Governance of Digital Platforms in the Modern Age of AI: How to Preserve the Social Media Platform of the Future from the Widespread of Fake Content

Abdallah Musmar

College of Business, University of South Florida, Tampa, FL 33620, [musmar@mail.usf.edu](mailto:musmar@mail.usf.edu)

He Zhang

College of Business, University of South Florida, Tampa, FL 33620, [hezhang@usf.edu](mailto:hezhang@usf.edu)

Sunil Mithas

College of Business, University of South Florida, Tampa, FL 33620, [smithas@usf.edu](mailto:smithas@usf.edu)

Balaji Padamanbhan

College of Business, University of South Florida, Tampa, FL 33620, [bp@usf.edu](mailto:musmar@usf.edu)

**Abstract**

The reliability of information has been an increasing concern with the new technology capabilities that can create fake information that is hard to distinguish from real information given the new era of AI. We study the effect of incorporating fake news in a social network on the perception of real news in social media. Our study uses agent-based modeling to create an environment that resembles a social media site where users consume and react to news articles from different sources. Each user will determine if the news articles they get in their feed are fake or real by training a classifier that simulates the decision each user will make when they read news articles. Our preliminary results highlight the effect of inducing fake news articles on the perception of real news. We noticed a high rate of misclassifications as many agents incorrectly classify fake and real news articles. Finally, we propose a governance framework that reduces misinformation concerns that arise from unreliable information being distributed in social media.

**Keywords:** AI Governance; Social Media; Fake News; Agent-based Modeling

# Introduction

There have been many advancements in the field of artificial intelligence in the last few years that affects the way we process information. Mainly, AI has contributed in the generation and distribution of content that would potentially be shared among millions of users around the globe. Among the many benefits that can be realized from this advancement comes, there is a concern about how it would become increasingly difficult to distinguish real information from fake information. The resulting ambiguity could potentially affect the perception of the end user on how they perceive the real world. And, on a larger scale, a fake event or a series of fake events can potentially affect the overall perception of reality for a larger community or sub-community of users.

One category of fake content that can affect people is the generation of fake news where a news source is incorporating them along with real news that has actually happened in a user’s news feed. We distinguish this type of misinformation from other types of misinformation such as rumors, clickbaits, satire, as they do not serve the same purpose. The intention of distributing fake news is to mislead users and create an alternative sense of reality. The outcomes of from the spread can be mostly attributed to manipulate users to influence their decisions they make based on this information.

Before the era of online news consumption, the task of checking the quality of information would have been challenging and tedious given the scarcity of resources available at the time to the individual and the laborious amount of effort required to assess truthfulness of news articles. Recent evolutionary advances in technology have transformed the methods of digital communication which has consequently changed the methods of news distribution and consumption. We observed over the last twenty years how modern information distribution channels repositioned their strategy to distribute news on online sources as a response to the migration of the users who became less interested in consuming news via newspapers, radio, or television. In this paper we analyze the relationship of incorporating fake news on the processing of real news. Along with the analysis, we propose a governance framework that will attempt to resolve some of the potential issues that arise from spreading fake information in a digital platform. We think this research contributes to the discussion of discovering the relationship between processing fake news and the accumulated perceived trustworthiness of news.

The main contributors to the change in news distribution and consumption are social media sites. Nowadays, a person can almost effortlessly know what happens in other areas of the world and become emotionally affected by what happens in them on a faster and larger scale than ever before. Therefore, we noticed many researchers attempted to explain how news travels on social media in a follower-followee network (Bastos et al. 2012). To highlight the effect of news distribution on social media, one study shows how the unusual activity on Twitter pertained to a certain stock ticker has a significant correlation with the same stock market activity (Tafti et al. 2016).

One of the great advantages of social media is the ability to spread content on high scale and speed across different communities. However, information quality is an issue that has been associated with social media outlets since these sites combine informal content which has questionable reliability along with content generated from news channels that rely on higher standards of truthfulness. Therefore, there is an undeniable amount of “noise” that distorts the “signal” which is real news being shared social media. Researchers have focused on different angles of this problem to explain part of the “noise” that can be reduced. One area of studies focused on detecting events on Twitter (Atfeh et al. 2015) while another popular area was to explore the network of users and explain the diffusion of rumors and users who might be spammers(Lee et al. 2015; Benevenuto et al. 2010).

In an attempt to find the difference between noise and signal. We argue that noise can be categorized into news articles that are: fake, rumor, clickbait, misleading information, and satire/comedy. Previous efforts have used a subset of these combinations (Waldrop 2017; Rubin et al. 2015; Tandoc Jr. et al. 2017). There are potentially more categories of noise that can be added to this list. To separate between these terms, we have searched in previous literature for each of these definitions and tool them into account when designing our simulated environment. “Fake news” on social media is a term used to describe an act of disinformation and can be defined as information intentionally fabricated for the purpose of deception (Volkova et al. 2017). In alignment with the previous definition, we define an event with fake information in this sense as an event that has been generated but did not actually happen, however, this event can be perceived in a way that is considered to be very close and almost indistinguishable from an event that have actually happened. In other words, the information being shared is completely false and has been intentionally made up as a real news article to deceive users. We define a rumor as a news article that has not been verified and lacks a source of credibility. Therefore, its truthfulness can neither be confirmed nor denied. We define clickbait as an article that masquerades as an interesting news article in the title, however, the content of the article has much less interesting information than what the title suggests. We define misleading information as a news article that mixes a real event with information that is incorrect, therefore, it is designed to drive the consumers to conclude and obtain false information. Finally, we define satire news as news that are intentionally designed to be fake, but the audience in general are aware that it is fake and like to read it for the sake of entertainment.

Since social media outlets will distribute all of these different types of news articles along with real news, we expect to see an effect on how users perceive real news. One potential outcome we argue will happen is that the users will no longer accept news articles that even are real and will question the truthfulness of all news articles being distributed if it is being shared on social media. In order to model this behavior, we have designed an agent-based model that will define the users, news channels, and social media outlets as agents. The users will read news articles and then determine if it is fake or real. We will observe a spill-over effect when we see a large number of users who interact and decide if the news is fake or real. As for the news channels, we will design the news channels to distribute news given that they have a certain threshold of credibility as a source, some of the news channels will have higher thresholds than others as there are online sites that have the intention to deceive users. We would like to examine how changing the number of fake news induced can affect the overall perception of news. Also, we would like to investigate if there is a news channel with a high threshold yet still has a decreasing source credibility score because of other news channels that do not have high thresholds and share news on social media. Finally, we design the social media site as a place that shares news articles from news channels.

The next step in our paper is to implement a governance framework where certain policies can govern the social media platform in order to prevent the widespread of fake information or at least contain the issues that arise from it. In our framework, we propose possible rewards or penalties that can be administered to news sources in order to hinder the propagation of fake information. Therefore, the aim of the framework is to shape the social media environment of the future. We envision this platform as a platform that can deliver reliable news and allow the users to engage in with less interference from malicious news sources that have ulterior motive when distributing fake information.

The rest of the paper is structured as follows: A literature review that summarizes previous research in fake news characteristics, detection methods, and governance platforms. Then we present our research method along with questions and hypotheses. Then we present the design of the agent-based model in detail and highlight the parameters of the model. In the next section we present our proposed governance framework. In the following section we discuss the results of the model along with the proposed governance policies. Finally, we discuss future steps that can be included in upcoming research.

# Literature Review

Many research articles investigated the spread of rumors and disinformation on social media. A recent article by (Vosoughi et. al. 2018) argues that fake news spread farther, faster, deeper, and more broadly than the truth in all categories of information. The authors also concluded that contrary to conventional wisdom, humans and robots accelerate the spread of real and fake news at the same rate implying that fake news spreads more than real news because humans, not robots are more likely to spread it. On the other hand, we have found another study that contradicts with the previous one and highlights the impact of social bots. The study shows that bots have significantly contributed in the spread of low-quality content on social media (Shao et al. 2018).

From previous research, we found three constructs related to disinformation in general and more specifically fake news. These constructs are user’s cognitive biases, analytical ability, source credibility, and the content features of the new article. As for cognitive bias, we have noticed that it is defined as where the individual seeks to confirm their own beliefs and interpret information from their own view of the world (Lazer et al. 2019). One study from the University of North Carolina at Charlotte has examined the effect of anchoring-bias on the individual’s decision making. The study found that providing visual anchors and strategy cues can greatly affect the user’s confidence but has mixed effects on the user’s speed and decision making (Wesslen et al. 2018). Another short paper shares preliminary results where there is an influence of cognitive style on the attitude of sharing content on Twitter (Haug and Gewald 2018). This implies that many users are affected by their beliefs when they want to share content.

Burkhardt (2017) argues that the structure of social networks allows a user to be surrounded by a group of friends and followers that share high similarity between their beliefs, hence, the information that is presented to the user is not objective not to mention that it was provided without them making much effort to search for it. Another study from Yale University has shown that the illusory truth effect has a significant impact on the believability of news. According to the study, even a small amount of plausibility is adequate for repeated efforts to achieve believability of the news (Pennycook et al. 2018). This effect can be easily applied to the context of social media where the user will repeatedly receive fake news on their channel hence it will affect their overall judgement. All these psychological factors contribute the engineering of false information which can be customized to the user’s beliefs. To highlight the importance of cognitive bias even when the user is aware of the probability of disinformation, Valdez and Ziefle (2018) show that there is little effect of showing the user the result of fact-checking badges on the user’s believability of the news.

Although not mentioned in previous studies as a literal construct, we have observed that the user’s analytical ability has an impact on their perception of news. In another study from the University of Wisconsin-Madison, the researchers have found that being misinformed is a function the person’s ability to detect falsehood. They also found that the group and society factors can increase the individual’s chances of being exposed to correct information (Scheufele and Krause 2019).

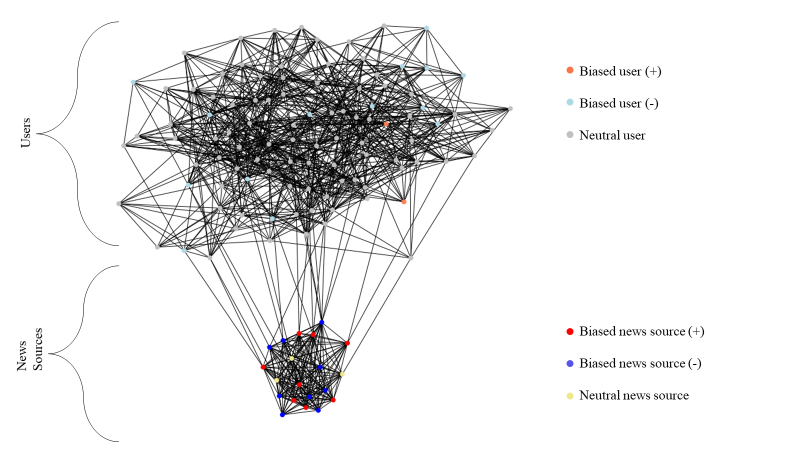
As for source credibility, many research articles have used this construct for detecting false information on the web. For example, Kumar and Geethakumari (2014) have used this construct to find a general acceptability score for tweets. Another study has found that source credibility affects the presence of echo chambers in social media which creates groups that eventually agree on their polarized opinions (Shu et al. 2017). This polarization of social groups was also demonstrated in another study that have found that the internet creates balkanized communities, so it is an inherent property in the structure of social networks (Van Alstyne and Brynjolfsson 1996). Source credibility is also used as a feature in fake news detection algorithms along with content features. For example, a short text classification study shows how user attributes along with content features can be used to detect the category of a tweet (Sriram et al. 2010). Another study was able to detect hoaxes on Facebook with classification accuracy of 99% using the information of the accounts that ‘liked’ the posts (Tacchini et al. 2017).

As for content features, we have seen a few studies contribute to the fake news research by providing a dataset that has labeled news as fake and real. One example is Wang (2017) who provided the LIAR dataset. Horne and Adalı (2017) differentiated between fake news and real news by highlighting the differences in the structure of the articles. The researchers found significant differences in the structure, title, and repetition of words in fake news and real news. Kumar and Shah (2018) have summarized multiple research efforts and have found common characteristics attributed to fake news. These characteristics are twofold: user characteristics and text characteristics. As for the user characteristics, the researchers found that fake news usually have an incoherent body and title, tends to be shorter than real news, packs the majority of information in the title, has a more repetitive nature, contains less technical words, contains smaller words, and are generally easier to read. As for the user characteristics, the researchers have found that the users have generally adopted more recent accounts that were called “throw-away accounts” and that bot accounts that spread false information usually have close distances to each other when modeled in a graph and appear as groups with significant overlap. Another study has used linguistic features such as absurdity, grammar, and humor to detect potentially misleading news (Rubin et al. 2016). Finally, a study from Pacific Northwest National Laboratory has used tweet content and social network interactions to classify news as verified or suspicious (Volkova et al. 2017).

# Research Method

The research method is based on creating a simulation that resembles real-life scenarios when users react to news from different sources on a social media site. The reactions of the agents to the news posted on the social media site will be that they determine if a news article is fake or real. The classifications of news events that happen in the mind of the user are captured in our model by incorporating previous constructs that were introduced in prior literature. We were able to estimate these classifications by incorporating the important features that were used to build fake news classifiers using machine learning algorithms that are trained to classify news articles based on the most important features found in previous literature.

The figure below illustrates the concept of recreating a social network in the interest of learning about news events from the users. We have divided the users and news sources each into three types: one neutral type and two types that are biased in opposing directions. In this division we are assuming that all users and news sources are interested in the news topic. For example, if we were to create a simulation of financial news, then we assume that every node created in the network is interested in financial news. The users and news sources that are not interested will simply be not considered since both news sources and users are not involved in that particular subject and do not have an interest to participate. Then, we considered the bias of opinion as a main contributor to the structure of the network. Following the financial news example, a user who has beliefs that are strongly negative in the financial market will probably have a different take and news feed than someone whose beliefs are extremely positive. We take this into consideration by capturing the bias in both news sources and user along with the ability for any of them to not have a bias and have a neutral stand.



In our simulation design, we can control the percentage at which fake news is being generated. We attempt to gradually increase the number of fake news articles in the environment and observe the number of misclassifications. We expect to find an increasing rate of misclassifications that can be explained by inducing fake news into the environment.

Although there are many prior research articles that had major contributions in defining fake news, characterizing fake news articles, distinguishing them from other types of disinformation and misinformation, explaining the spread of fake news, and detecting fake news using many different classification approaches, we believe that the effect of fake news on processing real news in a complex system has not been fully investigated in prior literature.

By designing the environment to resemble the essence of a digital platform that contains news sharing and user reactions on social media, we aim to find the effect of increasing the number of fake news being induced in the environment on how users react to real news. Therefore, our first research question is investigating this effect and can be formally defined as:

***Research Question 1:*** *What is the effect of injecting “fake news” in a user’s news feed on the users’ decisions when processing “real news”?*

To validate the hyperparameter selection, we studied a large number of tweets we collected from Twitter that are regarding news on the cryptocurrency market from May 2018 to April 2019. During that time, we collected more than 1.5 million tweets querying four popular cryptocurrencies: “Bitcoin” “Ethereum” “Ripple XRP” and “Litecoin” to collect the data. Then we performed Latent Dirichlet Allocation on samples from different time periods to understand the topics that were discussed during the 11-month period. We also used “SentiStrength” which provides sentiment analysis on short and informal text to find the differences in sentiment values (Thelwall et al. 2010).

In order to address this question, we aim to construct a social media environment that contains news sources and users. In this simulated network, a news sources could have the intent to shape all of users’ perceptions or a subset of them. To test this effect in detail, we have proposed the following hypotheses:

***Hypothesis 1:*** *Increasing the percentage of “fake news” shared on a social media platform to a certain configuration increases the percentage of users incorrectly classifying real news as fake news*

***Hypothesis 2:*** *Increasing the percentage of “fake news” shared on a social media platform to a certain configuration increases the percentage of users incorrectly classifying fake news as real news*

***Hypothesis 3:*** *Increasing the percentage of “fake news” shared on a social media platform to a certain configuration leads to creating more polarized communities and sub-communities*

After designing an environment that resembles a digital platform and ensuring that it mimics reality, we want to use the environment as a mechanism to investigate configurations of governance policies that will aid into reducing the effect of fake news on the perception of real news in the network. Therefore, our second research question can be formally defined as:

***Research Question 2:*** *What possible configurations can be implemented on a digital platform in order to reduce the user’s perception of real news as fake news?*

It is a common convention nowadays for a digital platform to provide their users with news updates from a variety of news sources without much restrictions or limitations. Therefore, there is an expectation of heterogeneity in the quality of the news being shared on the users’ feed. We think some of the reasons behind the heterogeneity in news quality could be due to the difference in journalism standards practiced on social media and the standards practiced among traditional news environments. In social media, reporting news events could be much simpler as it does not require thorough source validation. This means that reporting unreliable news events can happen easily since there are many news sources who will not choose to follow a rigorous journalism standard. Another reason why we think of the heterogeneity in news quality is because of the lower barrier to grow in the network and create an audience that will tune in for news in social media whereas traditional journalism venues do not gain mass followers easily and quickly. The two previous reasons could result in many news sources fabricating news events, and the fake news events will be shared on the network and could potentially spread through multiple communities on the social network. Therefore, we can conclude that unreliable news sources have the potential to become popular and spread viral news events that are questionable or might even be completely false. Nowadays, the digital platforms are not following guidelines or policies to govern fake news and real news sharing. For example, in October 2019, in the midst of a discussion regarding Libra, the new cryptocurrency project of Facebook, the CEO of Facebook Mark Zuckerberg pointed out that Facebook’s current direction on political ads is not to remove the ads that have disinformation in them. However, he had mentioned that there is some degree of monitoring shared content (Kang 2019). We think that this example implies that social media networks are not taking enough action to create a healthy environment for spreading more reliable news.

While there are current discussions on social media platforms considering paying for news sources to reduce the amount of unreliability in news updates, we still have not observed to this moment any action beyond the call for social media platforms to take responsibility of reliable news sharing instead of relying solely on the user’s preferences and judgement. We propose a governance framework that offers an incentive or penalty to news sources based on user perception of how the proportion of fake news to real news that comes from each news source. In our simulation, the users’ perception will be estimated by the classifier that mimics the behavior of the user.

Hypothetically, when a news source is consistently sharing content that is not fake, the social network will provide an incentive to the news source in order to reward the practice of reliable journalism. On the other hand, if another news source shares fake content either because for it not following a reliable journalism standard or having malicious intent and deliberately creating fake content, then the network penalizes the news source in a manner that reduces its effect on the network and limit the spread of fake content. The incentive can be applied after sharing a certain number of news events so it can be addressing a collective behavior and not treat honest mistakes as bad practice or malicious intent to disinform the public. We argue that this mechanism should be implemented in a way that does not interfere with freedom of expression, but rather provide a higher chance for the user to have real news on their feed and less chance to spread fake news across the social network.

In designing our governance framework, we are considering three main areas to focus on in order to obtain our goal which is a digital platform that enables the spread of reliable news and maintains the users’ ability to express themselves. The three main areas of consideration are allowing news sources the ability to improve their reporting, consider a relative component for the rewards and penalties as news sources will be compared to each other in terms of news quality, and finally the impact of sources on each other. As an example of how a news source can impact another news source, if we consider two news sources A and B sharing two different news events on a user’s news feed in the same time. And we assume that news source A has precedence in a user’s news feed with no control from neither the news sources nor the user, then we argue that fake news shared by news source A has an effect on the perception of the news shared by news source B.

|  |  |
| --- | --- |
| **Guideline** | **Research** |
| Design as an Artifact | The artifact designed in this paper is the social network environment which will be tested under several governance frameworks |
| Problem Relevance | The artifact is designed and instantiated to address a current problem on social media with the spread of fake news and its effects on the users. There is also potential for this problem to grow in the near future with more advances in AI. |
| Design Evaluation | The evaluation of the model is determined by comparison of the data that is generated with real data that was obtained from social media |
| Research Contributions | The first contribution is the simulation environment which can generate new data on user activity given model settings of the social media network dynamics. The second contribution is the governance framework that ensures that the mass effect of fake news is contained or remediated with some configurations from the network. |
| Research Rigor | Building on complex systems, fake news classification, and AI governance literature. This research proposes a framework to illustrate the effects of disinformation on a social network and the potential remediations to reduce the negative effects associated with disinformation. |

## Agent-Based Model Design

In our design, the first step we have made was to define the elements of the simulation. Therefore, we have defined three components that will be important for the simulation to capture reality: news articles, news sources, and users. We will explain the characteristics and actions of each component in this section along with the overall steps needed to start the simulation.

As for news articles, we found during our review of the literature that focused on fake news characteristics four important features for determining if a news article is fake. We have found that source credibility is an important construct that is associated with believability (Valdez and Ziefle 2018). We also found the user’s analytical ability along with their motivation to have an impact on the user being able to determine if a news article is fake or real (Scheufele and Krause 2019). Lastly, another important component of determining if a news article is fake or real is the content itself mainly the sentiment of the title headline along with the number of shares for the article. Generally, fake news has a higher tendency to include a news headline with an extreme sentiment that would be much higher than the sentiment of real news articles. We also found that fake news articles typically have a significantly lower number of shares than real news articles.

The table below contains a summary of the most important features we found in prior literature that would be required to classify a news article as fake or real:

|  |  |
| --- | --- |
| **Feature** | **Feature details** |
| Sentiment of the news article headline | Score is on a scale of (-5,-1) for negative sentiment and (1,5) for positive sentiment |
| Number of shares | Count of other users who have shared the news article. It must be a natural number. |
| Source credibility | A percentage that reflects the user’s perception about the source of the news article being a reliable reporting agency or not. |
| Analytical ability and motivation | A binary value where 1 means the user is analytical and motivated, and 0 means the user does not have the ability or motivation. |

In our design, the user predicts an article to be either fake or real. In order to achieve this, we train a classifier for each user to in order for them to distinguish fake news based on the features that are associated with the news article, news source, and the users internal features. Another aspect of our design is that a news article will be known to the users that it was fake or real after an amount of time. We think of this setting as the most probable case in reality. This can be implemented in our design as follows: after a certain number of steps where multiple news events have taken place, the user develops an understanding about the events that happened in recent history and consequently trains their fake news classifier on the new data they have observed and what they have believed to be fake or real. Only a small fraction of users will continue to believe a completely fabricated news article did actually take place after an amount of time passes by.

Based on our analysis of a the data we captured from Twitter on the cryptocurrency market, we have initialized our model title sentiment to follow a normal distribution with a mean of 4 and standard deviation of 1 for fake news articles while real news articles were assigned a normal distribution with a mean of 2 and standard deviation of 1. The sentiment scores can have a maximum value of 5 and minimum value of 1. For the moment, we regard negative and positive sentiment to have the same effect on the article being fake or real. As for the number of shares, we have initialized our model to follow a normal distribution with a mean of 5 and standard deviation of 2 for fake news articles while real news articles were assigned a normal distribution with a mean of 50 and standard deviation of 2. As for the number of shares, we have initialized our model to follow a normal distribution with a mean of 5 and standard deviation of 2 for fake news articles while real news articles were assigned a normal distribution with a mean of 50 and standard deviation of 2.

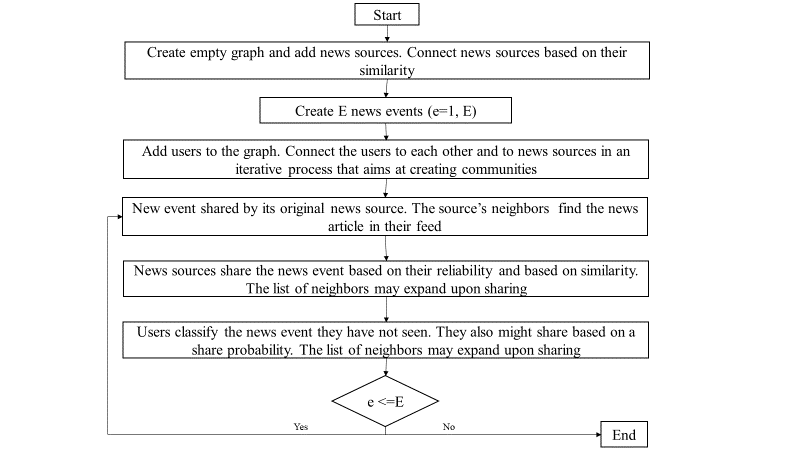
As for the news sources, every news source has a binary flag that captures reliability of the news source. When this flag is set to one, it means that the news source is practicing journalism in a responsible manner and would only post a news article after it has been vetted thoroughly. In our model, we expect a network to have more unreliable news sources. Generally, finding credible news sources is a difficult task for every user. To capture this distribution in our model, we create news sources with a probability of 10% to be reliable. After determining if a news source is reliable, we create a number of news articles for the news source to be posted. If the news source is not reliable, the articles will have a higher probability to include fake news. This means that the fake articles shared by that source will probably have a higher score for the title sentiment and lower number of shares on average.

As for the users, we have designed the user to have a score that resembles their biases between negative one and one. For our simulation, we have initialized the values of a user bias to follow a normal distribution with a mean of 0.01 and a standard deviation of 2. After generating the user’s bias, we have assigned each user with a score for their analytical ability. We assigned the probability of a user to have the required analytical ability to 5%. We believe that the majority of the users on social media that follow a certain news topic do not have both the analytical ability and the necessary motivation to assess news articles on their own. We think that from prior research, the majority of the users are casually browsing the network and are not focused enough and motivated enough to read news articles thoroughly. The maximum value of analytical ability is one and the minimum analytical ability is zero.

In the next step we create a source credibility score between zero and one for each news source by the user. We initialize this score to be a weighted difference in bias between the user and news source. Following this computation, a user will consider a news source to be more credible if they share the same biases as their core beliefs. One potential drawback of this calculation is that a user that is extremely biased will not accept a news source to be credible unless it was close to them on the spectrum of bias. Therefore, this could lead to extreme users dismissing credible news source that are neutral or have a bias on the other end of the bias spectrum. A higher score in bias will result in more extreme source credibility scores.

In our final step of designing the user, we include a fake news classifier specific to that user. For the classifier, we have initialized all users to have a default classifier that was trained on a balanced dataset of news articles, news sources, and users. The reason behind this decision was to avoid misclassifications due to training issues with the classifier and narrow the reason for misclassification to the features that the classifier is trained on. As for the type of classifier, we have chosen a random forest classifier. The reasons why we chose random forest was due to it being an ensemble method and generally a good performer on this type of a machine learning task.

The flow chart below explains the steps of running our simulation on a high level. We consider each news event generated as a new step in the simulation. :



The simulation models two main decisions the users and news sources make during every step. These decisions are whether they want to spread it or not and their own judgement of the news event that is being shared on their news feed. The decisions made inherently affect other nodes in the network which makes the problem of understanding the mass effect of fake news more complex. However, after a small period of when the step is over, the nature of previously shared news articles is revealed, and it will become common knowledge across the mass. However, we do not expect this knowledge to have an effect on the bias of a user or a news sources at least on the short term. But we do observe the changes in beliefs over many steps and expect a change over a relatively long period of time.

## Governance

Filler

# Preliminary Results

Our first simulation highlights some interesting results in the misclassification rates. We have fixed all the parameters and only increased the rate of which a news source with deception intent will generate fake news from 5% to 60% by 5% increments. In order to ensure the robustness of the logistic regression model classifications, we have sampled the training dataset to have a balanced size for each value of the target variable which is the fake news variable. We have undersampled the training dataset to have the number of observations with real news match the number of observations with fake news. The results are the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Percentage of fake news | Training size | Testing size | F1 score | Misclassifications rate |
| 5% | 400,000 | 120,000 | 0.57 | 43% |
| 10% | 440,000 | 132,000 | 0.53 | 46% |
| 15% | 560,000 | 168,000 | 0.64 | 35% |
| 20% | 1,200,000 | 360,000 | 0.56 | 43% |
| 25% | 980,000 | 294,000 | 0.70 | 30% |
| 30% | 440,000 | 132,000 | 0.59 | 41% |
| 35% | 1,800,000 | 540,000 | 0.56 | 40% |
| 40% | 1,520,000 | 456,000 | 0.70 | 29% |
| 45% | 1,780,000 | 534,000 | 0.67 | 30% |
| 50% | 1,760,000 | 528,000 | 0.80 | 19% |
| 55% | 1,660,000 | 498,000 | 0.80 | 19% |
| 60% | 2,140,000 | 642,000 | 0.75 | 24% |

Table . Logistic Regression Results

We notice from the results that the misclassifications rates are at least 19%. This means even if the population in general is able to distinguish between fake news and real news. There is still a considerable amount of users who will misinterpret real news as fake and fake news as real. This can be highlighted further when we compare the misclassification rates by category to see if there is a difference between fake news predicted as real news and real news predicted as fake news. The results are in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Percentage of fake news | Rate of misclassifications | Fake news predicted as real news | Real news predicted as fake news |
| 5% | 43% | 21% | 22% |
| 10% | 46% | 16% | 30% |
| 15% | 35% | 13% | 23% |
| 20% | 43% | 17% | 27% |
| 25% | 30% | 10% | 20% |
| 30% | 41% | 18% | 23% |
| 35% | 40% | 6% | 35% |
| 40% | 29% | 4% | 25% |
| 45% | 30% | 1% | 29% |
| 50% | 19% | 3% | 16% |
| 55% | 19% | 8% | 12% |
| 60% | 24% | 3% | 21% |

Table . Misclassification Rates by Category

We can see from the table that the percentage of real news classified as fake news is almost always higher than fake news classified as real news. This can be explained as that the effect a news source with an intention to deceive the public is aiming for is to generally let the public trust real news less by masquerading as a reputable news source and induce fake news that will confuse the user’s perception of real news. This effect can be best demonstrated in this simulation when the percentage of fake news within a deceiving news sources is set to 35% of their generated content.

# Further Research

For further research we plan to extend the simulation to include multiple time steps, more machine learning algorithms, and investigate network effects. As for including multiple time steps, we think that this step will demonstrate how fake news can have an effect on processing real news over time. By determining how users react to news sources and change their perception of the credibility of a source, we think that this might be an interesting avenue for further analysis. As for the machine learning algorithm, we are implementing other well-known classification methods such as k-Nearest Neighbor, Random Forest, and Support Vector Machines to compare the misclassification results and see if the type of classifier has an effect on our analysis. Finally, for the network effects, we think that including a network of users in the model design will add a feature that is currently limiting our analysis which is the effect of the group on a user’s perceptions. By adding the network graph, we can be able to study how user’s share fake news and real news if they have believed them and add another layer of analysis to the model.

# Limitations

This research has several limitations which are in part due to how the model will mimic a real social network. In our study, we are limiting our analysis to one hypothetical social network. We also designed the model to distribute news only on a certain topic that we assume all the users were interested in and have read the news article titles at the very least. We also have not included network effects thus not capturing the effect of the group on the user, but we plan to do so in the near future. We also have limitations on the machine learning model as we have designed the logistic regression model to mimic all user’s classifications given the constructs that were used in the model design.

# Conclusion

In this research, we aim to study the effect of inducing fake news in a social media site on the perception of real news for the users on the social network. Preliminary results show that there are higher misclassification rates for real news predicted as fake news which means that inducing fake news affects the user’s believability of real news. The results also show that misclassification rates are the highest when fake news is induced by deceiving news sources at a rate of 35%. This means that news sources that have an intention for deceiving the public can mix fake news with real news and generally confuse a considerable amount of the users.

This research contributes to the modeling of disinformation on social media as it combines research on fake news characteristics from a psychological perspective along with mining fake news using machine learning from a data science perspective. We have also created an environment that simulates social media from basic Python libraries which is a contribution to agent-based modeling in the sense of modeling fake news on social media. For the future, we aim to add multiple time steps in our simulation to capture a changes in source credibility. We also aim to add more machine learning algorithms to compare classification rate results. Finally, we will add networking effects to investigate the effect of a user’s group of friends and followers on their perception of fake news and real news.

# References

Atfeh, F., & Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence, 31*(1), 132-164.

Bastos, M. T., Travitzki, R., & Puschmann, C. (2012). *What Sticks With Whom? Twitter Follower-Followee Networks and News Classification.* Association for the Advancement of Artificial Intelligence.

Benevenuto, F., Magno, G., Rodrigu, T., & Almeida, V. (2010). Detecting Spammers on Twitter. *7th annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference.* Redmond, Washington.

Burkhardt, J. M. (2017). *Combating Fake News in the Digital Age* (Vol. 53). Chicago: ALA TechSource, a unit of the publishing dept. of the American Library Association.

Calero Valdez, A., & Ziefle, M. (2018). Believability of News Understanding Users Perceptions of Fake News and the Effectiveness of Fact-Checking Badges. *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, (pp. 469-477). Florence, Italy.

Haug, M., & Gewald, H. (2018). Are you the Reason Fake News exists? Investigating News Sharing Attitude on Twitter. *Thirty Ninth International Conference on Information Systems.* San Francisco.

Horne, B. D., & Adalı, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *CoRR.*

Kateb, F., & Kalita, J. (2015). Classifying Short Text in Social Media: Twitter as Case Study. *International Journal of Computer Applications, 111*(9).

Kang, C., Isaac, M., & Popper, N. (2019). Charm Offensive Meets House Bombardment. *The New York Times,* https://www.nytimes.com/2019/10/23/technology/facebook-zuckerberg-libra-congress.html.

Kumar, K. K., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences, 4*, 1-22. doi:10.1186/s13673-014-0014-x

Kumar, S., & Shah, N. (2018). False Information on Web and Social Media: A Survey. *Computing Research Repository (CoRR), 1*, 1-35. Retrieved from https://arxiv.org/abs/1804.08559

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., . . . Zittrain, J. (2018). The science of fake news. *Science, 359*, 1094-1096. doi:10.1126/science.aao2998

Lee, J., Agrawal, M., & Rao, H. (2015). Message diffusion through social network service: The case of rumor and non-rumor related tweets during Boston bombing 2013. *Information Systems Frontiers, 17*, 997-1005.

Pennacchiotti, M., & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter Classification. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, (pp. 281-288). Sunnyvale, USA.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology. General, 147*(12), 1865-1880. doi:10.1037/xge0000465

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception Detection for News: Three Types of Fakes. *The Annual Meeting of the Association for Information Science and Technology.* S.t. Louis, Missouri.

Rubin, V. L., Conroy, N. J., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *Proceedings of NAACL-HLT*, (pp. 7-17). San Diego, California.

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences of the United States of America*, 1-8. doi:10.1073/pnas.1805871115

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 1-16. doi:10.1038/s41467-018-06930-7

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorations, 19*, 22-36.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short Text Classification in Twitter to Improve Information Filtering. *Special Interest Group on Information Retrieval.* Geneva, Switzerland.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). *Some Like it Hoax: Automated Fake News Detection in Social Networks.* Technical Report UCSC-SOE-17-05 School of Engineering, University of California, Santa Cruz.

Tafti, A., Zotti, R., & Jank, W. (2016). Real-Time Diffusion of Information on Twitter and the Financial Markets. *PLoS ONE, 11*(8).

Tandoc Jr., E. C., Lim, Z. W., & Ling, R. (2017). Defining “Fake News”. *Digital Journalism, 6*(2), 137-153. doi:10.1080/21670811.2017.1360143

Thai, M. T., Wu, W., & Xiong, H. (2016). *Big Data in Complex and Social Networks.* New York: CRC Press. doi:10.1201/9781315396705

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558.

Van Alstyne, M., & Brynjolfsson, E. (1996). Electronic communities: Global village or cyberbalkans. *Proceedings of the International Conference on Information Systems*, (pp. 80-98).

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. O. (2017). Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 647–653). Vancover, Canada: Association for Computational Linguistics (ACL). doi:10.18653/v1/P17-2102

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*, 1146-1151. doi:10.1126/science.aap9559

Waldrop, M. M. (2017). News Feature: The genuine problem of fake news. *Proceedings of the National Academy of Sciences, 114*(48), 12631-12634. doi:10.1073/pnas.1719005114

Wang, W. Y. (2017). *“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection.* Santa Barbara, CA: Department of Computer Science, University of California, Santa Barbara.

Wesslen, R., Santhanam, S., Karduni, A., Cho, I., Shaikh, S., & Dou, W. (2018). Anchored in a Data Storm: How Anchoring Bias Can Affect User Strategy, Confidence, and Decisions in Visual Analytics. *CoRR*.