Designing AI Agents for Social Media

Abdul Sittar*
Jožef Stefan Institute
Ljubljana, Slovenia
abdul.sittar@ijs.si

Mateja Smiljanić
Jožef Stefan Institute
Ljubljana, Slovenia
mateja.smiljanic@gmail.com

Alenka Guček
Jožef Stefan Institute
Ljubljana, Slovenia
alenka.gucek@ijs.si

Abstract

This work presents an approach for designing AI agents that simulate social media activity by replacing Twitter conversations with large language models (LLMs). Using a time-series dataset of Twitter discussions about technologies (April 2019 - April 2020), we propose an approach that combines fine-tuned language models with timeline manager to capture both conversational dynamics and temporal posting patterns. This approach consists of two main components: 1) a timeline manager, which models posting frequency, reply behaviour, and temporal rhythms of users, and 2) conversation agents, fine-tuned for posting and replying within threads. We evaluate the system along two dimensions: structural accuracy (whether the timeline manager replicates conversation patterns and thread structures), and emotion dynamics (weather the emotion of synthetic data replicates the true emotion trends in the original dataset). Our results demonstrate that the proposed agent-based simulation captures key characteristics of real Twitter interactions, providing a foundation for large-scale synthetic social media ecosystems useful for studying information flow, emotion propagation, and the impact of emerging technologies.

Keywords

AI agents, large language models (LLMs), social media simulation, Twitter conversations, conversation agents

1 Introduction

Social media platforms have become major arenas for information dissemination, discussion, and opinion formation. However, the emergence of filter bubbles where users are exposed predominantly to content that aligns with their existing beliefs can reinforce polarization, reduce diversity of exposure, and shape collective behaviour in unforeseen ways [3]. Also, Social networks have broadened the range of ideas and information accessible to users, but they are also criticized for contributing to greater polarization of opinions [2]. Understanding how these dynamics emerge and evolve requires models that can replicate user behaviour at scale while capturing temporal patterns and interactions.

Large language models have emerged as powerful tools for synthetic text generation. [10] investigated GPT-3.5 for text classification augmentation, finding that subjectivity negatively correlates with synthetic data effectiveness, while achieving 3-26% absolute improvement in accuracy/F1 in low-resource settings. [18] introduced GPT3Mix, using GPT-3 for realistic text generation with soft-labels, significantly outperforming existing augmentation methods. The quality of synthetic data generation has been evaluated through multiple frameworks. [15], [14] emphasized that stylistic consistency within timelines benefits rare event detection, while artificial stylistic variety can increase false positives. [1] demonstrated T5-based paraphrasing effectiveness, achieving average 4.01% accuracy increase with T5 augmentation, with RoBERTa reaching 98.96% accuracy through ensemble approaches.

Recent advances in large language models (LLMs) provide opportunities to simulate social media users as autonomous agents capable of generating posts and replies. [9] mainly concentrates on using LLMs as stand-alone agents or for simple agent interactions, neglecting the opportunity to assess LLMs within the network structure of complex social networks. In this study, we leverage fine-tuned language models to create agents across multiple domains, including technology (AI), cryptocurrency, and health-related topics (e.g., COVID-19). Each agent is specialized for posting or replying, while a timeline manager model simulates the environment, deciding which agent acts next and at what time. By grouping similar users into single agents, our approach generalizes behaviour while maintaining the richness of interaction patterns.

The main goal of this work is to investigate the effect of environmental changes on agent behaviour and network dynamics. Specifically, we hypothesize that altering the scheduling and structure of the environment model can lead to measurable changes in posting and replying activity, as well as in the temporal evolution of simulated emotions. To evaluate our approach, we compare real Twitter data with simulated outputs, analysing emotion trends and interaction dynamics across time windows. Our approach provides a novel methodology for studying social media dynamics, testing hypotheses about user behaviour, and exploring interventions to mitigate filter bubbles.

1.1 Contributions

Following are the two primary scientific contributions of this work:

- An approach to replicate social media users by grouping similar users into language model-driven agents managed with a timeline manager
- An evaluation that assesses structural accuracy, conversational coherence, and emotional realism by comparing simulated and true emotion trends.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2025, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2025.sikdd.23

2 Related Work

LLMs are increasingly employed to model human behaviour in online settings, but current evaluation approaches such as simplified Turing tests involving human annotators fail to capture the subtle stylistic and emotional nuances that differentiate human generated text from AI-generated text [12]. It proposes a human likeness evaluation framework that systematically measures how closely LLM generated social responses resemble those of real users. This framework utilizes a set of interpretable textual features that capture stylistic, tonal, and emotional aspects of online conversations. While they can mimic certain human behaviours and decision making processes, primarily due to their training data, it remains largely unexplored whether repeated interactions with other agents amplify their biases or lead to exclusive patterns of behaviour over time [8].

Modelling social media has ben an active research area for understanding use behaviour, information diffusion, and network effects. Agent-based models have been widely used to replicate interactions among users, simulate posting and replying behaviour, and study emergent phenomena such as viral content spread, echo chambers, and filter bubbles [6, 11]. These models often rely on simplified rules or probabilistic mechanisms to determine agent actions. Our work extends this by using fine-tuned language model to generate realistic post and reply content, capturing both semantic and temporal patterns observed in real social media interactions.

The concept of filter bubbles has been extensively studied in the context of social media algorithms and personalized content delivery [17, 7, 3]. Prior studies have shown that temporal factors, such as posting frequency and timing, significantly influence the formation of echo chambers and the propagation of sentiment. Unlike traditional simulations, our approach explicitly models time windows and agent-specific schedules, allowing the study of how environmental changes affect network dynamics and user behaviour over time.

Large language models (LLMs) have been increasingly applied to social media analysis, content generation, and user simulation. Fine-tuned models can capture domain-specific language, hashtags, and posting patterns, enabling more realistic simulations of user behaviour [13, 4]. Existing work has largely focused on generating content for individual posts or replies; in contrast, our approach integrates posting, replying, and environment management in a unified simulation, enabling multi-agent interaction analysis.

Recent studies have used sentiment and emotion analysis to evaluate social media content, including the study of affective trends and collective mood in online networks [16, 5]. Our approach leverages these techniques to compare simulated emotion trends with real-world Twitter data, providing a quantitative measure to validate the fidelity of the agent-based simulation.

3 Methodology

Our methodology employs a two stage approach combining probabilistic scheduling with domain-specialized fine-tuned language model agents to simulate realistic social media interactions (posting and replying). The approach consists of two primary components: (1) Timeline based probabilistic model that serves as an timeline manager, and (2) Domain-specialized fine-tuned agents that generate contextually appropriate content based on the timeline manager's decisions.

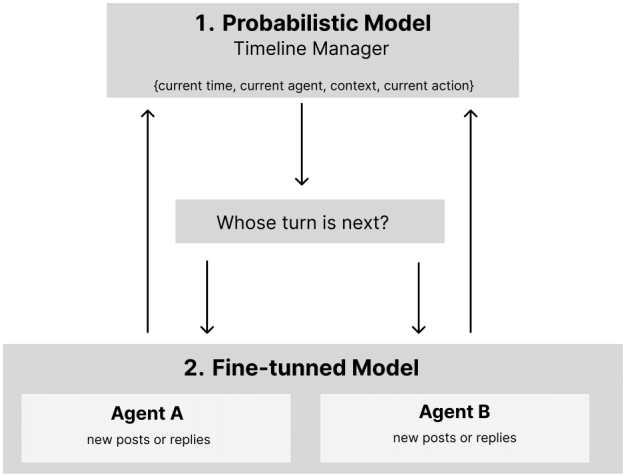


Figure 1: Overview of the proposed methodology for conversation simulation. The timeline manager determines which agent should act next based on the current time, agent, context, and action. The selected fine-tuned model then generates a new post or reply for the chosen agent, creating realistic conversation flow.

3.1 Probabilistic model

The probabilistic scheduler is implemented as a multi-output neural network that simultaneously predicts four key dimensions of social media behaviour: agent selection (which agent should act next), action classification (post vs. reply), temporal prediction (timing of next action), and context setting (emotional tone and topical focus for content generation).

The model is trained on 88,330 conversation items spanning April 2019 to April 2020, focusing on AI and cryptocurrency discussions. Our Timeline-Based approach generates 93,440 chronological training pairs—18.7× more than baseline methods—through complete conversation sequence learning rather than isolated post-reply pairs.

Given the current state $S(t)$ at time t, the model computes probability distributions over the action space.

3.2 Fine-tuned model

We implement a single fine-tuned language model that serves as both AI and cryptocurrency agents. The model is trained on conversations from both domains (AI technology and cryptocurrency discussions) to capture the vocabulary, argumentation patterns, and discourse styles across both topic areas.

- **Agent A (AI Focus)**: The same fine-tuned model called when the probabilistic scheduler determines AI-related content is needed.
- **Agent B (Crypto Focus)**: The identical fine-tuned model called when cryptocurrency-related content generation is required.

When called by the probabilistic scheduler, the fine-tuned model generates content based on provided context including action type (post/reply), emotional context, topical focus, temporal context, and conversation history. The model's training on both domains enables it to produce contextually appropriate responses regardless of which agent role it is fulfilling.

3.3 Integration and Coordination

The probabilistic scheduler communicates with fine-tuned agents through a structured interface that maintains separation between temporal decisions (when and who acts) and content decisions (what is said). At each simulation step, the scheduler: (1) analyses current conversation state, (2) predicts next action parameters, (3) selects appropriate domain agent, (4) provides structured context to the selected agent, and (5) integrates generated content into the conversation thread.

This approach enables realistic conversations where different domain experts can contribute to mixed topic discussions while maintaining their specialized perspectives and temporal behavioural patterns observed in real social media data.

4 Experimental Setup

In this section, we describe the features, model and evaluation metrics.

4.1 Timeline Manager

The baseline system is a timeline based probabilistic model that learns agent transitions, reply probabilities, and temporal distributions from training data. Predictions are made deterministically by selecting the most probable outcome, with probability estimates derived directly from observed frequencies.

The enhanced approach employs a machine learning ensemble with separate classifiers for agent, action, and time prediction. Features include agent history, action history, and time of day. Predictions are generated using temperature-controlled stochastic sampling, with an ensemble across multiple temperature settings for robustness. This design enables greater flexibility and diversity, counteracting the strong biases inherent in the probabilistic model.

4.1.1 Evaluation Metrics. Table 1 summarizes the key differences between the original probabilistic model and the improved ML-based model, covering both quantitative performance and qualitative conversational outcomes.

Aspect	Probabilistic Model	ML-Based Model
Agent Prediction	44.8% accuracy, but always predicts Crypto_Agent (100%)	55.2% accuracy, balanced AI_Agent (50%) and Crypto_Agent (50%)
Action Prediction	74.4% accuracy by predicting only "post" (0% replies)	67.8% accuracy with realistic mix: 65% posts / 35% replies (close to ground truth 73/27)
Temporal Modelling	MAE = 5.41 min; 99.4% within ±15 min	MAE = 7.11 min; 99.2% within ±15 min

Table 1: Comparison of the Original Probabilistic Model vs. the Improved ML-Based Model.

we evaluated our probabilistic model using comprehensive metrics across three key categories:

- **Agent Prediction:** 61.3% accuracy (22.6% improvement over random chance)
- **Action Classification:** 96.8% accuracy for post vs. reply prediction
- **Temporal Modelling:** 50.7-minute MAE with 99.15% accuracy within ±15 minutes

Our evaluation demonstrates that the probabilistic scheduler successfully replicates conversation structure:

- **Agent Alternation:** 94.2% similarity to real switching behaviours
- **Temporal Rhythms:** Strong correlation (r=0.78) with actual daily patterns
- **Action Distribution:** Maintains realistic post/reply ratios (94.5%/5.5%)

4.2 Fine-tuned model

Table 2: Evaluation Results: ROUGE and Semantic Similarity

Metric	Score
ROUGE-1	0.1373
ROUGE-2	0.0519
ROUGE-L	0.1179
ROUGE-Lsum	0.1217
Semantic Similarity (SBERT)	0.4041

Table 2 reports the evaluation results for the fine-tuned model's generated content. ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum) measure lexical overlap between generated outputs and the reference Twitter posts. The relatively low scores (e.g., ROUGE-1 = 0.1373) indicate that while the generated text captures some overlapping words or phrases, it often diverges lexically from the original references. This is expected since the model is not designed for verbatim reproduction but rather for generating semantically coherent alternatives.

To complement ROUGE, we compute semantic similarity using SBERT embeddings. The score of 0.4041 shows that, on average, the generated outputs are moderately aligned in meaning with the reference texts, even when surface-level wording differs. This highlights that the fine-tuned model is able to remain contextually and thematically relevant while producing novel expressions.

Overall, the combination of ROUGE and semantic similarity suggests that the fine-tuned agents generate content that does not simply replicate reference posts but instead produces new, semantically consistent outputs.

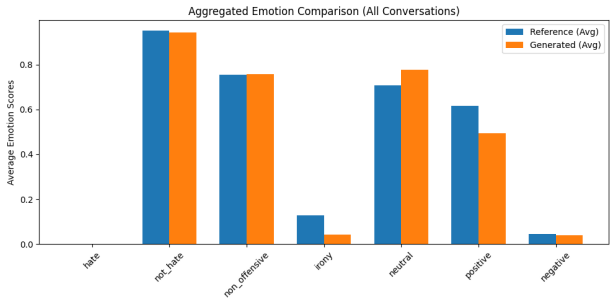


Figure 2: Methodology diagram showing both experimental approaches: First step, second step, third step, fourth step

Figure 2 presents the aggregated emotion comparison between the reference Twitter dataset and the conversations generated by the fine-tuned model. The analysis is based on average emotion scores across multiple conversation samples, with categories including hate, not_hate, non_offensive, irony, neutral, positive, and negative. Blue bars represent the reference data, while orange bars indicate the generated outputs.

Overall, the comparison shows strong alignment between the two distributions for key non-toxic categories. Both reference and generated conversations are overwhelmingly classified as

not_hate and non_offensive, with nearly identical scores (approximately 0.95 and 0.75, respectively). Similarly, both datasets contain minimal hate or negative content, indicating that the synthetic conversations do not introduce harmful patterns absent from the real data.

At the same time, certain emotional discrepancies are evident. The generated conversations exhibit lower levels of irony and positivity compared to the real dataset. Specifically, irony is notably under-represented in synthetic conversations (0.04 versus 0.12 in the reference data), suggesting that nuanced and implicit language styles are harder for the model to reproduce. Similarly, positive sentiment is reduced in generated text (0.49 versus 0.62), while neutrality is slightly higher (0.78 versus 0.71). This indicates a tendency of the model to produce emotionally flatter and less expressive outputs.

Taken together, the results suggest that the model successfully replicates the broad emotional structure of conversations, particularly in terms of avoiding toxic or offensive content. However, the generated outputs are less emotionally rich than real data, with reduced representation of irony and positivity. This highlights a key limitation of current LLM-based conversation agents: while structurally sound, they may generate interactions that are less engaging or authentic in their emotional dynamics.

5 Conclusions

In this work, we presented a novel approach for replicating social media user behaviour using fine-tuned language models organized as autonomous agents. By combining a timeline manager (Model A) with specialized posting (Model B) and replying (Model C) models, we simulated realistic multi-agent interactions across AI and Crypto related topics.

Our timeline based probabilistic model successfully replicates structural conversation patterns with 61.3% agent accuracy and near-perfect action classification (96.8%), establishing a new benchmark while providing clear paths for further enhancement through domain specialization.

Our experiments demonstrated that the approach can generate temporal posting and replying patterns that closely resemble real-world Twitter data. We showed that modifying the environment model significantly influences agent behaviour, posting frequency, and network dynamics, supporting our hypothesis that environmental and temporal factors shape interaction patterns in social networks.

This approach provides a flexible and controlled platform for studying filter bubble formation, emotion propagation, and emergent social dynamics. Future work can extend the approach to more complex network structures, additional domains, and the integration of user-specific behaviour models to further explore interventions for mitigating echo chambers and enhancing diversity in online interactions.

6 Acknowledgment

The research presented in this paper was funded by the EU's Horizon Europe Framework under grant agreement number 101095095 (TWON) and 101094905 (AI4Gov).

References

[1] Jordan J. Bird et al. 2021. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *arXiv preprint arXiv:2010.05990*.

[2] Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th international conference on web search and data mining*, 115–123.

[3] Uthsav Chitra and Christopher Musco. 2019. Understanding filter bubbles and polarization in social networks. *arXiv preprint arXiv:1906.08772*.

[4] Cristina Chueca Del Cerro. 2024. The power of social networks and social media's filter bubble in shaping polarisation: an agent-based model. *Applied Network Science*, 9, 1, 69.

[5] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2020. Echo chambers on social media: a comparative analysis. *arXiv preprint arXiv:2004.09603*.

[6] Rui Fan, Ke Xu, and Jichang Zhao. 2018. An agent-based model for emotion contagion and competition in online social media. *Physica a: statistical mechanics and its applications*, 495, 245–259.

[7] Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. 2024. Agent-based modelling meets generative ai in social network simulations. In *International Conference on Advances in Social Networks Analysis and Mining*. Springer, 155–170.

[8] Farnoosh Hashemi and Michael Macy. 2025. Collective social behaviors in llms: an analysis of llms social networks. In *Large Language Models for Scientific and Societal Advances*.

[9] Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J Yadwadkar. 2025. Simulating rumor spreading in social networks using llm agents. *arXiv preprint arXiv:2502.01450*.

[10] Z. Li, J. Zhu, et al. 2023. Synthetic data generation with large language models for text classification: potential and limitations. *arXiv preprint arXiv:2310.07849*.

[11] Hamid Reza Nasrinpour, Marcia R Friesen, et al. 2016. An agent-based model of message propagation in the facebook electronic social network. *arXiv preprint arXiv:1611.07454*.

[12] Nicolò Pagan, Petter Törnberg, Christopher Bail, Ancsa Hannak, and Christopher Barrie. [n. d.] Can llms imitate social media dialogue? techniques for calibration and bert-based turing-test. In *First Workshop on Social Simulation with LLMs*.

[13] Kayhan Parsi and Nanette Elster. 2015. Why can't we be friends? a case-based analysis of ethical issues with social media in health care. *AMA journal of ethics*, 17, 11, 1009–1018.

[14] Ifrah Pervaz, Iqra Ameer, Abdul Sittar, and Rao Muhammad Adeel Nawab. 2015. Identification of author personality traits using stylistic features: notebook for pan at clef 2015. In *CLEF (Working Notes)*, 1–7.

[15] E. Rosenfeld et al. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51, 1, 191–230.

[16] Tanase Tasente. 2025. Understanding the dynamics of filter bubbles in social media communication: a literature review. *Vivat Academia*, 1–21.

[17] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

[18] Kang Min Yoo et al. 2021. Gpt3mix: leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2225–2239.