) in reducing the sharing of fake news. Interestingly, low credibility evidence source nudge showed a high significance over the combination of low credibility evidence source nudge and supporting viewpoints($p=0.001$). H2e was not supported. For sharing fake news, the Wald tests showed that opposed fear viewpoints did not differ significantly ($p>0.05$) on users' sharing intention of fake news compared with using low credibility evidence source nudge.H2f was not supported.

Table 24: Wald Tests of Different Types of Nudges.

Fake News Credibility Assessment Performance	**$\chi 2$**
Opposed viewpoint (-0.177*) Vs Opposed-Fear viewpoints (-0.263**)	7.631*
Low Credibility Evidence Source (-0.232**) vs High and Low Credibility Evidence Source (0.130)	**12.710****
Low Evidence Credibility Evidence Source (-0.232**) vs Low Credibility Evidence Source + Opposed viewpoints (-0.044)	5.963
Low Evidence Credibility Evidence Source (-0.232**) vs Low + Opposed-Fear viewpoints(0.295)	**22.414*****
Low Evidence Credibility Evidence Source (-0.232**) vs Low Credibility Evidence Source + Supporting Viewpoints (0.310)	**29.985*****
Low Evidence Credibility Evidence Source (-0.232**) vs Low Credibility Evidence Source + Mixed Viewpoints (-0.112)	5.671
Fake News Sharing Intentions Performance	
Opposed viewpoint (-0.177*) Vs Opposed-Fear viewpoints (-0.263**)	2.302

Low Credibility Evidence Source (-0.232**) vs High and Low Credibility Evidence Source (0.207)	14.232
Low Evidence Source Credibility (-0.232**) vs Low Credibility Evidence Source + Opposed Viewpoints (-0.024)	3.181
Low Evidence Credibility Evidence Source (-0.232**) vs Credibility Evidence Source + Opposed-Fear Viewpoints –(0.118)	3.180
Low Evidence Credibility Evidence Source (-0.232**) vs Low Credibility Evidence Source + Supporting Viewpoints (0.225)	**17.040*** **
Low Evidence Credibility Evidence Source (-0.232**) vs Low Credibility Evidence Source + Mixed Viewpoints (0.074)	5.698

5.9 Discussion

First, our findings shed light on the efficiency of evidence credibility nudges and viewpoint consistency nudges on credibility assessment or sharing intention of fake news on SMS. Our finding showed that evidence source credibility nudges and particularly low evidence source credibility nudge was the only nudge type that was both influential in enhancing the credibility assessment and sharing intention of fake news on SMS. On the other hand, viewpoint consistency nudge results showed the effectiveness of mixed, opposed and opposed-fearful viewpoints in enhancing the performance of credibility assessment of fake news in comparison with other nudges in the same category. Finally, the combination of two nudge categories(i.e., evidence source credibility and viewpoint point consistency) have shown an effect on credibility assessment of fake news when only combining a low credibility evidence source nudge with the mixed viewpoints.

The findings showed that high credibility evidence source nudge and combining the two types of nudges did not lead to further performance improvement in credibility assessment were surprising. Our results suggested that combining nudges from the two categories(i.e., evidence source credibility and viewpoint consistency) may undermine the effectiveness of nudge categories on the credibility assessment or sharing intention of fake news. We provide the following explanations. First, for the high credibility evidence source nudge, accessing the factual information requires the user to click and access the evidence on a different page. Our results from the viewpoint consistency nudge showed that providing a nudge with information on the same page rather than directing the user to another page has shown its effect in several viewpoint nudges. Second, we analyzed the number of clicks on news articles and found that around 84% of the users did not click on the associated articles provided by high credibility evidence source nudge. Third, previous research showed that the design aspect of a nudge is critical when combining different nudge categories[93]. Our findings confirmed such suggestion and we found that the majority of combined nudges did outperform using one of the nudges and only combining low credibility evidence source nudge with mixed nudge produces an effect on credibility assessment. We suspect that maybe producing a balanced viewpoint design by showing mixed articles (i.e., supporting and opposing)within the nudge may complement the features of the low evidence-based nudge. Although several research conducted to analyze user behavior when seeing fake news have asked their participants credibility assessment and sharing intention questions consecutively, recent research nudging related studies on SMS [95], [266]suggested asking participants either sharing intention question or credibility assessment questions and not both.

This research provides the following contributions. First, we developed theoretical-based nudges based on the credibility heuristic approaches, taking into consideration two types of nudging designs: a) automatic mode of thinking(evidence source credibility nudge) and reflective thinking (viewpoint consistency nudge). As far as we know, this is the first work that considers three evidence credibility metrics (i.e., profitability, bias, and expertise)into the design of nudges(evidence source credibility nudge) on SMS. In addition, our viewpoint nudge is the first nudge that considers evidence from different evidence types of news organizations on SMS. This work also assesses and compares different types of viewpoint nudges on the health-related context on SMS. We evaluated when multiple types of evidence produce different types of viewpoints (supporting, opposed, mixed). In addition, we design and assess a new type of viewpoint nudge that considers opposed- fearful viewpoints from different evidence types of news organizations. Our work designed and assessed opposed-fearful viewpoint nudge compared with using opposed viewpoints. Adding an emotional aspect to the nudge have not been considered by previous work. Using emotions to combat misinformation has been suggested by multiple authors[1], [259] in the area of fake news detection on SMS. Our work is the first that combines the use of emotion in the design of nudging to enhance the credibility assessment of fake news. Finally, we designed and evaluated the effect of combining the two nudge designs (evidence source credibility and viewpoint consistency) compared with using each nudge category individually.

Our work proposed the following implications on the design and development of nudges on SMS. First, it showed the automatic- transparent mode of thinking design is effective in the development of nudges to enhance the credibility assessment and sharing intention of fake news. Specifically, using a low evidence-based nudge that takes into consideration three credibility metrics have shown a significant effect in reducing the credibility scores and limiting the intention to share of fake news on SMS. In addition, for the reflective thinking, our results call. Second, in terms of providing related article under the SMS posts. Our evaluation of the viewpoint consistency nudging design highlights the importance of providing multiple types of evidence types from multiple news organization that follows different political affiliation types in the design of viewpoint nudging to combat misinformation of SMS. In addition, our results suggest the use of mixed viewpoint and opposed-fearful viewpoint in developing viewpoint nudges on SMS. Despite using opposed viewpoints have showed efficiency in enhancing the credibility of fake news assessment, our work showed that using opposed-fearful viewpoint have shown greater effect in evaluating fake news on SMS. Thus, we believe that using emotions of other types should be evaluated in the future. Finally, our work showed that combining nudges is not always beneficial and supports the finding of previous literature[83], [111],[272], that highlights the design aspect of the nudges before combining them to combat misinformation on SMS.

CHAPTER 6: CONCLUSIONS

This dissertation applied evidently theory to evaluate the credibility of misinformation on SMS. We highlighted that health misinformation not only exists in news headlines but also exists from users' arguments toward health-related topics. Therefore, we applied evidence-based categorization to detect arguments supportive and against health misinformation. In addition, based on the evidentiality theory, we developed several design artifacts to augment human credibility assessment of health misinformation on SMS.

6.1 Summary

Chapter 3 aims to show the type of evidence sources discussed during health controversies. We showed that there is a clear manifestation of non-authoritative and non-scientific evidence sources within UGCs on SMS. Also, our results showed that evidence sources were prevalent in the conspiracy dataset. We developed a customized scraper that accessed Twitter to collect tweets and specific types of features tweets related to COVID-19 for a period of 7 months. In addition, based on a variety of work in the area of journalism, we developed a large lexicon of information sources (evidence sources) to aid our analysis; we collected around 700,000 tweets about COVID-19 and conducted a qualitative and quantitative analysis. Although news publisher evidence sources constitute most of the evidence sources across both conspiracy and non-conspiracy datasets, the evidence sources of partisan news media are primarily used in discussing conspiracy topics. In addition, bot analysis reveals the frequent involvement of automated accounts in discussing COVID-19 topics and, more importantly, the higher tendency of bot accounts in discussing conspiracy than non-conspiracy topics. Therefore, the current research calls for developing a computation-based approach to signaling misinformation such as conspiracies within UGCs on SMS. Another implication is the need for highlighting misinformation evidence sources when discussing critical topics such as health-related topics.Drawing on the evidentiality theory, in chapter 4, we developed an approach that combines the evaluation of evidence sources with stance analysis to detect arguments leaning toward health misinformation on SMS. Unlike existing studies that are focused on detecting UGCs in the form of fake news on SMS [43],[28], [239], we address claims which contain arguments that show a favoring stance toward misinformation. The literature in the areas of linguistics [162], [148] and argumentation mining [273] shows the importance of distinguishing arguments(which contain stance and evidence) from opinions (which only contain stance) for information credibility assessment. In our research, we addressed such differentiation and developed an annotation scheme for health-related arguments and related evidence in UGCs. Based on the schema, we built a dataset of tweets that contain user arguments toward COVID-19 health-related topics. Based on the annotated datasets, we built machine learning models to classify the stance within arguments' that take support or against arguments toward COVID-19 misinformation on SMS. The findings of this chapter showed the importance of evidence-based features in identifying arguments related to COVID-19 vaccination. In addition, such features improved the performance of detecting supportive and against stance toward COVID-19 vaccination.

Based on the elaboration likelihood model (ELM) [240], peripheral routes such as the source's authority and expertise are one of the routes in analyzing information credibility online. Drawn on the evidentiality theory and the ELM model, in chapter 5, we propose EvidencEval, an evidence-based nudging system that equips users with evidence credibility nudge and viewpoint consistency nudge to evaluate the evidence within UGCs. Based on these two nudges categories, we developed 11 types of nudges and assessed their effectiveness in enhancing the credibility of fake news on SMS. Our experiment showed the importance of two design artifacts that formed the design of these sub-nudges: a) automatic-transparent mode of thinking b) and reflective mode of thinking for enhancing credibility assessment of health information on SMS. In addition, the system's design triggers users with information retrieved from authoritative-based websites of health-related information provided to users when analyzing UGCs on Twitter.

6.2 Contributions

This dissertation made the following contributions to the literature:

First, this dissertation extends an evidentiality theory [22] and credibility cognitive heuristics provided by health experts to analyze the types of evidence incorporated with health-related UGCs, including favoring or opposing stance toward health misinformation on SMS. Applying an evidence-based approach to detect health misinformation in SMS within arguments has not been investigated in previous works[22].

Second, this dissertation constructs an evidence-based schema for categorizing evidence within UGC and presents a categorization scheme of different types of evidence within UGC online. This work provides a semi-automated approach for classifying evidence within UGC on SMS.

Third, this dissertation expands the foundations of stance detection on SMS by differentiating between opinion and argument-based UGCs on SMS. In addition, it creates a dataset of opinions and arguments that includes favoring or opposing stance toward health misinformation on SMS.

Fourth, it uses evidentiality theory as the kernel theory to guide the design of digital nudges. In particular, it shows how an evidence-based design artifacts can be used to support augmented intelligence for mitigating the spread of health-related misinformation on SMS. This work presents 11 nudge types designed to analyze the evidence within UGC on SMS. To the best of our knowledge, this dissertation is the first that incorporates evidence categorization into the design of digital nudges to support users' credibility judgment of health-related information on SMS.

Fifth, the dissertation combines cognitive heuristics to the design of digital nudges. Specifically, it uses credibility and consistency to analyze UGCs on SMS. As far as we know, this dissertation is the first that integrates dimensions of the three information credibility metrics (trustworthiness, homophily, expertise) to the development of digital nudges.

6.3 Implications

Our work has several implications for misinformation detection on SMS.

The first implication is that identifying arguments is important in evaluating the credibility of UGC on SMS. Specifically, identifying the stance position toward health-related topics can be improved when considering the argumentative nature of the UGC. For developers, recognizing supportive or against arguments that are incorporated with evidence sources can enhance the level of persuasion between users in their discussions by considering stance with pieces of evidence (i.e., arguments). In addition, incorporating argument detection may decrease the spread of health information that do not contain evidence on SMS. In other words, users can identify UGCs with accompanied evidence toward health-related topics. In addition, identifying arguments on SMS will incorporate stronger evidence when discussing health-related topics on SMS.

The second implication, our work highlights the importance of evidence credibility nudges and viewpoint consistency nudging on SMS. This research will help developers of nudges on SMS rely on evidence nudges more than the widely used labor and time-consuming approaches such as fact-checking websites. For users, using evidence source credibility nudges will assist them in choosing information from an evidence source that satisfies the credibility metrics. In addition, using consistency nudges will help them to determine how different evidence sources' viewpoints are aligned with the viewpoint of evidence sources within UGC. Moreover, it will highlight the importance of evaluating different viewpoints to users when assessing the credibility of information or sharing information on SMS.

Third, the findings of this work emphasize the importance of the automatic-transparent mode of thinking design and reflective design on users' credibility assessment and sharing intentions of UGCs on SMS. Such results pave the way for developers to inspire their designs from these two design dimensions for the goal of combating misinformation on SMS. For users, relying on nudges will inspire them to rely more on other types of digital nudges on SMS.

Fourth, our work shows that incorporating links to authoritative websites might not be sufficient to reduce the credibility assessment or sharing intentions of fake news. Such findings will help the developers of current SMS platforms that incorporate links to authoritative websites to recognize the efficiency of such an approach on misinformation spread.

6.4 Limitations

Like other studies, this dissertation has several limitations. First, due to the traffic the scraper could generate in collecting posts from Twitter, we limited the data collection of COVID-19 vaccination tweets to 9 months. Such action restricted our analysis to the early stages of vaccine distribution. So, the findings may vary to other stages of COVID-19 vaccination distribution. Second, we aimed to build detection models by incorporating arguments that have a neutral stance position; however, we found their frequency was very low, and building machine learning models on too low post frequency is challenging. Third, our evidence-based digital nudges cover two sub-topics of COVID-19 vaccination. Other subtopics related to vaccination may follow a different pattern. Lastly, the developed nudges rely heavily on the evidence sources within a UGC, and those UGCs without evidence sources would not be analyzed.

6.5 Future Work

The scope of this research can be extended in several ways. First, we aim to increase the size of the COVID-19 vaccination tweets of argumentation types to cover multiple timeframes. Analyzing the polarity of arguments and the type of evidence disseminated between an early stage of the pandemic with a later stage will shed light on how people discuss an ongoing health crisis on SMS. Second, to expand our context, we plan to investigate the author of arguments on SMS to evaluate evidence categorization to the account level. Such a step will be beneficial in the case a stance is expressed without evidence in the UGC. Third, we plan to further analyze and annotate arguments toward other targets, such as other types of COVID-19 vaccinations. Such effort will enable our machine learning models to detect arguments and stance positions toward multiple COVID-19 vaccination types. Fourth, our nudging design highlights the importance of examining nudge designs that are categorized under the two categories (i.e., mode of thinking and transparency). Experimenting with other different designs under the categories will shed light on the most reliable nudges designs to combat misinformation spread on SMS.

Fifth, we intend to use our nudging design to another misinformation context, such as the political context. Unlike other nudges that rely on context-specific fact-checking services to verify the information within UGC, our nudging design relies on the type of evidence source which can be extended to other misinformation contexts. Moreover, examining our nudging designs in other information contexts will generalize the reliability of our nudges on the credibility assessment of information on SMS. Sixth, our nudging design can be expanded to examine the author of the UGC on SMS. Conducting such a step will enhance the analysis of evidence sources from focusing only on evidence sources within UGC to the author of UGC. In addition, extending the scope to analyze the author of the UGC will contribute to digital nudges on SMS by examining UGC that does not contain evidence. Finally, although the number of UGCs with two or more evidence sources are scarce based on our analysis, we intend to extend the scope of evidence credibility nudges where two types of evidence (with different credibility) exist within UGC.

REFERENCES

[1] D. M. J. Lazer *et al.*, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, Mar. 2018, doi: 10.1126/science.aao2998.

[2] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[3] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran, “Falling for Fake News: Investigating the Consumption of News via Social Media,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada, Apr. 2018, pp. 1–10. doi: 10.1145/3173574.3173950.

[4] B. Kang, T. Höllerer, and J. O’Donovan, “Believe it or Not? Analyzing Information Credibility in Microblogs,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM ’15*, Paris, France, 2015, pp. 611–616. doi: 10.1145/2808797.2809379.

[5] M. J. Metzger and A. J. Flanagin, “Credibility and trust of information in online environments: The use of cognitive heuristics,” *Journal of Pragmatics*, vol. 59, pp. 210–220, Dec. 2013, doi: 10.1016/j.pragma.2013.07.012.

[6] D. A. Broniatowski *et al.*, “Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate,” *Am J Public Health*, vol. 108, no. 10, pp. 1378–1384, Oct. 2018, doi: 10.2105/AJPH.2018.304567.

[7] D. Spohr, “Fake news and ideological polarization: Filter bubbles and selective exposure on social media,” *Business Information Review*, vol. 34, no. 3, pp. 150–160, Sep. 2017, doi: 10.1177/0266382117722446.

[8] 1615 L. St NW, Suite 800Washington, and D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “The Social Life of Health Information,” *Pew Research Center: Internet, Science & Tech*, Jun. 11, 2009. https://www.pewresearch.org/internet/2009/06/11/the-social-life-of-health-information/ (accessed Feb. 14, 2020).

[9] I. Seaman and C. Giraud-Carrier, “Prevalence and Attitudes about Illicit and Prescription Drugs on Twitter,” in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, Chicago, IL, USA, Oct. 2016, pp. 14–17. doi: 10.1109/ICHI.2016.98.

[10] D. Scanfeld, V. Scanfeld, and E. L. Larson, “Dissemination of health information through social networks: Twitter and antibiotics,” *American Journal of Infection Control*, vol. 38, no. 3, pp. 182–188, Apr. 2010, doi: 10.1016/j.ajic.2009.11.004.

[11] S. Shi, A. R. Brant, A. Sabolch, and E. Pollom, “False News of a Cannabis Cancer Cure,” *Cureus*, Jan. 2019, doi: 10.7759/cureus.3918.

[12] A. Nastasi, T. Bryant, J. K. Canner, M. Dredze, M. S. Camp, and N. Nagarajan, “Breast Cancer Screening and Social Media: a Content Analysis of Evidence Use and Guideline Opinions on Twitter,” *J Canc Educ*, vol. 33, no. 3, pp. 695–702, Jun. 2018, doi: 10.1007/s13187-017-1168-9.

[13] M. Mendoza, B. Poblete, and C. Castillo, “Twitter under crisis: can we trust what we RT?,” in *Proceedings of the First Workshop on Social Media Analytics - SOMA ’10*, Washington D.C., District of Columbia, 2010, pp. 71–79. doi: 10.1145/1964858.1964869.

[14] J. Agley and Y. Xiao, “Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science,” *BMC Public Health*, vol. 21, no. 1, p. 89, Jan. 2021, doi: 10.1186/s12889-020-10103-x.

[15] W.-Y. S. Chou, A. Oh, and W. M. P. Klein, “Addressing Health-Related Misinformation on Social Media,” *JAMA*, vol. 320, no. 23, p. 2417, Dec. 2018, doi: 10.1001/jama.2018.16865.

[16] X. Lin, P. R. Spence, and K. A. Lachlan, “Social media and credibility indicators: The effect of influence cues,” *Computers in Human Behavior*, vol. 63, pp. 264–271, Oct. 2016, doi: 10.1016/j.chb.2016.05.002.

[17] “Judgment of information quality and cognitive authority in the web,” *J. Am. Soc. Inf. Sci.*.

[18] R. Savolainen, “Judging the quality and credibility of information in Internet discussion forums,” *J. Am. Soc. Inf. Sci.*, vol. 62, no. 7, pp. 1243–1256, Jul. 2011, doi: 10.1002/asi.21546.

[19] E. K. Vraga and L. Bode, "Using Expert Sources to Correct Health Misinformation in Social Media," *Science Communication*, vol. 39, no. 5, pp. 621–645, Oct. 2017, doi: 10.1177/1075547017731776.

[20] J. Stanford, E. R. Tauber, B. J. Fogg, and L. Marable, "Experts vs. Online Consumers: A Comparative Credibility Study of Health and Finance Web Sites," p. 60, 2002.

[21] A. E. Lillie and E. R. Middelboe, "Fake News Detection using Stance Classification: A Survey," *arXiv:1907.00181 [cs]*, Jun. 2019, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/1907.00181

[22] W. Chafe, "Evidentiality in English conversation and academic writing," *Evidentiality: The linguistic coding of epistemology*, vol. 20, pp. 261–272, 1986.

[23] C. Wardle and H. Derakhshan, "INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making," 2017.

[24] M. Viviani and G. Pasi, "Credibility in social media: opinions, news, and health information-a survey: Credibility in social media," *WIREs Data Mining Knowl Discov*, vol. 7, no. 5, p. e1209, Sep. 2017, doi: 10.1002/widm.1209.

[25] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, Jun. 2015, doi: 10.1126/science.aaa1160.

[26] A. Ghenai and Y. Mejova, "Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter," *arXiv:1707.03778 [cs]*, Jul. 2017, Accessed: Sep. 11, 2020. [Online]. Available: http://arxiv.org/abs/1707.03778

[27] A. M. Jamison, D. A. Broniatowski, and S. C. Quinn, "Malicious Actors on Twitter: A Guide for Public Health Researchers," *Am J Public Health*, vol. 109, no. 5, pp. 688–692, May 2019, doi: 10.2105/AJPH.2019.304969.

[28] W. Y. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," *arXiv:1705.00648 [cs]*, May 2017, Accessed: Mar. 05, 2020. [Online]. Available: http://arxiv.org/abs/1705.00648

[29] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News," in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, California, 2016, pp. 7–17. doi: 10.18653/v1/W16-0802.

[30] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Systems with Applications*, vol. 116, pp. 209–226, Feb. 2019, doi: 10.1016/j.eswa.2018.09.009.

[31] A. Addawood, "Usage of Scientific References in MMR Vaccination Debates on Twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 971–979. doi: 10.1109/ASONAM.2018.8508385.

[32] R. M. Palau and M.-F. Moens, "Argumentation mining: the detection, classification and structure of arguments in text," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law - ICAIL '09*, Barcelona, Spain, 2009, p. 98. doi: 10.1145/1568234.1568246.

[33] R. Rinott, L. Dankin, C. Alzate Perez, M. M. Khapra, E. Aharoni, and N. Slonim, "Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Sep. 2015, pp. 440–450. doi: 10.18653/v1/D15-1050.

[34] G. Rowe and C. Reed, "Argument diagramming: The araucaria project," in *Knowledge cartography*, Springer, 2008, pp. 164–181.

[35] K. Ashley and V. Walker, "From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study," 2013. doi: 10.3233/978-1-61499-359-9-29.

[36] M. Lippi and P. Torroni, "Argumentation Mining: State of the Art and Emerging Trends," *ACM Trans. Internet Technol.*, vol. 16, no. 2, pp. 1–25, Apr. 2016, doi: 10.1145/2850417.

[37] A. Gruzd and P. Mai, "Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter," *Big Data & Society*, vol. 7, no. 2, p. 2053951720938405, Jul. 2020, doi: 10.1177/2053951720938405.

[38] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web - WWW '11*, Hyderabad, India, 2011, p. 675. doi: 10.1145/1963405.1963500.

[39] J. M. Leung and D. D. Sin, "Smoking, ACE-2 and COVID-19: ongoing controversies," *Eur Respir J*, vol. 56, no. 1, p. 2001759, Jul. 2020, doi: 10.1183/13993003.01759-2020.

[40] A. Nguyen and D. Catalan-Matamoros, "Digital Mis/Disinformation and Public Engagment with Health and Science Controversies: Fresh Perspectives from Covid-19," *MaC*, vol. 8, no. 2, pp. 323–328, Jun. 2020, doi: 10.17645/mac.v8i2.3352.

[41] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and Sentiment in Tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, pp. 1–23, Jul. 2017, doi: 10.1145/3003433.

[42] K. Kucher, T. Schamp-Bjerede, A. Kerren, C. Paradis, and M. Sahlgren, "Visual analysis of online social media to open up the investigation of stance phenomena," *Information Visualization*, vol. 15, no. 2, pp. 93–116, Apr. 2016, doi: 10.1177/1473871615575079.

[43] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," *arXiv preprint arXiv:1812.00315*, 2018.

[44] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38–55, 2019.

[45] A. Guess, B. Nyhan, and J. Reifler, "Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign," p. 49.

[46] M. M. Bhuiyan, K. Zhang, K. Vick, M. A. Horning, and T. Mitra, "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*, 2018, pp. 205–208.

[47] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A Platform for Tracking Online Misinformation," in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, Montr&#233;al, Qu&#233;bec, Canada, 2016, pp. 745–750. doi: 10.1145/2872518.2890098.

[48] S. Finn, P. T. Metaxas, and E. Mustafaraj, "Investigating Rumor Propagation with TwitterTrails," *arXiv:1411.3550 [cs]*, Nov. 2014, Accessed: Mar. 31, 2020. [Online]. Available: http://arxiv.org/abs/1411.3550

[49] P. Resnick, S. Carton, S. Park, Y. Shen, and N. Zeffer, "RumorLens: A System for Analyzing the Impact of Rumors and Corrections in Social Media," p. 5.

[50] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-Time Credibility Assessment of Content on Twitter," in *Social Informatics*, vol. 8851, L. M. Aiello and D. McFarland, Eds. Cham: Springer International Publishing, 2014, pp. 228–243. doi: 10.1007/978-3-319-13734-6_16.

[51] A. Karduni, "Human-Misinformation interaction: Understanding the interdisciplinary approach needed to computationally combat false information," *arXiv:1903.07136 [cs]*, Mar. 2019, Accessed: Feb. 09, 2020. [Online]. Available: http://arxiv.org/abs/1903.07136

[52] A. Karduni *et al.*, "Can You Verifi This? Studying Uncertainty and Decision-Making About Misinformation Using Visual Analytics," p. 10.

[53] A. Karduni *et al.*, "Vulnerable to misinformation?: Verifi!," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray California, Mar. 2019, pp. 312–323. doi: 10.1145/3301275.3302320.

[54] V. Narwal *et al.*, "Automated Assistants to Identify and Prompt Action on Visual News Bias," *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*, pp. 2796–2801, 2017, doi: 10.1145/3027063.3053227.

[55] S. Vosoughi and D. Roy, "A Human-Machine Collaborative System for Identifying Rumors on Twitter," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Atlantic City, NJ, USA, Nov. 2015, pp. 47–50. doi: 10.1109/ICDMW.2015.221.

[56] M. Weinmann, C. Schneider, and J. vom Brocke, "Digital Nudging," *Bus Inf Syst Eng*, vol. 58, no. 6, pp. 433–436, Dec. 2016, doi: 10.1007/s12599-016-0453-1.

[57] V. G. Cerf, "Augmented Intelligence," *IEEE Internet Computing*, vol. 17, no. 5, pp. 96–96, Sep. 2013, doi: 10.1109/MIC.2013.90.

[58] T. Kroll and S. Stieglitz, “Digital nudging and privacy: improving decisions about self-disclosure in social networks,” *Behaviour & Information Technology*, pp. 1–19, 2019.

[59] Y. Wang, P. G. Leon, K. Scott, X. Chen, A. Acquisti, and L. F. Cranor, “Privacy nudges for social media: an exploratory Facebook study,” in *Proceedings of the 22nd international conference on world wide web*, 2013, pp. 763–770.

[60] Sofia University “St. Kliment Ohridski”, Bulgaria *et al.*, “Fully Automated Fact Checking Using External Sources,” in *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, Nov. 2017, pp. 344–353. doi: 10.26615/978-954-452-049-6_046.

[61] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[62] M. R. Pinto, Y. O. de Lima, C. E. Barbosa, and J. M. de Souza, “Towards Fact-Checking through Crowdsourcing,” in *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, May 2019, pp. 494–499. doi: 10.1109/CSCWD.2019.8791903.

[63] C. Castillo, M. Mendoza, and B. Poblete, “Predicting information credibility in time-sensitive social media,” *Internet Research*, vol. 23, no. 5, pp. 560–588, Oct. 2013, doi: 10.1108/IntR-05-2012-0095.

[64] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, “Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy,” in *Proceedings of the 22nd International Conference on World Wide Web - WWW ’13 Companion*, Rio de Janeiro, Brazil, 2013, pp. 729–736. doi: 10.1145/2487788.2488033.

[65] A. Gupta and P. Kumaraguru, “Credibility ranking of tweets during high impact events,” in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media - PSOSM ’12*, Lyon, France, 2012, pp. 2–8. doi: 10.1145/2185354.2185356.

[66] Z. Jin, J. Cao, Y. Zhang, and J. Luo, “News Verification by Exploiting Conflicting Social Viewpoints in Microblogs,” p. 7.

[67] F. Yang, Y. Liu, X. Yu, and M. Yang, “Automatic detection of rumor on Sina Weibo,” in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 2012, pp. 1–7.

[68] S. Sun, H. Liu, J. He, and X. Du, “Detecting event rumors on sina weibo automatically,” in *Asia-Pacific Web Conference*, 2013, pp. 120–131.

[69] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, “Prominent Features of Rumor Propagation in Online Social Media,” in *2013 IEEE 13th International Conference on Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 1103–1108. doi: 10.1109/ICDM.2013.61.

[70] Z. Zhao, P. Resnick, and Q. Mei, “Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts,” in *Proceedings of the 24th International Conference on World Wide Web - WWW ’15*, Florence, Italy, 2015, pp. 1395–1405. doi: 10.1145/2736277.2741637.

[71] G. Giasemidis *et al.*, “Determining the Veracity of Rumours on Twitter,” *arXiv:1611.06314 [cs, stat]*, vol. 10046, pp. 185–205, 2016, doi: 10.1007/978-3-319-47880-7_12.

[72] A. Addawood and M. Bashir, “‘What Is Your Evidence?’ A Study of Controversial Topics on Social Media,” in *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, Berlin, Germany, 2016, pp. 1–11. doi: 10.18653/v1/W16-2801.

[73] A. Addawood, J. Schneider, and M. Bashir, “Stance Classification of Twitter Debates: The Encryption Debate as A Use Case,” in *Proceedings of the 8th International Conference on Social Media & Society - #SMSociety17*, Toronto, ON, Canada, 2017, pp. 1–10. doi: 10.1145/3097286.3097288.

[74] B. Ghanem, P. Rosso, and F. Rangel, “An Emotional Analysis of False Information in Social Media and News Articles,” *arXiv:1908.09951 [cs]*, Aug. 2019, Accessed: Mar. 09, 2020. [Online]. Available: http://arxiv.org/abs/1908.09951

[75] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, “Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, 2017, pp. 647–653. doi: 10.18653/v1/P17-2102.

[76] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, “News Credibility Evaluation on Microblog with a Hierarchical Propagation Model,” in *2014 IEEE International Conference on Data Mining*, Dec. 2014, pp. 230–239. doi: 10.1109/ICDM.2014.91.

[77] T. Hua, C.-T. Lu, and I.-R. Chen, “A topic-focused trust model for Twitter,” *Computer Communications*, vol. 76, pp. 1–11, Feb. 2016, doi: 10.1016/j.comcom.2015.08.001.

[78] R. Oshikawa, J. Qian, and W. Y. Wang, “A Survey on Natural Language Processing for Fake News Detection,” *arXiv:1811.00770 [cs]*, Nov. 2018, Accessed: Feb. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.00770

[79] A. Acerbi, “Cognitive attraction and online misinformation,” *Palgrave Commun*, vol. 5, no. 1, p. 15, Dec. 2019, doi: 10.1057/s41599-019-0224-y.

[80] T. Lucassen and J. M. Schraagen, “The influence of source cues and topic familiarity on credibility evaluation,” *Computers in Human Behavior*, vol. 29, no. 4, pp. 1387–1392, Jul. 2013, doi: 10.1016/j.chb.2013.01.036.

[81] M. A. Stefanone, M. Vollmer, and J. M. Covert, “In News We Trust?: Examining Credibility and Sharing Behaviors of Fake News,” in *Proceedings of the 10th International Conference on Social Media and Society - SMSociety ’19*, Toronto, ON, Canada, 2019, pp. 136–147. doi: 10.1145/3328529.3328554.

[82] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, “Tweeting is believing?: understanding microblog credibility perceptions,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW ’12*, Seattle, Washington, USA, 2012, p. 441. doi: 10.1145/2145204.2145274.

[83] S. M. Shariff, X. Zhang, and M. Sanderson, “On the credibility perception of news on Twitter: Readers, topics and features,” *Computers in Human Behavior*, vol. 75, pp. 785–796, Oct. 2017, doi: 10.1016/j.chb.2017.06.026.

[84] T. Vaidya, D. Votipka, M. L. Mazurek, and M. Sherr, “Does Being Verified Make You More Credible?: Account Verification’s Effect on Tweet Credibility,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ’19*, Glasgow, Scotland Uk, 2019, pp. 1–13. doi: 10.1145/3290605.3300755.

[85] R. Li and A. Suh, “Factors Influencing Information credibility on Social Media Platforms: Evidence from Facebook Pages,” *Procedia Computer Science*, vol. 72, pp. 314–328, 2015, doi: 10.1016/j.procs.2015.12.146.

[86] S. Sundar, “The MAIN Model : A Heuristic Approach to Understanding Technology Effects on Credibility,” 2007. https://www.semanticscholar.org/paper/The-MAIN-Model-%3A-A-Heuristic-Approach-to-Technology-Sundar/de80aa094f380342a632eadb0ee8d4221e8920ba (accessed May 07, 2020).

[87] J. Simko, M. Hanakova, P. Racsko, M. Tomlein, R. Moro, and M. Bielikova, “Fake news reading on social media: an eye-tracking study,” in *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 2019, pp. 221–230.

[88] A. Pal, A. Y. K. Chua, and D. Hoe-Lian Goh, “Debunking rumors on social media: The use of denials,” *Computers in Human Behavior*, vol. 96, pp. 110–122, Jul. 2019, doi: 10.1016/j.chb.2019.02.022.

[89] K. Shu, S. Wang, and H. Liu, “Understanding User Profiles on Social Media for Fake News Detection,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Miami, FL, Apr. 2018, pp. 430–435. doi: 10.1109/MIPR.2018.00092.

[90] J. B. Long and J. M. Ehrenfeld, “The Role of Augmented Intelligence (AI) in Detecting and Preventing the Spread of Novel Coronavirus,” *J Med Syst*, vol. 44, no. 3, pp. 59, s10916-020-1536–6, Mar. 2020, doi: 10.1007/s10916-020-1536-6.

[91] T. Mirsch, C. Lehrer, and R. Jung, “Digital Nudging: Altering User Behavior in Digital Environments,” p. 15.

[92] A. Acquisti *et al.*, “Nudges for Privacy and Security: Understanding and Assisting Users’ Choices Online,” *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–41, Oct. 2017, doi: 10.1145/3054926.

[93] P. L. Moravec, A. Kim, and A. R. Dennis, “Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media,” *Information Systems Research*, vol. 31, no. 3, pp. 987–1006, Sep. 2020, doi: 10.1287/isre.2020.0927.

[94] O. Ajao, D. Bhowmik, and S. Zargari, “Sentiment Aware Fake News Detection on Online Social Networks,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 2507–2511. doi: 10.1109/ICASSP.2019.8683170.

[95] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, “Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention,” PsyArXiv, preprint, Mar. 2020. doi: 10.31234/osf.io/uhbk9.

[96] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, “Systematic Literature Review on the Spread of Health-related Misinformation on Social Media,” *Social Science & Medicine*, vol. 240, p. 112552, Nov. 2019, doi: 10.1016/j.socscimed.2019.112552.

[97] H. S. Wald, C. E. Dube, and D. C. Anthony, “Untangling the Web—The impact of Internet use on health care and the physician–patient relationship,” *Patient Education and Counseling*, vol. 68, no. 3, pp. 218–224, Nov. 2007, doi: 10.1016/j.pec.2007.05.016.

[98] A. Ghenai and Y. Mejova, “Fake Cures: User-centric Modeling of Health Misinformation in Social Media,” *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–20, Nov. 2018, doi: 10.1145/3274327.

[99] A. Stechemesser, L. Wenz, and A. Levermann, “Corona crisis fuels racially profiled hate in social media networks,” *EClinicalMedicine*, p. 100372, May 2020, doi: 10.1016/j.eclinm.2020.100372.

[100] K. M. Douglas *et al.*, “Understanding conspiracy theories,” *Political Psychology*, vol. 40, pp. 3–35, 2019.

[101] C. R. Sunstein and A. Vermeule, “Conspiracy theories: Causes and cures,” *Journal of Political Philosophy*, vol. 17, no. 2, pp. 202–227, 2009.

[102] J. E. Uscinski *et al.*, “Why do people believe COVID-19 conspiracy theories?,” *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, Apr. 2020, doi: 10.37016/mr-2020-015.

[103] A. Addawood, “UNDERSTANDING MISINFORMATION ON TWITTER IN THE CONTEXT OF CONTROVERSIAL ISSUES,” p. 184.

[104] M. J. Paul and M. Dredze, “Discovering Health Topics in Social Media Using Topic Models,” *PLoS ONE*, vol. 9, no. 8, p. e103408, Aug. 2014, doi: 10.1371/journal.pone.0103408.

[105] M. L. Antheunis, K. Tates, and T. E. Nieboer, “Patients’ and health professionals’ use of social media in health care: Motives, barriers and expectations,” *Patient Education and Counseling*, vol. 92, no. 3, pp. 426–431, Sep. 2013, doi: 10.1016/j.pec.2013.06.020.

[106] S. Priya, R. Sequeira, J. Chandra, and S. K. Dandapat, “Where should one get news updates: Twitter or Reddit,” *Online Social Networks and Media*, vol. 9, pp. 17–29, Jan. 2019, doi: 10.1016/j.osnem.2018.11.001.

[107] D. Allington, B. Duffy, S. Wessely, N. Dhavan, and J. Rubin, “Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency,” *Psychol. Med.*, pp. 1–7, Jun. 2020, doi: 10.1017/S003329172000224X.

[108] J. S. Brennen, F. M. Simon, P. N. Howard, and R. K. Nielsen, “Types, Sources, and Claims of COVID-19 Misinformation,” p. 13.

[109] E. Ferrara, “What types of COVID-19 conspiracies are populated by Twitter bots?,” *FM*, May 2020, doi: 10.5210/fm.v25i6.10633.

[110] H. Lyu, L. Chen, Y. Wang, and J. Luo, “Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19,” *IEEE Trans. Big Data*, pp. 1–1, 2020, doi: 10.1109/TBDATA.2020.2996401.

[111] W. Ahmed, F. López Seguí, J. Vidal-Alaball, and M. S. Katz, “COVID-19 and the ‘Film Your Hospital’ Conspiracy Theory: Social Network Analysis of Twitter Data,” *J Med Internet Res*, vol. 22, no. 10, p. e22374, Oct. 2020, doi: 10.2196/22374.

[112] M. Motta, D. Stecula, and C. Farhart, “How Right-Leaning Media Coverage of COVID-19 Facilitated the Spread of Misinformation in the Early Stages of the Pandemic in the U.S.,” *Can J Pol Sci*, pp. 1–8, May 2020, doi: 10.1017/S0008423920000396.

[113] J. Uyheng and K. M. Carley, “Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines,” *J Comput Soc Sc*, Oct. 2020, doi: 10.1007/s42001-020-00087-4.

[114] M. J. Wood, “Propagating and Debunking Conspiracy Theories on Twitter During the 2015–2016 Zika Virus Outbreak,” *Cyberpsychology, Behavior, and Social Networking*, vol. 21, no. 8, pp. 485–490, Aug. 2018, doi: 10.1089/cyber.2017.0669.

[115] L. Atlani-Duault, A. Mercier, C. Rousseau, P. Guyot, and J.-P. Moatti, “Blood Libel Rebooted: Traditional Scapegoats, Online Media, and the H1N1 Epidemic,” *Culture Medicine and Psychiatry*, vol. Vol 39, pp. 43–61, Mar. 2015, doi: 10.1007/s11013-014-9410-y.

[116] R. Kouzy *et al.*, “Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter,” *Cureus*, vol. 12, no. 3, doi: 10.7759/cureus.7255.

[117] L. Chen *et al.*, “A Social Media Study on the Associations of Flavored Electronic Cigarettes With Health Symptoms: Observational Study,” *Journal of Medical Internet Research*, vol. 22, no. 6, p. e17496, 2020, doi: 10.2196/17496.

[118] Y. Wang, J. Luo, R. Niemi, Y. Li, and T. Hu, “Catching Fire via ‘Likes’: Inferring Topic Preferences of Trump Followers on Twitter,” p. 4.

[119] X. Che, D. Metaxa-Kakavouli, and J. T. Hancock, “Fake News in the News: An Analysis of Partisan Coverage of the Fake News Phenomenon,” in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2018, pp. 289–292.

[120] D. Plotkina, A. Munzel, and J. Pallud, “Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews,” *Journal of Business Research*, vol. 109, pp. 511–523, Mar. 2020, doi: 10.1016/j.jbusres.2018.12.009.

[121] L. Singh *et al.*, “A first look at COVID-19 information and misinformation sharing on Twitter,” *arXiv preprint arXiv:2003.13907*, 2020.

[122] M. Zimdars, “False, misleading, clickbait-y, and satirical ‘news’ sources,” *Google Docs*, 2016.

[123] D. M. Blei, “Latent Dirichlet Allocation,” p. 30.

[124] “LIWC | Linguistic Inquiry and Word Count.” https://liwc.wpengine.com/ (accessed Dec. 09, 2020).

[125] D. F. Larcker and A. A. Zakolyukina, “Detecting deceptive discussions in conference calls,” *Journal of Accounting Research*, vol. 50, no. 2, pp. 495–540, 2012.

[126] R. Lamsal, “Design and analysis of a large-scale COVID-19 tweets dataset,” *Appl Intell*, Nov. 2020, doi: 10.1007/s10489-020-02029-z.

[127] “COVID: Top 10 current conspiracy theories,” *Alliance for Science*. https://allianceforscience.cornell.edu/blog/2020/04/covid-top-10-current-conspiracy-theories/ (accessed Nov. 25, 2020).

[128] J. M. Miller, “Do COVID-19 Conspiracy Theory Beliefs Form a Monological Belief System?,” *Can J Pol Sci*, vol. 53, no. 2, pp. 319–326, Jun. 2020, doi: 10.1017/S0008423920000517.

[129] E. Milani, E. Weitkamp, and P. Webb, “The Visual Vaccine Debate on Twitter: A Social Network Analysis,” *MaC*, vol. 8, no. 2, pp. 364–375, Jun. 2020, doi: 10.17645/mac.v8i2.2847.

[130] K. Sharma, S. Seo, C. Meng, S. Rambhatla, and Y. Liu, “COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations,” *arXiv:2003.12309 [cs]*, Oct. 2020, Accessed: Nov. 28, 2020. [Online]. Available: http://arxiv.org/abs/2003.12309

[131] “5G-Coronavirus-Report.pdf.” http://center-profilaktika.ru/wp-content/uploads/2020/05/5G-Coronavirus-Report.pdf (accessed Nov. 25, 2020).

[132] “List of fake news websites,” *Wikipedia*. Nov. 29, 2020. Accessed: Dec. 07, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=List_of_fake_news_websites&oldid=991405432

[133] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nat Commun*, vol. 9, no. 1, p. 4787, Dec. 2018, doi: 10.1038/s41467-018-06930-7.

[134] D. A. Broniatowski *et al.*, “The COVID-19 Social Media Infodemic Reflects Uncertainty and State-Sponsored Propaganda,” *arXiv:2007.09682 [physics]*, Jul. 2020, Accessed: Dec. 06, 2020. [Online]. Available: http://arxiv.org/abs/2007.09682

[135] “Media Bias/Fact Check - Search and Learn the Bias of News Media.” https://mediabiasfactcheck.com/ (accessed Dec. 07, 2020).

[136] "COVID-19 Vaccine Misinformation Super-spreaders," *NewsGuard*. https://www.newsguardtech.com/special-report-covid-19-vaccine-misinformation/ (accessed Dec. 06, 2020).

[137] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016, doi: 10.1145/2818717.

[138] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and Generalizable Social Bot Detection through Data Selection," *AAAI*, vol. 34, no. 01, Art. no. 01, Apr. 2020, doi: 10.1609/aaai.v34i01.5460.

[139] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis," p. 11.

[140] A. Chakraborty, S. Ghosh, N. Ganguly, and K. P. Gummadi, "Dissemination Biases of Social Media Channels: On The Topical Coverage of Socially Shared News," *New York*, p. 4.

[141] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 171–175.

[142] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, Maryland, USA, 2014, pp. 63–70. doi: 10.3115/v1/W14-3110.

[143] "Detecting bursts in sentiment-aware topics from social media | Elsevier Enhanced Reader." https://reader.elsevier.com/reader/sd/pii/S0950705117305282?token=31559866ADB2AC48431299F36E335DE6C71CC0D902B59CBB9F3B482DFB4A3F3D3B095B7A6A3E96D44BB045272237C4B0&originRegion=us-east-1&originCreation=20210423170449 (accessed Apr. 23, 2021).

[144] H.-J. Kim, Y. K. Jeong, Y. Kim, K. Kang, and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," *Journal of Information Science*, vol. 42, Oct. 2015, doi: 10.1177/0165551515608733.

[145] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," 2015.

[146] A. Gruzd and M. Goertzen, "Wired Academia: Why Social Science Scholars Are Using Social Media," in *2013 46th Hawaii International Conference on System Sciences*, Jan. 2013, pp. 3332–3341. doi: 10.1109/HICSS.2013.614.

[147] T. Bosc, E. Cabrio, and S. Villata, "Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media.," *COMMA*, vol. 2016, pp. 21–32, 2016.

[148] J. Gough, "The Differences between Opinion and Argumentation," p. 12.

[149] A. J. Freeley and D. L. Steinberg, *Argumentation and debate*. Cengage Learning, 2013.

[150] E. W. Pamungkas, V. Basile, and V. Patti, "Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure," *arXiv:1901.01911 [cs]*, Jan. 2019, Accessed: Jul. 24, 2020. [Online]. Available: http://arxiv.org/abs/1901.01911

[151] N. Green, K. Ashley, D. Litman, C. Reed, and V. Walker, Eds., *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, Maryland: Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-21.

[152] O. of the Commissioner, "Coronavirus Update: FDA and FTC Warn Seven Companies Selling Fraudulent Products that Claim to Treat or Prevent COVID-19," *FDA*, Mar. 27, 2020. https://www.fda.gov/news-events/press-announcements/coronavirus-update-fda-and-ftc-warn-seven-companies-selling-fraudulent-products-claim-treat-or (accessed Aug. 02, 2020).

[153] V. A. Ketcham, *The theory and practice of argumentation and debate*. Macmillan, 1914.

[154] E. J. MacEwan, *The essentials of argumentation*. DC Heath & Company, 1898.

[155] C. Stab and I. Gurevych, "Annotating Argument Components and Relations in Persuasive Essays," p. 10.

[156] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 Task 6: Detecting Stance in Tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, Jun. 2016, pp. 31–41. doi: 10.18653/v1/S16-1003.

[157] A. Lytos, T. Lagkas, P. Sarigiannidis, and K. Bontcheva, “The evolution of argumentation mining: From models to social media and emerging tools,” *Information Processing & Management*, vol. 56, no. 6, p. 102055, Nov. 2019, doi: 10.1016/j.ipm.2019.102055.

[158] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho, “Detecting discussion communities on vaccination in twitter,” *Future Generation Computer Systems*, vol. 66, pp. 125–136, Jan. 2017, doi: 10.1016/j.future.2016.06.032.

[159] R. Boodoosingh, L. O. Olayemi, and F. A.-L. Sam, “COVID-19 vaccines: Getting Anti-vaxxers involved in the discussion,” *World Dev*, vol. 136, p. 105177, Dec. 2020, doi: 10.1016/j.worlddev.2020.105177.

[160] C. M. Pulido, B. Villarejo-Carballido, G. Redondo-Sama, and A. Gómez, “COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information,” *International Sociology*, vol. 35, no. 4, pp. 377–392, Jul. 2020, doi: 10.1177/0268580920914755.

[161] R. Villa-Cox, S. Kumar, M. Babcock, and K. M. Carley, “Stance in Replies and Quotes (SRQ): A New Dataset For Learning Stance in Twitter Conversations,” *arXiv:2006.00691 [cs]*, Jun. 2020, Accessed: Jul. 24, 2020. [Online]. Available: http://arxiv.org/abs/2006.00691

[162] M. Paz, G. Villalba, and P. Saint-dizier, *Some Facets of Argument Mining for Opinion Analysis*.

[163] S. Y. Rieh, “Credibility and cognitive authority of information,” 2010.

[164] E. Aharoni *et al.*, “A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics,” in *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland, Jun. 2014, pp. 64–68. doi: 10.3115/v1/W14-2109.

[165] “3.1 What is Argument?,” *Humanities Libertexts*, Jun. 30, 2019. https://human.libretexts.org/Bookshelves/Composition/Introductory_Composition/Book%3A_Let's_Get_Writing_(Browning%2C_DeVries%2C_Boylan%2C_Kurtz_and_Burton)/3%3A_Argument/3.1_What_is_Argument (accessed Aug. 12, 2020).

[166] “Argument Versus Opinion.” https://www.bradford.ac.uk/t4-ssis/ruski-files/academic-voice/page_07.htm (accessed Aug. 11, 2020).

[167] J. Schneider, K. Samp, A. Passant, and S. Decker, “Arguments about deletion: how experience improves the acceptability of arguments in ad-hoc online task groups,” in *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW ’13*, San Antonio, Texas, USA, 2013, p. 1069. doi: 10.1145/2441776.2441897.

[168] D. N. Walton, *Dialog theory for critical argumentation*. John Benjamins Pub., 2007.

[169] I. Habernal, J. Eckle-Kohler, and I. Gurevych, “Argumentation Mining on the Web from Information Seeking Perspective,” p. 14.

[170] O. Biran and O. Rambow, “IDENTIFYING JUSTIFICATIONS IN WRITTEN DIALOGS BY CLASSIFYING TEXT AS ARGUMENTATIVE,” *Int. J. Semantic Computing*, vol. 05, no. 04, pp. 363–381, Dec. 2011, doi: 10.1142/S1793351X11001328.

[171] J. Park and C. Cardie, “Identifying Appropriate Support for Propositions in Online User Comments,” in *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland, Jun. 2014, pp. 29–38. doi: 10.3115/v1/W14-2105.

[172] K. A. Alnemer *et al.*, “Are Health-Related Tweets Evidence Based? Review and Analysis of Health-Related Tweets on Twitter,” *JOURNAL OF MEDICAL INTERNET RESEARCH*, p. 6.

[173] M. Cinelli *et al.*, “The COVID-19 Social Media Infodemic,” *arXiv:2003.05004 [nlin, physics:physics]*, Mar. 2020, Accessed: Jun. 18, 2020. [Online]. Available: http://arxiv.org/abs/2003.05004

[174] “The Coding Manual for Qualitative Researchers - Google Scholar.” https://scholar.google.com/scholar?hl=en&as_sdt=0%2C34&q=The+Coding+Manual+for+Qualitative+Researchers&btnG= (accessed Apr. 14, 2021).

[175] C. Nath, J. Huh, A. K. Adupa, and S. R. Jonnalagadda, “Website Sharing in Online Health Communities: A Descriptive Analysis,” *J Med Internet Res*, vol. 18, no. 1, p. e11, Jan. 2016, doi: 10.2196/jmir.5237.

[176] J. F. George, A. Mirsadikov, and B. E. Mennecke, “Website credibility assessment: an empirical-investigation of prominence-interpretation theory,” *AIS Transactions on Human-Computer Interaction*, vol. 8, no. 2, pp. 40–57, 2016.

[177] H. Slimi, I. Bounhas, and Y. Slimani, “URL-Based Tweet Credibility Evaluation,” in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Abu Dhabi, United Arab Emirates, Nov. 2019, pp. 1–6. doi: 10.1109/AICCSA47632.2019.9035255.

[178] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[179] M. Dias and K. Becker, “Inf-ufrgs-opinion-mining at semeval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 378–383.

[180] H. Elfardy and M. Diab, “Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 434–439.

[181] A. Misra, B. Ecker, T. Handleman, N. Hahn, and M. Walker, “A semi-supervised approach to detecting stance in Tweets,” *arXiv preprint arXiv:1709.01895*, 2017.

[182] L. Liu, S. Feng, D. Wang, and Y. Zhang, “An empirical study on Chinese microblog stance detection using supervised and semi-supervised machine learning methods,” in *Natural Language Understanding and Intelligent Applications*, Springer, 2016, pp. 753–765.

[183] J. Xu, S. Zheng, J. Shi, Y. Yao, and B. Xu, “Ensemble of feature sets and classification methods for stance detection,” in *Natural Language Understanding and Intelligent Applications*, Springer, 2016, pp. 679–688.

[184] C. Janze and M. Risius, “Automatic detection of fake news on social media platforms,” 2017.

[185] D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo, “CLPsych 2016 Shared Task: Triaging content in online peer-support forums,” in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, San Diego, CA, USA, 2016, pp. 118–127. doi: 10.18653/v1/W16-0312.

[186] S. Somasundaran and J. Wiebe, “Recognizing stances in ideological on-line debates,” in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 2010, pp. 116–124.

[187] Q. Sun, Z. Wang, Q. Zhu, and G. Zhou, “Exploring Various Linguistic Features for Stance Detection,” in *Natural Language Understanding and Intelligent Applications*, Cham, 2016, pp. 840–847. doi: 10.1007/978-3-319-50496-4_76.

[188] A. Badawy, E. Ferrara, and K. Lerman, “Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign,” *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258–265, Aug. 2018, doi: 10.1109/ASONAM.2018.8508646.

[189] J. K. Burgoon, J. P. Blair, and R. E. Strom, “Cognitive Biases and Nonverbal Cue Availability in Detecting Deception,” *Human Communication Research*, vol. 34, no. 4, pp. 572–599, Oct. 2008, doi: 10.1111/j.1468-2958.2008.00333.x.

[190] J. S. Blumenthal-Barby and H. Krieger, “Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy,” *Medical Decision Making*, vol. 35, no. 4, pp. 539–557, 2015.

[191] L. Cen, G. Hilary, K. C. J. Wei, and J. Zhang, “The Role of Anchoring Bias in the Equity Market,” *SSRN Journal*, 2010, doi: 10.2139/ssrn.1572855.

[192] C. R. Mynatt, M. E. Doherty, and R. D. Tweney, “Confirmation bias in a simulated research environment: An experimental study of scientific inference,” *Quarterly Journal of Experimental Psychology*, vol. 29, no. 1, pp. 85–95, 1977.

[193] C. R. Sunstein, “Nudging: a very short guide,” *Journal of Consumer Policy*, vol. 37, no. 4, pp. 583–588, 2014.

[194] R. H. Thaler and C. R. Sunstein, “Libertarian paternalism,” *American economic review*, vol. 93, no. 2, pp. 175–179, 2003.

[195] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, “Science vs Conspiracy: Collective Narratives in the Age of Misinformation,” *PLOS ONE*, vol. 10, no. 2, p. e0118093, Feb. 2015, doi: 10.1371/journal.pone.0118093.

[196] “Twitter link shortener (t.co) and how it works | Twitter Help.” https://help.twitter.com/en/using-twitter/url-shortener (accessed Apr. 25, 2021).

[197] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, “TweetCred: Real-Time Credibility Assessment of Content on Twitter,” *arXiv:1405.5490 [physics]*, Jan. 2015, Accessed: Jul. 03, 2020. [Online]. Available: http://arxiv.org/abs/1405.5490

[198] E. Nekmat, “Nudge Effect of Fact-Check Alerts: Source Influence and Media Skepticism on Sharing of News Misinformation in Social Media,” *Social Media + Society*, vol. 6, no. 1, p. 2056305119897322, Jan. 2020, doi: 10.1177/2056305119897322.

[199] “Flagging fake news on social media: An experimental study of media consumers’ identification of fake news | Elsevier Enhanced Reader.” https://reader.elsevier.com/reader/sd/pii/S0740624X21000277?token=72246664C3960E03C997CFA389E48EFB0D309C38D56B011724E42352668FA7AC1EA5FF2559F870DD39C8ADEDBF97DE6F&originRegion=us-east-1&originCreation=20210711203959 (accessed Jul. 11, 2021).

[200] S. W. Schuetz, T. A. Sykes, and V. Venkatesh, “Combating COVID-19 fake news on social media through fact checking: antecedents and consequences,” *European Journal of Information Systems*, vol. 30, no. 4, pp. 376–388, Jul. 2021, doi: 10.1080/0960085X.2021.1895682.

[201] L. Graves, “Understanding the Promise and Limits of Automated Fact-Checking,” p. 8.

[202] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.

[203] F. J. Zuiderveen Borgesius, D. Trilling, J. Möller, B. Bodó, C. H. de Vreese, and N. Helberger, “Should we worry about filter bubbles?,” *Internet Policy Review*, vol. 5, no. 1, Mar. 2016, doi: 10.14763/2016.1.401.

[204] G. Pennycook and D. G. Rand, “Fighting misinformation on social media using crowdsourced judgments of news source quality,” *Proc Natl Acad Sci USA*, vol. 116, no. 7, pp. 2521–2526, Feb. 2019, doi: 10.1073/pnas.1806781116.

[205] I. Dylko, I. Dolgov, W. Hoffman, N. Eckhart, M. Molina, and O. Aaziz, “The dark side of technology: An experimental investigation of the influence of customizability technology on online political selective exposure,” *Computers in Human Behavior*, vol. 73, pp. 181–190, Aug. 2017, doi: 10.1016/j.chb.2017.03.031.

[206] B. Ulicny and K. Baclawski, “New Metrics for Newsblog Credibility,” p. 2.

[207] B. Suh, L. Hong, P. Pirolli, and H. Chi, “Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network,” in *in Proceedings of the IEEE Second International Conference on Social Computing (SocialCom*, 2010, pp. 177–184.

[208] K. J. Stewart and Y. Zhang, “Effects of hypertext links on trust transfer,” in *Proceedings of the 5th international conference on Electronic commerce*, 2003, pp. 235–239.

[209] S. Kinsella, M. Wang, J. G. Breslin, and C. Hayes, “Improving categorisation in social media using hyperlinks to structured data sources,” in *Extended Semantic Web Conference*, 2011, pp. 390–404.

[210] G. Eysenbach and C. Köhler, “How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews,” *BMJ*, vol. 324, no. 7337, pp. 573–577, Mar. 2002.

[211] S. Park *et al.*, “The Source and Credibility of Colorectal Cancer Information on Twitter:,” *Medicine*, vol. 95, no. 7, p. e2775, Feb. 2016, doi: 10.1097/MD.0000000000002775.

[212] H. Song *et al.*, “Trusting Social Media as a Source of Health Information: Online Surveys Comparing the United States, Korea, and Hong Kong,” *J Med Internet Res*, vol. 18, no. 3, p. e25, Mar. 2016, doi: 10.2196/jmir.4193.

[213] T. Toivonen, I. Jormanainen, and M. Tukiainen, “Augmented intelligence in educational data mining,” *Smart Learning Environments*, vol. 6, no. 1, pp. 1–25, 2019.

[214] W. B. Rouse and J. C. Spohrer, “Automating versus augmenting intelligence,” *Journal of Enterprise Transformation*, vol. 8, no. 1–2, pp. 1–21, Apr. 2018, doi: 10.1080/19488289.2018.1424059.

[215] D. Bollier, *Artificial Intelligence Comes of Age: The Promise and Challenge of Integrating AI into Cars, Healthcare and Journalism*. Aspen Institute, 2017.

[216] D. Galeon and K. Houser, *IBM's watson AI recommends same treatment as doctors in 99% of cancer cases*. Futurism, 2016.

[217] E. J. Johnson and D. Goldstein, *Do defaults save lives?* American Association for the Advancement of Science, 2003.

[218] A. S. Gerber, D. P. Green, and C. W. Larimer, "Social pressure and voter turnout: Evidence from a large-scale field experiment," *American political Science review*, pp. 33–48, 2008.

[219] "Augmented Intelligence." http://lcfi.ac.uk/projects/kinds-of-intelligence/augmented-intelligence/ (accessed Mar. 09, 2021).

[220] L. Ling and C. W. Tan, "Human-Assisted Computation for Auto-Grading," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Singapore, Singapore, Nov. 2018, pp. 360–364. doi: 10.1109/ICDMW.2018.00059.

[221] D. Wightman, "Crowdsourcing human-based computation," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10*, Reykjavik, Iceland, 2010, p. 551. doi: 10.1145/1868914.1868976.

[222] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, Vancouver, BC, Canada, 2011, p. 1403. doi: 10.1145/1978942.1979148.

[223] E. J. Johnson, R. Hassin, T. Baker, A. T. Bajger, and G. Treuer, "Can consumers make affordable care affordable? The value of choice architecture," *PloS one*, vol. 8, no. 12, p. e81521, 2013.

[224] M. Thaler, "The 'Fake News' Effect: An Experiment on Motivated Reasoning and Trust in News," p. 111.

[225] C. Thornhill, Q. Meeus, J. Peperkamp, and B. Berendt, "A Digital Nudge to Counter Confirmation Bias," *Front. Big Data*, vol. 2, 2019, doi: 10.3389/fdata.2019.00011.

[226] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news," *J Exp Psychol Gen*, vol. 147, no. 12, pp. 1865–1880, Dec. 2018, doi: 10.1037/xge0000465.

[227] M. M. Bhuiyan, M. Horning, S. W. Lee, and T. Mitra, "NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges," *arXiv:2108.01536 [cs]*, Aug. 2021, doi: 10.1145/3479571.

[228] P. G. Hansen and A. M. Jespersen, "Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy," *European Journal of Risk Regulation*, vol. 4, no. 1, pp. 3–28, 2013.

[229] J. S. B. Evans, "In two minds: dual-process accounts of reasoning," *Trends in cognitive sciences*, vol. 7, no. 10, pp. 454–459, 2003.

[230] R. H. Thaler and C. R. Sunstein, "Nudge: Improving decisions about health, wealth, and happiness." HeinOnline.

[231] A. T. Adams, J. Costa, M. F. Jung, and T. Choudhury, "Mindless computing: designing technologies to subtly influence behavior," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 719–730.

[232] J. K. Phillips, G. Klein, and W. R. Sieck, "Expertise in judgment and decision making: A case for training intuitive decision skills," *Blackwell handbook of judgment and decision making*, vol. 297, p. 315, 2004.

[233] P. M. Todd and G. Gigerenzer, "Précis of Simple heuristics that make us smart," *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 727–741, Oct. 2000, doi: 10.1017/S0140525X00003447.

[234] W. L. Oroh, "THE INFLUENCE OF CONSUMERS'TIE STRENGTH, HOMOPHILY AND SOURCE CREDIBILITY TOWARD ELECTRONIC WORD-OF-MOUTH (EWOM) BEHAVIOR," *Jurnal EMBA: Jurnal Riset Ekonomi, Manajemen, Bisnis dan Akuntansi*, vol. 2, no. 3, 2014.

[235] S. Sinaga and C. Callison, "Credibility of PR practitioners: The impact of professional journalism background on trustworthiness, expertness, and homophily evaluations," *Public Relations Review*, vol. 34, no. 3, pp. 291–293, Sep. 2008, doi: 10.1016/j.pubrev.2008.04.002.

[236] Y. Sun, Y. Zhang, J. Gwizdka, and C. B. Trace, “Consumer Evaluation of the Quality of Online Health Information: Systematic Literature Review of Relevant Criteria and Indicators,” *J Med Internet Res*, vol. 21, no. 5, p. e12522, May 2019, doi: 10.2196/12522.

[237] I. Kareklas, D. D. Muehling, and T. J. Weber, “Reexamining Health Messages in the Digital Age: A Fresh Look at Source Credibility Effects,” *Journal of Advertising*, vol. 44, no. 2, pp. 88–104, Apr. 2015, doi: 10.1080/00913367.2015.1018461.

[238] Z. Wang, J. B. Walther, S. Pingree, and R. P. Hawkins, “Health Information, Credibility, Homophily, and Influence via the Internet: Web Sites Versus Discussion Groups,” *Health Communication*, vol. 23, no. 4, pp. 358–368, Aug. 2008, doi: 10.1080/10410230802229738.

[239] H. Gimpel, J. Kasper, S. Heger, and R. Schäfer, “The Power of Related Articles – Improving Fake News Detection on Social Media Platforms,” p. 10.

[240] R. E. Petty and J. T. Cacioppo, “The elaboration likelihood model of persuasion,” in *Communication and persuasion*, Springer, 1986, pp. 1–24.

[241] E. E. Griffith, C. J. Nolder, and R. E. Petty, “The Elaboration Likelihood Model: A Meta-Theory for Synthesizing Auditor Judgment and Decision-Making Research,” *AUDITING: A Journal of Practice & Theory*, vol. 37, no. 4, pp. 169–186, Jan. 2018, doi: 10.2308/ajpt-52018.

[242] “Routes to Persuasion, Central and Peripheral,” in *The SAGE Encyclopedia of Political Behavior*, 2455 Teller Road, Thousand Oaks, California 91320: SAGE Publications, Inc., 2017. doi: 10.4135/9781483391144.n330.

[243] R. E. Petty, J. T. Cacioppo, and D. Schumann, “Central and peripheral routes to advertising effectiveness: The moderating role of involvement,” *Journal of consumer research*, vol. 10, no. 2, pp. 135–146, 1983.

[244] “11.3 Attitudes & Persuasion – Introductory Psychology.” https://opentext.wsu.edu/psych105/chapter/11-4-attitudes-persuassion/ (accessed Dec. 03, 2021).

[245] M. J. Metzger, A. J. Flanagin, and R. B. Medders, “Social and Heuristic Approaches to Credibility Evaluation Online,” *Journal of Communication*, vol. 60, no. 3, pp. 413–439, Sep. 2010, doi: 10.1111/j.1460-2466.2010.01488.x.

[246] J. B. Gotlieb and D. Sarel, “Effects of price advertisements on perceived quality and purchase intentions,” *Journal of Business Research*, vol. 22, no. 3, pp. 195–210, 1991.

[247] C. I. Hovland, I. L. Janis, and H. H. Kelley, “Communication and persuasion.,” 1953.

[248] K. Clayton *et al.*, “Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media,” *Political Behavior*, vol. 42, Dec. 2020, doi: 10.1007/s11109-019-09533-0.

[249] P. Bordia, N. DiFonzo, R. Haines, and E. Chaseling, “Rumors Denials as Persuasive Messages: Effects of Personal Relevance, Source, and Message Characteristics1,” *Journal of Applied Social Psychology*, vol. 35, no. 6, pp. 1301–1331, 2005, doi: https://doi.org/10.1111/j.1559-1816.2005.tb02172.x.

[250] Y. Tanaka, Y. Sakamoto, and T. Matsuka, “Toward a social-technological system that inactivates false rumors through the critical thinking of crowds,” in *2013 46th Hawaii International conference on system sciences*, 2013, pp. 649–658.

[251] A. Y. K. Chua and S. Banerjee, “Intentions to trust and share online health rumors: An experiment with medical professionals,” *Computers in Human Behavior*, vol. 87, pp. 1–9, Oct. 2018, doi: 10.1016/j.chb.2018.05.021.

[252] A. Caraban, E. Karapanos, D. Gonçalves, and P. Campos, “23 ways to nudge: A review of technology-mediated nudging in human-computer interaction,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.

[253] C. G. Lord, M. R. Lepper, and E. Preston, “Considering the opposite: a corrective strategy for social judgment.,” *Journal of personality and social psychology*, vol. 47, no. 6, p. 1231, 1984.

[254] G. Pennycook, T. Cannon, and D. G. Rand, “Implausibility and illusory truth: Prior exposure increases perceived accuracy of fake news but has no effect on entirely implausible statements,” *Unpublished Paper Manuscript, December*, vol. 11, p. 2017, 2017.

[255] Q. V. Liao, W.-T. Fu, and S. S. Mamidi, “It is all about perspective: An exploration of mitigating selective exposure with aspect indicators,” in *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*, 2015, pp. 1439–1448.

[256] C. Hang, T. Ono, and S. Yamada, “Designing nudge agents that promote human altruism,” *arXiv preprint arXiv:2110.00319*, 2021.

[257] V. Bakir and A. McStay, “Fake News and The Economy of Emotions: Problems, causes, solutions,” *Digital Journalism*, vol. 6, no. 2, pp. 154–175, Feb. 2018, doi: 10.1080/21670811.2017.1345645.

[258] D. A. Scheufele and N. M. Krause, “Science audiences, misinformation, and fake news,” *Proc Natl Acad Sci USA*, vol. 116, no. 16, pp. 7662–7669, Apr. 2019, doi: 10.1073/pnas.1805871115.

[259] C. Martel, G. Pennycook, and D. G. Rand, “Reliance on emotion promotes belief in fake news,” *Cognitive Research: Principles and Implications*, vol. 5, no. 1, p. 47, Oct. 2020, doi: 10.1186/s41235-020-00252-3.

[260] R. Zaalberg and C. Midden, “Enhancing human responses to climate change risks through simulated flooding experiences,” in *International conference on persuasive technology*, 2010, pp. 205–210.

[261] A. Öhman and S. Mineka, “Fears, phobias, and preparedness: toward an evolved module of fear and fear learning.,” *Psychological review*, vol. 108, no. 3, p. 483, 2001.

[262] C. Salvi *et al.*, “Going Viral: How Fear, Socio-Cognitive Polarization and Problem-Solving Influence Fake News Detection and Proliferation During COVID-19 Pandemic,” *Frontiers in Communication*, vol. 5, p. 127, 2021, doi: 10.3389/fcomm.2020.562588.

[263] Z. R. Steelman, B. I. Hammer, and M. Limayem, “Data collection in the digital age,” *Mis Quarterly*, vol. 38, no. 2, pp. 355–378, 2014.

[264] M. Freeze *et al.*, “Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect,” *Polit Behav*, Feb. 2020, doi: 10.1007/s11109-020-09597-3.

[265] A. Kim, P. L. Moravec, and A. R. Dennis, “Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings,” *Journal of Management Information Systems*, vol. 36, no. 3, pp. 931–968, Jul. 2019, doi: 10.1080/07421222.2019.1628921.

[266] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, “Shifting attention to accuracy can reduce misinformation online,” PsyArXiv, preprint, Nov. 2019. doi: 10.31234/osf.io/3n9u8.

[267] B. Bago, D. G. Rand, and G. Pennycook, “Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines.,” *Journal of experimental psychology: general*, vol. 149, no. 8, p. 1608, 2020.

[268] A. Appelman and S. S. Sundar, “Measuring message credibility: Construction and validation of an exclusive scale,” *Journalism & Mass Communication Quarterly*, vol. 93, no. 1, pp. 59–79, 2016.

[269] M. D. Ward and J. S. Ahlquist, *Maximum likelihood for social science: Strategies for analysis*. Cambridge University Press, 2018.

[270] C. Tong, H. Gill, J. Li, S. Valenzuela, and H. Rojas, “‘Fake News is Anything They Say!’ – Conceptualization and Weaponization of Fake News Among the American Public,” *Mass Communication and Society*, vol. 23, Jul. 2020, doi: 10.1080/15205436.2020.1789661.

[271] A. Hussein, “Sequential Comparison of Two Treatments Using Weighted Wald-Type Statistics,” *Communications in Statistics - Theory and Methods*, vol. 34, no. 7, pp. 1631–1641, Jul. 2005, doi: 10.1081/STA-200063206.

[272] P. L. Moravec, A. Kim, and A. R. Dennis, “Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media,” p. 40.

[273] R. Mochales and M.-F. Moens, “Argumentation mining,” *Artificial Intelligence and Law*, vol. 19, no. 1, pp. 1–22, 2011.

[274] T. Buchanan, “Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation,” *PLoS ONE*, vol. 15, no. 10, p. e0239666, Oct. 2020, doi: 10.1371/journal.pone.0239666.

APPENDIX A: SEARCH KEYWORDS AND TERMS

Support Vaccination	**Against Vaccination**
#VaccinesSaveLives #VaccinesWork #WorldImmunizationWeek #VaxWithMe #HealthForAll #HealingStartsHere #WakeUpAmerica #WhyIVax #immunization #lasvacunasfuncionan #Vaccinate #vaccinateyourkids #vaccinatedandhappy #WorldImmunizationWeek #imtheherd #ThisIsOurShot #GetVaccinatedNow #vaxxed #ivax2protect #NationalImmunizationAwarenessMonth #iwillgetvaccinated #TrustFauci #ScienceMatters #KnowTheScience #provax #vaxwithm	#LearnTheRisk #VaccineInjury #VaccineDeath #VaccineDamage #VaccinesCauseAutism #CDCFraud #CDCWhistleBlower #CDCTruth #WakeUpAmerica #HearUs #HealthFreedom #sonotoneedles #wedonotconsent #novaccine #nocivilwar #Nonewnormal #wakeup #nocovidvaccine #nomandatoryvaccinations #justsaynotovaccines #makethisgoviral #idonotconsent #idonotcomply #wedonotconsent #novaccine #nocivilwar #nonewnormal #cd-cwhistleblower #hearthiswell #novax #cd-cfraud #HearThisWell #HHSlawsuit #vaccineskill # vaccinedamage #Billgatesvaccine #antivaxxer #b1less #vaccinescauseAIDS #vaccinescauseautism #justasking #mybodymychoice

APPENDIX B: CODEBOOK FOR STANCE-BASED COVID-19 TWEETS

- For each tweet, your task is to choose the labeling 1 or 0 of each of the questions/features (**columns F to L in** the dataset provided to you)of the tweets based on your interpretation of the **tweet content** column.

- You should record each of the selected features from the **dropdown menu** in the correspondent cell in the dataset provided to you.

- You have the option of suggesting any features in the comment last column (M).

❖ Topic Relevancy

- **Topic** means the topic nature of the tweets and their relation to the topic we consider in this research.
    - **This research focuses** on vaccination (for any disease).

- **Vaccination:** The tweet is vaccination related. **Note :** vaccination for all kinds of diseases
    - 1=**Yes 0=No**

- **Non-vaccination_OR_NO_content:** Not related to vaccination and mentions other topics (e.g., social issues, economic issues, politics, OR other health issues of COVID-19, etc..), OR there is no content or content is not comprehendible OR Only URLs posted without any content. **Note:** if yes, please record your answer and **move** to the next tweet.
    - 1=**Yes 0=No**

❖ Stance Presence and Evidence Source Consistency

- **Stance Presence:** checking **if there is a viewpoint** expressed on vaccination.

- **Evidence Source Consistency:** checking if the content of the evidence source in the tweet **is consistent with** the **viewpoint** expressed by the tweet.

- **Stance_with_Consistent_Evidence_Source:** The tweet is about COVID-19 and explicitly OR implicitly expresses a point of view on vaccination AND shares an evidence with an article/content that is consistent with the tweet's viewpoint toward vaccination.

  Note: Evdience source consistency with the tweet's viewpoint occurred if at least **ONE** of the following points(s) is discussed within the content in the evidence source:

  - Any type of content (news , personal opinions , reports, etc.) about vaccination administration ,
  - Any type of content (news , personal opinions , reports, etc.) about Vacciantin manufacturing ,
  - Any type of content (news , personal opinions , reports, etc.) about vaccine mandates,
  - Any type of content (news , personal opinions , reports, etc.) about conspiracies or involvement of public figures in relation to vaccination,
  - Any type of content (news , personal opinions , reports, etc.) about side effects, usage, precautions, symptoms about vaccination
    - 1=**Yes** **0=No**

- **Stance_Not_Consistent_Evidence_Source:** The tweet is about COVID-19 and explicitly OR implicitly expresses a point of view on vaccination AND shares a evdience that do not seem to be consistent with the tweet's viewpoint.

  - 1=**Yes** **0=No**

- **Stance_Evidence_Source_Not_Clear:** The tweet is about COVID-19 and explicitly OR implicitly expresses a viewpoint on vaccination, but the evidence's viewpoint is not clear**.**

- **Stance_No_Evidence_Source_Access:** The tweet is about COVID-19 and explicitly OR implicitly expresses a point of view on vaccination AND shares evidence source(s) that cannot be accessed.
  - 1=**Yes** **0=No**

- **No_Stance_or_Only_Questions:** The tweet is related to vaccination but just ask questions or no_viewpoint expressed. **Note:** if yes please record your answer and **move** to the next tweet.
  - 1=**Yes** **0=No**

❖ Evdience Source Entity

- **Evidence Source Entity:** if the author of the source is an individual or organization

- The evidence source belongs to an individual.
    - 1=**Yes 0=No**
- The evidence source belongs to an organization
    - 1=**Yes 0=No**
- The evidence source's entity is not clear.
    - 1=**Yes 0=No**

❖ Evidence Source Type

- **Evidence Source Type means** if the source(s) mentioned in the tweet belongs to one of the following evdience source types:
    - government agency, medical association, hospital, charity, news agency, magazine, educational/research institution commercial site, blog, scientific publisher/peer-review journal, an account of individual on social media platforms (Twitter, Facebook, YouTube, Reddit, etc.).

- **Government:** The evidence mentioned in the tweet is a government agency
    - 1=**Yes 0=No**
- **Medical Association :** The evidence mentioned in the tweet is a medical association
    - 1=**Yes 0=No**
- **Hospital**: The evidence mentioned in the tweet is a hospital
    - 1=**Yes 0=No**

- **Charity:** The evidence mentioned in the tweet is a charity
    - 1=**Yes 0=No**

- **News agency** : The evidence mentioned in the tweet is a news agency
    - 1=**Yes 0=No**
- **Magazine :** The evidence mentioned in the tweet is a magazine
    - 1=**Yes 0=No**
- **Education/research institution :** The evidence mentioned in the tweet is an educational/research institution

    - 1=**Yes 0=No**

- **Commercial site:** The evidence mentioned in the tweet is a commercial site
    - 1=**Yes 0=No**

- **Blog:** The evidence mentioned in the tweet is a blog
    - 1=**Yes 0=No**

- **Scientific publisher:** The evidence mentioned in the tweet is a scientific publisher/peer-review journal
    - 1=**Yes 0=No**

- **Other_evidence_source:**The evidence is **not** from the above categories of sources or it belongs to an individual user on Twitter or an individual on other social media platforms.
    - 1=**Yes 0=No**

❖ Evidence Source Bias and Expertise

- **Biased evidence source:** If the source mentioned in the tweet has **a political affiliation**

    - You can check bias ratings by seeing the following sites:
        - https://mediabiasfactcheck.com/
        - https://www.allsides.com/media-bias/media-bias-ratings

- **Expert evidence source:** If the evidence source mentioned in the tweet has expertise in health-related issues.
    - You can check the **Wikipedia** page for the area of expertise**:** https://en.wikipedia.org/wiki/Main_Page
- **Note:** if the evidence source belongs to a hospital, medical association, non-profit organization, peer-reviewed journal, or magazine in health-related issues (but the bias was not found for it) then label all these source types as **non-biased** and **expert**
- **Note:** if the evidence source is a news agency with a specialty in politics, economy,etc..(other than health) then directly label it as **not expert**

- **non-biased :** The evidence source is not-biased.

  - 1=**Yes** **0=No**
- **Left :** The evidence source bias rating is left.
  - 1=**Yes** **0=No**
- **Extreme_left:** The evidence source bias rating is extreme left.
  - 1=**Yes** **0=No**

- **Left_center:** The evidence source bias rating is left-center.
  - 1=**Yes** **0=No**

- **Center:** The evidence source bias rating is center or mixed.
  - 1=**Yes** **0=No**
- **Least_biased:** The evidence source bias rating is least biased.
  - 1=**Yes** **0=No**

- **Right:** The evidence source bias rating is right.
  - 1=**Yes** **0=No**
- **Right_center:** The evidence source bias rating is right-center.
  - 1=**Yes** **0=No**
- **Extreme_right:** The evidence source bias rating is extreme right.
  - 1=**Yes** **0=No**

- **Pro_science:** The evidence source bias rating is a pro science.
  - 1=**Yes** **0=No**

- **Does_not_exist:**The evidence source bias rating does not exist.
  - 1=**Yes** **0=No**

- **Expert:**The evidence source in the tweet is an expert in health-related issues.
  - 1=**Yes** **0=No**
- **Not_expert:** The evidence source in the tweet is not an expert in health-related issues.
  - 1=**Yes** **0=No**
- **Cant_ assess_expertise** : I can't assess the expertise of the evidence source.
  - 1=**Yes** **0=No**

❖ Stance

- Tweet stance means the position (support or against, neutral, or not clear) the tweet holds about COVID-19 vaccination. The definition of each position is as follows:

  - **Support:** The tweet **explicitly** or **implicitly** vaccination.
  - **Against:** The tweet **explicitly** or **implicitly** against vaccination.

- ○ **Neutral:** The tweet **supports** vaccination **BUT** at the same time shows a claim **against** vaccination.
- ○ **Not clear:** Not clear if the tweet supports or is against vaccination, jokes.

➢ Vaccination

- **Support:** The tweet supports vaccination
  - **1=Yes 0=No**
- **Against:** The tweet is against vaccination
  - **1=Yes 0=No**
- **Neutral** : The tweet shows a neutral claim toward vaccination
  - **1=Yes 0=No**
- **Not Clear**: The tweet shows a not clear claim toward vaccination
  - **1=Yes 0=No**

APPENDIX C: MACHINE LEARNING MODELS SETUP

A. Support Vector Machine (SVM)

- For kernel :we used ‘RBF’
- For C (Regularization parameter): we used according to multiple performed tests the best value for C in same cases was 1 and other 10.
- For gamma: a test was performed using this list of values [0.001, 0.01, 0.1, 1, 10, 100]] and 1 was the best pick for all the tests.

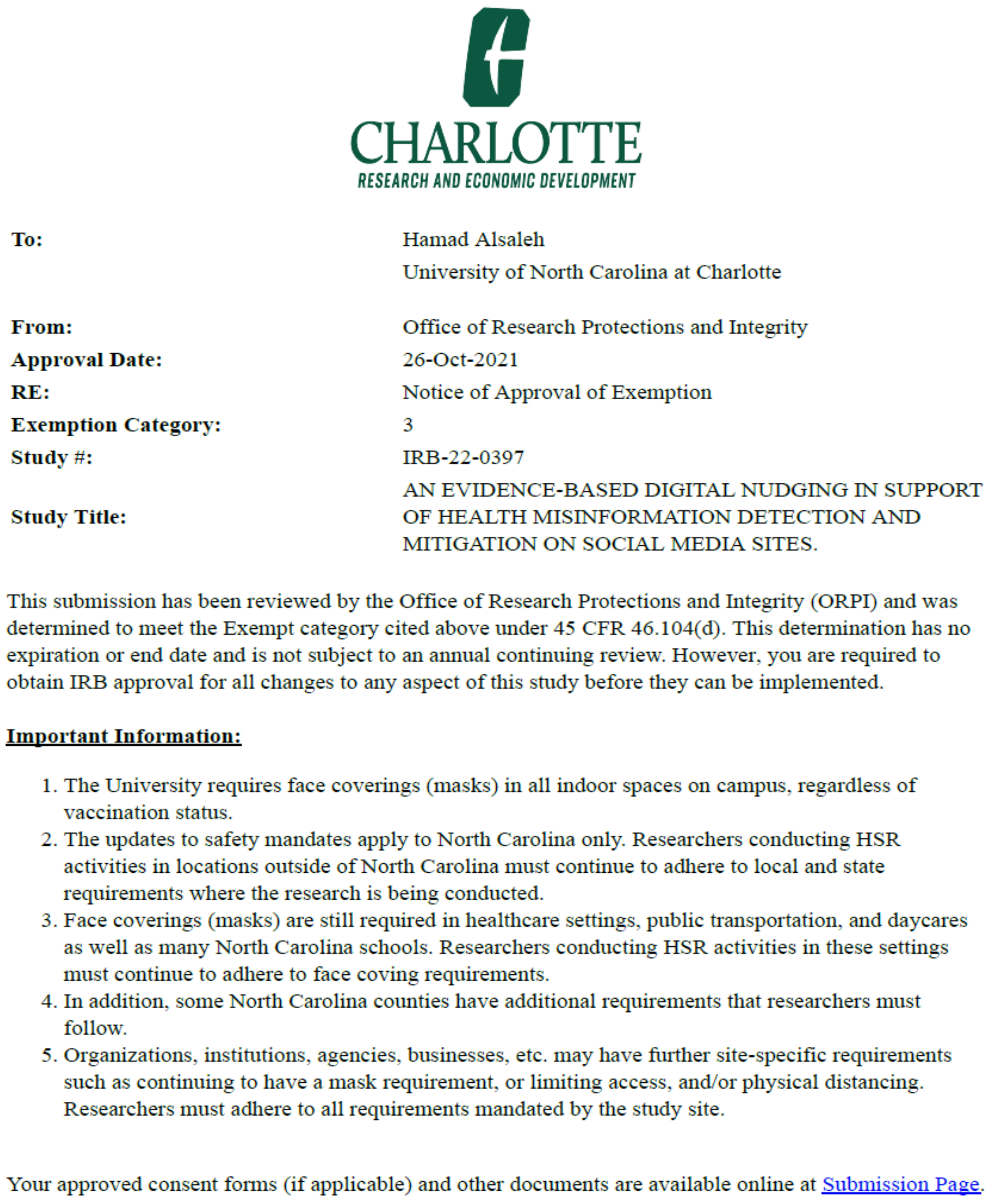
B. Decision Tree (DT):

- max_depth: 32

C. Random Forest (Rf)

- n_estimators = 500 , max_depth = 32

APPENDIX D: IRB APPROVAL FOR SURVEY STUDY



CHARLOTTE
RESEARCH AND ECONOMIC DEVELOPMENT

To:	Hamad Alsaleh University of North Carolina at Charlotte
From:	Office of Research Protections and Integrity
Approval Date:	26-Oct-2021
RE:	Notice of Approval of Exemption
Exemption Category:	3
Study #:	IRB-22-0397
Study Title:	AN EVIDENCE-BASED DIGITAL NUDGING IN SUPPORT OF HEALTH MISINFORMATION DETECTION AND MITIGATION ON SOCIAL MEDIA SITES.

This submission has been reviewed by the Office of Research Protections and Integrity (ORPI) and was determined to meet the Exempt category cited above under 45 CFR 46.104(d). This determination has no expiration or end date and is not subject to an annual continuing review. However, you are required to obtain IRB approval for all changes to any aspect of this study before they can be implemented.

Important Information:

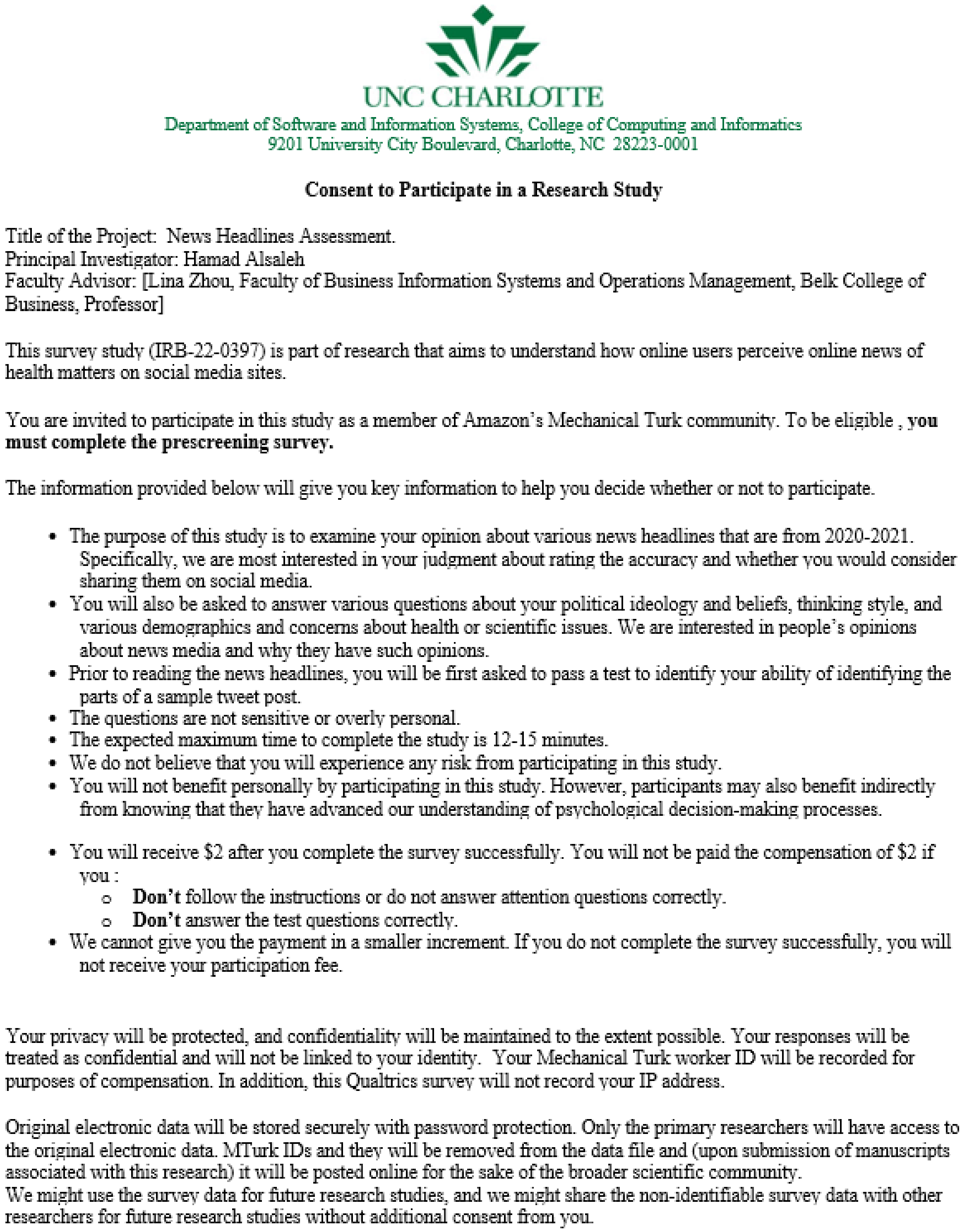
1. The University requires face coverings (masks) in all indoor spaces on campus, regardless of vaccination status.
2. The updates to safety mandates apply to North Carolina only. Researchers conducting HSR activities in locations outside of North Carolina must continue to adhere to local and state requirements where the research is being conducted.
3. Face coverings (masks) are still required in healthcare settings, public transportation, and daycares as well as many North Carolina schools. Researchers conducting HSR activities in these settings must continue to adhere to face coving requirements.
4. In addition, some North Carolina counties have additional requirements that researchers must follow.
5. Organizations, institutions, agencies, businesses, etc. may have further site-specific requirements such as continuing to have a mask requirement, or limiting access, and/or physical distancing. Researchers must adhere to all requirements mandated by the study site.

Your approved consent forms (if applicable) and other documents are available online at Submission Page.

APPENDIX E: SELECTED HEADLINES

Headline Veracity	**Headline**
True	1. *it is safe for those with a history of sever food allergies to get the Pfizer and Moderna vaccines*
	2. *Covid-19 vaccine rollout for ages 12 to 15 is 'better than expected,' health officials say*
	3. *FDA greenlights COVID-19 booster vaccine for some immunocompromised patients*
	4. *FDA recommends COVID-19 booster vaccine during pregnancy*
	5. *No proof COVID-19 vaccine affects menstruation or fertility*
	6. *CDC strengthens recommendation for pregnant women to get vaccination against COVID-19*
	7. *Long-term side effects from COVID-19 vaccine unlikely, data shows*
	8. *No conclusion that facial paralysis cases were caused by vaccination*
False	9. *People with diabetes (type 1 or type 2) have been not prioritized for vaccination across the U.S*
	10. *COVID-19 vaccine linked to cognitive decline, acceleration of Alzheimer's like symptoms, research finds*
	11. *Moderna Vaccine may cause deadly reactions in people with facial fillers*
	12. *COVID-19 vaccine maybe less effective in some people with cancer*
	13. *Death and destruction continue to follow Pfizer covid-19 vaccines-now approved by the FDA*
	14. *Nearly 1 million and 600,000 adverse events were reported after COVDFI-19 vaccine boosters*
	15. *CDC says J&J vaccine has a common 'risk' of neurological disorder*
	16. *FDA says there is common side effects of heart inflammation have been reported in the United States after mRNA COVID-19 vaccination(Pfizer - BioNTech and Moderna)*

APPENDIX F: DATA COLLECTION MATERIALS



UNC CHARLOTTE

Department of Software and Information Systems, College of Computing and Informatics
9201 University City Boulevard, Charlotte, NC 28223-0001

Consent to Participate in a Research Study

Title of the Project: News Headlines Assessment.
Principal Investigator: Hamad Alsaleh
Faculty Advisor: [Lina Zhou, Faculty of Business Information Systems and Operations Management, Belk College of Business, Professor]

This survey study (IRB-22-0397) is part of research that aims to understand how online users perceive online news of health matters on social media sites.

You are invited to participate in this study as a member of Amazon's Mechanical Turk community. To be eligible , **you must complete the prescreening survey.**

The information provided below will give you key information to help you decide whether or not to participate.

- The purpose of this study is to examine your opinion about various news headlines that are from 2020-2021. Specifically, we are most interested in your judgment about rating the accuracy and whether you would consider sharing them on social media.
- You will also be asked to answer various questions about your political ideology and beliefs, thinking style, and various demographics and concerns about health or scientific issues. We are interested in people's opinions about news media and why they have such opinions.
- Prior to reading the news headlines, you will be first asked to pass a test to identify your ability of identifying the parts of a sample tweet post.
- The questions are not sensitive or overly personal.
- The expected maximum time to complete the study is 12-15 minutes.
- We do not believe that you will experience any risk from participating in this study.
- You will not benefit personally by participating in this study. However, participants may also benefit indirectly from knowing that they have advanced our understanding of psychological decision-making processes.

- You will receive $2 after you complete the survey successfully. You will not be paid the compensation of $2 if you :
    - **Don't** follow the instructions or do not answer attention questions correctly.
    - **Don't** answer the test questions correctly.
- We cannot give you the payment in a smaller increment. If you do not complete the survey successfully, you will not receive your participation fee.

Your privacy will be protected, and confidentiality will be maintained to the extent possible. Your responses will be treated as confidential and will not be linked to your identity. Your Mechanical Turk worker ID will be recorded for purposes of compensation. In addition, this Qualtrics survey will not record your IP address.

Original electronic data will be stored securely with password protection. Only the primary researchers will have access to the original electronic data. MTurk IDs and they will be removed from the data file and (upon submission of manuscripts associated with this research) it will be posted online for the sake of the broader scientific community.
We might use the survey data for future research studies, and we might share the non-identifiable survey data with other researchers for future research studies without additional consent from you.

Survey Questionnaire

Start of Block: Concerns About COVID-19 news+Attention_1

Q5 How concerned are you about COVID-19 ?

	Not concerned at all	Extremely concerned
	0 10 20 30 40 50	60 70 80 90 100
1 ()		

Q6 How often do you proactively check the news regarding COVID-19 ?

○ 1-Never (1)

○ 2-Rarely (2)

○ 3-Sometimes (3)

○ 4-Often (4)

○ 5-Very often (5)

Q7 When a big news story breaks, people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. Please select Google News website and USA Today website as your two answers.

Q8 When there is a big news story, which is the one news website you would visit first? (Please choose two options):

□ Huffington Post

□ Google News

□ Yahoo! News

□ USA Today website

□ Other ____

End of Block: Concerns About COVID-19 news+Attention_1

Start of Block: Scenario+ Inst

Q9 **-You will be presented** with a series of news headlines about the Coronavirus (COVID-19) one by one.

-We are interested in whether you think :

- **the claim** of the news headline is **accurate.**

-**In addition** , we are interested in whether:

- you **will share** this news headline on social media sites such Twitter or Facebook.

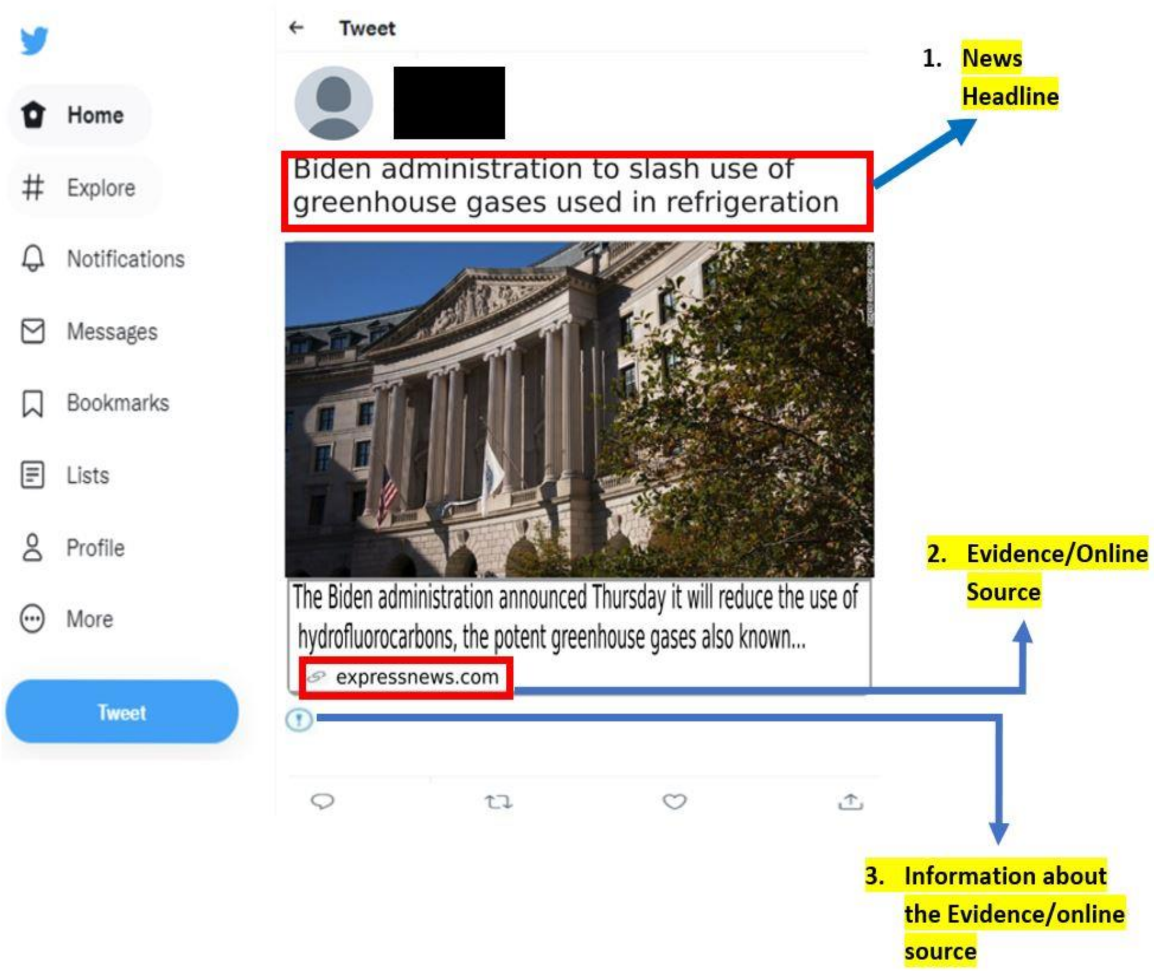
Please do not reference any outside source or leave this survey window.

Note: The news headlines may take a moment to load.

Page Break

Training:

- The following page is an illustration of a sample Twitter page of news headline that you will be reading. You are required to use the information from the three different sections of the page as highlighted in red boxes to respond to the survey questions. To make sure that you understand the type of contents in different sections of the page, you will be asked to take a test by answering a few questions. Only those participants who have answered all the questions correctly will be able to proceed with this study. Each participant can take at most two attempts , otherwise, they will be disqualified from the study.



[redacted]

Biden administration to slash use of greenhouse gases used in refrigeration

The Biden administration announced Thursday it will reduce the use of hydrofluorocarbons, the potent greenhouse gases also known...

expressnews.com

1. News Headline

2. Evidence/Online Source

3. Information about the Evidence/online source

Q9.1 What is the news headline in the above figure ?

○ Joe Biden's poll numbers are plummeting at exactly the wrong time (1)

○ President Joe Biden is mobilizing the federal government to deal with the effects of extreme heat (2)

○ Biden administration to slash use of greenhouse gases used in refrigeration (3)

Q9.2 What is the online evidence/online source in the above figure ?

○ expressnew.com

○ HotNews.com

○ NewsUnion.com

Q9.3 What is the mark/symbol that highlights information about the evidence/online source in the above figure ?

○ ()

○ !

○ #

End of Block: Scenario+ Inst

Start of Block: D1-16

Q11

Q12 Based on the claim in the news headline posted by the author, How likely do you think the claim is accurate and truthful?

	Very unlikely (1)	Unlikely (2)	Somewhat unlikely (3)	Undecided (4)	Somewhat likely (5)	Likely(6)	Very likely (7)
(1)	○	○	○	○	○	○	○

Q13 I think that the online source that generated the news headlines is accurate and truthful

	Very unlikely (1)	Unlikely (2)	Somewhat unlikely (3)	Undecided (4)	Somewhat likely (5)	Likely(6)	Very likely (7)
(1)	○	○	○	○	○	○	○

Did you rely on the information from the ⓘ to answer the previous question ?

○ No (0)

○ Yes (1)

Q14 How likely would you be to share this claim online (for example, through Facebook or Twitter)?

	Very unlikely (1)	Unlikely (2)	Somewhat unlikely (3)	Undecided (4)	Somewhat likely (5)	Likely(6)	Very likely (7)
(1)	○	○	○	○	○	○	○

Q15 How likely do you think it is that you have seen or heard the above claim before today?

	Very unlikely (1)	Unlikely (2)	Somewhat unlikely (3)	Undecided (4)	Somewhat likely (5)	Likely(6)	Very likely (7)
(1)	○	○	○	○	○	○	○

End of Block: D1-16

Post-Experiment Survey Questionnaire

Start of Block: Page 4 CRT-section_1+ Attention(2)

Q1 **In the following section**, you will be asked questions about your style of thinking. Please do your best to answer as accurately as possible.

Press the button below to continue.

Page Break

Q2 The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam ?

____

Q3 If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take 50 printers to print out 50 pages of paper?

____

Q4 On a loaf of bread, there is a patch of mold. Every day, the patch doubles in size. If it takes 40 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half of the loaf of bread?

____

Attention_2 Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of those options.

Q 6 What is your favorite color?

□ White (1)

□ Black (2)

□ Red (3)

□ Pink (4)

□ Green (5)

□ Blue (6)

End of Block: Page 4 CRT-section_1+ Attention(2)

Start of Block: Page 5 General Science Knowledge Quiz

Q7 Following, you will be asked a series of True/False questions. Please answer the following questions to the best of your ability.

Please do not consult any outside information or leave this survey window.

Press the button below to continue.

Q8 it is the father's gene that decides whether the baby is a boy or girl.

○ True (1)

○ False (2)

○ Don know (3)

Q9 The center of the Earth is very hot.

○ True (1)

○ False (2)

○ Don't know (3)

Q10The universe began with a huge explosion.

○ True (1)

○ False (2)

○ Don't know (3)

Q11 Lasers work by focusing sound waves.

○ True (1)

○ False (2)

○ Don't know (3)

Q12 Antibiotics kill viruses as well as bacteria.

○ True (1)

○ False (2)

○ Don't know (3)

Q13 The earliest humans lived at the same time as the dinosaurs.

○ True (1)

○ False (2)

○ Don't know (3)

Q14 Nuclear plants destroy the ozone layer.

○ True (1)

○ False (2)

○ Don't know (3)

Q15 Ordinary tomatoes, the ones we normally eat, do not have genes whereas genetically modified tomatoes do.

○ True (1)

○ False (2)

○ Don't know (3)

Q16 The continents on which we live have been moving their locations for millions of years and will continue to move in the future.

○ True (1)

○ False (2)

○ Don't Know (3)

Q17 For the first time in recorded history, some species of plants and animals are dying out and becoming extinct.

○ True (1)

○ False (2)

○ Don't know (3)

Q18 Does the Earth go around the Sun or does the Sun go around the Earth?

○ The Earth goes around the Sun (1)

○ The Sun goes around the Earth (2)

○ Don't know (3)

Q19 More than half of the human genes are identical to those of mice.

○ True (1)

○ False (2)

○ Don't know (3)

Q20 Human beings, as we know them today, developed from earlier species of animals.

○ True (1)

○ False (2)

○ Don't know (3)

Q21 Which travels faster: light or sound?

○ Lights (1)

○ Sound (2)

○ Don't know (3)

Q22 The primary human activity that causes global warming is the burning of fossil fuels, such as oil and coal.

○ True (1)

○ False (2)

○ Don't know (3)

Q23 Electrons are smaller than atoms.

○ True (1)

○ False (2)

○ Don't know (3)

Q24 All radioactivity is man-made.

○ True (1)

○ False (2)

○ Don't know (3)

End of Block: Page 5 General Science Knowledge Quiz

Start of Block: Medical Maximizer (MMS)

Q25 It is important to treat a disease even when it does not make a difference in survival.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q26 It is important to treat a disease even when it does not make a differences in quality of life.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q27 Doing everything to fight a disease is always the right choice

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q28 When it comes to health care, the only responsible thing to do is actively seek medical care

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q29 When it comes to health care, watching and waiting is never an acceptable option.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q30 When it comes to medical treatment, more is usually better.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q31 If I have a medical problem, my preference is to go straight to a doctor and ask his or her opinion.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q32 I often suggest that friends and family see their doctor.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q33 If I have a health issue, my preference is to wait to see if the problem gets better on its own before going to a doctor.

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

Q34 If I feel unhealthy. The first thing I do is to go to the doctor and get a prescription

○ Strongly disagree (1)

○ Disagree (2)

○ Somewhat disagree (3)

○ Neither agree nor disagree (4)

○ Somewhat agree (5)

○ Agree (6)

○ Strongly agree (7)

End of Block: Medical Maximizer (MMS)

Start of Block: Attention_3

Q35 In the grid below, you will see a list of statements. Please tell us whether you agree or disagree with each of the statements.

	Agree strongly(1)	Agree(2)	Somewhat agree(3)	Neither agree nor disagree(4)	Somewhat disagree(5)	Disagree(6)	Disagree strongly(7)
People convicted of murder should be given the death penalty. (1)	○	○	○	○	○	○	○
Please click the "neither agree nor disagree" response. (2)	○	○	○	○	○	○	○
Gays and lesbians should have the right to legally marry. (3)	○	○	○	○	○	○	○
World War I came after World War II (4)	○	○	○	○	○	○	○
In order to reduce the budget deficit, the federal government should raise taxes on people that make more than $250,000 per year. (5)	○	○	○	○	○	○	○
The Affordable Care Act passed by Congress in 2010 should be repealed. (6)	○	○	○	○	○	○	○

The government should require all electricity power plants to significantly reduce their greenhouse gas emissions even if it might increase electricity bills a few dollars a month. (7)	○	○	○	○	○	○	○

End of Block: Attention_3

Start of Block: Digital Media Literacy_Detailed

Q36 The term “media” used in the following items, unless otherwise specified, refers to current digital technology platforms including but not limited to web sites, online forums, social networks, video sharing sites and virtual worlds in which anyone can share any digital content. Please indicate how you feel about your knowledge and skills for each of the following statements.

	Agree strongly(1)	Agree(2)	Somewhat agree(3)	Neither agree nor disagree(4)	Somewhat disagree(5)	Disagree(6)	Disagree strongly(7)
I know how to use searching tools to get information needed in the media. (1)	○	○	○	○	○	○	○
I am good at catching up with the changes in the media. (9)	○	○	○	○	○	○	○
It is easy for me to make use of various media environments to reach information. (10)	○	○	○	○	○	○	○
I realize explicit and implicit media messages (12)	○	○	○	○	○	○	○
I notice media contents containing mobbing and violence. (4)	○	○	○	○	○	○	○
I understand the political, economical, and social dimensions of media content. (29)	○	○	○	○	○	○	○
I perceive different opinions and thoughts in the media. (13)	○	○	○	○	○	○	○

	Agree strongly(1)	Agree(2)	Somewhat agree(3)	Neither agree nor disagree(4)	Somewhat disagree(5)	Disagree(6)	Disagree strongly(7)
I can distinguish different functions of media (communication, entertainment, etc.). (8)	○	○	○	○	○	○	○
I am able to determine whether or not media contents have commercial messages. (9)	○	○	○	○	○	○	○
I manage to classify media messages based on their producers, types, purposes and so on. (10)	○	○	○	○	○	○	○
I can compare news and information across different media environments. (11)	○	○	○	○	○	○	○
I can combine media messages with my own opinions. (12)	○	○	○	○	○	○	○
I consider media rating symbols to choose which media contents to use. (13)	○	○	○	○	○	○	○

It is easy for me to make decision about the accuracy of media messages. (14)	○	○	○	○	○	○	○

	Agree strongly(1)	Agree(2)	Somewhat agree(3)	Neither agree nor disagree(4)	Somewhat disagree(5)	Disagree(6)	Disagree strongly(7)
I am able to analyze positive and negative effects of media contents on individuals. (8)	○	○	○	○	○	○	○
I can evaluate media in terms of legal and ethical rules (copyright, human rights, etc.). (9)	○	○	○	○	○	○	○
I can assess media in terms of credibility, reliability, objectivity and currency. (10)	○	○	○	○	○	○	○
I manage to defend myself from the risks and consequences caused by media contents. (11)	○	○	○	○	○	○	○
It is easy for me to create user accounts and profiles in media environments. (12)	○	○	○	○	○	○	○
I can use hardware necessary for developing media contents (text, image, video, etc.). (13)	○	○	○	○	○	○	○

I am able to use software necessary for developing media contents (text, image, video, etc.). (14)	○	○	○	○	○	○	○

Q40

	Agree strongly(1)	Agree(2)	Somewhat agree(3)	Neither agree nor disagree(4)	Somewhat disagree(5)	Disagree(6)	Disagree strongly(7)
I can use basic operating tools (button, hyperlinks, file transfer etc) in the media. (8)	○	○	○	○	○	○	○
I am good at sharing digital media contents and messages on the Internet. (9)	○	○	○	○	○	○	○
I can make contribution or comments to media contents shared by others. (10)	○	○	○	○	○	○	○
I am able to rate or review media contents based on my personal interests and liking. (15)	○	○	○	○	○	○	○
I manage to influence others' opinions by participating to social media environments. (16)	○	○	○	○	○	○	○
I can make contribution to media by reviewing current matters from different perspectives (social, economical, ideological etc.). (17)	○	○	○	○	○	○	○
I am able to collaborate and interact with diverse media users towards a common purpose. (18)	○	○	○	○	○	○	○

Q41

	Agree strongly(1)	Agree(2)	Somewhat agree(3)	Neither agree nor disagree(4)	Somewhat disagree(5)	Disagree (6)	Disagree strongly(7)
It is easy for me to construct online identity consistent with real personal characteristics. (8)	○	○	○	○	○	○	○
I can make discussions and comments to inform or direct people in the media. (9)	○	○	○	○	○	○	○
I am skilled at designing media contents that reflect critical thinking of certain matters. (10)	○	○	○	○	○	○	○
I am good at producing opposite or alternative media contents. (15)	○	○	○	○	○	○	○
I produce media contents respectful to people’s different ideas and private lives. (16)	○	○	○	○	○	○	○
it is important for me to create media contents that comply with legal and ethical rules. (17)	○	○	○	○	○	○	○
I am able to develop original visual and textual media contents (video clips, web page, etc.) (18)	○	○	○	○	○	○	○

Q42 How important is it to you that you share news articles on social media (such as Facebook and Twitter) only if they are accurate?

○ Not at all important (1)

○ Mostly not important (2)

○ Slightly important (3)

○ Moderately important (4)

○ Important (5)

○ Very important (6)

○ Extremely important (7)

Q43 Some people think that by criticizing leaders, news organizations keep political leaders from doing their job. Do you agree with this position ?

	I do not agree at all	**I extremely agree**
	0 10 20 30 40 50 60 70 80 90 100	
1 ()		

Q44 Some people think that by criticizing leaders is worth it because it keeps political leaders from doing things that should not be done. Do you agree with this position?

	I do not agree at all								**I extremely agree**		
	0	10	20	30	40	50	60	70	80	90	100
1 ()											

Q45 In presenting the news dealing with political and social issues, do you think that news organizations deal fairly with all sides?

	Not at all likely								**Extremely likely**		
	0	10	20	30	40	50	60	70	80	90	100
1 ()											

Q46 In presenting the news dealing with political and social issues, do you think that news organizations tend to favor one side?

	Not at all likely								**Extremely likely**		
	0	10	20	30	40	50	60	70	80	90	100
1 ()											

Q47 To what extent do you trust the information that comes from the following?

	Not at all (1)	Mostly not (2)	Slightly (3)	A moderate amount (4)	A little more (5)	A lot(6)	A great deal (7)
National news organizations (1)		○	○	○	○	○	○
Local news organizations (2)		○	○	○	○	○	○
Friends and family (3)		○	○	○	○	○	○
Social networking sites (e.g., Facebook, Twitter) (4)		○	○	○	○	○	○
3rd party fact-checkers (e.g., snopes.com, factcheck.org) (5)		○	○	○	○	○	○

End of Block: Digital Media Literacy_Detailed

Start of Block: Page 2 Demographics_Detailed

Q48 What is your gender ?

○ Male (1)

○ Female (2)

○ other (3)

Q49 What is the highest degree of school you have completed ?

○ Did not graduate from high school (1)

○ High school diploma or the equivalent (2)

○ some college (3)

○ Associate's Degree (4)

○ Bachelor Degree (5)

○ Master's Degree (6)

○ Professional or Doctorate Degree (7)

+eligibility

Q50 In what state you currently reside ?

▼ Alabama (1) ... I do not reside in the United States (54)

Q51 Please choose whichever ethnicity that you identify with

○ White (1)

○ Black or African American (2)

○ American Indian or Alaska Native (3)

○ Asian (4)

○ Native Hawaiian or Pacific Islander (5)

○ Other (6)

Q52 Are you of Hispanic/Latino origin ?

○ Yes (1)

○ No (2)

Q53 Which of the following best describes your political preference?

○ Strongly Democratic (1)

○ Democratic (2)

○ Lean Democratic (3)

○ Independent (4)

○ Lean Republican (5)

○ Republican (6)

○ Strongly Republican (7)

○ Other (8) ____

Q54 Who did you vote for in the 2016 Presidential Election?

Reminder: This survey is anonymous.

○ Hillary Clinton (1)

○ Donald Trump (2)

○ Other candidate (such as Jill Stein or Gary Johnson) (3)

○ I did not vote for reasons outside of my control (4)

○ I did not vote, but I could have (5)

○ I did not vote out of protest (6)

End of Block: Page 2 Demographics_Detailed

Start of Block: Page CRT-section_2

Q55 If you’re running a race and you pass the person in second place, what place are you in?

____

Q56 A farmer had 15 sheep and all but 8 died. How many are left?

____

Q57 Emily’s father has three daughters. The first two are named April and May. What is the third daughter’s name?

____

Q58 Have you seen any of the last 7 word problems before?

○ Yes (1)

○ Maybe (2)

○ No (3)

End of Block: Page 4 CRT-section_2

Start of Block: Random+Questions

Q59 Did you search the internet (via Google or otherwise) for any of the news headlines?

Note: Please be honest! You will get your HIT regardless of your response.

○ Yes (1)

○ No (2)

Q60 Did you receive or intend to take COVID-19 vaccination ?

○ Yes (1)

○ Maybe (2)

○ No (3)

End of Block: Random+Questions

Start of Block: Final_Page

Thank you for your time, here is your Amazon Mechanical Turk Confirmation Code: SM-${e://Field/Random_AMT_Code}-CC
Please submit **the code to the Amazon Mechanical Turk Page**.
Note: If you did not **submit the code to Amazon Mechanical Turk Page, we can not** reimburse you for taking the survey.

Note: If you had any psychological burden during the study please consult your doctor.
Please click the button below to save your answers.

End of Block: Final_Page

APPENDIX G: STUDY ITEMS(CONTROL VARIABLE)

Study Measure	Item	Scale	Reference
News Reading Frequency	*How often do you read news in English on social media over the past 6 months?*	(1= Several times a day, 2= On a daily basis,3= Several times a week, 4= Once a week, 5= Once a month, 6= Several times a year, 7=Once a year)	[93]
News Sharing Frequency	*How often do you share news in English on social media over the past 6 months?*	(1= Several times a day, 2= On a daily basis,3= Several times a week, 4= Once a week, 5= Once a month, 6= Several times a year, 7=Once a year)	[93]
Political Leaning	*Which of the following best describes your political preference?*	(1=Strongly Democratic, 2= Democratic,3= Lean Democratic, 4= Independent, 5=Lean Republican, 6= Republican,7= Strongly Republican)	[266], [274]
Trust in News Organizations	*To what extent do you trust the information that comes from the following?*	(1= not at all to 7=a great deal)	[266]
COVID-19 Concerns	*How concerned are you about COVID-19?*	(0= not concerned at all to 100=extremely concerned)	[95]
COVID-19 News Checking	*How often do you proactively check the news regarding COVID-19 ?*	(1= never to 5=very often).	[95]