**Building Immunity to Online Deception: A New Approach Using Active Inoculation**

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Austin Baney
Spring, 2021

Technical Project Team Members
Austin Baney
Eric Stoloff

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date 5/5/2021
Austin Baney

Approved _____ Date _____
Raymond Pettit, Assistant Professor, Department of Computer Science

Approved _____ Date _____
Rich Nguyen, Assistant Professor, Department of Computer Science

**Abstract**

The spread of false information on social media is a dangerous threat to modern society because it can lead people to take misguided and even violent actions. Social media is an incredibly important information source for people and it allows for the propagation of content around the world. Despite this, users have an inability to recognize false information and fake news. Current solutions to the problem of false information spreading have the drawbacks of taking too much time for debunking and failing to address the misinformation before it takes root. We aim to provide a new user interface feature that aids the users themselves in recognizing misleading information and helps them be more informed and conscious social media users. When integrated with social media platforms, this feature will occasionally prompt users to complete a brief exercise in which they must identify misinformation and common strategies for its dissemination. This feature is meant to improve users' ability to discern the veracity of social media posts, which will help reduce the proliferation of false information on these networks. It will enable users to ask questions and think more critically about the content they encounter, while creating a norm in which credibly-sourced information is valued.

## Introduction

This project seeks to address the problem of false information circulating on social media and specifically aims to mitigate the degree to which false information can influence users. It is clear that false information can affect the beliefs of social media users and can even be used to modify their behavior (Allcott & Gentzkow, 2017; Bastick, 2021). Currently, social media companies use a variety of methods in their attempts to solve this problem. Facebook and Instagram combine fact-checking and machine learning approaches in order to identify misinformation and then flag false information for users (Guo et al., 2020; Facebook, 2021). Twitter does not divulge how they detect false information, but they classify it and label it to users (Roth & Pickles, 2020). Reddit is very confident in its community's ability to prevent the spread of false information with the upvoting and downvoting of posts, and Reddit is actually not a large source for false information at all, with more interaction between differing political groups than within the groups themselves for example (De Francisci Morales, Monti, & Starnini, 2021). YouTube uses its newly improved recommendation system to recommend authoritative sources and monitors its media for content that violates its guidelines and removes it (Google, 2019). Although a wide array of methods is used, the solutions to mitigating the spread of false information can mostly be characterized as reactive. This is because they wait for false information to appear and respond to it after the fact. This is dangerous because false information spreads faster than true information, and efforts to respond to it may occur after the information has already reached millions of users (Vosoughi, Roy, & Aral, 2018). There is an average 12-hour delay between false information's release and its debunking, so solutions to this problem must work very quickly (Kumar & Shah, 2018).

**Background**

There are many reasons that false information disseminates not only more rapidly, but farther and with more shares than true information on social media (Vosoughi, Roy, & Aral, 2018). Human vulnerability to it is the largest factor in its circulation, and this is exemplified by the research that shows humans are more likely to spread false information than bots (Vosoughi, Roy, & Aral, 2018). Humans are so susceptible in part due to their tendency to first accept information as truth, but also due to the many manipulation strategies that are wielded by false information (Dafonte-Gomez, 2018; Lewandowski et al., 2012). The first of these manipulation strategies is impersonation, or pretending to be someone else. Impersonation is used on social media platforms to steal the credibility of those being impersonated and trick users into thinking that the fake account contains reliable information (Roozenbeek & van der Linden, 2019). The second manipulation technique commonly employed by misinformation is using emotionally-charged content to entice the user to interact with a post. This is a very powerful tool that can be used to sway the user's opinion and increase the likelihood of content sharing (Roozenbeek & van der Linden, 2019). Furthermore, awareness of this technique is important because of strong emotion's ability to increase the user's memory of a post (Dafonte-Gomez, 2018). The third technique that false information uses to manipulate is polarization, which can be defined as intentionally increasing the fractionalization of groups and aggravating existing tensions. Polarization emphasizes the differences between people and creates the illusion for one side of an issue that the other side supports or believes something they do not (Roozenbeek & van der Linden, 2019). The fourth method of manipulation that false information draws from is conspiracy theories, which is best defined as “Creating or amplifying alternative explanations for traditional news events which assume that these events are controlled by a small, usually

malicious, secret elite group of people" (Roozenbeek & van der Linden, 2019). The fifth technique used to aid in the multiplication of false information on social media is discrediting people. This can either be done through denying accusations or by deflecting attention to the opposition by attacking them (Roozenbeek & van der Linden, 2019). The final strategy commonly used in conjunction with misinformation is trolling, which aims to entice a reaction out of the user that seduces them into participating in tangential discussions. Trolling uses the other strategies, sometimes in combination, to provoke users and get them to engage the troll, which results in the troll's post receiving more attention (Roozenbeek & van der Linden, 2019). In order to prevent false information from proliferating on social media platforms, it is important to consider these strategies and how humans can bypass them.

**Related Work**

Reacting to false information that flows through social media has been the predominant approach of people trying to solve this problem. There are many methods that are used to find the false information in the ocean of true information that exists on social media. There are multiple machine and deep learning systems that have shown very high success rates in detecting fake news, fake reviews, and other false information on social media platforms (Atodiresei, Tanaselea, & Iftene, 2018; Kumar & Shah, 2018). These solutions are feature-based and develop a model based on post content, user data, and metadata which they later use to classify or rate posts on their veracity (Guo et al., 2020; Kumar & Shah, 2018). Another class of solutions to finding false information on social media involves modeling its spread, which works because false information spreads differently than true information (Kumar & Shah, 2018; Yu et al., 2019). Community-based solutions are also very widely used such as fact checking and Twitter's new Birdwatch system (deBeer & Matthee, 2020; Guo et al., 2020). This system allows for users

to add notes to tweets they think are misleading as well as view the collection of these notes. Whatever the method of discovery may be, the conventional approach to dealing with known false information has been to flag it as false or to issue some kind of correction to the information (Chou, Gaysynsky, & Vanderpool, 2021; Kumar & Shah, 2018). This seems intuitive, as users are warned not to trust information or provided with additional explanation as to why some content is false or misleading. However, it is already too late. Research has shown that misinformation is very resistant to being corrected, and even when users believe and understand a correction, it has little effect of eliminating misinformation from users' minds (Lewandowsky et al., 2012). Therefore, the only viable option is to find a solution that can work proactively rather than reacting to instances of false information.

**System Design**

This system is a tool designed to be built into social media user interfaces to help users recognize false information and prevent its propagation. The architecture of the system would not be very complex, as it would mostly reside in the front-end of social media applications. Our system is not intended to be a stand-alone entity that connects to social media applications' APIs. Rather, it is intended to be a tool that can easily be integrated into existing social media platforms by the respective companies in order to combat the spread of misinformation.

**Active Inoculation Against False Information Techniques**

This tool uses the concept of active inoculation in order to help users improve their ability to recognize false information. Active inoculation in psychological terms works similarly to inoculation against a virus, where a patient is exposed to a weakened copy of a virus in order to trigger protective immune responses. It involves presenting someone with a weakened copy of a challenge, like a radical conspiracy theory, in order to trigger enhanced critical thinking and

improved resistance to false information (van der Linden, Roozenbeek, & Compton, 2020). The only studied implementation of inoculating against misinformation has been shown to be successful in increasing people's ability to discern whether or not some information is false and improving their confidence in this ability as well (van der Linden, Roozenbeek, & Compton, 2020). In addition, the "immunity" to false information that it provides has been shown to last for multiple months. This implementation comes in the form of the online game *Bad News*, which can be easily accessed in a web browser. Players of the game take on the role of a false-information creator and learn about how it spreads on social media. Another new implementation of active inoculation is a similar game, called *Go Viral!*, which is designed to specifically target misinformation about Covid-19, but it has not been researched much yet (van der Linden, Roozenbeek, & Compton, 2020). The issue with these online games is that though effective and played by millions of people, they have only reached a small fraction of the social media users worldwide (Chaffey, 2020; van der Linden, Roozenbeek, & Compton, 2020). Furthermore, social media companies could not conceivably require all their users to complete these games in order to use their platforms because they take a fair amount of time and users could simply use other platforms if they were greatly inconvenienced.

Our system combines the idea of active inoculation with use of social media platforms. By requiring users to complete activities that will inoculate them inside the applications, every social media user can hopefully be more immune to misinformation. These activities are extremely short, with the user being able to return to their normal perusal through their social media content after correctly answering one question. This is intended to decrease the intrusiveness into the user's normal operation of the application. Instead of administering larger "doses" of inoculation at once, like in *Bad News*, the intention of the system is to achieve the same effect by spreading

the inoculation out across the user's time spent on the platform, with more brief, but more frequent activities. An added benefit to the system is that it does not address unique instances of false information individually, but rather aids users in recognizing the manipulation *techniques* employed by false information. Because the user will be more aware of these techniques, they can even be ready to handle false information that has not been created yet.

**UX Design**

To address the various manipulation techniques, there are a wide variety of questions that could be asked of the user in the activity. Large amounts of these questions could be created relatively easily with the help of the many publicly available social media datasets that exist, some of which are dedicated to true information and some of which focus on false information (Kumar & Shah, 2018). Using the User Interface Design tool *Balsamiq*, we have created example mockups that address two of the manipulation techniques that false information utilizes. In these mockups, we made sure to use the appropriate principles of UX Design. We used the color red to indicate a wrong answer and green for a correct answer, and we chose the neutral color blue to indicate which answer choice the user has selected. Also, bubble-shaped text boxes were chosen to signify the question, while all the answer choices are contained in rectangular boxes. These mockups are described in the figures below.

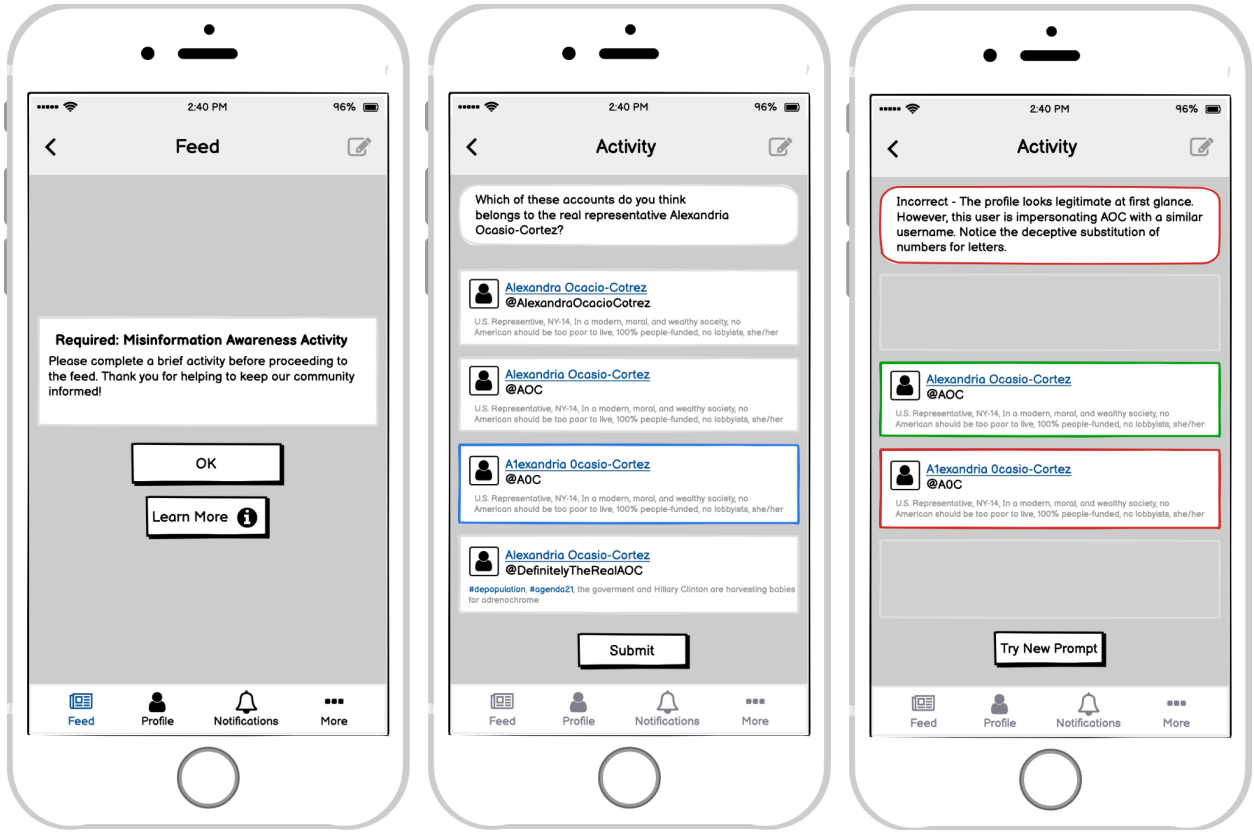


**Figure 1:** This activity is designed to help users recognize fake accounts that intend to impersonate someone and exploit their credibility. Users' ability to navigate the app is restricted until the activity is complete (see gray icons).

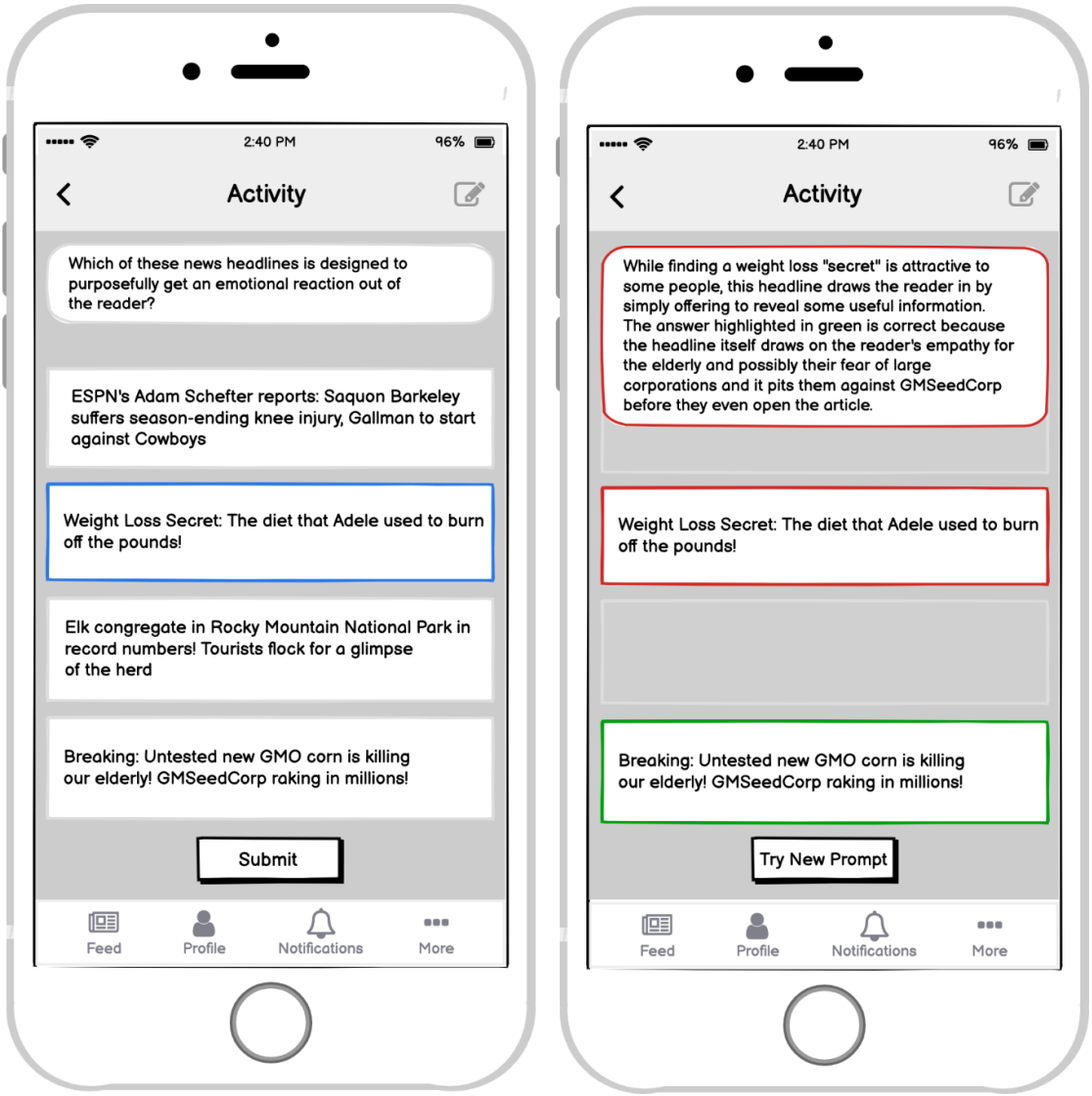


**Figure 2:** This activity is designed to aid users in recognizing language specifically intended to evoke an emotional reaction from the reader to increase their desire to interact with the information (Note: initial prompt screen from Figure 1 will be used to start every activity).

An additional goal of our mockups was to keep them relatively neutral, so that users would not feel like they are being forced to take certain positions on different issues. This is to greatly reduce the effect that personal biases, worldview, and other subjective sentiments may have on the user's willingness to participate and the effectiveness of the activities themselves. The overall goal of this project is to bring active inoculation to as many social media users as possible and eliminating these biases by keeping the activities free of controversial topics will

allow us to reach users regardless of political beliefs or other opinions. Also, focusing on the manipulation techniques instead of actual content will provide a further buffer to our system from biases. By integrating inoculation-based activities into the User Interface for social media platforms, this system will reach more people than the game *Bad News*, while being minimally invasive into users' normal browsing of posts. It will use proven techniques to bolster users' ability to discern false information from true information and ideally make social media platforms more informed, safer places.

**Procedure**

The usage of our tool, once integrated into a social media platform, follows a simple workflow. When a user begins a new session with the platform, they will be prompted by our system to complete a brief "Misinformation Awareness Activity". The user will not have access to the platform's content feed, posts, etc. until completing the activity.

The activity consists of a quiz, presenting a prompt with multiple choices; each choice serves as an example of digital content one might encounter, such as a social media post or a blog headline. To find the correct answer, users must be able to identify misinformation and common strategies used in its propagation. If the user answers incorrectly, they are corrected with the right answer and an explanation. They must then attempt to answer a new prompt before proceeding. Once the user answers one prompt correctly, they will gain access to the platform as usual.

**Results**

Three participants (hereafter referred to as A, B, and C) agreed to view mockups of our system, answer two prompts as if they were users, and provide their feedback. Participants were shown the initial prompt screen before the actual quiz questions. Participants A and B had no

significant reaction to this first view, and proceeded to the activity. Participant B, however, expressed a negative initial reaction to this prompt screen, expecting the following activity to consist of a topical lecture rooted in some divisive issue. Referencing his distrust for social media platforms like Facebook and their political biases, he believed that many users would share these sentiments and react with concern. However, once he viewed and participated in the actual quizzes, his perception of the activity became quite positive. He explained how it defied his expectations by focusing on generified misinformation and the tactics used by its perpetrators, instead of specific instances of established real-world political issues.

Indeed, the content populating our example quizzes was intended to be detached from concrete or partisan instances of political discourse, but participant B felt that the reference to GMO giant Monsanto in one of our second prompt's answers "makes it political." Participant C also expressed approval of a non-partisan approach. We have since altered our mockup accordingly, replacing the reference with a fictional company name (see Figure 2).

Participants A and C chose the correct answers for both prompts. Participant B answered one prompt incorrectly, admitting he "didn't look closely enough," but expressed understanding after reading the explanation. All participants understood and approved of our approach and example activity and agreed that completing such an activity at roughly a daily interval would not hinder their user experience. Ultimately, they believed that using our system could benefit their ability to recognize misinformation online.

**Conclusion**

We designed a system intended to combat the dissemination of false information on social media and its disastrous consequences for society. The system uses the concept of active inoculation to educate users on the manipulation techniques that drive the spread of

misinformation. This concept has significant support in the psychological context, as well as the game *Bad News*, which is the only tested implementation of it with regards to false information. *Bad News* research showed improvement between the participants ability to recognize false information in all of the technique categories and based on other persuasion resistance research, the improvements will be multiplied across users over time (van der Linden, Roozenbeek, & Compton, 2020). By applying active inoculation to the user experience of social media platforms, our system can increase the number of users inoculated against misinformation as well as provide them with more frequent "booster shots." This will increase the ability of social media's user base as a whole to recognize false information and therefore curtail its sharing. In addition, the feedback towards our demonstration was overwhelmingly positive, especially given the diverse perspectives of the volunteers. Universally, they agreed that this system would not impede their normal routines of browsing social media and would help them be more aware of the manipulation techniques employed by misinformation. Overall, this system applies proven strategies to mitigate the issue of false information circulating on social media, and it shows promise to make a positive impact on the resistance of users to deception.

**Future Work**

Given a reasonable amount of time and resources, fully implementing our system is certainly feasible. Social media companies could work with our team to integrate the system into their platform in a tailor-made fashion. Furthermore, we could expose a simple API to allow independent developers to integrate our Misinformation Awareness Activity into their software.

Useful data could be collected from users' performance in the activity. Over time, especially if the system is integrated into any popular platforms, these metrics could provide much insight into the perspectives of various end-user demographics, especially towards digital

deception, the strategies behind it, and our campaign against it. This data could be an additional incentive for digital platforms to integrate our tool, since companies may find value in gauging its users' susceptibility to misinformation.

## References


Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Atodiresei C-S, Tănăselea A, Iftene A. (2018). Identifying fake news and fake users on twitter. Procedia Comput. Sci. 2018;126:451–461. doi: 10.1016/j.procs.2018.07.279.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. Computers in Human Behavior. 116. 106633. 10.1016/j.chb.2020.106633.

Chaffey, D. (2020, August). Global Social Media Research. Smart Insights. Retrieved from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Chou, W.-Y. S., Gaysynsky, A., & Vanderpool, R. C. (2021). The COVID-19 Misinfodemic: Moving Beyond Fact-Checking. Health Education & Behavior, 48(1), 9–13. https://doi.org/10.1177/1090198120980675

Dafonte-Gómez, A. (2018). Audience as Medium: Motivations and Emotions in News Sharing. International Journal of Communication (19328036), 12, 2133–2152.

de Beer, D., & Matthee, M. (2020). Approaches to Identify Fake News: A Systematic Literature Review. Integrated Science in Digital Age 2020, 136, 13–22. https://doi.org/10.1007/978-3-030-49264-9_2

De Francisci Morales, G., Monti, C., & Starnini, M. (2021). No echo in the chambers of political interactions on Reddit. Scientific Reports, 11(1), 1–12. https://doi.org/10.1038/s41598-021-81531-x

Facebook. (2021). Working to Stop Misinformation and False News. https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news

Google. (2019). How Google Fights Disinformation. https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf?hl=en

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020, August 1). The Future of False Information Detection on Social Media: New Perspectives and Trends. ACM Computing Surveys, 53(4), 68 - 103.

Kumar, S. & Shah, N. (2018). False Information on Web and Social Media: A Survey. Social Media Analytics: Advances and Applications, by CRC press, 2018. https://arxiv.org/abs/1804.08559

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Roth, Y. & Pickles, N. (2020). Updating our Approach to Misleading Information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

Roozenbeek, J., van der Linden, S. Fake news game confers psychological resistance against online misinformation. Palgrave Commun 5, 65 (2019). https://doi.org/10.1057/s41599-019-0279-9

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating Against Fake News About COVID-19. Frontiers in Psychology, 11, N.PAG.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science 359, 6380 (2018), 1146–1151.

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2019). Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. Computers & Security, 83, 106–121. https://doi.org/10.1016/j.cose.2019.02.003