IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 October 2022, accepted 1 December 2022, date of publication 12 December 2022,
date of current version 15 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228258

RESEARCH ARTICLE

Evaluating the Influence of Twitter Bots via Agent-Based Social Simulation

ALDO AVERZA, KHALED SLHOUB, AND SIDDHARTHA BHATTACHARYYA, (Senior Member, IEEE)
College of Engineering and Science, Florida Institute of Technology, Melbourne, FL 32901, USA

Corresponding author: Aldo Averza (aaverza2020@my.fit.edu)

This work was supported in part by the Open Access Subvention Fund, and in part by the John H. Evans Library.

ABSTRACT Social Media is used by many as a source of information for current world events, followed by publicly sharing their sentiment about these events. However, when the shared information is not trustworthy and receives a large number of interactions, it alters the public's perception of authentic and false information, particularly when the origin of these stories comes from malicious sources. Over the past decade, there has been an influx of users on the Twitter social network, many of them automated bot accounts with the objective of participating in misinformation campaigns that heavily influence user susceptibility to fake information. This can affect public opinion on real-life matters, as previously seen in the 2020 presidential elections and the current COVID-19 epidemic, both plagued with misinformation. In this paper, we propose an agent-based social simulation environment that utilizes the social network Twitter, with the objective of evaluating how the beliefs of agents representing regular Twitter users can be influenced by malicious users scattered throughout Twitter with the sole purpose of spreading misinformation. We applied two scenarios to compare how these regular agents behave in the Twitter network, with and without malicious agents, to study how much influence malicious agents have on the general susceptibility of the regular users. To achieve this, we implemented a belief value system to measure how impressionable an agent is when encountering misinformation and how its behavior gets affected. The results indicated similar outcomes in the two scenarios as the affected belief value changed for these regular agents, exhibiting belief in the misinformation. Although the change in belief value occurred slowly, it had a profound effect when the malicious agents were present, as many more regular agents started believing in misinformation.

INDEX TERMS Agent-based modeling, agent-based social simulation, multi-agent systems, social media, twitter, twitter bot.

I. INTRODUCTION

The influence of social media in our daily lives has increased significantly over the last decade. What began as a social network to share thoughts with individuals has evolved into something bigger, with more active users than ever, including politicians, multi-million dollar companies, and advertising agencies. According to Dean [1], 40% of internet users utilize social media for work. In its current landscape, the impact of what is posted on these social media networking platforms has proven to have real political and economic consequences that cannot be overlooked. Twitter is one of the largest and most popular social media networks, with over 186 million daily active users, generating over $3 billion in revenue per year [2]. It allows individuals to share their thoughts through tweets using the 280-character text editor, alongside images, polls, videos, and links. It implements a follower-following system, where individuals follow other accounts to read a compilation of their tweets in a streamlined format.

In 2016, Twitter launched the Twitter API, providing developers with tools to implement many of the regular Twitter functions in their code. This allows the creation of many third-party apps and *Twitter bots*. Twitter bots can post content, respond to others, retweet content, create relationships

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Guidi.

with other users, and direct message to other accounts. They are autonomous accounts that can function with no human interference and interact in the social media environment alongside human users. While many Twitter bots have beneficial objectives, such as automatically informing their followers about earthquakes,[1] there is still a significant problem with bots designed with malicious intent. Many artificially inflate a tweet's user interaction or an account's follower count, while others are part of misinformation campaigns to spread false information on the social media networking site.

In 2011, many extensively used social media to locate victims of the earthquake that hit the east coast of Japan and reached 9.0 on the Richter scale. A study conducted by Takayasu et al. [3] reported that malicious users took advantage of the event to spread misinformation during the state of the emergency. Another study by Cresci et al. [4] reported that Twitter bots were disrupting the stock market by using the *piggybacking* technique, where bots are used to increase the stock price of a large company, increasing the price of subsidiaries as well. Also, Takacs and McCulloh in [5] studied the influence of social media in the 2018 U.S. Senate Election and reported that many Twitter accounts engaging with political tweets were bot users, inflating a politicians follower count and the number of user interactions to shape public opinion of candidates (e.g. more followers leading to think they were more popular). Also, a recent study by Allyn [6] at Carnegie Mellon University found that 45% of the accounts discussing the COVID-19 pandemic indicated signs of bot behavior. It is important to notice that even when the misinformation originates from a bot, the spread is mainly done by human users. A study by Zhao et al. [7] elucidated how the propagation of misinformation tends to reach more communities in comparison to real ones. This begs the question: If human users cause most of the propagation, how much influence do bots have on the beliefs of people?

In this paper, we propose an Agent-Based Social Simulation (ABSS) using agent-based systems to develop a social network similar to the Twitter network to study and analyze the influence of malicious bots in spreading fake information. In this regard, we developed three types of agents: *Deceptive agents*, representing bots or malicious human users; *neutral agents*, representing most Twitter users; and *news agents*, official sources who share accurate, verified information. Through this simulation, we investigated the differences between two distinct scenarios, one with and one without deceptive agents. The main objective behind this approach was to understand the impacts that bots have on the neutral agents when encountering false information. This was accomplished by implementing a belief value attribute, which allows studying and comparing how susceptible neutral agents are towards misinformation. To provide an abstract representation of misinformation, we depict the validity of information through three types of tweets that an agent can generate: neutral, real or fake.

During our research, we realized that the published articles in the literature discussing this subject of research tend to focus on the spread of misinformation through agent simulations; these attempts fail to address certain characteristics of social media, such as communities (large groups of accounts following similar topics) and different types of agents with distinct behaviors. The significance of our research paper has been in the development of a model that closely mirrors the behavior seen in social media, alongside the belief value. The belief value provides a quantitative way to simulate human beliefs. Our main contributions in this paper are as follows:

- The representation of characteristics of social media networks, such as communities, variable number of followers and tweets per agent.
- The validation of the proposed simulation by comparing real-life user interactions with those generated by the agents in the simulation.
- The creation of an approach for comparison between two scenarios to explore how neutral agents interact in a social network with and without deceptive agents.

The rest of this paper is organized as follows: Section II discusses previous related work. Section III provides background information regarding Twitter, bots and Multi-Agent Systems. Section IV gives an overview of the proposed methodology. In Section V we provide the results of the simulation and discuss them in Section VI. In Section VII we validate the data generated by the simulation and, finally, Section VIII concludes and outlines future work.

II. RELATED WORK

Several rumor diffusion models exist based on the Susceptible-Infected-Recovered (SIR) model designed by Daley and Kendall [8] where they compare the rumor spreading to epidemic models, where each agent can be in one of three states: Susceptible, Infected or Recovered. While in the Susceptible state a node is open to be infected by those nodes in the Infected state and, after some time, a cure would spread as well, provided by the Recovered state. While it is tempting to relate the SIR model to rumor spreading in social media, several changes have to be made to accommodate different means of communication and spread of information.

Many authors have expanded on the idea of the SIR model to apply it to Social Media, where misinformation spreading is a large unregulated problem. Serrano and Iglesias proposed a model to validate viral marketing strategies through an ABSS. They explored rumor propagation techniques and created a model to simulate it, called BigTweet [9], which they then released as an open-source software. The main limitation encountered was the ability to handle more than one type of agent, as this is critical for our research focused on malicious users.

Research by Ikeda et al. also use as a base the SIR model to generate an Agent-Based Information Diffusion Model to evaluate data from the previously mentioned 2011 East Japan earthquake. To make the model more robust, they

[1] https://twitter.com/earthquakesLA

introduced the idea of diversity and multiplexing of information paths [10]. Similarly, research conducted by Okada et al. also provide a more robust version of the SIR model, with data from the same event [11]. The results showed by the latter were compared with real Twitter data to evaluate the susceptible, infected and recovered users using metrics such as number of retweets. Similar to other related works, the rumor propagation is the focus, rather than the origin of that false information, which is something we address in our research.

Kundu et al. approach the information propagation model in their work by developing a novel fuzzy relative willingness model. The diffusion model was able to successfully utilize the external influence factor, as well as the susceptibility of individual nodes to quantify human willingness [12]. While the objective of the paper differs from our research, the implementation of the external influence factor contributed to our research as we applied a similar function as well.

Ross et al. validate the concept of the spiral of silence in their research paper [13] through an agent-based model. The spiral of silence explains how the influence of surrounding negative opinions can affect the spread of positive opinions in a social media environment. The way the simulation was built differs from ours, but it provided insight into how users tend to react while consuming negative media. Wang et al. study information entropy, which incorporates new types of variables to the simulation, including the degrees of trusts agents can set between each other [14]. This work, similar to others, centers on rumor spreading models on social media, but focuses on the interaction between two given agents by adding weights between each node to represent trust.

Research by Yan et al. dives deep into the concept of how *retweets* and *quote-tweets* influence the behavior of users in social networks by using the concepts of game theory and developing a reward mechanism [15]. The objective of the research is to study agent behavior given goals, such as a higher number of retweets. This work does not center on rumor spreading or malicious agents, but it proved to be insightful in providing more information about the retweet cycle, which is the means by which information spreads on social media sites, such as Twitter.

Research highlighting bot behavior includes work by Carley [16] outlining the BEND framework as a way of identifying misinformation maneuvers in social media environments. Other studies related to bot behavior can be reviewed in Cresci et al. paper [17] where they compare human behavior in social media similar to DNA sequences and identify bot behavior based on a predictable set of actions. They compare the actions of several Twitter human users and bot users, which they call their digital DNA, and try to identify similarities between them.

Later work by Beskow and Carley highlights two bot misinformation maneuvers: backing and bridging. The former focused on bots interacting with agents with high influence (larger number of edges) to spread false information, while the latter focuses on bridging two communities, with the bots at the center of it. For the model, the authors developed a *twitter_sim* framework as a way of simulating the Twitter social network. They implemented a belief value assigned to each agent in a way of representing the belief of a real user when encountering misinformation on its environment, along with a similarity weight between agents that has an impact on each agent's beliefs. This belief value defines how dedicated each agent is in believing misinformation based on the type of tweet presented [18].

By looking at the literature, we have become aware of the lack of research underlining the intentional spread of false information by malicious users. While many rumor-spreading models were proposed, most of the work did not provide information on the origin of those rumors nor insight into the effects this misinformation can have on the platform users. We address these issues in our ABSS model by defining different types of agents with different roles, behaviors, and objectives.

III. BACKGROUND

A. TWITTER SOCIAL NETWORK

The Twitter network site works similar to other social media platforms, where the contents users consume are derived from the accounts they follow. The primary posting system is called a *tweet*, described as "any message posted to Twitter which may contain photos, videos, links, and text."[2] Twitter generates a *Timeline* showing tweets from all followed accounts and offers the option to engage with a tweet by selecting a *Like* button, commenting on it, sharing it through a *retweet* to forward the tweet to other users, or using the *Quote retweet* feature to add additional text while sharing. Twitter calls these user interactions *engagements*, which are used to measure the total number of users interacting with an individual tweet.

One of the offered Twitter's features is *hashtags*, symbolized by the character "#".[3] This categorizes users' tweets to be part of a much larger conversation by writing a specific tag attached to their tweets. For example: using the "#Covid19" hashtag can direct users to tweets related to the Coronavirus pandemic. Similarly, tweets can also contain *cashtags* (using the character "$") as a way of categorizing conversations related to the stock market, such as "$AMZN" when referring to Amazon. Another Twitter feature is the *Verified* mark used to differentiate those users whose identities have been verified officially by the company. This mark commonly applies to to public figures, such as celebrities and politicians, to avoid impersonators on their platforms.

Twitter implements an algorithm to display contents relevant to its users, called *Top tweets*. While the inner features of this algorithm are not open source, Twitter claims that the targeted tweets are chosen by the tweet's popularity (high engagement) and keywords relevant to the users.[4]

[2] https://help.twitter.com/en/resources/new-user-faq
[3] https://help.twitter.com/en/using-twitter/how-to-use-hashtags
[4] https://help.twitter.com/en/using-twitter/top-search-results-faqs

B. TWITTER BOTS

As stated by Storey and Zagalsky [19], a software bot's objective is to automate actions by either engaging in repetitive tasks, providing assistance through a chat function, or simply simulating a human user. Lebeuf [20] categorizes bots based on their intelligence level and purpose. When it comes to intelligence, a bot can be characterized based on its adaptability to the environment, reasoning to rules, and autonomy from human users. Based on their purpose, they can be categorized as Generalist bots that have a wide range of abilities; Transnational bots that work for users; Informational bots that gather information for users; Productive bots that execute time-consuming tasks on behalf of users; Collaborative bots that improve users communication.

Roth and Pickles [21] describe Twitter bots as automated accounts with the same privileges as regular Twitter users, such as creating, sharing, or engaging with tweets. Their use in the Twitter social network is not inherently malicious or prohibited, and it is even encouraged by the company for customer service applications.

Twitter bots are developed through the Twitter API, allowing developers to automate the creation of tweets, respond to comments, and engage with other tweets through likes and retweets. What sets bots apart are based on the behaviors developers implement in them. As mentioned by Roth and Pickles, the real concern is platform manipulation, which is described as an alteration of the public conversation happening on the social network through fake engagements due to their impact on the human user's perception of public opinion.

Jamison et al. [22] study the common behaviors of Twitter bots and categorize them based on objectives. Some common characteristics of a malicious Twitter bot include extreme marketing campaigns, an inflated number of fake followers, a large number of scheduled tweets, or high levels of retweeted content. However, malicious Twitter users can also be human and present behaviors such as engaging heavily in political content, negative/opposing views, or spreading (retweeting) similar content from other sources.

C. MULTI-AGENT SYSTEMS

The simulation presented in this paper, at the base level, is developed as a Multi-Agent System (MAS) [23]. As described by Kinny and Georgeff [24], MAS environments consist of a series of agents: reactive, autonomous, and internally motivated entities with a defined behavior that can be altered during its execution based on internal or external factors. The simplest example of an agent-based system can be seen in a thermostat, activating heating or cooling systems based on what temperature it captures from the environment.

The interaction between agents is what sets MAS apart from other object-oriented methodologies. These agents can interact with one another, with the environment or directly with a human user, which can lead to changes in its behavior. Figure 1, demonstrates a high-level structure of multi-agent systems. In MAS environments, agents require a series of characteristics that define their behavior in an environment. As explained by Slhoub et al. [25], these are: Autonomy (able to act on their own), Reactivity (acts based on what it perceives from the environment), Proactivity (works to achieve a specific goal) and Sociability (interacts with other agents)

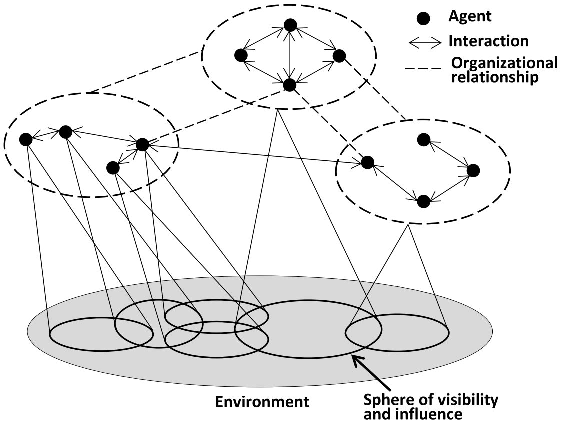

● Agent
<-> Interaction
--- Organizational relationship
Environment
Sphere of visibility and influence

FIGURE 1. High-level structure of multi-agent systems [23].

By definition, many common factors exist among agents and bots, but what sets bots apart is that they are typically implemented in environments where humans interact, as mentioned by Lebeuf et al. [20]. For this reason, non-human users are typically described as "bots" when it comes to social media, while the "agent" term is used more predominately in simulation-based experiments.

D. GRAPH THEORY

A graph, in its simplest form, is a set of points (vertices) connected by a set of lines (edges), as defined by Gibbons [26]. Any graph G, regardless of what they represent, is given by the formula $G = (V, E)$, with V representing the number of vertices and E the number of edges.

Graphs are commonly used to represent social networks due to similarities in their design; vertices represent people, and edges depict the connection between them. For this reason, graphs are widely used when simulating epidemic spreading models and social media environments. Graphs are also used in other fields, such as contact tracing, cybersecurity (interaction between IPs), fraud detection, recommendation algorithms, and network routing. These are connected by the relationships between the nodes and are perfect examples of how information can be presented.

Several types of graphs can be used to display the data based on what is being evaluated. It can be as simple as presenting only nodes and connections, while it can also be weighted (looking for the shortest/longest path). There are also bipartite graphs (for recommendation systems), hypergraphs, trees, and property graphs.

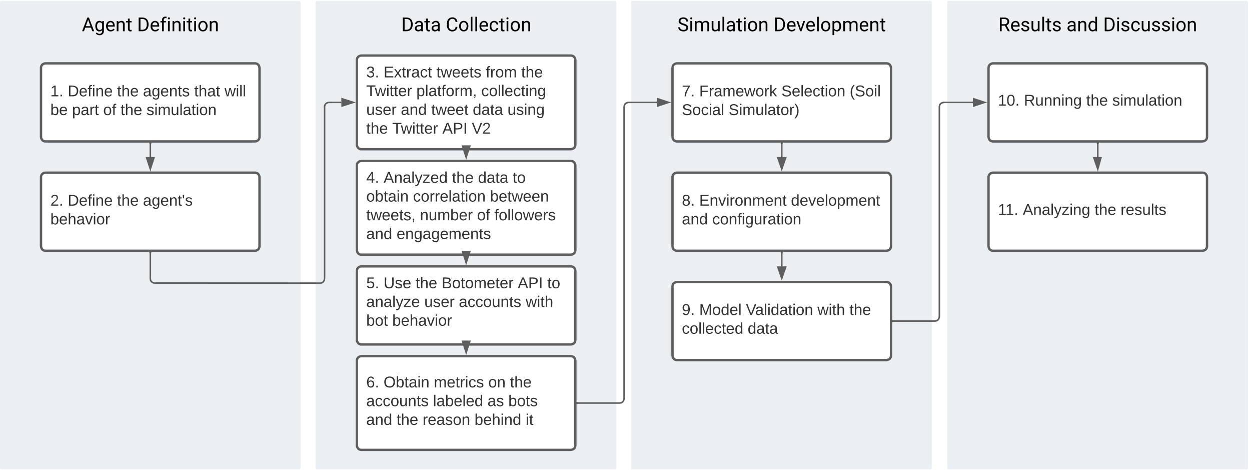

Agent Definition
1. Define the agents that will be part of the simulation
2. Define the agent's behavior
Data Collection
3. Extract tweets from the Twitter platform, collecting user and tweet data using the Twitter API V2
4. Analyzed the data to obtain correlation between tweets, number of followers and engagements
5. Use the Botometer API to analyze user accounts with bot behavior
6. Obtain metrics on the accounts labeled as bots and the reason behind it
Simulation Development
7. Framework Selection (Soil Social Simulator)
8. Environment development and configuration
9. Model Validation with the collected data
Results and Discussion
10. Running the simulation
11. Analyzing the results

FIGURE 2. **The high-level structure view of the proposed model.**

IV. THE SIMULATION DEVELOPMENT

As shown in Section II, related work focused on the spread of gossip and fake information through an environment built only from neutral agents, and how a single tweet propagates through the network. This study will focus on deceptive agents representing malicious users and automated bot accounts created with malicious intentions, and how much their actions affects the beliefs of the neutral agents. To achieve this, we introduce a belief value assigned to each neutral agent as a way of modeling its susceptibility from unofficial information sources. At the end of the timed execution of our simulation study, we inferred the influence based on the difference in the shift in the belief value from the initial parameters set for the environment.

To develop this ABSS, we defined the specifications of the system based on recommended practices defined by Slhoub et al. [25], where we assign each agent a series of properties based on their expected behavior. Next, we collected real data from Twitter to provide validation to our model. With the definition of all agents complete, we developed the model through the use of the *Soil* Social Simulator [27]. Figure 2 demonstrates a high-level structural view of the proposed model.

A. AGENT DEFINITION

Through the use of an Agent-Oriented Modeling technique introduced by Slhoub et al. in [25], we manage to describe the relationships between the agents to define their behavior. Here, we begin by classifying each agent that will be part of our ABSS and then defining their behavior based on their current state.

1) BELIEF VALUE

One of the fundamental pillars of this simulation is the belief value. Similar to its use by Beskow and Carley [18], we implement this property to determine the susceptibility of neutral agents to fake tweets, **ranging from 0** (believes only in official tweets) **to 1** (susceptible to fake tweets). Alongside its susceptibility, we also utilize the belief value in changing neutral agent states. When a threshold is met, the behavior of the neutral agent changes. For example, a belief value over 0.7 does not only indicate the user is more prone to sharing misinformation, but also it has a higher chance of generating fake tweets as well.

The belief value shifts after a neutral agent engages with a tweet. An official tweet causes the belief value to be reduced, while a fake tweet increases it. Neutral tweets can also be generated, causing no impact on the neutral agent's belief value. Whenever a neutral agent engages with a tweet, its effect on belief value is applied, regardless of the fact, whether it is neutral, fake or official. Through the belief value attribute, we simulate a human's beliefs.

2) AGENT CLASSIFICATION

All agents are modeled after a parent class, with each subclass representing a type of agent. This is due to the three types having access to the same functions, including sending, engaging, and sharing tweets.

a: NEUTRAL AGENTS

This agent represents the average Twitter user. It interacts in the environment through tweets created or shared by other agents. Each neutral agent has a belief value property that determines how susceptible it is to the tweets posted in the environment. This may change depending on the agent's perception of the content; the higher the belief value is, the higher chance of engaging and creating fake tweets. The characteristics of this agent are Autonomy, Reactiveness, and Sociability.

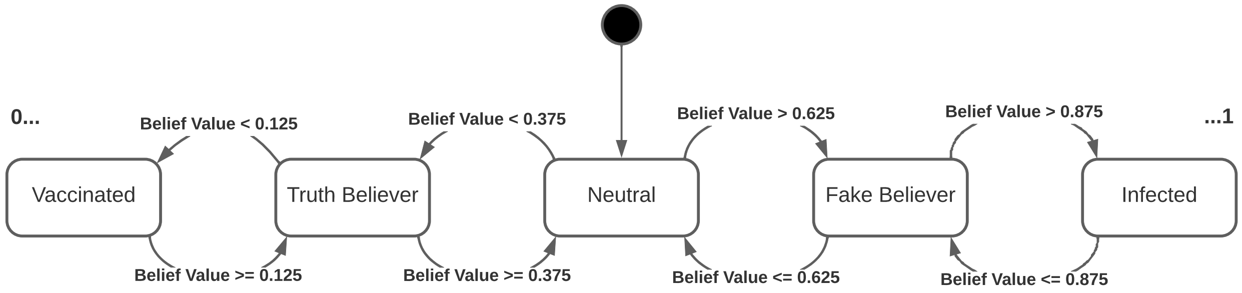


FIGURE 3. State diagram showing how the belief value states are distributed.

b: DECEPTIVE AGENTS

Represents malicious users in the Twitter network. They aim to post fake tweets and distribute them to the utmost extent in the environment. Neutral agents do not differentiate between neutral and deceptive agents. The characteristics of this agent are Autonomy, Proactiveness, and Sociability.

c: NEWS AGENTS

These agents represent news outlets on social media. These only generate official tweets, which represent verified news stories. Additionally, these agents do not engage with other tweets. The characteristics of this agent are Autonomy and Sociability.

3) AGENT BEHAVIOR

Each neutral agent is given a set of states that can shift during the duration of the execution, presenting different behaviors on each time step of the simulation. The state of each agent depends on the belief value or the type of agent. We developed a scale utilizing 5 different behaviors for the neutral agents, ranging between 0 and 1. As the neutral agents interact in the environment, their belief value shifts, and once it reaches a threshold, the neutral agent changes states. Figure 3 illustrates the different states that can be reached by neutral agents based on their current belief value.

a: NEUTRAL STATE

A neutral agent is in the *Neutral State* when its belief value is between 0.375 and 0.625. While in this state, the agent can only send neutral tweets.

b: FAKE BELIEVER STATE

A neutral agent is in the *Fake Believer State* when its belief value is between 0.625 and 0.875. While in this state, the agent has an increased probability of sending fake tweets itself.

c: INFECTED STATE

A neutral agent is in the *Infected State* when its belief value is between 0.875 and 1. While in this state, the agent can only share and post fake tweets.

d: TRUTH BELIEVER STATE

A neutral agent is in the *Truth Believer State* when its belief value is between 0.125 and 0.375. While in this state, the agent has an increased probability of sending and sharing official tweets.

e: VACCINATED STATE

A neutral agent is in the *Vaccinated State* when its belief value is between 0 and 0.125. While in this state, the agent has an increased probability of sending and sharing official tweets.

f: DECEPTION STATE

This state is exclusive to deceptive agents. During this state, the agent can only post fake tweets.

g: NEWS STATE

This state is exclusive to news agents. While in this state, the agent can only send official tweets, with a high probability of posting.

B. DATA COLLECTION

The Twitter data collection was carried out in four steps, as referenced in Figure 2: first, we extracted a tweets dataset through Twitter's API V2 containing the word "vaccine" from December 2021. We then analyze the data to obtain metrics regarding tweets, engagements and follower count. Next, we take a sample from the accounts listed in the dataset and processed them through the *Botometer* API [28] and, finally, obtain metrics regarding bot behavior in the accounts analyzed.

1) TWEET EXTRACTION

Utilizing the Twitter API V2,[5] we compiled a dataset with over 200,000 tweets containing the word "vaccine" from the 9th to 18th of December 2021. Each tweet was extracted 3 days after its initial posting date to collect the number of *retweets*, *likes*, *quotes-tweets* and generated comments. Additionally, we collected public user information for each user, including the number of followers, following accounts, verification status and tweets. A description of the structure of the collected data is shown in Table 1

[5] https://developer.twitter.com/en/docs/api-reference-index

TABLE 1. Structure of tweets collected.

Column	Description
Id	Unique tweet identification number
Text	Tweet text
Author Id	Unique author Id
Like Count	Number of likes received by the tweet
Retweet Count	Number of retweets received by the tweet
Quote Count	Number of quote retweets received by the tweet
Reply Count	Number of replies received by the tweet
Followers	Number of accounts that follow the author
Following	Number of accounts the author follows
Tweet Count	Number of tweets the author has sent
Account Creation	Date the account was created
Verified	Verification status

Our data collection process was created due to unavailability of datasets that provide the esential parameters needed for the proposed methodology, such as engagement metrics. The dataset was collected over several days with publicly available information regarding the accounts extracted and numbers related to engagements. In addition, since part of our analysis focused on collecting engagement metrics, we could not extract tweets as soon as they were posted, as a result, each collected tweet had to be at least 3 days old.

We were able to collect 220,085 tweets from 150,692 accounts. Tweets were filtered only by the keyword "Vaccine", excluded retweets and, as a result, were not part of the same thread. Our motivation was to obtain a set of metrics from tweets that we could use to compare with the results given in the simulation. In essence, this comparison would validate our simulation. We aligned our simulation to provide similar results to the data we collected, which can be seen in Section VII.

2) DATASET ANALYSIS

We calculated the correlation coefficient (r) between the number of followers and the number of engagements each tweet received. We noticed a weak correlation between both, with $r = 0.03913$. As a result, we did not consider this a deterministic factor in our simulation. When considering the number of tweets per agent, the dataset gave a moderate correlation between the number of followers and the number of tweets, with $r = 0.374347$.

During our analysis we found that the average number of followers for verified and non-verified accounts were highly different, with 336,191.3748 and 15,371.5414 followers, respectively. Although, some outliers were found, such as, follower counts going to as low as 14, the data showed that a high number of followers associated with verified accounts. We used these findings as the basis for the news agents in our simulation.

3) BOTOMETER

We took a sample of 10,420 accounts from the initial tweet dataset to be analyzed through the *Botometer* API [28]. This API analyzes the public information of each provided account and gives a score based on its activity on a scale of 0 to 5. The scores generated by the *Botometer* API are: Echo-chamber, for accounts that participate in follow-back groups (used by accounts wanting to increase their follower number by following other users back) and engage heavily with political tweets. Fake Follower, for accounts with a high number of bot followers. Financial, for bots using cashtags. Self-declared, for accounts that appear in botwiki.org. Spammer, for accounts labeled as bots in large Twitter data sets. Other, which takes into account miscellaneous manual reports or user feedback from *Botometer* users.

4) BOT ANALYSIS

Out of the 10,420 accounts analyzed, 164 were private or deleted accounts that could not be accessed, and 1,637 displayed an overall score of 3 or more. The highest scores were observed from the Other category, with 993 accounts over a score of 3, followed by Echo Chamber with 627 and Fake Followers over 299. While the authors of *Botometer* do not recommend binary classifications due to the nature of how scores are generated, we decided to focus on those with a score higher than 3 as a starting point for the number of deceptive agents in our simulation.

C. SIMULATION ENVIRONMENT

1) FRAMEWORK SELECTION

Based on the agent requirements established in Section IV-A, we decided to research existing frameworks to simulate social network environments that allow us to configure the agents in the simulation and their behavior. The frameworks BigTweet by Serrano et al. [29] and *Soil* by Sánches et al. [27] were tested, but the latter was chosen due to its customization options.

Soil is an ABSS developed in Python by Jesús Sánches, Carlos Iglesias and Fernando Sánches-Rada at Universidad Politécnica de Madrid, focused on simulating social network environments. The *Soil* Framework has three main elements: The network topology, the agents, and the environment. Each agent will have an assigned behavior based on its current state. In *Soil*, each agent is given a default state, and then based on the actions it can perform while on that behavior, it can change states. As an example, while simulating a SIR environment, agents can face a probability of infection while on the susceptible state; if this happens, the behavior of the agent changes, as it can now infect neighboring agents.

Soil uses JSON or YAML files for configuring the simulation. In this configuration, developers can set a series of values that will affect the execution of the simulation, such as the number of intervals, types of agents present in the simulation, and global environment parameters (e.g. the probability of spreading). By executing the configuration file, Python begins simulating the environment depending on the number of time-steps decided by the user. After it completes the simulation, the output is presented through a CSV file and a dynamic graph that can be displayed using a graph analysis application, such as Gephi.[6]

[6] https://gephi.org/

2) INITIALIZATION

The simulation initialization depends on the configuration provided in the YAML file, which denotes what types of agents are part of the simulation, their weight, the total number of agents, and hyper-parameters that affect the execution. The multi-agent system is given by the graph $G = (N, E)$, where N represents the user agents and E represents the indirect edges representing the following/follower of each node.

When the simulation is executed, *Soil* generates agents based on the weights assigned to each type in the configuration file. Following their creation, the agents are assigned one of seven topics, with the objective of creating communities (or clusters). When selecting each neutral agent's following accounts, we base it mainly on proximity to the chosen topic and making sure each agent follows between 1% and 5% of the total number of nodes to maintain an average path length between 2.5 and 3.0. This allows information to travel fast within a community, where users are more tightly coupled, but slower between agents on different topics. While focusing on maintaining clusters, agents can follow users outside their topics to allow communication between them.

Each neutral and deceptive agent was assigned a probability of tweeting, which correlates with the number of followers. The probability of *tweeting* is calculated by (1), where P_t represents the probability of *tweeting*, i shows the number of followers of a specific agent, and j presents the number of followers of every agent:

$$P_t(i) = \frac{N_i - min_j}{max_j - min_j} \tag{1}$$

Finally, when starting the simulation, each neutral agent is assigned a belief value between 0.375 and 0.625. This is performed to allow all agents to be initialized on the neutral state, which provides a good comparison point as the belief value shifts throughout the simulation.

3) RUNNING THE SIMULATION

During each time step, each agent can send a tweet based on the probability discussed in the previous section. Depending on the type of agent and its current state, the tweet can be of three types: Neutral, Official, or Fake. Regardless of the agent's type, it is sent to the *retweet* function, where the spread is calculated and counted. This is defined using two variables: *Impressions* and *Engagements*. The former focuses on the number of agents that have processed the tweet either by being a direct neighbor of the account or in contact with it after a neighbor shared it. The latter focuses on the number of agents that engage with it.

We decided to include the *retweet* function to simulate how quickly information spreads over Twitter. In Section V, we take a closer look at the numbers each tweet generates. This function is based on a recursion procedure, so every time a user shares a tweet, the same function will be called, and it will work under the assumption that every single agent will read all tweets presented in their timeline.

For the retweet function, we calculate two values per agent: the probability of engagement (P_e) and the probability of retweet (P_r). Both require the environmental variables $Prob_f$ and $Prob_r$ as a way determining the chance of spreading, based on it being fake or real information, respectively, and with x representing the agent.

The engagement probability values are used to determine if an agent believes the tweet presented. This variable can represent a *like*, *comment* or *retweet*, which indicates that the agent agrees with the information presented. The use of the environmental variables $Prob_f$ and $Prob_r$ is necessary to maintain a controlled number of engagements per tweet, as only considering the belief value will have different results from the dataset collected in section IV-B. The formulas for calculating the probability of engagement are shown in (2) and (3).

Fake news engagement:

$$P_e(x) = Prob_f * BeliefValue(x) \tag{2}$$

Real news engagement:

$$P_e(x) = Prob_r * (1 - BeliefValue(x)) \tag{3}$$

After engaging with a tweet, each agent can use the *retweet* function to forward the same tweet to its followers. This will result in the same function being called, but with the sharing agent as the parameter. The formula for the retweet probability is shown in (4) and (5)

Fake news retweet:

$$P_r(x) = P_t * BeliefValue(x) \tag{4}$$

Real news retweet:

$$P_r(x) = P_t * (1 - BeliefValue(x)) \tag{5}$$

A neighboring agent *affected* by a tweet will result in the agent's *BeliefValue* increasing or decreasing if the tweet was fake or authentic, respectively. After this, based on the probability of retweeting, agents might share that tweet with their neighbors, repeating the cycle. When the *BeliefValue* reaches a threshold, the agent changes to one of the states defined in section IV-A3. A graphic representation of the retweet function is demonstrated in Figure 4.

Inside the configuration, we implemented an external influence function. Given the same probabilities explained above, at the start of every time step, each neutral agent's belief value can be affected from outside the Twitter simulated environment, as it does not involve tweets. This allowed us to simulate external influences affecting human users outside the Twitter social network.

The final step when sending a tweet involves the *Targeted Tweet* function, working as a way of applying the *Top Tweets* feature used in the Twitter network. If the tweet reaches a large number of engagements in the simulation, it will be deemed a *viral* tweet. When this happens, the tweet will be shown to all agents in the network that share the same topics with the sender agent, even if an edge does not connect them.

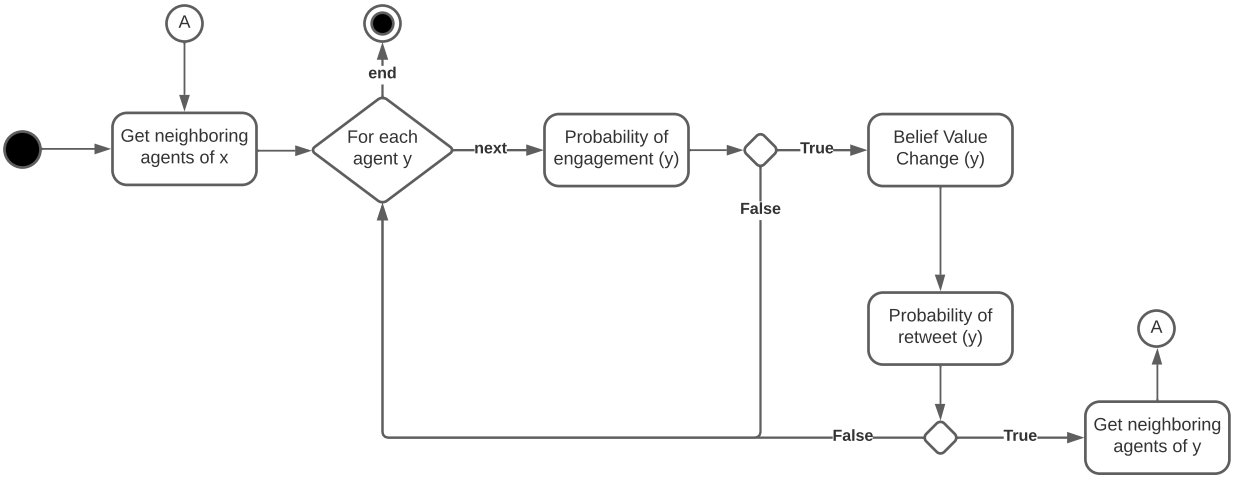

A
end
Get neighboring agents of x
For each agent y
next
Probability of engagement (y)
True
False
Belief Value Change (y)
Probability of retweet (y)
A
False
True
Get neighboring agents of y

FIGURE 4. High-level representation of the retweet function.

Details on how the Twitter algorithm works are unknown to the public, so we work under the assumption that all users use this feature, and the factors taken into account are the number of engaged users and topic similarity.

To summarize, we've included the pseudocode for the decision making algorithm for neutral agents in the neutral state in Algorithm 1 and neutral agents in the fake believer state in Algorithm 2, as an example of how the different states modify the behavior of neutral agents.

Algorithm 1 Neutral Agent Pseudocode for Neutral State

```
if $P_t(i)$ is True then
    Sending Neutral Tweet
    Retweet Function
    Targeted Tweet Function
end if
if $0.625 <= BeliefValue < 0.875$ then
    Set State = Fake Believer
else if $0.125 <= BeliefValue < 0.375$ then
    Set State = Truth Believer
end if = 0
```

V. RESULTS

We designed two scenarios to execute the simulation, with and without deceptive agents. Through this comparison, we were able to study the effect malicious users had over the neutral agents and evaluate the differences between external and internal influences.

A. SCENARIO 1: SIMULATION WITH DECEPTIVE AGENTS

The simulation consisted of 1,000 agents through 250 time-steps, with 70.9% neutral agents, 19.9% deceptive agents, and 9.2% news agents along with an average path length of 2.67. Figure 5 presents a sample of the distribution of agents at the end of the simulation, with each cluster representing the seven topics. We chose this distribution because we wanted to approximate our simulation as much as possible to the data extracted from Twitter. Starting with deceptive agents, we chose a 15% distribution as that was the percentage of accounts with bot behavior from our findings with the *Botometer* API. For news agents, we started with 5%, as those were approximately the number of verified accounts from the collected dataset. However, as we validate in Section VII, we had to fine-tune these numbers to remove bias and provide a better comparison point with the number of tweets created.

Algorithm 2 Neutral Agent Pseudocode for Fake Believer State

```
if $P_t(i)$ is True then
    if $P_e(x)$ is True then
        Sending Fake Tweet
        Retweet Function
        Targeted Tweet Function
    else if $P_e(x)$ is False then
        Sending Neutral Tweet
        Retweet Function
        Targeted Tweet Function
    end if
end if
if $0.375 <= BeliefValue < 0.625$ then
    Set State = Neutral
else if $0.875 <= BeliefValue < 1$ then
    Set State = Infected
end if = 0
```

130,255 tweets were sent throughout the 250 time-steps, most of them of were neutral tweets that do not alter the belief value. While the official and fake tweets reached a similar number of agents, the latter had more engagements per tweet. Table 2 and Figure. 6 illustrate the distributions per type of tweet, with the average number of impressions and engagements.

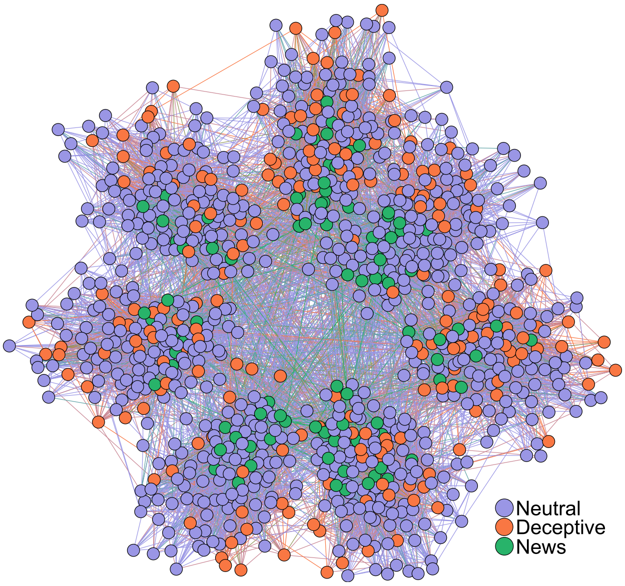


FIGURE 5. Distribution of agents in the network for scenario 1.

TABLE 2. Sample of tweets sent during the simulation for Scenario 1.

Tweets	Count	Avg. Impressions	Avg. Engagements
Neutral	85,591	115.08	3.63
Fake	23,954	59.28	1.42
Official	20,710	55.26	0.90

Even with most tweets being neutral, the effects of the fake tweets altered the belief value of the neutral agents, with 9.66% of agents in the *Fake Believer* state and 4.01% in the True Believer state at the end of the simulation. Figure 7 and Table 3 exhibit the average belief value per time step.

B. SCENARIO 2: SIMULATION WITHOUT DECEPTIVE AGENTS

In scenario 2, deceptive agents were not considered for the execution, keeping only neutral and news agents active in the environment. We implemented the external influence function that simulates changes in belief outside the Twitter network, with the same probability of being affected as a regular fake tweet. Our intention was to simulate an external face-to-face human interaction, as those can affect the beliefs of users as well.

We decided to modify the scenario by removing deceptive agents, keeping 800 agents instead of maintaining the same distribution, and replacing them with neutral agents. The official distribution for this scenario was 88.75% neutral and 11.25% news agents. The average path length was 2.72. An example of this distribution at the end of the simulation can be seen in Figure 8

In the chosen sample, 106,133 tweets were sent, with the majority in number and engagements being neutral. Only 1 fake tweet was sent by a neutral agent during this time, which reached 31 users, but no engagements. The distribution in scenario 2 is illustrated in Table 4 and Figure 9.

TABLE 3. Sample of belief values from scenario 1.

Time Step	Mean	Median	Mode	Highest	Lowest	Stand. Dev.
0	0.4952	0.4941	0.3856	0.6246	0.3752	0.0719
10	0.4960	0.4968	0.4549	0.6294	0.3749	0.0724
20	0.4968	0.4968	0.3821	0.6333	0.3740	0.0729
30	0.4976	0.4971	0.6091	0.6403	0.3750	0.0734
40	0.4983	0.4980	0.4940	0.6453	0.3720	0.0738
50	0.4990	0.4984	0.5317	0.6483	0.3710	0.0743
60	0.4997	0.4988	0.5337	0.6533	0.3680	0.0748
70	0.5004	0.4997	0.3912	0.6563	0.3600	0.0753
80	0.5012	0.4998	0.5994	0.6663	0.3580	0.0759
90	0.5021	0.5015	0.4569	0.6713	0.3590	0.0762
100	0.5027	0.5025	0.4510	0.6753	0.3570	0.0768
110	0.5033	0.5041	0.6014	0.6803	0.3510	0.0772
120	0.5042	0.5048	0.5795	0.6873	0.3480	0.0778
130	0.5052	0.5043	0.6052	0.6973	0.3460	0.0783
140	0.5062	0.5062	0.4096	0.7003	0.3440	0.0789
150	0.5071	0.5072	0.4379	0.7033	0.3440	0.0795
160	0.5078	0.5088	0.5147	0.7083	0.3420	0.0802
170	0.5087	0.5098	0.4868	0.7173	0.3370	0.0809
180	0.5095	0.5098	0.5766	0.7183	0.3330	0.0816
190	0.5105	0.5105	0.4870	0.7213	0.3310	0.0823
200	0.5116	0.5117	0.5412	0.7333	0.3320	0.0832
210	0.5126	0.5130	0.4794	0.7463	0.3300	0.0839
220	0.5136	0.5130	0.5337	0.7513	0.3300	0.0847
230	0.5146	0.5150	0.4814	0.7593	0.3260	0.0855
240	0.5158	0.5157	0.4580	0.7683	0.3220	0.0865
249	0.5166	0.5157	0.5369	0.7753	0.3180	0.0873

TABLE 4. Sample of tweets sent during the simulation for scenario 2.

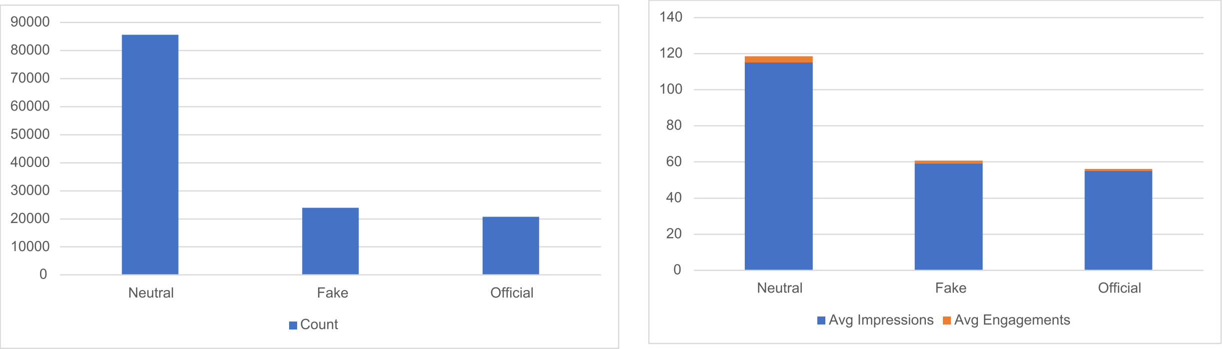
Tweets	Count	Avg. Impressions	Avg. Engagements
Neutral	85760	88.114	3.403
Fake	1	31	0
Official	20372	45	0.98

Despite no deceptive agents present in the simulation, 0.32% managed to reach the *Fake Believer* state and the engagements with official tweets only improved slightly compared to the previous scenario. Figure 10 shows the average belief value per time step, which curved on the opposite direction, alongside Table 5 displays a sample of time steps.

VI. DISCUSSION

We implemented the simulation with two distinct scenarios: scenario (1) utilizes deceptive agents representing malicious users in the Twitter social network and sending false tweets to misguide neutral users. Scenario (2) removes deceptive agents and adds an external influence function to represent external factors that might alter a neutral agent's beliefs outside the Twitter network. By implementing the belief value, we were able to quantify the susceptibility of neutral agents encountering misinformation in the environment.

The majority of tweets generated by the agents during the simulation were neutral tweets, with the highest number of both impressions and engagements, described in



(a)

(b)

FIGURE 6. (a) Number of tweets per type (b) Impressions and Engagements per type of tweet for scenario 1.

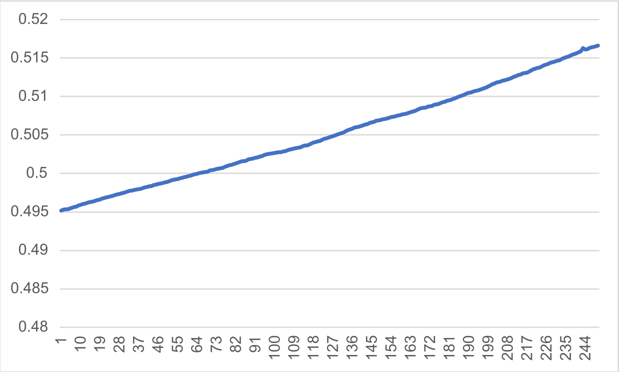


FIGURE 7. Average belief value per time step for scenario 1.

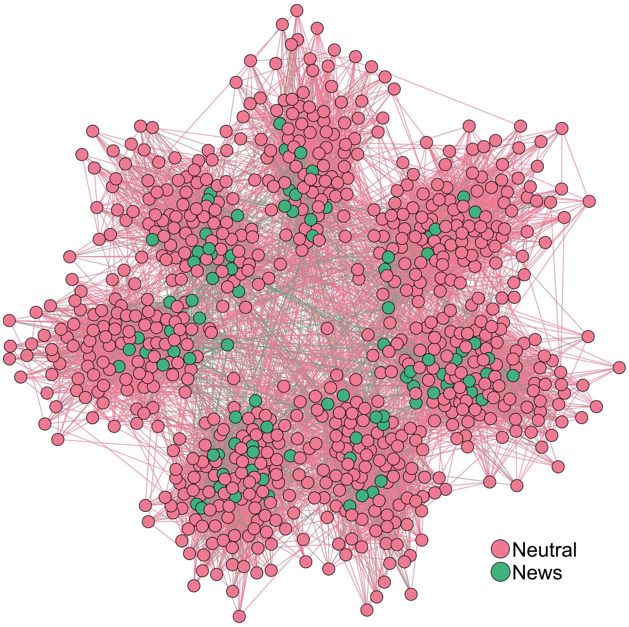


FIGURE 8. Distribution of agents in the network for scenario 2.

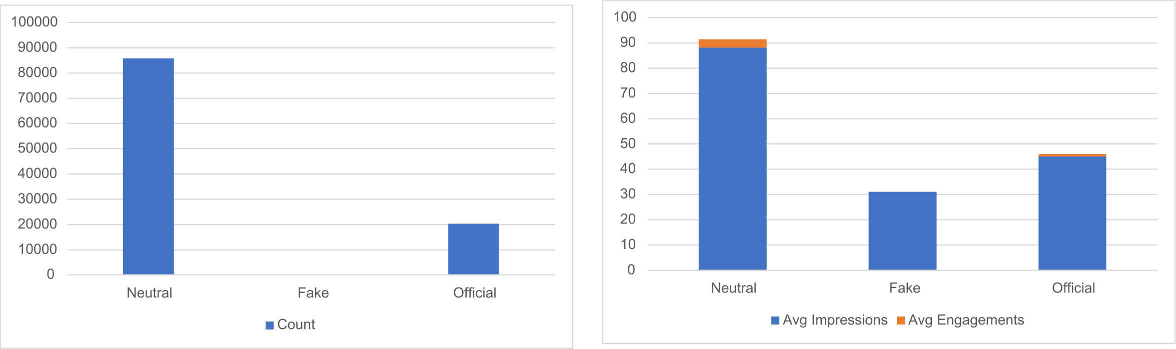
section IV-C3. The number of fake tweets and official tweets generated by agents remains very similar, but for different reasons as illustrated below:

TABLE 5. Sample of belief values from scenario 2.

Time Step	Mean	Median	Mode	Highest	Lowest	Stand. Dev.
0	0.4975	0.4988	0.4448	0.6243	0.3751	0.0746
10	0.4965	0.4979	0.4968	0.6244	0.3715	0.0749
20	0.4955	0.4970	0.4267	0.6254	0.3694	0.0752
30	0.4946	0.4954	0.5964	0.6250	0.3674	0.0754
40	0.4936	0.4943	0.5666	0.6247	0.3634	0.0757
50	0.4926	0.4942	0.3767	0.6260	0.3604	0.0760
60	0.4916	0.4929	0.4102	0.6242	0.3594	0.0764
70	0.4906	0.4923	0.4315	0.6242	0.3554	0.0766
80	0.4896	0.4917	0.4072	0.6232	0.3514	0.0769
90	0.4886	0.4903	0.5075	0.6232	0.3474	0.0772
100	0.4875	0.4888	0.4888	0.6223	0.3444	0.0776
110	0.4865	0.4878	0.3890	0.6233	0.3428	0.0779
120	0.4855	0.4859	0.3880	0.6223	0.3368	0.0781
130	0.4845	0.4851	0.4187	0.6223	0.3328	0.0785
140	0.4835	0.4838	0.4275	0.6223	0.3298	0.0788
150	0.4823	0.4824	0.5451	0.6213	0.3286	0.0791
160	0.4812	0.4809	0.5630	0.6213	0.3216	0.0796
170	0.4801	0.4801	0.3721	0.6213	0.3176	0.0799
180	0.4790	0.4796	0.4918	0.6233	0.3166	0.0803
190	0.4780	0.4777	0.3497	0.6223	0.3118	0.0807
200	0.4770	0.4770	0.5577	0.6223	0.3088	0.0811
210	0.4759	0.4761	0.5567	0.6223	0.3058	0.0814
220	0.4748	0.4738	0.4878	0.6223	0.3048	0.0818
230	0.4736	0.4726	0.4659	0.6223	0.2998	0.0823
240	0.4724	0.4710	0.5547	0.6213	0.2968	0.0827
249	0.4715	0.4705	0.3829	0.6214	0.2908	0.0830

- News agents exist in smaller quantities than the other types of agents inside the simulation, but their rate of official tweets were very high.
- Deceptive agents, on the other hand, had a similar rate of sending fake tweets as neutral agents, but there were less deceptive agents than neutral agents.
- Neutral agents could also send official tweets and fake tweets, but at a lower probability. As a result, neutral agents didn't alter the number of fake or authentic tweets in an significant way.

During the execution of scenario (1), fake tweets sent by deceptive agents never reached a *viral tweet* status (described



(a)

(b)

FIGURE 9. **(a) Number of tweets per type. (b) Impressions and Engagements per type of tweet for scenario 2.**

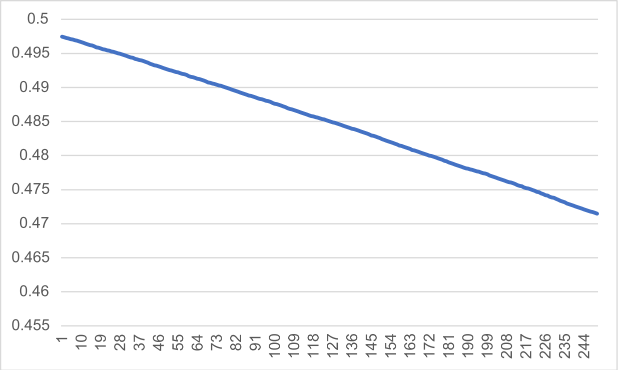


FIGURE 10. **Average belief value per time step for scenario 2.**

in section IV-C3), although several managed to generate over 300 impressions out of the 1000 total agents in the simulation. Deceptive agents focused their efforts on spreading large quantities of fake tweets to affect the belief value of neutral agents, instead of aiming to get a high number of engagements.

Despite the removal of the deceptive agents from the simulation in scenario (2), the results between the two scenarios had a similar number of engagements, impressions, tweets, and the overall standard deviation in belief value, mainly over the early stages. This indicates that the effects deceptive agents had on neutral agents were more evident the more time passed in the simulation.

The average belief value of the neutral agents in scenario (1) increased exponentially during the later stages of the 250 time-step simulation due to the large number of the fake tweets plaguing the simulation by the deceptive agents, with the highest belief value reaching 0.7753. As explained in section IV-A, neutral agents in the *Fake Believer* state have the option to also send fake tweets, which resulted in 256 fake tweets sent by neutral agents, generating over 33,000 engagements total. This spotlights the significant influence deceptive agents had on the neutral agents, altering their beliefs, as fake tweets sent by neutral agents also affect other neutral agents, increasing the average belief value throughout.and turning them into malicious users.

Throughout the execution of scenario (2), we witnessed a decrease in the average belief value due to the removal of the deceptive agents. However, with the external influence function into effect, an increase in the belief value of the neutral agents was still occurring at a lower rate. The influence of fake tweets on the neutral agents in scenario (1) was more substantial than the external influence due to the higher rate at which neutral agents encountered fake tweets. This enables further comparison between the influence in beliefs between face-to-face human interactions and social network interactions. Information travels faster on social networks, so beliefs have a greater chance of being altered.

One key element observed in scenario (2), despite the removal of the deceptive agents, was the presence of fake tweets. During our tests, the neutral agents ended up in the *Fake Believer* state, being able to create fake tweets, although in fewer quantities than in scenario (1). We believe that removing malicious users from the Twitter network will reduce the number of fake tweets but not resolve the problem of misinformation being spread.

During the simulation, we worked under the assumption that the neutral agents perceive all tweets in their surrounding environment. Despite this, we observed a slow shift in belief value, as tweets did not display high number of engagements on average. Not a single agent reached the *Infected* or *Vaccinated* state (discussed in section IV-A3) during the 250 time-step simulation, which indicates that it takes longer for the neutral agent's beliefs to be altered.

VII. VALIDATION

This section aims to compare the simulation results with those from the dataset presented in Section IV-B to validate our configuration and ensure it is closely related to the data extracted from the Twitter social network. Due to the limitation of being unable to measure a human user's beliefs in a verifiable way, we focused our efforts on validating engagement numbers related to the collected tweets.

TABLE 6. Data validation using different configurations of the Global Probability Variable.

Configuration	0 Engagements	Viral Tweets
Dataset	**40%**	**0.35%**
no global probability	0%	99.964%
0.2 global prob.	7.190%	52.166%
0.1 global prob.	21.708%	04.181%
0.07 global prob. (final)	35.434%	0.32%

Our efforts were centered around accurately representing the number of tweets generated during the simulation, alongside an accurate number of engagements to simulate how the neutral agents would react to tweets displayed in their timeline. Throughout our analysis, we found that 40% of tweets did not have a single engagement, regardless of their verified status or the number of followers. To scale it accordingly, we worked under the assumption that a *viral* tweet is represented by those with over 500 engagements in the Twitter social network and those with over 2.5% of the total number of agents in the simulation. We included this comparison to bring validity to the conducted simulation.

As mentioned in Section IV-C2, we included a global probability variable alongside the belief value to manage the rate at which the neutral agents would engage with tweets. Throughout our tests, utilizing only the belief value would result in a more significant number of engagements compared to the collected dataset. For this reason, we included this variable to correct the engagement flow and maintain it closer to reality. Our validation results can be observed in Table 6, displaying a comparison of the tweets collected with the ones generated by the social simulation with different configurations.

VIII. CONCLUSION

We successfully developed a multi-agent system simulating the Twitter social media network through the *Soil* Framework. With it, we studied the differences in the social network when bots and malicious users are inserted into the simulation and how the belief of the users are affected by the misinformation shared through it. We simulated two distinct scenarios, one with malicious agents and one without them, to compare how much influence they have on the neutral agents. Also, to evaluate it alongside an external factor in order to find which one was more influential towards the agents. The two scenarios were executed and showed similar outcomes regarding agent interactions with tweets. However, the removal of deceptive agents drastically reduced the average belief value, which leads us to conclude that removing bots from the Twitter platform will not solve the misinformation issue, but would greatly reduce it. The fake tweets' influence over the neutral agents was weak but still effective with the passing of time in the simulation.This is true regardless of the applied scenario, as fake tweets were created by neutral agents in an environment eliminating deceptive agents.

By extracting the Twitter dataset, we ensured the validity of the proposed model by comparing key elements, such as the number of engagements per tweet and bots in the environment. This contributed to the creation of an accurate simulation that we were able to verify against the actual tweets from the Twitter social network.

For future work, validating the belief value in the natural Twitter social network, based on how susceptible users of the social network are when presented with false information, would significantly improve the accuracy of the simulation. In this paper we used the umbrella term *engagements* to simulate a user's perception being affected by a tweet; future work can approach with more granularity, defining differences in *Likes*, *retweets*, *Comments* and *Quote-tweets*, as the fourth one is typically used in the social media networking site to correct false information. Additionally, implementing a likelihood function to neutral agents interacting with tweets on similar topics can improve the simulation's accuracy and more realistically measure the neutral agent's probability of engagement.

While reviewing the simulation results in scenario (1), we observed several neutral agents shifting towards lower belief values despite the average belief value trend going upwards. For future work, a deeper psychological study could provide a better understanding of the behavior of the neutral agents maintained firm beliefs on official tweets, despite being in an environment plagued with fake tweets.

REFERENCES

[1] B. Dean. (Nov. 21, 2021). *Social Network Usage & Growth Statistics: How Many People Use Social Media in 2022?* Accessed: Jan. 19, 2022. [Online]. Available: https://backlinko.com/social-media-users

[2] M. Iqbal. *Twitter Revenue and Usage Statistics*. Accessed: Nov. 13, 2021. [Online]. Available: https://www.businessofapps.com/data/twitter-statistics/

[3] M. Takayasu, K. Sato, Y. Sano, K. Yamada, W. Miura, and H. Takayasu, ''Rumor diffusion and convergence during the 3.11 earthquake: A Twitter case study,'' *PLoS ONE*, vol. 10, pp. 1–18, Apr. 2015.

[4] S. Cresci, F. Lillo, D. Regoli, S. Tardelli, and M. Tesconi, ''Cashtag piggybacking,'' *ACM Trans. Web*, vol. 13, no. 2, pp. 1–27, Apr. 2019.

[5] R. Takacs and I. McCulloh, ''Dormant bots in social media: Twitter and the 2018 senate election,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2019, pp. 796–800.

[6] B. Allyn. (May 20, 2020). *Researchers: Nearly Half of Accounts Tweeting About Coronavirus are Likely Bots*. Accessed: Nov. 13, 2021. [Online]. Available: https://www.npr.org/sections/coronavirus-live-updates/2020/05/20/859814085/researchers-nearly-half-of-accounts-tweeting-about-coronavirus-are-likely-bots

[7] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin, ''Fake news propagates differently from real news even at early stages of spreading,'' *EPJ Data Sci.*, vol. 9, no. 1, p. 7, Dec. 2020.

[8] D. J. Daley and D. G. Kendall, ''Stochastic rumours,'' *IMA J. Appl. Math.*, vol. 1, no. 1, pp. 42–55, 1965.

[9] E. Serrano and C. A. Iglesias, ''Validating viral marketing strategies in Twitter via agent-based social simulation,'' *Expert Syst. Appl.*, vol. 50, pp. 140–150, May 2016.

[10] K. Ikeda, Y. Okada, F. Toriumi, T. Sakaki, K. Kazama, I. Noda, K. Shinoda, H. Suwa, and S. Kurihara, ''Multi-agent information diffusion model for Twitter,'' in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, vol. 1, Aug. 2014, pp. 21–26.

[11] Y. Okada, K. Ikeda, K. Shinoda, F. Toriumi, T. Sakaki, K. Kazama, M. Numao, I. Noda, and S. Kurihara, ''SIR-extended information diffusion model of false rumor and its prevention strategy for Twitter,'' *J. Adv. Comput. Intell. Intell. Informat.*, vol. 18, no. 4, pp. 598–607, 2014.

[12] S. Kundu, T. Kajdanowicz, P. Kazienko, and N. Chawla, "Fuzzy relative willingness: Modeling influence of exogenous factors in driving information propagation through a social network," *IEEE Access*, vol. 8, pp. 186653–186662, 2020.

[13] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, "Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks," *Eur. J. Inf. Syst.*, vol. 28, no. 4, pp. 394–412, Jul. 2019, doi: 10.1080/0960085X.2018.1560920.

[14] C. Wang, Z. X. Tan, Y. Ye, L. Wang, K. H. Cheong, and N.-G. Xie, "A rumor spreading model based on information entropy," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, Dec. 2017.

[15] Y. Yan, F. Toriumi, and T. Sugawara, "Understanding how retweets influence the behaviors of social networking service users via agent-based simulation," *Comput. Social Netw.*, vol. 8, no. 1, pp. 1–21, Sep. 2021.

[16] K. M. Carley, "Social cybersecurity: An emerging science," *Comput. Math. Org. Theory*, vol. 26, pp. 365–381, Nov. 2020.

[17] S. Cresci, R. di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Exploiting digital DNA for the analysis of similarities in Twitter behaviours," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2017, pp. 686–695.

[18] D. M. Beskow and K. M. Carley, "Agent based simulation of bot disinformation maneuvers in Twitter," in *Proc. Winter Simulation Conf. (WSC)*, Dec. 2019, pp. 750–761.

[19] M.-A. Storey and A. Zagalsky, "Disrupting developer productivity one bot at a time," in *Proc. 24th ACM SIGSOFT Int. Symp. Found. Softw. Eng. (FSE)*. New York, NY, USA: Association for Computing Machinery, Nov. 2016, pp. 928–931.

[20] C. Lebeuf, M.-A. Storey, and A. Zagalsky, "Software bots," *IEEE Softw.*, vol. 35, no. 1, pp. 18–23, Jan./Feb. 2018.

[21] Y. Roth and N. Pickles. (May 18, 2020). *Bot or Not? The Facts About Platform Manipulation on Twitter*. Accessed: Jan. 20, 2022. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2020/bot-or-not

[22] A. M. Jamison, D. A. Broniatowski, and S. C. Quinn, "Malicious actors on Twitter: A guide for public health researchers," *Amer. J. Public Health*, vol. 109, no. 5, pp. 688–692, May 2019.

[23] K. Slhoub and M. Carvalho, "Towards process standardization for requirements analysis of agent-based systems," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 3, no. 3, pp. 80–91, May 2018.

[24] D. Kinny and M. Georgeff, "Modelling and design of multi-agent systems," in *Intelligent Agents III Agent Theories, Architectures, and Languages*, J. P. Müller, M. J. Wooldridge, and N. R. Jennings, Eds. Berlin, Germany: Springer, 1997, pp. 1–20.

[25] K. Slhoub, M. Carvalho, and W. Bond, "Recommended practices for the specification of multi-agent systems requirements," in *Proc. IEEE 8th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2017, pp. 179–185.

[26] A. M. Gibbons, *Graph Theory*. Hoboken, NJ, USA: Wiley, 2003, pp. 755–759.

[27] J. M. Sánchez, C. A. Iglesias, and J. F. Sánchez-Rada, "Soil: An agent-based social simulator in Python for modelling and simulation of social networks," in *Advances in Practical Applications of Cyber-Physical Multi-Agent Systems: The PAAMS Collection*, Y. Demazeau, P. Davidsson, J. Bajo, and Z. Vale, Eds. Cham, Switzerland: Springer, 2017, pp. 234–245.

[28] K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48–61, Jan. 2019.

[29] E. Serrano, C. Iglesias, and M. Garijo, "A survey of Twitter rumor spreading simulations," in *Computational Collective Intelligence*. Jul. 2015, pp. 113–122.

ALDO AVERZA was born in Panama City, Panama, in 1995. He received the B.S. degree in systems and computer engineering from the Technological University of Panama, in 2017, and the M.S. degree in software engineering from the Florida Institute of Technology, in 2022. From 2018 to 2020, he worked as a RPA Consultant at Bprosys Inc. He is currently working as an Automation Specialist at Dell Technologies. He was awarded a Fulbright Scholarship by IIE and the United States Embassy in Panama. His research interest includes agent-based systems, with an emphasis on social network behavior. He is a part of the Phi Kappa Phi Honor Student Society.

KHALED SLHOUB received the B.S. degree in computer science from the University of Benghazi, Libya, in 1999, the M.S. degree in computer science from the University of New Brunswick, Canada, in 2008, and the Ph.D. degree in computer science from the Florida Institute of Technology, USA, in 2018. He is currently an Assistant Professor at the Department of Computer Engineering and Sciences, Florida Institute of Technology. His main research interests include software engineering and agent-based systems. His research has centered on developing standard frameworks for formalizing the development processes of agent-based systems. He is currently focusing on studying and analyzing the quality of existing agent-oriented methodologies to provide unified agent-oriented development approaches that can deliver practical means in industrial settings. He is also working on developing a framework to enable understanding and flagging disruptive behavior of distributed social agents by deploying policing agents to determine risks. In addition, his research is focused on finding effective testing approaches to verify autonomous systems. He mainly focuses on finding ways to verify the irregular behavior of agent-based systems.

SIDDHARTHA BHATTACHARYYA (Senior Member, IEEE) is currently an Associate Professor at the Florida Institute of Technology. His Assured Safety Security and Intent with Systematic Tactics (ASSIST) Research Laboratory, primarily conducts research in the area of formal methods for the design, verification, and validation of intelligent autonomous systems, avionics, cyber security, and trust. Previously, he was a Sr. Research Engineer at Rockwell Collins' Advanced Technology Center, where he worked on research programs for assurance of safety critical systems. He was also a Summer Faculty Research Fellow at Oak Ridge National Laboratory and a Summer Research Assistant at the Applied Research Laboratory, Penn State.