Simulating Misinformation Diffusion on Social Media Through CoNVaI: A Textual- and Agent-Based Diffusion Model

Raquel Rodríguez-García, **Roberto Centeno** and **Álvaro Rodrigo**
NLP & IR Group, UNED
{rrodriguez, rcenteno, alvarory}@lsi.uned.es

Abstract

Misinformation has experienced increased online diffusion, leveraging strategies, such as emotional manipulation, to influence users' opinions. Efforts are underway to develop tools to mitigate its effects, such as misinformation propagation models used to simulate the diffusion of information. There are different approaches within these models, although, they show a significant limitation by disregarding the content of the information shared, crucial to the diffusion. We consider it the central aspect of modeling information dissemination. To this end, we focus on Agent-Based Modeling due to its suitability to simulate the complex interactions and heterogeneous behaviors observed on social media. We base our approach on a state-of-the-art Agent-Based Model that we modify and extend to account for the texts of the messages shared, focusing on two aspects that influence agents' decisions: *i)* the novelty of the content and; *ii)* its diffusion and behavior over time. To determine whether this content proves informative, we conduct an empirical evaluation using social media data from Twitter. Based on our experimental results, we observe that our textual-based approach reflects information diffusion more realistically than the state of the art, reducing the error regarding real diffusion.

1 Introduction

Fake news and misinformation have interfered with fundamental foreign affairs or spread dangerous health-related advice [Cuan-Baltazar *et al.*, 2020]. Detecting the diffusion of this content proves a significant challenge to mitigate its spread [Raza and Ding, 2022]. From current efforts, we notice a general absence of a holistic perspective, studying misinformation through separate components, from a local [Hu *et al.*, 2024] to a global perspective [Caldarelli *et al.*, 2020].

Evaluating detection models or mitigation strategies before implementation is also relevant to these efforts, which rely on understanding information diffusion processes. In these terms, propagation diffusion models are a powerful tool to study information cascades [Lotito *et al.*, 2021]. While not the only approaches, epidemiology-based models are the most widespread [Muhammad and Kasahara, 2024], mainly focused on user behavior without considering shared content. Whenever textual content is included [Kumar *et al.*, 2021; Milli, 2021], it is limited to user profiling, not as part of the communicative device. As such, texts with distinct characteristics (e.g. empty string, emotionally manipulative, or unintelligible) would have the same diffusion, not addressing why fake and real information differ [Vosoughi *et al.*, 2018].

Contrary to these models, we consider content a relevant aspect of information diffusion modeling. In this area of epidemiology-based models, we base our approach on a state-of-the-art agent-based model [Serrano and Iglesias, 2016], that we modify and extend to propose the *Textual* ***Co****ntent-based* ***N****eutral-****Va****ccinated-****I****nfected* (CoNVaI) model to simulate the diffusion of textual content on social media. We exploit textual characteristics from two perspectives: the novelty of the content; and the diffusion and behavior over time. We empirically validate and compare our approach to the base state-of-the-art agent-based model [Serrano and Iglesias, 2016] that we considered representative of similar epidemiology-based models that ignore textual content.

We also correct another standard limitation in evaluation processes: the lack of realistic evaluation environments. Current approaches rely on synthetic networks [Coates *et al.*, 2021] or real topologies that do not match the information being propagated [Zehmakan *et al.*, 2023], disregarding their impact on engagement [Karnstedt *et al.*, 2011].

With this paper, we make the following three contributions: *i)* We propose the CoNVaI model[1], where each agent is characterized based on a unique user and provided with a decision mechanism to determine when and how to disseminate information. *ii)* We consider the textual content of the information shared, mainly ignored in epidemiology-based models, from different perspectives. *iii)* We validate our model with data from real scenarios and compare it to a state-of-the-art model, highlighting the importance of modeling the content.

The paper is structured as follows: Section 2 reviews related work. Section 3 presents the fundamentals of CoNVaI. Section 4 covers the components of our model and its behavior. Experimental results are discussed in Section 5, and Section 6 details our findings and future work.

[1]The code, supplementary material, and experimentation results are available in https://github.com/Kasdeyael/ABSS_CoNVaI

2 Related Work

Many research efforts have been dedicated to studying information cascades and predicting their spread [Zhong *et al.*, 2023]. These approaches exploit various characteristics, from network topologies to temporal dynamics or the content of the messages [Liu *et al.*, 2023; Sun *et al.*, 2023; Zhong *et al.*, 2023], with a focus on deep learning. Related to these efforts, we have propagation diffusion models. Besides the potential prediction of the information cascades, the focus veers to modeling the users affected by the information and its diffusion, where the emphasis is placed on their decision-making abilities and behaviors [Coates *et al.*, 2021].

Propagation models originate in deterministic compartmental epidemiological models for viral contagion [Kermack and McKendrick, 1927], which use ordinary differential equations to reflect transition rates. Infected individuals are introduced into a group, and the virus spreads to susceptible individuals until they get removed. Individuals are compartmentalized into *Susceptible*, *Infected*, or *Removed*, creating the SIR model. An early application to information propagation considers it an "intellectual epidemic" [Goffman and Newill, 1964], where the virus is the information. Other initial variations, such as the Daley-Kendall model [Daley and Kendall, 1964], adopted elements from information diffusion.

The limitations of epidemic models applied to information diffusion eventually become apparent, such as assuming homogeneous behaviors [Nekovee *et al.*, 2007]. Some models include behaviors from social media, reflecting a belief system and a hesitancy stage [Xia *et al.*, 2015]. The Emotion-based SIS (ESIS) [Wang *et al.*, 2015] introduces the concepts of emotion within the information by categorizing them, making some emotions more effective for propagation. These models still present limitations, such as compartmentalization, to compensate for the lack of individual behavior [Zhang *et al.*, 2018]. Other approaches have been inspired by physical phenomena, such as the *Forest Fire Model* [Kumar *et al.*, 2021], influenced by a fire spreading in a forest, which also introduces user-based similarity leveraging shared topics. Still, these textual characteristics only model users' relationships without giving relevance to the information itself.

Agent-based simulation has been used to overcome limitations regarding user and topology homogeneity. Epidemiology-based models have been implemented [Serrano and Iglesias, 2016] while differentiating user behavior [Gausen *et al.*, 2021]. Social theories have also been researched through skepticism and gullibility [Tambuscio *et al.*, 2018], or user-based similarity [Li *et al.*, 2019]. The Big Five Personality traits model has been proposed to study the effect of political beliefs [Coates *et al.*, 2021] or to model agent-based trust [Muhammad and Kasahara, 2024]. Once again, the content is ignored in favor of the users, sometimes characterized based on psychological models, without considering specific social media behavior and personality traits may lack correlation [Azucar *et al.*, 2018] or be time-dependent.

3 Preliminaries

This section defines and formalizes the fundamental components of our proposal.

3.1 Information Diffusion Fundamentals

To model information diffusion, we first introduce the formal definitions of the elements that shape it from the standpoint of our agent-based framework.

We adopt and extend the definition of a multi-agent system (MAS) provided by Centeno *et al.* [2009]. Thus, we define an Agent-based Simulation Diffusion System as follows:

Definition 1. *An Agent-based Simulation Diffusion System is a tuple $\langle U, \mathcal{X}, \mathcal{A}, \Phi, x_0, \varphi, t\rangle$, where:*

- *U is a set of social agents, where $|U|$ denotes the total number of social agents within the system.*
- *$\mathcal{X}$ is the environmental state space. As an attribute of $\mathcal{X}$, we consider the set of conversations $\mathcal{C}$ the social agents create, where $|\mathcal{C}|$ denotes the number of conversations.*
- *$\mathcal{A}$ is the action space formed by the 3 actions that agents can perform. In our system, these are: starting a new conversation as a_{new}, replying to a conversation as $a_{reply}(c)$, or doing nothing as a_{skip}, where $c \in \mathcal{C}$.*
- *$\Phi : \mathcal{X} \times \mathcal{A}^{|U|} \times \mathcal{X} \to [0..1]$ is the system's transition probability distribution, reflecting how $\mathcal{X}$ evolves with the agents' actions.*
- *$x_0 \in \mathcal{X}$ establishes the initial state of the system.*
- *$\varphi : U \times \mathcal{X} \times \mathcal{A} \to \{0, 1\}$ is the agent's capability function, which determines whether an agent can perform an action at a given environmental state.*
- *t reflects the time, discretized in steps, which represents the execution time of the system.*

We have extended the MAS definition to consider conversations a part of $\mathcal{X}$. We deem the time an explicit part of the system, enabling the agents to perform actions that affect the environment within each time unit.

Following the definition of an agent provided by Centeno *et al.* [2009], we define a social agent as follows:

Definition 2. *A Social Agent is a tuple $\langle \mathcal{S}, \mathcal{O}, U_{in}, g, f, per, s_0\rangle$, where:*

- *$\mathcal{S}$ defines the set of internal states of an agent.*
- *$\mathcal{O}$ is the set of observations the agent perceives from its environment. As part of $\mathcal{O}$, the agent has a set of conversations $\mathcal{C}$ that they perceive.*
- *U_{in} is a subset of social agents such that the agent can read their conversations (their followees).*
- *$g : \mathcal{O} \times \mathcal{S} \to \mathcal{S}$ is the transition function of the agent's states.*
- *$f : \mathcal{S} \to \mathcal{A}$ is the decision function, representing the agent's diffusion model.*
- *$per : \mathcal{C} \times \mathcal{X} \to \mathcal{O}$ is a perception function of the agents, allowing them to assign an observation in an environmental state. For an agent u_i, $\mathcal{O}$ is composed of conversations $C_i \subseteq \mathcal{C}$ such that u_i is already part of a conversation or $\exists u_j \in U$ where $u_i \in U_{in}(u_j)$ and u_j has participated in a conversation.*
- *s_0 is the initial internal state of the agent.*

We have extended the definition to reflect the social setting of the agent, introducing U_{in}, such that $U_{in}(u_i) = \{u_j \in U \mid (u_j, u_i) \in E\}$, where E is a set of connections between the agents such that $E \subseteq \{(x, y) \mid x, y \in U^2 \text{ and } x \neq y\}$. We also consider the conversations to be part of the observations of an agent, which affect the perception function. From this definition, social agents can participate in the conversations they perceive. These conversations are defined as follows:

Definition 3. *A conversation is a tuple $\langle m_0, p, \mathcal{M}\rangle$, where:*

- *m_0 is the initial message that starts a conversation.*
- *p is the textual content that is being discussed.*
- *$\mathcal{M}$ is a set of messages that reply to the conversation.*

A conversation is a diffusion process that encapsulates other messages, allowing agents to maintain discussions about some information[2] $p \in P$. Messages are defined as:

Definition 4. *A message is a tuple $\langle u_i, t_j, s_k\rangle$, where:*

- *u_i is the agent that sent the message.*
- *t_j is the time step at which the message was sent.*
- *s_k is the message's state, reflecting the agent's opinion. It can manifest their agreement or disagreement.*

Through their actions, agents create conversations and messages to interact with each other. These interactions are determined by the actions $\mathcal{A} = \{a_{new}, a_{reply}(c), a_{skip}\}$ they can take. Formally, their behavior involves:

- a_{new} starts a conversation c about a content p with a message m_0 and includes c in the conversations of the system $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$.
- $a_{reply}(c)$ replies to a conversation c with a message m_i, such that $\mathcal{M} \leftarrow \mathcal{M} \cup \{m_i\}$, where $\mathcal{M} \in c$.
- a_{skip} does nothing.[3]

3.2 Base AB-SIR Model

Most widespread information diffusion simulation models are designed based on epidemiological principles and employ minor to no textual features. Since our contribution focuses on these approaches, we chose a sophisticated MAS proposed by Serrano and Iglesias [2016] for our base model, which leverages epidemiological concepts for information diffusion. Thus, we consider it to represent these models, posing an option to simulate diffusion processes while being complex and validated with empirical data, guaranteeing it reflects reality.

This model, which we refer to as Agent-Based SIR (AB-SIR) for this paper, extends the definition of social agent as:

Definition 5. *A SIR-agent is a tuple $\langle \mathcal{S}, \mathcal{O}, U_{in}, g, f, per, s_0\rangle$, where:*

- *$\mathcal{O}, U_{in}, f, s_0$ are defined as in Definition 2.*
- *$\mathcal{S}$ has an attribute that labels the agent as one of four states regarding $c \in \mathcal{C}$, such that $S = \{Neu, Inf, Vac, Cu\}$. A Neutral ($Neu$) agent is unaware of c. Infected (Inf) and Vaccinated (Vac) agents spread it by agreeing or disagreeing, respectively. A Cured (Cu) agent was Infected stops spreading.*
- *the transition function $g : \mathcal{O} \times \mathcal{S} \times P_{TR} \rightarrow \mathcal{S}$ is also dependent on a set of transition parameters $P_{TR} = \{P_{INF}, P_{MD}, P_{AD}\}$, that determine the infection, vaccination and cure rates.*
- *$per : \mathcal{C} \times \mathcal{X} \rightarrow \mathcal{O}$ limits the conversations they perceive. For each agent u_i such that $\exists u_j \in U, u_i \in U_{in}(u_j)$, and $\forall \langle m_0, p, \mathcal{M}\rangle \in \mathcal{C}$, they perceive a conversation $\langle m_0, p, \mathcal{M}_k\rangle$ where $\mathcal{M}_k \leftarrow \{m \in \mathcal{M} \mid u_j \in \mathcal{M}_k\}$.*

Algorithm 1 AB-SIR Transition function g

Input: $o_j, s_i, P_{TR} = \{P_{INF}, P_{MD}, P_{AD}\}$

$\langle m_0, p, M_i\rangle \leftarrow$ ExtractConversation(o_j)
$s_j \leftarrow s_i$
for $m = \langle u, t, s_k\rangle$ in M_i **do**
  Let $U_1, U_2, U_3, U_4 \leftarrow \mathcal{U}(0, 1)$ be random values
  if $s_k = Vac \wedge s_i = Inf \wedge U_1 \leq P_{AD}$ **then**
    $s_j \leftarrow Cu$
  else if $s_k = Vac \wedge s_i = Neu \wedge U_2 \leq P_{AD}$ **then**
    $s_j \leftarrow Vac$
  else if $s_k = Inf \wedge s_i = Neu$ **then**
    if $U_3 \leq P_{INF}$ **then**
      $s_j \leftarrow Inf$
    else if $U_4 \leq P_{MD}$ **then**
      $s_j \leftarrow Vac$
    end if
  end if
end for
return s_j

The g function is detailed in Algorithm 1. Let $u_1 = \langle \mathcal{S}, \mathcal{O}, U_{in}, g, f, per, s_0\rangle$ represent an AB-SIR agent, and let x_j represent the current environmental state, where $o_j = per(x_i)$ is its current observation space, and s_i is its current state. Transitions are probabilistic and depend on random values drawn from uniform distributions $\mathcal{U}$ and the configurable probabilities P_{TR}. When reading a message $s_k = Inf$ from a perceived conversation (ExtractConversation(o_j)), a *Neutral* agent might get infected with a probability P_{INF}, or vaccinated with a probability $(1 - P_{INF}) \cdot P_{AD}$. An *Infected* agent might turn *Cured* with a probability P_{AD} from reading a message $s_k = Vac$. In contrast, a *Neutral* agent would turn *Vaccinated* with the same probability.

After updating the state s_j according to the changes in the environment, an agent-SIR would determine its action based on f, which can be defined as follows:

$$f(s_j) = \begin{cases} a_{new} & \text{if } s_j \in \{Inf, Vac\} \wedge \nexists c \\ a_{reply}(c) & \text{if } s_j \in \{Inf, Vac\} \wedge \exists c \\ a_{skip} & \text{otherwise} \end{cases}$$

An agent-SIR would start a conversation (a_{new}) or reply to the one that updated its state ($a_{reply}(c)$) if $s_j \in \{Inf, Vac\}$, otherwise they would do nothing (a_{skip}). This cycle would repeat until the maximum time limit for the simulation is reached at T, or the users stop spreading the information

[2] In this work, we focus only on textual information, leaving multimedia information such as images or videos.

[3] It allows us to model and simulate asynchronous simulations.

after a time X. For this last condition, let t denote the current time in the simulation and $t_l(u_i)$ the last time the agent u_i modified s_i, $\forall c \in C$. The simulation stops when $\forall u_i \in U, t - t_l(u_i) > X$.

4 CoNVaI: Textual Content-Based Neutral-Vaccinated-Infected

To correct the shortcomings of most epidemiology-based models where the textual component is disregarded, we introduce the CoNVaI model following the fundamentals introduced in Section 3.1. We propose an extension of a SIR-agent where the f decision and g transition functions rely on two additional components (compared to AB-SIR): *i)* user profiles with individual characteristics, which are common in state-of-the-art approaches; and *ii)* the textual content shared through two approaches: the novelty; and the influence and engagement.

4.1 User Characteristics

In terms of the characterization of the users, we determine their influence over others. Previous approaches exploit user similarity measures and metrics based on the user profile [Kumar *et al.*, 2021; Milli, 2021]. To this end, we decided to employ the social-based information from their profiles for our study. We extract three of the most relevant measures: the follower count (*followers*), the number of listed posts (*posts*), and the verified status (*verified*). These values have previously been utilized to characterize users and indicate a user's influence. We also consider one that has not yet been fully explored: the number of followees (*followees*). When studied with the number of followers, this value has been positively associated with engagement [Peng and Lu, 2024].

Based on the previous information, we set the probability of a user $u \in U$ to influence others with the formula:

$$P_{usr}(u) = F_{INFL} * Infl(u)$$

where F_{INFL} is a configurable parameter to set the relevance of $Infl(u)$ to sway opinions, and $Infl(u)$ is the influence of a user, calculated with the previous four factors as follows:

$$Infl(u) = 0.4 \cdot sc(\frac{followers}{followees}) + 0.4 \cdot sc(posts) + 0.2 \cdot verified$$

where:

- $sc(X)$ represents a logarithmic scaling function that allows us to normalize the factor X as follows:

$$sc(X) = 1 - e^{-\alpha \cdot X} \tag{1}$$

  where α sets the middle value of the scaling function. For each metric, we select α based on the mean.
- The weights (0.4; 0.4; and 0.2) have been adjusted to assign higher importance to graded metrics since they provide more information about a user.

4.2 News Novelty

Regarding the shared textual content, we introduce the novelty of a piece of information $p \in P$. Following works where the novelty of a news piece has proven relevant to the spread of information [Photiou *et al.*, 2021; Ulloa *et al.*, 2023], we decided to model its effects using Gaussian distributions. Let $ini \in T$ be the initial time when p was sent within a conversation $c \in \mathcal{C}$, and $t \in T$ be the current time. The novelty of p is obtained as:

$$P_{nov}(p,t) = F_{NOV} \cdot (Nov(p) \cdot e^{\frac{-(t-ini)^2}{2*10^2}})$$

where F_{NOV} is a configurable parameter to set the relevance of the novelty $Nov(p)$. We consider two processes for $Nov(p)$. The first one would be for users to be cognizant of the information. It would take some time for any information to reach them. The second one would be the increase of the novelty, as the forgetting and the need to keep up with the information starts acting again. To model the novelty of each content p, we use the following formula:

$$Nov(p) = \begin{cases} (1 - e^{\frac{-(t_m-60)^2}{2*20}}) \cdot Entr(p) & \text{if } t_m \leq 60 \\ (1 - e^{\frac{-(t_m-60)^2}{2*110}}) \cdot Entr(p) & \text{otherwise} \end{cases}$$

where:

- t_m refers to the time in minutes when the last information related to p was sent.
- $Entr$ represents the Entropy obtained through the Kullback-Leibler Divergence (KLD) to evaluate the information gain. We consider an initial window of the six previous hours before the content was sent for the forgetting mechanism. We explored intervals from two to 24 hours to account for different speeds in information diffusion and time zones, and settled on six to favor the fast-paced news environment. Since KLD measures the probability of an event, we obtain the probability of each word as the relationship between its frequency in a text and the rest of the frequencies for the other words. We use the formula:

$$Entr(p) = sc(\sum_k s_k \cdot ln(\frac{s_k}{q_k}))$$

  where $s_k = [s_{k1}, s_{k2}, \ldots, s_{kn}]$ is the vector of the probabilities of the k words in the content p for a time t, and $q_k = [q_{k1}, q_{k2}, \ldots, q_{kn}]$ is the vector of the probabilities for $t-1$. We perform a previous smoothing step, introducing a small ϵ when probabilities are zero to avoid indeterminations, and adjust the others accordingly. A higher entropy corresponds with new information, while a value of 0 would indicate the information has been seen before. To keep $Entr(p)$ normalized, we have applied Equation 1, with α being the median of the values.

4.3 News Influence and Engagement

For the second textual dimension, we study the influence of news and its engagement over time with two variables. One is the cumulative engagement a piece of information has (*News Influence*), and the other is the distribution of that engagement over time (*Engagement Over Time*). Our model considers both, since two pieces of information could have the same total engagement with different temporal dynamics.

For these components, we used two regression models to predict engagement of $p \in P$, represented as $\hat{y}_p = f_r(p)$, where $\hat{y}_p$ is the prediction and $f_r(p)$ represents the function used by the regression model. Since the predicted $\hat{y}_p$ differs depending on the component, we cover them individually:

- *News Influence*. For this component, the prediction $\hat{y}_p$ would be a scalar value $\hat{y}_p \in \mathbb{R}$, as the prediction of the aggregated engagement of p. To use within the decision mechanism of our model, we apply Equation 1 to scale the value, producing:

$$P_{nw}(p) = sc(\hat{y}_p)$$

- *Engagement Over Time*. This model would predict the timeline of the diffusion per time unit, and it is represented by a one-dimensional vector $\hat{y}_p = (\hat{y}_{p_1}, \hat{y}_{p_2}, \ldots, \hat{y}_{p_n})$. For the model, we turn the prediction into a distribution as follows:

$$P_{rpl}(p) = \frac{\sum_{j=i}^{i+w-1} \hat{y}^*_{p_j}}{\sum_j \hat{y}_{p_j}} \tag{2}$$

where $\hat{y}^*_p$ is a repeated array from the second element onward of $\hat{y}_p = (\hat{y}_{p_1}, \ldots, \hat{y}_{p_n})$, to disregard the initial comment that started the conversation, for w times:

$$\hat{y}^*_p = (\underbrace{\hat{y}_{p_2}, \ldots, \hat{y}_{p_2}}_{\text{w times}}, \ldots, \underbrace{\hat{y}_{p_L}, \ldots, \hat{y}_{p_L}}_{\text{w times}})$$

and L is the maximum output size, determined by the 99th percentile of the real diffusion during training. For each index $i = 1, \ldots, M$, where $M = len(\hat{y}^*_p) - w + 1$ in Equation 2, we have computed a rolling window and scaled the values, obtaining the one-dimensional vector $P_{rpl}(p)$ to reflect likelihood intervals for sent messages.

For these regressor models, we employ some standard algorithms for the *News Influence* component: Random Forest, AdaBoost, and Gradient Boosting, with the Scikit-learn library[4]. For the *Engagement Over Time*, we selected KNeighbors and Decision Trees since they support multiple outputs natively. In terms of the input, we explored a selection of characteristics through a set of tools: MultiAzterTest [Bengoetxea, 2021] and Empath [Fast *et al.*, 2016], to explore syntactic and semantic measures (lexical diversity, readability, polysemic index, verbs in passive voice...), as well as the emotional dimensions, such as joy or anger. We considered three approaches: using only the characteristics from MultiAzterTest and Empath, using the texts directly with two bag-of-words representations: frequency-based and TF-IDF, or combining both texts and characteristics. We select the more significant categories through Recursive Feature Elimination (RFE) since they cannot be generalized due to being dataset and social media-dependent [Aldous *et al.*, 2019].

4.4 Runtime Behavior

We conclude the section by defining a ConVaI-agent, which extends Definition 2 as follows:

[4]https://scikit-learn.org/

Algorithm 2 CoNVaI Transition function g

Input: $o_j, s_i, t_1, P_{TR} = \{P_{INF}, P_{MD}, P_{AD}, P_{RD}, P_{OPI}\}$ $P_{TX} = \{P_{nov}, P_{rpl}, P_{nw}\}, P_{usr}$

$t \leftarrow$ current time
Let $U_1 \leftarrow \mathcal{U}(0, 1)$
if $U_1 \leq P_{read}(u_i, t)$ **then**
    $\langle m_0, p, M_i \rangle \leftarrow$ ExtractConversation(o_j)
    $m = \langle u, t, s_k \rangle \leftarrow$ ExtractMessage(M_i)
    if UnknownConversation(c) **then**
        $s_j \leftarrow$ ReadSc$(m_0, p, t, P_{TX}, P_{INF}, P_{MD}, P_{usr})$
    else
        $s_j \leftarrow$ ReadMs$(m, s_i, p, t, P_{TX}, P_{AD}, P_{OPI}, P_{usr})$
    end if
end if
return s_j

Definition 6. *A CoNVaI-agent is a tuple* $\langle \mathcal{S}, \mathcal{O}, U_{in}, g, f, per, s_0 \rangle$, *where:*

- $\mathcal{O}, U_{in}, f, per, s_0$ *are defined as in Definition 2.*
- $\mathcal{S}$ *considers three labels* $S = \{Neu, Inf, Vac\}$, *similar to the SIR-agent. In AB-SIR,* Cu *is introduced as a sink state to limit diffusion. A CoNVaI-agent only interacts in response to a message, and would merely stop engaging.*
- *the transition function* $g : \mathcal{O} \times \mathcal{S} \times t \times P_{TR} \times P_{TX} \times P_{usr} \rightarrow \mathcal{S}$ *is dependent on a set of transition parameters* P_{TR}, *as well as the parameters from the textual* $P_{TX} = \{P_{nov}, P_{rpl}, P_{nw}\}$ *and user characteristics* P_{usr}, *as well as the time t.*

The transition function of a CoNVaI-agent is conditioned by a set of $P_{TR} = \{P_{INF}, P_{AD}, P_{MD}, P_{RD}, P_{OPI}\}$ where we consider a rate to read messages per time unit (P_{RD}) and to share an opinion (P_{OPI}), besides the parameters from AB-SIR. We also include the previously defined textual $P_{TX} = \{P_{nov}, P_{rpl}, P_{nw}\}$ and user-based characteristics P_{usr}.

The transition function g would behave as in Algorithm 2. Let $u_1 = \langle \mathcal{S}, \mathcal{O}, U_{in}, g, f, per, s_0 \rangle$ represent a CoNVaI-agent, and let x_j represent the current environmental state, where $o_j = per(x_i)$ is its observation space, and s_i is its state. After determining whether they can read a message based on a probability $P_{read}(u_i, t) = {}^{P_{RD}}/_{m_r(u_i,t)}$ where $m_r(u_i, t)$ reflect the messages a user u_i has read at the current t, they extract a message (ExtractMessage(M_i)) from a conversation they perceive. Their actions depend on whether they know said information (UnknownConversation(c)).

If the information is new, their behavior follows Algorithm 3. They would reply to c based on whether the information is relevant, determined by P_{TX}, and turn *Infected* or *Vaccinated* based on the poster's influence (P_{usr}), and the rates to infect (P_{INF}) or vaccinate (P_{MD}). If the information is not new, their behavior follows Algorithm 4. In this case, agents determine whether p is relevant based on P_{TX} and P_{usr}, and their state is conditioned to their previous s_i and that of the message. If they both agree, agents might feel validated and affirm their posture [Ballara, 2023]. If they disagree, they could change their mind and share the other agent's state or

Algorithm 3 ReadSc

Input: $m_0 = \langle u_k, t_0, s_k \rangle, p, t, P_{TX} = \{P_{nov}, P_{rpl}, P_{nw}\}$ P_{INF}, P_{MD}, P_{usr}

$s_j \leftarrow Neu$
Let $U_1 \leftarrow \mathcal{U}(0, 1 - P_{nov}(p,t))$
if $U_1 \leq P_{rpl}(p)[t]$ **then**
  $Replying()$
  Let $U_2, U_3 \leftarrow \mathcal{U}(0, 1 - P_{usr}(u_k) - P_{nw}(p))$
  if $U_2 \leq P_{INF}$ **then**
    $s_j \leftarrow Inf$
  else if $U_3 \leq P_{MD}$ **then**
    $s_j \leftarrow Vac$
  end if
end if
return s_j

trigger a confirmation bias or a "backfire effect," making the agent defend and reinforce their opinion [O'Boyle, 2022].

Finally, the diffusion function f is defined as follows:

$$f(s_j) = \begin{cases} a_{new} & \text{if } s_j \in \{Inf, Vac\} \wedge \nexists c \\ a_{reply}(c) & \text{if } s_j \in \{Inf, Vac\} \wedge Replying \wedge \exists c \\ a_{skip} & \text{otherwise} \end{cases} \tag{3}$$

Our model considers real-time dynamics, social behavior, and the message's content as part of the decision-making. The diffusion is now conditional on a desire to reply ($Replying$), contrary to AB-SIR, which assumes agents always engage.

5 Experimental Evaluation

After presenting the model, we proceed with the evaluation to establish its ability to reflect information diffusion processes.

5.1 Selected Dataset

For the empirical evaluation, we aim to recreate the observed scenarios of a news piece's diffusion on social media, so we prioritized realistic scenarios. After evaluating the most commonly used datasets, we selected PHEME-9 [Zubiaga *et al.*, 2016]. As far as we know, this is the only readily available dataset that contains the information to recreate those scenarios: textual data of the shared news, temporal information, user characteristics and their topology, and the stance of the messages, which we use to evaluate diffusion. We avoided synthetic data, such as the network topology, because it would add noise or biases.

PHEME-9 centers around nine 2014/2015 events, with 66k tweets and retweets organized into 297 threads for 55k users, containing their ego-networks and the HTML of external links. We employed the Wayback Machine to extract unrecovered articles. We decided on an 80:20 ratio for the train and test partitions for the evaluation, choosing the Ottawa shooting event for testing. We employ the train partition to tune the *News Influence* and *Engagement Over Time* components. Since the information these components exploit for this tuning differs from the one for the evaluation, data leakage is not a concern.

Algorithm 4 ReadMs

Input: $m = \langle u_k, t_i, s_k \rangle, s_i, p, t, P_{TX} = \{P_{nov}, P_{rpl}, P_{nw}\}$ P_{AD}, P_{OPI}, P_{usr}

$s_j \leftarrow s_i$
Let $U_1 \leftarrow \mathcal{U}(0, 1 - P_{nov}(p,t) - P_{usr}(u_k))$
if $U_1 \leq P_{rpl}(p)[t]$ **then**
  Let $U_2, U_3, U_4 \leftarrow \mathcal{U}(0,1)$
  if $s_i = s_k \wedge U_2 \leq P_{OPI}$ **then**
    $Replying()$
    $s_j \leftarrow s_i$
  else if $s_i \neq s_k$ **then**
    if $U_3 \leq P_{AD}$ **then**
      $Replying()$
      $s_j \leftarrow s_k$
    else if $U_4 \leq P_{OPI}$ **then**
      $Replying()$
      $s_j \leftarrow s_i$
    end if
  end if
end if
return s_j

5.2 Metrics

For the evaluation, we follow the methodology used to validate AB-SIR, where the accumulated real diffusion is compared to the model's output [Serrano and Iglesias, 2016]. The message's stance is used to establish the user's state. We assume a retweet or comment *supporting* reflects their *Infected* state. If users are *denying*, they would be *Vaccinated*. We also assume *underspecified* comments are *supporting*.

We selected $RMSE$ to express our results. Based on the stances of the users, we have both *Vaccinated* and *Infected* states. We generate two metrics, $RMSE_{deb}$ and $RMSE_{spr}$, for the users denying and spreading information, respectively. We use the formula $RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_{state_i} - \hat{y}_{state_i})^2}$ where:

- T reflects the total number of time units per test instance, which we set as one-minute intervals. T is set based on the lack of engagement after 60 minutes.
- y_{state_i} is the number of users in the dataset with $state \in \{Vac, Inf\}$ for $RMSE_{deb}$ or $RMSE_{spr}$, respectively.
- $\hat{y}_{state_i}$ is the number of agents in the simulation, with $state \in \{Vac, Inf\}$, at each given time.

We linearly combine and normalize the metrics as follows:

$$NRMSE_{st} = \frac{RMSE_{spr} + RMSE_{deb}}{state_{max} - state_{min}} \tag{4}$$

where $state_{min}$ and $state_{max}$ are the minimum and maximum diffusion values, respectively. Similarly, we normalize $RMSE_{spr}$ and $RMSE_{deb}$ with their cumulative values.

5.3 Experimental Setup

CoNVaI and AB-SIR were implemented with *Repast Symphony*[5]. It is extensible, scalable, open source, and allows

[5]https://repast.github.io/

Metric	CoNVaI	AB-SIR
Average $NRMSE_{spr}$	1.7318	91.8037
SD $NRMSE_{spr}$	11.1759	599.9526
Median $NRMSE_{spr}$	0.1081	6.5244
Average $NRMSE_{deb}$	15.5952	29566.0260
SD $NRMSE_{deb}$	25.5216	13328.4132
Median $NRMSE_{deb}$	2.9861	35406.1663
Average $NRMSE_{st}$	1.5669	392.5724
SD $NRMSE_{st}$	9.4302	1348.8904
Median $NRMSE_{st}$	0.1832	188.5129
Average $NRMSE_{st}$ 90th percentile	0.2147	176.8263
SD $NRMSE_{st}$ 90th percentile	0.1427	119.5583
Median $NRMSE_{st}$ 90th percentile	0.1741	178.7382

Table 1: NRMSE metrics (Average, SD, and Median) for the CoNVaI and AB-SIR models

for complex behavior. For the components of CoNVaI, we used the training partition with a validation set to tune the regressor models and choose the best-performing ones. We observed differences depending on the text sources when applying RFE. Most features focus on the tweets for the *News Influence* component, while it is evenly split for the *Engagement Over Time*. For the first component, only 20% relate to the emotional aspect, while that percentage goes over 60% for the second one. Some of the common characteristics, such as readability, A1-C1 incidence, or passive voice, relate to the linguistic differences between real and fake news [Kasseropoulos and Tjortjis, 2021; Zhou *et al.*, 2020; Manikonda *et al.*, 2022], indicating that there are similarities despite the differences in intent.

Regarding the test set, we start by setting the simulator to create the agents depending on the model and the connections based on the ego-networks from the dataset. The user who started the conversation introduces the content in $t_1 = 0$. In each consecutive discretized step, agents that receive the information will choose how to act. The combinations for the adjustable parameters for each model are covered in the supplementary material and repeated for each test instance. We report the results for the best-performing combinations.

5.4 Experimental Results and Discussion

Table 1 covers the results for the test set. We include the average, standard deviation (SD), and median of $NRMSE_{st}$ and the normalized versions of $NRMSE_{spr}$ and $NRMSE_{deb}$. CoNVaI is more accurate at describing reality, where most differences are in lower orders of magnitude. The most striking variations are observed for $NRMSE_{deb}$. After analyzing the debunkers in the test set, we observed that the highest value for a test instance was under ten. These errors are smoothed in CoNVaI for $NRMSE_{st}$, indicating that debunkers are not a big concern in the overall diffusion.

We observed some outliers when comparing the SD and averages. Selecting the 90th percentile of $NRMSE_{st}$ shows a significant decrease for CoNVaI and AB-SIR, although the errors remain orders of magnitude over CoNVaI. To determine the reason behind the performance of AB-SIR, we further adjusted the parameters for a representative set of test instances.

From these additional experiments, we observed some interesting trends. Firstly, run time increased by several hours per run. Secondly, although the simulation was closer to the real diffusion, information still reached most of the network. It showed a linear growth, only modifying the curve gradient, affecting the whole population without reflecting real diffusion: fast in the beginning, and slowly losing momentum. Agent-SIR behavior determines that, to control diffusion, we need *Infected* agents to turn *Cured*. It requires *Vaccinated* users, which are uncommon and rarely engage [Zubiaga *et al.*, 2016; Vosoughi *et al.*, 2018]. Even limiting infection rates, *Vaccinated* users will continue spreading. This would be consistent with epidemiological models since the population is expected to recover, but not with information diffusion.

From previous experiments with AB-SIR, we considered how it was validated through empirical evaluation since information would affect the entire network. We theorize that a part of this could be due to the population and the evaluation framework. Typically, simulations are compared to the diffusion in a real dataset, which means it is being compared to actual engagement. We might see the distinction between the possible states (*Infected* or *Vaccinated*), but those users have engaged somehow. Our study has considered other non-affected users by employing ego-networks instead of synthetic [Serrano and Iglesias, 2016]. These evaluation scenarios would be biased by disregarding other users who decided not to engage. Another possible cause is how results are measured, comparing only spreaders [Gausen *et al.*, 2021], ignoring the rest. Our experiments show that this benefits the AB-SIR model without accurately reflecting diffusion. The state-based metric is also bounded by the population, limiting the evaluation scenario. Despite the popularity of epidemiological models, these effects have not been reported to the best of our knowledge. However, to study users' behavior, we need realistic scenarios with interacting and non-interacting users.

6 Conclusions and Future Work

We propose CoNVaI in the scope of epidemiology-based models, which are the most widespread in simulating information diffusion. Our contribution shows how information diffusion can be approached holistically. From a local perspective, we incorporate the textual content of the messages, which had been mainly ignored. From a global perspective, agents interact with each other based on their ego-networks.

From our experiments, we determined that incorporating textual content positively affected our simulations. AB-SIR, a representative model, overestimates how many users the information reaches. It highlights the need for models focused on online diffusion with realistic evaluation scenarios. We have also shown how the evaluation is bounded by the total users and does not reflect engagement. This simplification could be problematic when applying social science concepts to user behaviors, such as gullible or skeptical users.

In future work, we plan to study unbounded evaluation methodologies centered around messages. We also plan to enhance CoNVaI by considering how often users continue conversing and incorporating in-depth user profiling.

Acknowledgements

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación), through the DeepInfo project PID2021-127777OB-C22 (MCIU/AEI/FEDER, UE) and the HOLISTIC ANALYSIS OF ORGANISED MISINFORMATION ACTIVITY IN SOCIAL NETWORKS project (PCI2022-135026-2).

Raquel Rodríguez-García is supported by the UNED's FPI doctoral grant, Spain.

References

[Aldous *et al.*, 2019] Kholoud Khalil Aldous, Jisun An, and Bernard J. Jansen. Predicting Audience Engagement Across Social Media Platforms in the News Domain. In *Social Informatics*, pages 173–187. Springer, 2019.

[Azucar *et al.*, 2018] Danny Azucar, Davide Marengo, and Michele Settanni. Predicting the Big 5 Personality Traits from Digital Footprints on Social Media: A Meta-Analysis. *Personality and Individual Differences*, 124:150–159, 2018.

[Ballara, 2023] Noli Ballara. The Power of Social Validation: A Literature Review on How Likes, Comments, and Shares Shape User Behavior on Social Media. *IJRPR*, 4:3355–3367, 2023.

[Bengoetxea, 2021] Kepa Bengoetxea. MultiAzterTest: A Multilingual Analyzer on Multiple Levels of Language for Readability Assessment. *arXiv preprint*, 2021. arXiv:2109.04870.

[Caldarelli *et al.*, 2020] Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. The Role of Bot Squads in the Political Propaganda on Twitter. *Communications Physics*, 3(1), 2020.

[Centeno *et al.*, 2009] Roberto Centeno, Holger Billhardt, Ramón Hermoso, and Sascha Ossowski. Organising MAS: A Formal Model Based on Organisational Mechanisms. In *Proceedings of the 2009 ACM SAC*, pages 740–746, 2009.

[Coates *et al.*, 2021] Annabel Coates, Tim Muller, and Sean Sirur. Simulating the Impact of Personality on Fake News. In *TRUST@ AAMAS*, 2021.

[Cuan-Baltazar *et al.*, 2020] Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health and Surveillance*, 6(2), 2020.

[Daley and Kendall, 1964] Daryl J. Daley and David G. Kendall. Epidemics and Rumours. *Nature*, 204(4963):1118, 1964.

[Fast *et al.*, 2016] Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 ACM CHI*, pages 4647–4657. ACM, 2016.

[Gausen *et al.*, 2021] Anna Gausen, Wayne Luk, and Ce Guo. Can We Stop Fake News? Using Agent-Based Modelling to Evaluate Countermeasures for Misinformation on Social Media. In *ICWSM Workshops*, 2021.

[Goffman and Newill, 1964] William Goffman and Vaun A. Newill. Generalization of Epidemic Theory: An Application to the Transmission of Ideas. *Nature*, 204(4955):225–228, 1964.

[Hu *et al.*, 2024] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI*, 38(20):22105–22113, 2024.

[Karnstedt *et al.*, 2011] Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. The Effect of User Features on Churn in Social Networks. In *Proceedings of the WebSci Conference*. ACM, 2011.

[Kasseropoulos and Tjortjis, 2021] Dimitrios Panagiotis Kasseropoulos and Christos Tjortjis. An Approach Utilizing Linguistic Features for Fake News Detection. In Ilias Maglogiannis, John Macintyre, and Lazaros Iliadis, editors, *AIAI*, volume 627, pages 646–658. Springer, 2021.

[Kermack and McKendrick, 1927] William Ogilvy Kermack and Anderson G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772), 1927.

[Kumar *et al.*, 2021] Sanjay Kumar, Muskan Saini, Muskan Goel, and B. S. Panda. Modeling Information Diffusion in Online Social Networks Using a Modified Forest-Fire Model. *Journal of Intelligent Information Systems*, 56(2):355–377, 2021.

[Li *et al.*, 2019] Weihua Li, Quan Bai, and Minjie Zhang. A Multi-agent System for Modelling Preference-Based Complex Influence Diffusion in Social Networks. *The Computer Journal*, 62(3):430–447, 2019.

[Liu *et al.*, 2023] Xiaoyang Liu, Chenxiang Miao, Giacomo Fiumara, and Pasquale De Meo. Information propagation prediction based on spatial–temporal attention and heterogeneous graph convolutional networks. *IEEE Transactions on Computational Social Systems*, 11(1):945–958, 2023.

[Lotito *et al.*, 2021] Quintino Francesco Lotito, Davide Zanella, and Paolo Casari. Realistic Aspects of Simulation Models for Fake News Epidemics over Social Networks. *Future Internet*, 13(3), 2021.

[Manikonda *et al.*, 2022] Lydia Manikonda, Dorit Nevo, Benjamin Horne, Clare Arrington, and Sibel Adali. The Reasoning behind Fake News Assessments: A Linguistic Analysis. *AIS Transactions on Human-Computer Interaction*, 14(2):230–253, 2022.

[Milli, 2021] Letizia Milli. Opinion Dynamic Modeling of News Perception. *Applied Network Science*, 6(1), 2021.

[Muhammad and Kasahara, 2024] Radifan Fitrach Muhammad and Shoji Kasahara. Agent-Based Simulation of Fake News Dissemination: The Role of Trust Assessment and

Big Five Personality Traits on News Spreading. *Social Network Analysis and Mining*, 14(1), 2024.

[Nekovee *et al.*, 2007] Maziar Nekovee, Yamir Moreno, Ginestra Bianconi, and Matteo Marsili. Theory of Rumour Spreading in Complex Social Networks. *Physica A: Statistical Mechanics and its Applications*, 374(1), 2007.

[O'Boyle, 2022] Matthew O'Boyle. Digital Dilemmas: How the Backfire Effect and Echo Chamber Effect on Social Media Contribute to Political Polarization in the United States. *Gettysburg Social Sciences Review*, 6(1), 2022.

[Peng and Lu, 2024] Yi Peng and Liling Lu. Untangling Influence: The Effect of Follower-Followee Comparison on Social Media Engagement. *Journal of Retailing and Consumer Services*, 78, 2024.

[Photiou *et al.*, 2021] Antonis Photiou, Christos Nicolaides, and Paramveer S. Dhillon. Social Status and Novelty Drove the Spread of Online Information during the Early Stages of COVID-19. *Scientific Reports*, 11(1), 2021.

[Raza and Ding, 2022] Shaina Raza and Chen Ding. Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach. *International Journal of Data Science and Analytics*, 13(4), 2022.

[Serrano and Iglesias, 2016] Emilio Serrano and Carlos A. Iglesias. Validating Viral Marketing Strategies in Twitter via Agent-Based Social Simulation. *Expert Systems with Applications*, 50:140–150, 2016.

[Sun *et al.*, 2023] Xigang Sun, Jingya Zhou, Ling Liu, and Wenqi Wei. Explicit time embedding based cascade attention network for information popularity prediction. *Information Processing & Management*, 60(3):103278, 2023.

[Tambuscio *et al.*, 2018] Marcella Tambuscio, Diego F. M. Oliveira, Giovanni Luca Ciampaglia, and Giancarlo Ruffo. Network Segregation in a Model of Misinformation and Fact-Checking. *Journal of Computational Social Science*, 1(2):261–275, 2018.

[Ulloa *et al.*, 2023] Roberto Ulloa, Mykola Makhortykh, Aleksandra Urman, and Juhi Kulshrestha. Novelty in News Search: A Longitudinal Study of the 2020 US Elections. *Social Science Computer Review*, 42(3), 2023.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The Spread of True and False News Online. *Science*, 359(6380):1146–1151, 2018.

[Wang *et al.*, 2015] Qiyao Wang, Zhen Lin, Yuehui Jin, Shiduan Cheng, and Tan Yang. ESIS: Emotion-based Spreader–Ignorant–Stifler Model for Information Diffusion. *Knowledge-Based Systems*, 81:46–55, 2015.

[Xia *et al.*, 2015] Ling-Ling Xia, Guo-Ping Jiang, Bo Song, and Yu-Rong Song. Rumor Spreading Model Considering Hesitating Mechanism in Complex Social Networks. *Physica A: Statistical Mechanics and its Applications*, 437:295–303, 2015.

[Zehmakan *et al.*, 2023] Ahad N. Zehmakan, Charlotte Out, and Sajjad Hesamipour Khelejan. Why Rumors Spread Fast in Social Networks, and How to Stop It. In *Proceedings of the IJCAI*, pages 234–242. IJCAI, 2023.

[Zhang *et al.*, 2018] Yaming Zhang, Yanyuan Su, Li Weigang, and Haiou Liu. Rumor and Authoritative Information Propagation Model Considering Super Spreading in Complex Social Networks. *Physica A: Statistical Mechanics and its Applications*, 506, 2018.

[Zhong *et al.*, 2023] Chu Zhong, Fei Xiong, Shirui Pan, Liang Wang, and Xi Xiong. Hierarchical attention neural network for information cascade prediction. *Information Sciences*, 622:1109–1127, 2023.

[Zhou *et al.*, 2020] Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. Fake News Early Detection: A Theory-driven Model. *Digital Threats*, 1(2), 2020.

[Zubiaga *et al.*, 2016] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS ONE*, 11(3):e0150989, 2016.