Control Approach for Misinformation Mitigation via Recommender Systems on Social Network Platforms

Andreas Philippou (1663623)
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
a.philippou@student.tue.nl

***Abstract*—This paper presents a control framework for mitigating misinformation in social networks through a misinformation-sensitive recommender system. This is done by extending the closed-loop Friedkin-Johnsen opinion dynamics model to incorporate psychological factors driving misinformation spread, specifically negative emotional extremity and content novelty. The proposed approach modifies engagement-focused recommendation systems to penalize content. Both model-free and model-based control strategies are used, and both demonstrate significant reductions in misinformation propagation, up to 76% improvement. This is validated through simulations using the LIAR2 dataset. Practical limitations are identified and discussed.**

I. Introduction

Modern social networks have become a critical digital infrastructure for information, connecting billions of users globally. While social platforms enable unprecedented connectivity, they simultaneously create conditions that contribute to the rapid spread of misinformation [1]. The societal implications of this underlying negative effect have become increasingly evident in recent decades, with significant negative consequences for democratic processes, public health decisions, and social cohesion [2] [3]. In addition, the issue of viral misinformation is predicted to expand if measures are not taken, since recent research suggests that misinformation spreads faster and deeper than true information [4].

To date, the applied strategies to mitigate misinformation have primarily been a verdict on the truthfulness of shared content. Truthfulness verdict strategies rely on machine learning models, user reporting, official warnings and professional fact-checking evaluations [5]. These approaches have been somewhat effective in mitigating misinformation, however, they ignore the psychological and sociological aspect of content, and additionally ignore the social dynamic aspect of the social network platform. This allows for an underrated misinformation effect to take place: misinformation is more likely to be spread compared to true content due to a successful tendency in exploiting cognitive biases and emotional reactions in platform users. Specifically, misinformation content generates emotions, such as anger, fear, anxiety and shock [6], prompting users to further share the misinformation material. The virality effect is further enhanced for content that is novel [7]. Consequently, rapid spread of misinformation that exploits cognitive biases leads to the development of echo-chamber effects and social identity cohesion through social opinion dynamics. This in turn facilitates misinformation as a permanent problem for social networks [8].

A recent possible development is that of implementing recommender systems that can be incorporated into an efficient closed-loop control framework that takes into account the social dynamics of the social network platforms. Such an implementation can be done through the closed-loop Friedkin-Johnsen opinion dynamics model formulated extensively in [9], which allows for a recommender system to act as a control input. The specific closed-loop model represents individuals as nodes with internal states (adapted from the Friedkin-Johnsen model) that evolve based on interactions with individual neighbors and the recommender system. As highlighted in [9], the specific closed-loop control recommendation algorithm fundamentally alters opinion dynamics by selectively exposing users to content that maximizes engagement metrics. However, research specifically addressing misinformation propagation within this control feedback system remains limited.

Consequently, the contribution of this paper enhances the existing recommendation system - Friedkin-Johnsen control model from [9] to explicitly account for misinformation while facilitating the existing content engagement. The methodology extends prior work by additional constraints to the engagement maximization cost function, in theory reducing misinformation spread by reducing recommended content that exploits cognitive biases. Both model-free (MF) and model-based (MB) approaches are used and fully mathematically developed with accurate steady-state convergence proofs. A secondary aim of the paper is to test the enhanced model on real-life data for insight into relative practical effectiveness. This is done by the use of Large language models that can concisely analyze the psychological attributes of fact-checked datasets (LIAR2 dataset [10]) and formulate them into the control system.

Section II focuses on the theoretical model details, mathematical proofs for attainable stability, tailoring the LIAR2 dataset for the model use case, developing a synthetic simu-

lated representation of the model in Python and extending the simulated model by integrating it with the tailored dataset. Section III provides simulation results regarding model effectiveness in misinformation mitigation, with an additional deeper focus on opinion - time evolution and quantitative measures, assessing the propagation of misinformation directly facilitated by the recommendation system. This is done in order to prove model effectiveness and prior sociological research. Section IV and V discuss additional insights regarding the model and simulations, possible drawbacks, areas of improvement and future work.

II. Methods

This section begins with a theoretical overview of the recommender system model and its structure in II-A. The main cost function alterations for misinformation mitigation are shown in II-B. Both the model free and model based approaches are formulated in II-C, with mathematical feasibility and theoretical convergence proofs given in II-D. The implementation strategy of LLMs for content emotion approximation is given in II-E. The section concludes with II-F, describing code implementation and dataset tailoring.

The model dynamics and the general control-loop formulation in II-A are adapted from [9]. Thus, for additional details on model foundation and theoretical dependencies, see [9].

A. Model Dynamics & Optimization-Based Control Overview

The stand-alone Friedkin-Johnsen opinion dynamic model is a composition of user nodes with respective opinions and connections, in this case, the users and connections are taken as online accounts interacting on an online social network platform. The influence of connections between users can be represented as a row-(sub)stochastic adjacency matrix, given as $\mathbf{W_{total}} \in [0,1]^{(n+1)\times(n+1)}$. Defining the last user in the matrix as a "recommender" allows for the division of $\mathbf{W_{total}}$ into $\mathbf{W} \in [0,1]^{n\times n}$, corresponding to the user-to-user adjacency matrix and into $\mathbf{w_{rec}} \in [0,1]^n$, corresponding to the recommender influence on each user. Unlike conventional approaches, the user opinion state $\mathbf{x}(t) \in [0,1]^n$ is taken to qualitatively represent the negative emotional extremity of an opinion instead of thematic opinion agreeableness. Hence, $x_i(t) = 0$ corresponds to minimum negative emotional extremity, while $x_i(t) = 1$ corresponds to the maximum achievable negative emotional extremity. The main control problem perspective is found in the transformation of this Friedkin-Johnsen model into an inhomogeneous linear time invariant system. This can be shown as

$$\mathbf{x}(t+1) = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{W}\mathbf{x}(t) + (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{w_{rec}}u(t) + \mathbf{\Lambda}\mathbf{x}(0), \quad (1)$$

where $\mathbf{x}(t) \in [0,1]^n$ represents the user negative emotional extremity vector at time t, $\mathbf{\Lambda} \in [0,1]^{n\times n}$ is the user diagonal stubbornness matrix, $\mathbf{I}_n$ denotes the identity matrix, $u(t) \in [0,1]$ represents the control input at time t, and $\mathbf{x}(0)$ is the initial negative emotional extremity vector. The equation shown in (1) can be notationally simplified to give a clearer representation, resulting in

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) + \mathbf{\Lambda}\mathbf{x}(0), \quad (2)$$

Where $\mathbf{A} = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{W}$ and $\mathbf{B} = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{w_{rec}}$. The general closed-loop system can be seen in Figure 1. Through optimal control, a cost function definition can be formulated, the solution of which corresponds to the recommender opinion extremity $u^*(t)$ (controller output). Thus, defining the cost function adequately allows misinformation to be minimized by decreasing the recommendation of content that exploits human cognitive biases.

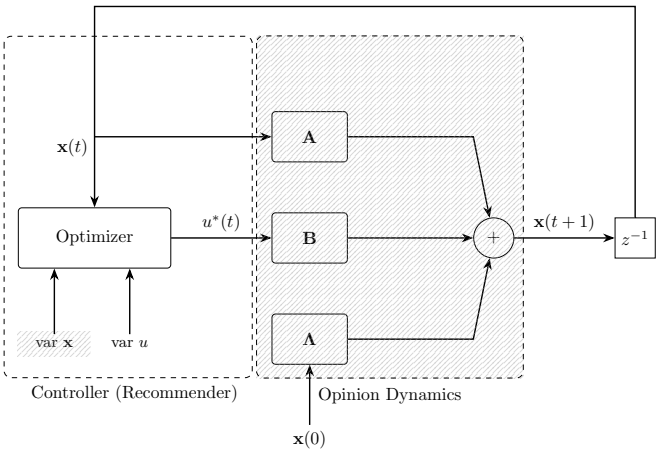


Fig. 1: Discrete-time closed-loop representation of the recommendation system control model. Shaded components indicate elements required only for the model-based approach found in II-C. var x and var u correspond to internal optimizer variables.

B. Cost Function Definition

A careful definition of the cost function allows for a mathematically tractable optimization problem. This can be done by defining a convex time-step specific cost function θ. The specific formulation of θ governs the platform's behavioral objectives. While the ultimate objective of this paper is misinformation mitigation, any practical solution must acknowledge that social network platforms fundamentally rely on user engagement. Therefore, successful misinformation mitigation strategies must be built upon existing engagement maximization frameworks rather than replacing them entirely. In the original model proposed by [9], user engagement maximization is achieved by defining θ as:

$$\theta(\mathbf{x}(t), u(t)) = \|\mathbf{x}(t) - u(t)\mathbf{1}_n\|_2^2 \quad (3)$$

With this representation, $\theta(\mathbf{x}(t), u(t))$ can be viewed as the squared Euclidean distance between the negative emotional extremity of the users $\mathbf{x}(t)$ and the negative emotional extremity of a proposed recommendation $u(t)$. Given this cost

function, the idealized recommender aims to minimize the infinite-horizon cost:

$$\min \sum_{t=0}^{\infty} \theta(\mathbf{x}(t), u) \tag{4}$$

While the original formulation in [9] applied (3) to thematic opinion agreement, the adaptation in this paper is for negative emotional extremity. This adaptation is valid because the original assumption of confirmation bias also applies to emotional content preferences. Building on this foundation, θ can be modified to penalize content exhibiting characteristics commonly associated with cognitive bias exploitation. Specifically, the characteristics of content negative emotional extremity, negative sentiment and novelty. This reformulation introduces appropriate disincentives for misinformation while maintaining fundamental engagement objectives. The individual components will be introduced in sequential order before the complete reformulated cost function.

1) Negative Emotional Extremity and Negative Sentiment: Negative emotional extremity measures the radical position of a statement, which contributes significantly to echo-chamber formation and is a characteristic often observed in misinformation content. The negative emotional extremity of a candidate social network recommendation can be penalized by adding:

$$E(u(t)) = \|u(t)\|^2, \tag{5}$$

to the existing cost function θ. The quadratic form is deliberately chosen to penalize increasingly extreme opinions while being lenient toward slightly biased content. This approach prevents overly censorial behavior for mild opinions while effectively discouraging highly extreme content.

2) Novelty: Misinformation research indicates that novel content contributes to virality and wider spreading of potentially harmful information. A novelty factor that modulates the emotional extremity penalty in (5) can be formulated as:

$$\begin{aligned} N(t, t_c) &= e^{-\lambda(t-t_c)} \\ \text{subject to:} \quad & t - t_c \leq z \end{aligned} \tag{6}$$

where λ controls the decay rate, t_c represents the creation time-step of the content and z is a fixed integer defining the eligible content window. This formulation ensures that extremely novel posts are initially penalized, but given a substantial amount of time, the penalty diminishes, allowing relevant information to persist in recommendations. Additionally, constraining $t - t_c$ to a bounded interval defined by z allows for mathematical tractability (discussed more in II-D), computational feasibility and a balance between content engagement with misinformation mitigation.

3) Combined Formulation: Incorporating both metrics with the engagement function θ yields the final modified objective function $\theta_m(\mathbf{x}(t), u(t))$ as:

$$\begin{aligned} \theta_m(\mathbf{x}(t), u(t)) &= \theta + a(n) \cdot \|u(t)\|^2 \cdot e^{-\lambda(t-t_c)} \\ \text{content subject to:} \quad & t - t_c \leq z \end{aligned} \tag{7}$$

where $a(n)$ is a scaling function and can be further expressed as $a(n) = \rho \cdot n$, where $\rho \geq 0$ is a rational number, controlling the overall strength of the misinformation penalty. The incorporation of n ensures that the effect of the penalty remains consistent regardless of the user network size. The full combined formulation in (7) maintains the original goal of engagement maximization while simultaneously discouraging the recommendation of content with characteristics commonly associated with misinformation. For θ_m, minimizing the infinite-horizon cost function as shown in (4) is still the idealized aim. Since it is not realistic to calculate the idealized infinite-horizon cost function, alternative methods with the same theoretical objective can be developed, specifically, model-free (MF) and model-based (MB) approaches.

C. Model-free and Model-based Approaches

1) Model-free Approach: The MF optimization problem can be given as

$$u_{MF}(t) = \arg \min_{u \in [0,1]} \theta_m(\mathbf{x}(t), u) \tag{8}$$

This approach attempts to minimize θ_m only for the time-step t by finding the most appropriate value for variable u. Consequently, the MF approach only requires $\mathbf{x}(t)$.

2) Model-based Approach: The MB optimization problem can be viewed as a Model predictive control (MPC) technique. The theoretical optimal negative emotional extremity steady-state is given by:

$$\begin{aligned} (\mathbf{x}^*_{MB}, u^*_{MB}) &= \arg \min_{\mathbf{x}, u} \theta_m(\mathbf{x}, u) \\ \text{subject to} \quad & \mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}u + \mathbf{\Lambda}x(0), \\ & u \in [0, 1]. \end{aligned} \tag{8}$$

where $\mathbf{x}^*_{MB}$ and u^*_{MB} are the theoretical optimal states. The optimal states in (8) always exist (see II-D for details). The steady-state $\mathbf{x}^*_{MB}$ and the general control loop dependencies are used to help solve the full MPC formulation, given as

$$\begin{aligned} O^*_t := \min_{\mathbf{x}_{\xi|t}, u_{\xi|t}} & \sum_{k=0}^{T-1} \theta_m(\mathbf{x}_{k|t}, u_{k|t}) \\ \text{subject to} \quad & \mathbf{x}_{k+1|t} = \mathbf{A}\mathbf{x}_{k|t} + \mathbf{B}u_{k|t} + \mathbf{\Lambda}\mathbf{x}(0), \\ & \mathbf{x}_{0|t} = \mathbf{x}(t), \\ & \mathbf{x}_{T|t} = \mathbf{x}_{MB}, \\ & u_{k|t} \in [0, 1], \\ \text{for all} \quad & k \in [0, T-1] \end{aligned} \tag{9}$$

where O^*_t is the general optimization cost function and T is the prediction horizon. The actual optimizer output $u_{MB}(t)$ at each time-step is the first output that satisfies the MPC formulation, thus, $u_{MB}(t) = u_{0|t}$. It is clear that the MF and MB approaches have significant differences and scopes of informational access. Unlike the MF approach, the MB approach must have access to the opinion dynamic dependencies $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{\Lambda}$.

D. Mathematical Analysis and Convergence Proofs

The new cost function definition in (7) requires updated analysis of the control strategies and their convergence properties. While the Friedkin-Johnsen model structure and graph properties from [9] remain unchanged, the modified optimization problem necessitates new theoretical results. In addition, all steady-state convergence values can be seen as a region of convergence rather than a single point, this is due to the reliance of all steady-state solutions on $t - t_c$.

1) MF Steady-state Recommender Output Value: The steady state recommender output u^*_{MF} can be found by first reformulating θ_m and then taking the instantaneous minimum by $\frac{\partial \theta_m}{\partial u} = 0$ and solving for $u(t)$ to find $u^*_{MF}(t)$, which gives:

$$u^*_{MF}(t) = \frac{\sum_{i=1}^{n} x_i(t)}{n(1 + \rho \cdot e^{-\lambda(t-t_c)})} \tag{10}$$

2) MF User Steady-state and Adjacency Matrix Analysis: Substituting u^*_{MF} into the system dynamics given in (2) and rewriting gives:

$$\mathbf{x}(t+1) = (\mathbf{I}_n - \mathbf{\Lambda})\mathbf{F}\mathbf{x}(t) + \mathbf{\Lambda}\mathbf{x}(0) \tag{11}$$

where:

$$\mathbf{F} = \mathbf{W} + \frac{\mathbf{w_{rec}}\mathbf{1}_n^T}{n(1 + \rho \cdot e^{-\lambda(t-t_c)})} \tag{12}$$

The $\mathbf{F}$ matrix can be seen as an adjacency graph. It is sub-row stochastic, meaning that it satisfies convergence guarantees and is well-posed, as preliminary required from [9].

To get $\mathbf{x}^*_{MF}$, the property $\mathbf{x}(t) = \mathbf{x}(t+1)$ must be satisfied. Substituting this property into (11) and solving for $\mathbf{x}(t)$ gives:

$$\mathbf{x}^*_{MF} = \left(\mathbf{I}_n - \mathbf{A} - \frac{\mathbf{B} \cdot \mathbf{1}_n^T}{n(1 + \rho \cdot e^{-\lambda(t-t_c)})}\right)^{-1} \mathbf{\Lambda}\mathbf{x}(0) \tag{13}$$

3) MB User Steady-state and Recommender Output: The steady-state values of u^*_{MB} and $\mathbf{x}^*_{MB}$ can be found by defining the Lagrangian and applying Karush-Kuhn-Tucker (KKT) conditions. The full derivation is left out due to space constraints. The final interior solution ($(0 < u < 1)$) gives:

$$u^*_{MB} = \frac{\mathbf{1}_n^T\mathbf{y} - \mathbf{v}^T\mathbf{y}}{-\mathbf{1}_n^T\mathbf{v} + n + \mathbf{v}^T\mathbf{v} - \mathbf{v}^T\mathbf{1}_n + \rho n e^{-\lambda(t-t_c)}} \tag{14}$$

$$\mathbf{x}^*_{MB} = \mathbf{v}u^*_{MB} + \mathbf{y} \tag{15}$$

where $\mathbf{v} = (\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}$ and $\mathbf{y} = (\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{\Lambda}\mathbf{x}(0)$.

4) MB Convergence Analysis: The proving process for u^*_{MB} and $\mathbf{x}^*_{MB}$ convergence is the same as in [9]. The only change is found in the definition of the new $\mathbf{H}$ matrix, which is defined as:

$$H = \begin{bmatrix} \mathbf{I}_n & -\mathbf{1}_n \\ -\mathbf{1}_n^T & n(1 + \rho e^{-\lambda(t-t_c)}) \end{bmatrix} \tag{16}$$

See [9] for further steps, since they remain identical.

E. Large Language Models For Negative Emotional Extremity

For the data-based simulations, real-life qualitative candidate content l must be transformed into a quantitative option $\mathcal{C}(l) \in [0, 1]$. This can be done using large language models (LLMs) to create a weighted approximation. Content can be processed through a pre-trained LLM to extract emotional features. The LLM produces scores across the relevant emotional dimensions: fear, disgust, anxiety and shock. A general assessment of negative sentiment and opinion subjectivity is also generated. Opinion subjectivity is measured because emotional extremity heavily correlates with subjective language [11]. These scores are combined using a matrix multiplication approach:

$$\mathcal{C}(l) = \mathbf{c}^T\mathbf{f}(l), \tag{17}$$

where the stochastic vector $\mathbf{c}^T = [w_{e1}\ w_{e2}\ w_{e3}\ w_{e4}\ w_{ng}\ w_{su}]$ represents the vector of weights and the vector $\mathbf{f}^T(l) = [p_{e1}(l)\ p_{e2}(l)\ p_{e3}(l)\ p_{e4}(l)\ p_{ng}(l)\ p_{su}(l)] \in [0, 1]^6$ represents the vector of features. Subscripts $e.1 \cdots, e.4$ refer to fear, disgust, anxiety and shock, while subscripts ng and su refer to negative sentiment and opinion subjectivity.

F. Dataset Tailoring and Simulation Details

The simulated social network environment was developed using Python. All used variables are determined using a random defined seed. Data for evaluating the model is taken from [10]. The dataset was tailored by extracting only the text content of all social media posts, comments, public announcements, speeches, articles and news. In addition, the corresponding truthfulness labels of 'Pants On Fire' (very false), 'False' 'Mostly-True' and 'True' were extracted for all the content in the dataset. For simplicity of evaluation, the 'Mostly-True' and 'True' were combined into a general 'True' label while the 'Pants On Fire' and 'False' label were combined into a general 'False' label. The textual content was then analyzed by the method described in (II-E) using the Mistral NeMo model through an API integration. The NeMo model was chosen due to its small size, open-source nature, accuracy and efficiency. For realism and time-step indication, the true and false content was randomly distributed among the available time-steps. This was done by creating a randomized "time of creation" (t_c) for all data points.

For the data-based simulation, the controller applies the same MF and MB logic as found in (II-C), but using discrete optimization. The optimizer selects the best content to share in the MF approach, and selects the best sequence of accessible content to share for the MB approach. The misinformation mitigation metric $\mathcal{M}$ is calculated at every run of the data-based simulation, given as:

$$\mathcal{M} = \frac{\sum_{i=1}^{\tau} \mathbf{1}_{[\text{label}_i = \text{false}]}}{l_t + l_f} \tag{18}$$

where l_t refers to the size of 'True'-labelled data and l_f refers to the size of the 'False'-labelled data.

III. Results

A. Parameters and Metrics

Table I shows all the set parameters for both the theoretical synthetic and data-based simulations. The parameter λ was set to 0 for all simulations because the LIAR2 dataset does not include social network shareability timestamps, only approximate creation dates. Thus, for realism, the effect of novelty is ignored by setting $\lambda = 0$. In addition, $\lambda = 0$ allows for a clearer comparison between synthetic and data-based results. Table II contains parameters for the data-based simulations, while Table III shows statistics after applying the LLM-based approximation described in II-E. The misinformation metric (18) is used to evaluate model effectiveness for the data-driven simulations.

TABLE I: General and Data-based Simulation Parameters

Parameter	Description	Value
n	Number of users	100
Λ_h	Highest stubbornness	0.05
Λ_l	Lowest stubbornness	0.00
κ_u	User-to-user connectivity	0.25
κ_r	Recommender-to-user connectivity	0.80
τ	Time-steps	100
ρ	Penalty strength regulator	[0.00 5.50 0.10]
T	Prediction horizon	50
z	Eligible content window	5
λ	Decay rate	0.00

TABLE II: Additional Dataset-based Simulation Parameters

Parameter	Description	Value
l_t	'True' dataset size	2000
l_f	'False' dataset size	2000
c_l	Content per time-step	40
$\mathbf{c^T}$	LLM weights	[0.15 0.15 0.15 0.15 0.2 0.2]

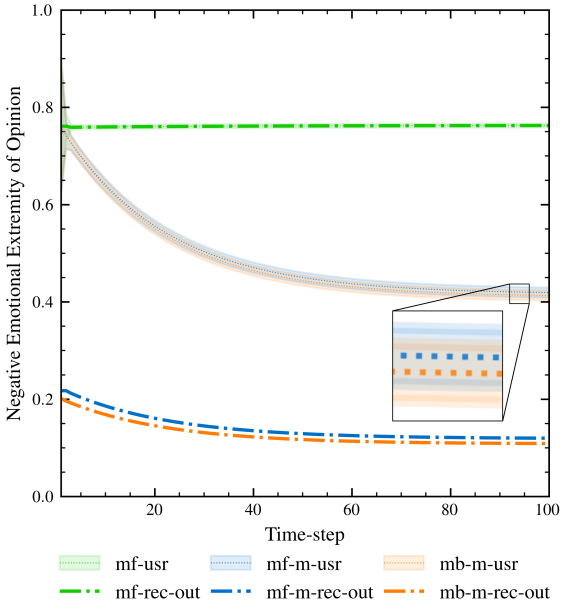
TABLE III: Post-LLM Dataset Content Statistics

Label	Mean	Min	Max	Median
False	0.537	0.000	1.000	0.564
True	0.379	0.000	0.920	0.372
Total Dataset Size: 4000				

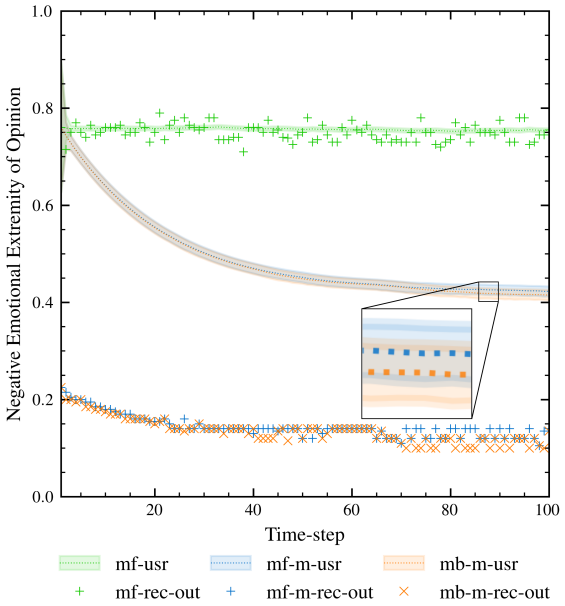
B. Synthetic and Data-based Results

Figure 2 shows the evolution of the negative emotional extremity of opinions using θ and θ_m for both synthetic and data-based simulations. The input values are $\rho = 2.5$ and all remaining values from Table I. For the data-based results shown in Figure 2b, the Misinformation metric $\mathcal{M}$ is calculated and subdivided by the MF and MB approach (only MF for θ). The results are shown in Table IV.

Figure 3 shows data-based simulation results for the misinformation mitigation metric $\mathcal{M}$ when varying ρ from 0.00 to 5.50 in steps of 0.10. Both MF and MB results are shown.



(a) Synthetic



(b) Data-based

Fig. 2: Comparison between models implementing θ and θ_m for both synthetic and data-based simulations, with $\rho = 2.50$. "-rec-out" represents $u^*(t)$ at each time-step and "-usr" represents the mean of $\mathbf{x}(t)$ with standard deviation at each time-step. "-m-" represents the model using θ_m

.

TABLE IV: Data-driven Misinformation Metric $\mathcal{M}$ Results

ρ	θ	θ_m	
	MF	MF	MB
2.50	0.02100	0.00700	0.00625

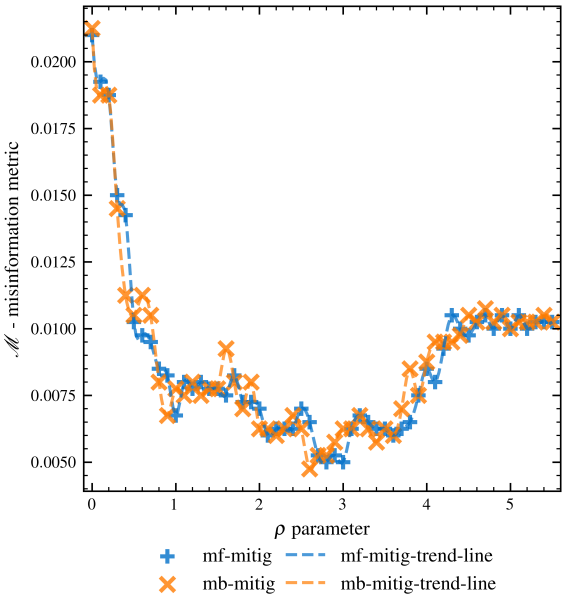


Fig. 3: Evolution of $\mathcal{M}$ with changing ρ implemented with parameters and values defined in Table I.

IV. Discussion

The modified engagement function θ_m consistently demonstrates lower misinformation spread compared to the original engagement function θ across all scenarios. The individual simulation with misinformation metric results from Table IV show a 67% decrease in misinformation spread for the MF approach, and a 70% decrease for the MB approach when compared to the θ-MF approach. Additional preliminary tests (not shown due to space constraints) further show theoretical and practical alignment for the convergence values found in II-D1.

Figure 3 indicates a misinformation decrease of up to 76% for the most ideal ρ value. However, Figure 3 additionally represents a discrepancy with theory for $\rho > 3$. Instead of misinformation decreasing or remaining the same, it increases. A possible reason for this is that some content in the LIAR2 dataset is described in the third person, creating an inconsistency for the LLM. A potential additional reason is that a minority of misinformation has well-designed objective tone for increased believability [12].

In hindsight, while altering the LIAR2 dataset to binary labels simplifies model evaluation, it might lead to indistinguishability. In Table III, the difference between the emotional extremity mean of "False" and "True" content is only 0.158. This could explain the increased misinformation effect found in Figure 3, since information can have varying levels of truthfulness.

Future work should focus on testing larger and more detailed alternative datasets to enhance the model effectiveness and realism. In particular, datasets that reflect real-time social network time-flows, as this would allow the novelty aspect to be data-tested and evaluated. The discretization of labels to varying levels of truthfulness, and the analogous change of the misinformation metric to account for it, is an additional possibility. Further enhancements could implement the inclusion of the variable ρ and metric $\mathcal{M}$ into the closed-loop. This could be beneficial by creating a self-regulating system that can dynamically react to disinformation-campaigns or sensitive periods (e.g. elections). Alternative custom trained LLMs or alternative sentiment analysis tools can also be developed for real-time content analysis, increasing accuracy and speed.

V. Conclusions

Despite the aforementioned limitations in IV, the results confirm that the enhanced recommendation system with θ_m can effectively reduce misinformation while maintaining reasonable user engagement. This is done by penalizing content with negative emotional extremity commonly found in misinformation. Simulations demonstrate up to 76% reduction in misinformation spread while maintaining user engagement. Both MF and MB approaches work successfully with minimal difference, thus, the MF approach is preferred due to easier and faster practical application. The framework and additional mathematical proofs provides a foundation for future development of social dynamics-aware recommendation systems on social network platforms.

References

[1] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

[2] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[3] N. Persily, "The 2016 us election: Can democracy survive the internet?" *Journal of democracy*, vol. 28, no. 2, pp. 63–76, 2017.

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[6] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7313–7318, 2017.

[7] J. Berger and K. L. Milkman, "What makes online content viral?" *Journal of marketing research*, vol. 49, no. 2, pp. 192–205, 2012.

[8] M. D. Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "Echo chambers in the age of misinformation," 2015. [Online]. Available: https://arxiv.org/abs/1509.00189

[9] B. Sprenger, G. De Pasquale, R. Soloperto, J. Lygeros, and F. Dörfler, "Control strategies for recommendation systems in social networks," *IEEE Control Systems Letters*, vol. 8, pp. 634–639, 2024.

[10] C. Xu and M.-T. Kechadi, "An enhanced fake news detection system with fuzzy deep learning," *IEEE Access*, vol. 12, pp. 88 006–88 021, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3418340

[11] M. D. Rocklage, D. D. Rucker, and L. F. Nordgren, "The evaluative lexicon 2.0: The measurement of emotionality, extremity, and valence in language," *Behavior Research Methods*, vol. 50, pp. 1327 – 1344, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:22785347

[12] D. A. Scheufele and N. M. Krause, "Science audiences, misinformation, and fake news," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7662–7669, 2019. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1805871115