Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TWON social media: a scalable MERN-Stack platform for experimental research in online social networks

Abdul Sittar
abdul.sittar@ijs.si
Jozef Stefan Institute
Michael Heseltine
University of Amsterdam
Francois t'Serstevens
University of Amsterdam
Natan Viteznik
Slovenian Press Agency
Corinna Oschatz
University of Amsterdam
Mateja Smiljanic
University of Ljubljana
Alenka Gucek
Jozef Stefan Institute
Damian Trilling
Vrije Universiteit Amsterdam
Marko Grobelnik
Jozef Stefan Institute

Research Article

Keywords: online social networks, experimental social media platforms, computational social science, information diffusion, agent-based simulation, misinformation research, MERN stack

Posted Date: April 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9332950/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Online social networks play a central role in shaping public discourse, yet con- ducting controlled experimental research on such platforms remains challenging due to limited access, lack of transparency in ranking algorithms, and restricted intervention capabilities. This paper presents TWON, a scalable and modular

social media platform designed to enable controlled, reproducible experimen- tation on user behavior, information diffusion, and algorithmic interventions. The platform further incorporates large language model (LLM) capabilities for content generation, moderation, and agent-based simulation, enabling hybrid experimental designs that combine human participants with automated agents. The system is validated through multiple empirical deployments across diverse research contexts, including (i) a disinformation study analyzing user engagement with manipulated news content (272 participants), (ii) a scientific communication study evaluating pre-bunking and uncertainty interventions (1200 participants),

(iii) a toxicity prevention study leveraging real-time AI-assisted comment rewrit- ing (574 participants), and (iv) large-scale agent-based simulations exploring conversational dynamics under different ranking strategies (54 agents). Across these studies, TWON supports flexible experimental configurations and cap- tures fine-grained behavioral data, enabling systematic analysis of engagement patterns and intervention effects.

1 Introduction

Online social networks (OSNs) have become central infrastructures for public dis- course, civic participation, and democratic engagement. Their large scale, algorithmic mediation, and complex interaction dynamics, however, make them difficult to analyze under controlled experimental conditions [1]. Empirical studies conducted directly on commercial platforms are often constrained by limited data access, restricted experimental control, and ethical or regulatory considerations.

Despite the growing importance of online social networks for public discourse, con- ducting controlled experimental research on these platforms remains challenging [2]. Researchers face limitations such as restricted API access, opaque ranking algorithms, and limited intervention capabilities [3]. These barriers impede causal analysis of information diffusion, algorithmic amplification, and user behavior in social media ecosystems.

While prior work has explored experimental web environments and agent-based simulations, these systems often lack scalability, modular architecture, or integration with modern AI-driven interaction models. As a result, researchers lack a unified infrastructure capable of supporting both human-subject experiments and large-scale agent-based simulations within a controlled social media environment [4, 5].

To address these limitations, this work introduces a set of experimental software platforms collectively referred to as TWONs. These platforms are designed as simu- lated digital replicas of OSNs, enabling controlled and reproducible experimentation on information diffusion, user behavior, and the dynamics of online public debate. By abstracting core functional elements of social networking systems—such as user

interaction mechanisms, content feeds, and engagement metrics—TWONs provide a flexible and modular research infrastructure.

Beyond experimental control, the platforms support methodological develop- ment, system validation, and public outreach activities. Their architecture enables configurable study designs, systematic data collection, and integration with com- putational models, thereby bridging the gap between empirical user studies and simulation-based analysis. As such, TWONs constitute a technical framework for inves- tigating socio-technical phenomena in online environments under rigorously controlled conditions.

The goal of this work is to design and deploy a scalable experimental social media infrastructure that enables controlled manipulation of platform dynamics while maintaining ecological realism. The proposed system allows researchers to study infor- mation diffusion, algorithmic interventions, and social interaction behavior under reproducible experimental conditions.

1.1 Contributions

- TWON social media platform provides a flexible and modular MERN-stack archi- tecture that enables researchers to simulate realistic social media environments while maintaining full experimental control.
- It supports diverse experimental interventions, including content ranking manipu-

lation, multi-round feed dynamics, misinformation exposure, uncertainty communi- cation, and AI-assisted paraphrasing.

- The platform has been validated across three independent case studies conducted

by different research teams in multiple linguistic and thematic contexts.

- It enables LLM-based fine-tuned agents to interact with each other within the plat- form, supporting fully simulated social media ecosystems while also allowing future integration of human–agent interactions and hybrid experimental designs.

2 Related work

Research on online social networks has traditionally relied on observational data collected from large-scale platforms such as Twitter, Facebook, and, more recently, decentralized systems such as Mastodon [1, 6–8]. These platforms provide valuable ecological validity but offer limited opportunities for controlled experimentation due to restricted API access, hidden ranking algorithms, privacy constraints, and gover- nance fragmentation [9]. Consequently, most studies conducted on such platforms focus on correlational analysis rather than causal inference [4].

Federated social networks, such as Mastodon, address issues of centralization and governance by distributing control across independent instances via the Activ- ityPub protocol [10, 11]. While this decentralization enhances transparency and community-level moderation, it complicates experimental control. Researchers cannot systematically manipulate ranking algorithms, enforce randomized exposure

con- ditions, or guarantee consistent treatment delivery across instances. As a result, Mastodon-based research typically relies on observational methods or externally implemented bot interventions [9].

To overcome these limitations, prior work has explored custom-built experimental platforms and simulated social environments [2, 3]. Controlled online experiments have been conducted using web applications to study misinformation diffusion, algorithm ranking effects, and behavior [6, 8]. However, many such systems lack modularity, scal- ability, or integration with modern AI tools, limiting their ability to support complex multi-round interventions or hybrid human-agent ecosystems.

Recent advances in large language models (LLMs) have introduced new possibilities for simulating social interaction [5, 12, 13]. Studies have demonstrated that LLM-based agents can generate realistic conversational behavior, enabling synthetic experimen- tation in social contexts [14]. Nevertheless, most implementations remain external overlays on existing platforms rather than integrated experimental infrastructure.

TWON advances this line of research by providing a modular MERN-stack archi- tecture explicitly designed for experimental control in online social networks [9]. Unlike observational platforms such as Mastodon, TWON enables full manipulation of feed ranking, randomized treatment assignment, and programmable interventions. Its integrated intelligence module supports LLM-based content generation, toxic- ity scoring, retrieval-augmented generation, and temporal scheduling. This enables researchers to conduct hybrid experiments that combine human participants and AI- driven agents in a reproducible, scalable environment. By combining experimental orchestration, scalable backend infrastructure, and integrated AI capabilities, TWON bridges the gap between ecological realism and methodological control. It provides a research-oriented alternative to both commercial social media platforms and fed- erated networks, enabling rigorous investigation of digital social interventions under controlled conditions.

3 Platform components

3 Platform components

The TWON experimental social media system is a modular, scalable platform designed to support controlled experiments on online social networks. It is struc- tured around several core components, each serving a distinct purpose in facilitating research, experimental control, and data collection.

1. Social media core

The social media core serves as the foundation of the TWON platform, emulating the behavior of a real social network. It provides user management functionality, includ- ing registration, authentication, and role-based access control. The component also handles content management, allowing the creation, storage, and retrieval of posts, comments, and media. Feed generation and ranking ensure that participants receive content according to experiment-specific algorithms, while interaction tracking logs user actions such as likes, dislikes, and comments. Finally, session handling main- tains participant state across interactions, enabling accurate temporal analysis and consistent experimental conditions.

2. Experiment manager

The experiment manager orchestrates all aspects of experimental design and execution. It handles randomization and group assignments to ensure statistically valid treatment and control conditions. Researchers can define experimental conditions for each study, specifying how participants interact with content and interventions. The component enforces treatments such as exposure to specific posts, algorithmic manipulations, or behavioral prompts, enabling reproducible, controlled investigations of online social behavior.

3. Intervention engine

The intervention engine provides a programmable interface for applying algorithmic and AI-based interventions within the platform. It integrates closely with the exper- iment manager to deliver randomized interventions to participants according to the study design. Typical interventions include feed manipulations, prompts, and content moderation. This engine allows researchers to systematically test the effects of different interventions on user behavior and engagement within the simulated social network environment.

4. Data & analytics

The data & analytics component is responsible for capturing and storing all research- relevant data. Event-level logging records participant actions with precise timestamps, enabling fine-grained behavioral analysis. Data is stored in both time-series storage and persistent databases, with experiment metadata maintained for reproducibility. TWON leverages an in-memory database for high-speed operations and MongoDB for scalable, document-oriented storage of experiments, content, and user interactions. This architecture ensures both performance and reliability for large-scale experiments.

5. Intelligence & AI module

The intelligence & AI module provides advanced computational capabilities for con- tent generation, analysis, and simulation. It includes LLM-based generation for the creation of synthetic posts, comments, and prompts. The module performs toxic- ity scoring and sentiment analysis to monitor and moderate participant content in real-time. It also supports retrieval-augmented generation (RAG) to provide contex-tually relevant information to users or agents. For simulations, the module includes machine-learning schedulers to orchestrate temporal dynamics and model agent behav- ior, enabling hybrid experiments that combine human participants with AI-driven agents.

4 Platform architecture

The TWON social media application follows a classic MERN stack architecture, which provides a robust and scalable foundation for modern web applications. This archi- tecture is designed around four main layers that work together seamlessly to deliver a comprehensive social media experience.

4.1 Frontend layer - user experience and interface

The frontend layer is built using React.js with Redux for state management, creating a dynamic and responsive user interface to handle all user interactions. The applica- tion includes several key components: user authentication modules for registration and login, social interaction features for creating and managing posts and comments, and a sophisticated UI system that supports both light and dark themes. The frontend also incorporates an admin panel that provides administrative users with compre- hensive control over the platform, including the ability to manage posts, users, and comments. Additionally, an explore page allows users to discover content from other users through a paginated interface, while profile pages display user information, posts, and social connections such as followers and following lists. The Redux store central- izes state management, ensuring consistent data flow throughout the application and maintaining user preferences and session information.

4.2 Communication layer - API and data transfer

Between the frontend and backend lies the HTTP communication layer, which facil- itates all data exchange through RESTful API calls. This layer implements JWT (JSON Web Token) authentication to ensure secure communication between the client and server. The system supports real-time features through push notifications, which require public and private keys configured in the environment variables. All communication is designed to be stateless and follows REST principles, making the API predictable and easy to maintain. The communication layer also handles error responses, data validation, and ensures that sensitive information is properly encrypted during transmission.

4.3 Backend layer - business logic and processing

The backend layer, built with Node.js and Express.js, serves as the application's core processing engine. This layer contains multiple specialized services that handle differ- ent aspects of the application's functionality. The authentication service manages user registration, login, and JWT token generation, while implementing secure password hashing with salt encryption to protect user credentials. User management services handle CRUD operations for user profiles, following and unfollowing functionality, and profile updates. The post management system allows users to create, edit, and delete posts while tracking views and reads, and managing likes and dislikes. A com- prehensive comment system supports nested comments with their own like and dislike functionality. The admin services provide dashboard analytics showing total counts of posts, users, and comments, along with content moderation capabilities. Various middleware components handle cross-origin requests (CORS), rate limiting to prevent abuse, and comprehensive error handling to ensure application stability.

4.4 Database layer - data persistence and management

MongoDB serves as the application's primary database, organizing data into sev- eral specialized collections. The Users collection stores comprehensive user profiles,

authentication credentials, and social connection data including followers and follow- ing relationships. The Posts collection maintains all user-generated content, associated metadata such as creation dates and author information, engagement metrics includ- ing likes and dislikes, and tracking data for views and reads.

The Comments collection stores comment content with references to their parent posts, engagement metrics, and supports the hierarchical structure required for nested comments. An Analyt- ics collection captures user activity patterns, performance metrics, and data required for the admin dashboard's statistical displays. Finally, a Sessions collection man- ages JWT tokens, active login sessions, and user preferences such as theme selection. This database structure supports efficient querying and maintains referential integrity across related data.

4.5 Security and deployment considerations

The application implements robust security measures throughout all layers. JWT authentication provides stateless session management, while password encryption with salt ensures that user credentials remain secure even if the database is compromised. The system is designed to communicate over HTTPS to prevent man-in-the-middle attacks and data interception. For deployment, the frontend is built and served from port 3000, while the backend runs on a separate Node.js server instance. The MongoDB database can be hosted locally during development or deployed to cloud services for production use. The application also integrates with external services for enhanced functionality, including push notification services for real-time user engage- ment and cloud storage services for handling media uploads and static assets. This comprehensive architecture ensures that the TWON social media application can scale effectively while maintaining security and performance standards expected of modern web applications.

5 Case studies and empirical deployments

5.1 TWON platform deployment for Serbia–Ukraine disinformation study

The TWON platform was deployed for a Serbian-language data generation study examining engagement with news and information related to the war in Ukraine [1]. This study was conducted in collaboration with the Slovenian Press Agency (STA) and fielded on May 28, 2025. The central objective was to analyze (1) which users click on disinformation articles, and (2) how users respond behaviorally when exposed to varying degrees of disinformation.

The experimental stimuli consisted of manipulated Serbian-language news articles concerning the war in Ukraine, alongside general Serbian-language social media posts covering sports and entertainment. Ukraine-related news articles contained varying levels of disinformation, enabling systematic observation of engagement patterns across content types.

[1]GitHub link: https://github.com/abdulsittar/TWON STA Case Study.git - STA Case Study

Table 1
Overview of TWON deployments across empirical case studies and simulation experiments.

Study	Partner / Context	N	Platform Features	Research Focus
STA Disinformation Study	Slovenian Press Agency (STA), Serbia	272	Multi-round feed refresh, adaptive ranking,disinformation scoring	User engagement with Ukraine-related disinformation
RKI Scientific Communication Study	Robert Koch Institute (Germany)	1200	Controlled feed exposure,pre-bunking interventions, sentiment analysis	Effects of pre-bunking and uncertainty communication
Toxicity Prevention Study	Online discussion platform	574	LLM-based comment rewriting, toxicity scoring, AI moderation	Reducing toxic language in online discussions
Large-scale Agent Simulations	Synthetic agent ecosystem	54	Ranking algorithms, agent-based interaction, automated content generation	Simulation of information diffusion and conversational dynamics

Participants were recruited through Latenta, a Serbian recruitment provider, with participation restricted to desktop users. In total, 272 participants successfully completed the study.

5.1.1 Multi-round feed design and adaptive ranking

This study implemented a multi-round experimental design to simulate users refreshing a social media feed over time. A refresh button was integrated at the top of the TWON interface, allowing participants to load new content across sequential rounds. This enabled observation of dynamic engagement behavior conditioned on prior interactions.

Upon entry, participants were randomized into one of three groups: a control group, a reinforcing group, and an opposing group. In Round 0, all participants viewed the same randomized feed. After refreshing:

- The control group received another randomized feed.
- The reinforcing group received a feed in which content aligned with their prior engagement in Round 0 (e.g., Ukraine-related or disinformation content) was up- ranked.
- The opposing group received a feed in which previously engaged content types were

down-ranked.

This adaptive ranking logic was repeated for two additional rounds, resulting in three total rounds of exposure and 19 possible feed pathways across conditions.

5.1.2 Content scoring and engagement metrics

Each post in the feed was assigned underlying scores on two dimensions: (1) Ukraine relevance and (2) disinformation presence. Likes contributed positively to a post's engagement score, dislikes were inversely weighted (multiplied by −1), and clicking on

a news article counted as positive engagement. For Ukraine-related news content, two out of three articles per round contained some level of disinformation, establishing a known baseline expectation for disinformation engagement.

Across the full study, the platform recorded 3,663 likes, 2,055 news article clicks, 1,860 dislikes, and 464 comments. The primary behaviors of interest were likes/dislikes and news article clicks, as participants were explicitly instructed to engage with posts and view articles.

5.1.3 Descriptive engagement patterns

Descriptive results indicate that non-Ukraine content consistently generated more news article clicks than Ukraine-related content across all rounds, despite equal baseline representation (three Ukraine and three non-Ukraine news stories per round). With respect to disinformation engagement, the initial round (Round 0) showed a higher-than-baseline proportion of disinformation link clicks among Ukraine articles. However, in subsequent rounds, disinformation click rates declined and aligned more closely with baseline expectations. This pattern suggests that disinformation may generate initial curiosity or novelty-driven engagement, but that this effect attenuates

with repeated exposure.

A similar pattern emerged for likes. Non-Ukraine content consistently received more likes than Ukraine-related posts. In Round 0, disinformation posts received a disproportionately high share of likes relative to baseline expectations, but this imbal- ance diminished in later rounds. Overall, disinformation was not consistently more engaging than non-disinformation content over time.

5.1.4 Engagement heterogeneity by gender

As in the German case study, engagement differences were examined by gender. Women comprised approximately 32% of the final sample.

For news article clicks, gender distributions were relatively balanced across Ukraine and non-Ukraine content. Differences in disinformation clicking behavior appeared across rounds, but these differences alternated in direction and did not display a stable pattern.

In terms of likes, engagement with Ukraine and non-Ukraine posts was also broadly similar across genders. However, in Round 0, women were somewhat more likely to like disinformation posts compared to men. This difference did not persist in subse- quent rounds, indicating that any initial gender-based divergence in disinformation engagement attenuated over time.

5.1.5 Implications for simulation modeling

Collectively, the descriptive findings suggest three main patterns: (1) Ukraine-related content was generally less engaging than non-Ukraine content, (2) disinformation con- tent may initially attract elevated engagement but does not sustain this advantage across repeated exposure, and (3) gender-based engagement differences are limited and unstable across rounds.

The multi-round adaptive design and resulting behavioral dataset provide a structured foundation for simulation modeling of content ranking dynamics, user adaptation, and disinformation exposure effects. These data support the development of more realistic agent-based or language model–driven simulations of online news engagement and algorithmic amplification mechanisms.

5.2 TWON platform deployment for scientific communication study

The TWON platform was deployed for a German experimental case study conducted in collaboration with the Robert Koch Institute between March 24 and March 28, 2025

[2]. The study focused on two key facets of scientific communication: (1) pre-bunking of misinformation, and (2) communication of statistical uncertainty. The experimental stimuli were based on authentic social media posts previously used by the Robert Koch Institute. Pre-bunking posts addressed mRNA vaccines, while uncertainty-related posts focused on the Monkeypox vaccine.

Participants (N = 1,200 completers; 1,295 initial entrants) were recruited via Pro- lific with pre-screening criteria including German residency, German language fluency, desktop participation, and a prior approval rate of at least 90%. Respondents were randomly assigned to one of five groups: a control group, a pre-bunking treatment group, a pre-bunking plus “truth sandwich” treatment group, an uncertainty control group, and an uncertainty treatment group.

5.2.1 Pre-survey and user onboarding

Before entering the TWON platform, participants completed a pre-survey collecting demographic and attitudinal variables, including self-rated ideology (1–11 left-right scale), age, gender, education level, and political interest (1–5 scale). To enhance ecological validity and platform realism, respondents selected one of four predefined avatar and username combinations prior to interacting with the simulated social media environment.

5.2.2 Experimental feed design and interaction requirements

After completing the pre-survey, users were directed to a treatment-specific version of the TWON main feed. Each participant viewed eight posts presented in randomized order. Seven posts were identical across conditions and derived from real institutional accounts (e.g., news outlets and brands), covering politics, sports, and entertainment. The eighth post contained the treatment manipulation corresponding to the

partici- pant's assigned condition. All experimental posts were visually standardized, differing only in minor textual and image manipulations to isolate treatment effects.

To encourage authentic engagement, users were prompted to scroll through all posts and interact as they would on a typical social media platform. To ensure active participation, respondents were required to like at least two posts and comment on at least three posts before proceeding.

[2]GitHub link: https://github.com/abdulsittar/TWON RKI Case Study.git - RKI Case Study

5.2.3 Post-survey measures and attitudinal outcomes

Following platform interaction, participants completed a post-exposure survey measur- ing key outcome variables related to: (1) perceived credibility and interpretation of the content viewed, (2) attitudes toward scientific institutions, and (3) attitudes toward vaccines. These measures were designed to assess potential attitudinal differences across treatment conditions.

5.2.4 Engagement metrics and sentiment analysis

User engagement data were recorded at the post level, including number of comments, likes, dislikes, and average comment sentiment. Comment sentiment was classified using a zero-shot prompting approach with Phi-3.5 on a three-point scale (0 = nega- tive, 1 = neutral, 2 = positive). Sentiment scores were aggregated to generate average sentiment values per post.

Descriptive results indicate that healthcare-related posts (including treatment and control vaccine posts) consistently received among the most positive sentiment scores. The uncertainty treatment post showed slightly lower sentiment compared to its con- trol counterpart. Outside healthcare content, sports and video game posts generated relatively positive engagement, while political posts—particularly those related to U.S. politics—elicited the most negative sentiment. Entertainment content (e.g., Netflix programming) also received comparatively low sentiment and high dislike rates.

Healthcare posts generally received fewer total likes and dislikes compared to enter- tainment or political posts. The video game post received the highest number of likes overall, while the Netflix and U.S. politics posts received the highest number of dislikes.

5.2.5 Engagement heterogeneity by gender

Gender-based subgroup analyses were conducted to explore engagement heterogeneity. Women constituted approximately 40% of the sample. Distinct engagement patterns emerged: female participants showed higher liking rates for entertainment content such as Too Hot to Handle, lower engagement with football and general political posts, greater engagement (likes and comments) with healthcare-related content, and reduced interaction with video game posts.

Dislike patterns largely mirrored like distributions. Women showed fewer dislikes for entertainment content, similar or higher dislike rates for political posts, and fewer dislikes for healthcare posts. Sentiment analysis revealed that women expressed signif- icantly lower comment sentiment for certain entertainment content

despite high like rates, and more negative sentiment on posts related to political protests. For most other posts, gender-based sentiment differences were modest.

Overall, the dataset generated through this deployment provides rich behavioral interaction data suitable for simulation modeling and language model training. It enables analysis of engagement dynamics across content types, treatment conditions, and user subgroups within a controlled yet ecologically valid social media environment.

5.3 TWON platform deployment for toxicity prevention

The TWON platform was deployed for a toxicity prevention study[3], which required extensive customization to support AI-assisted moderation. The platform integrated real-time LLM-generated paraphrasing using GPT−5. When participants submitted a comment, the system automatically generated a rewrite that preserved the original content and stance while reducing toxic language. Users could choose to accept the AI suggestion, manually edit it, or discard it entirely. All interactions and decisions were logged, including the original comment, the AI suggestion, the final posted comment, and the user's choice regarding the intervention.

5.3.1 Discussion thread management

To simulate realistic social media conditions, discussion threads were initialized with pre-set inflammatory comments to create a toxic environment. Each thread contained at most 5 users, which in most cases resulted in 5 user-made comments, with some users posting more than a single comment. After the 5-user limit was reached, a new thread was created with identical seed posts. This was done to ensure a continuous discussion while maintaining experimental control. The platform supported asynchronous multi-user interactions, allowing participants to self-select discussion topics and join threads at any point, mirroring real-world OSN dynamics.

5.3.2 Data logging and metrics

For data collection, TWON recorded both system-level metrics (timestamps, thread depth, comment sequences) and intervention-specific data (which comments received AI suggestions and whether they were adopted). Additionally, toxicity scores were computed using the Perspective API across six dimensions: toxicity, severe toxicity, insult, profanity, threat, and identity attack. These scores were stored in a structured format to facilitate downstream statistical analyses and evaluation of both the AI intervention and platform performance.

5.3.3 Technical features and architecture

The technical deployment highlights several important features of TWON: mod- ular design allowing case-specific customization, integration with external services (e.g., Prolific for participant recruitment), real-time LLM intervention pipelines, and detailed logging for reproducibility and analysis. The platform's architecture demon- strates how AI can be embedded into social network simulations to enable experimental interventions at scale, while maintaining flexibility to accommodate different research objectives or use cases.

[3]GitHub link: https://github.com/abdulsittar/toxicity-prevention-study.git - Toxicity Prevention Case Study

5.4 TWON platform deployment for multi-agent simulation study

The TWON platform was extended to support a multi-agent simulation study focused on modeling social media interaction dynamics using large language model (LLM)- driven agents. The objective of this deployment was to analyze how simulated agents generate posts and comments, how interaction patterns evolve over time, and how different ranking strategies influence network structure and information diffusion.

The simulation framework integrates persona-based agents, temporal interaction modeling, retrieval-based context grounding, and configurable ranking mechanisms [15]. These components enable agents to interact within a dynamic social network by generating posts, replying to other agents, and adapting to evolving conver- sational contexts. The system supports controlled experimentation on algorithmic exposure, interaction dynamics, and structural properties of simulated social media environments.

5.4.1 Simulation design and agent interaction

The simulation environment consists of multiple agents initialized using real-world data distributions derived from social media platforms[16]. Agents are assigned pro- files based on historical behavioral and linguistic patterns, allowing them to generate contextually consistent posts and comments.

Agent interactions are governed by a scheduler that predicts the sequence of events, including which agent acts, what action is performed (post or comment), and the target of the interaction. This enables the simulation of realistic temporal dynamics such as bursty activity, response patterns, and interaction sequences. Retrieved contextual information from prior interactions is used to ground content generation, ensuring continuity across conversations.

Simulations were conducted across multiple domains, including technology, cli- mate change, and COVID-19 discussions, allowing evaluation of the framework under different topical settings.

5.4.2 Ranking mechanisms and exposure conditions

To analyze the effects of algorithmic curation, the simulation incorporates multiple ranking strategies that determine content visibility:

- Chronological ranking, where content is displayed based on time order.
- Engagement-based ranking, where highly interacted content is prioritized.
- Hybrid ranking, combining temporal and engagement signals.

These ranking strategies were applied after initial baseline interactions, enabling the study of how exposure mechanisms influence cascade formation, interaction con- centration, and clustering within the network. Both global and local ranking settings were explored to simulate different levels of content personalization.

Table 2
Effects of ranking strategy and simulation scale on conversational structure and interaction patterns across domains. Global and local chronological, engagement, and hybrid rankers are compared against baseline simulations using qualitative and quantitative evaluation metrics.

Domain	Ranker	Interaction type distribution (synthetic vs real)	Agent population size	Influence concentration	Simulation time (minutes)
COVID	Chronological (G)	12/488	4/7	CovidTherapies/GeneralChat	379
COVID	Chronological (L)	13/487	4/7	CovidTherapies/GeneralChat	147
COVID	Engagement (G)	12/488	4/7	CovidTherapies/GeneralChat	296
COVID	Engagement (L)	7/493	4/7	CovidTherapies/GeneralChat	190
COVID	Hybrid (G)	8/492	4/7	CovidTherapies/GeneralChat	326
COVID	Hybrid (L)	7/493	4/7	CovidTherapies/GeneralChat	171
Climate	Chronological (G)	1/499	11/9	ClimateChange/GeneralChat	275
Climate	Chronological (L)	2/498	10/9	ClimateChange/GeneralChat	549
Climate	Engagement (G)	1/499	11/9	ClimateChange/GeneralChat	285
Climate	Engagement (L)	3/497	9/9	ClimateChange/GeneralChat	191
Climate	Hybrid (G)	0/500	9/9	ClimateChange/GeneralChat	544
Climate	Hybrid (L)	1/499	12/9	ClimateChange/GeneralChat	198
Technology	Chronological (G)	111/389	32/18	StreamingMedia/GeneralChat	274
Technology	Chronological (L)	215/784	32/18	StreamingMedia/GeneralChat	379
Technology	Engagement (G)	109/391	28/18	StreamingMedia/GeneralChat	281
Technology	Engagement (L)	81/419	32/18	StreamingMedia/GeneralChat	189
Technology	Hybrid (G)	86/414	28/18	StreamingMedia/GeneralChat	134

5.4.3 Implications for simulation-based experimentation

The multi-agent simulation deployment demonstrates that TWON can support controlled experimentation on social media dynamics without requiring real user pop- ulations. The platform enables systematic analysis of temporal interaction patterns, ranking effects, and structural properties of online discourse.

The generated datasets provide a foundation for studying information diffusion, interaction dynamics, and algorithmic exposure in a controlled setting. Additionally, the framework allows exploration of hypothetical scenarios, such as alternative ranking strategies or changes in interaction behavior, supporting simulation-driven analysis of online social systems.

6 Discussion

This work presents TWON, a scalable and modular platform designed to support experimental research on online social network dynamics. The system was developed to address a major limitation faced by computational social scientists and com- munication researchers: the difficulty of conducting controlled experiments within existing commercial social media platforms. By providing a fully configurable social media environment, TWON enables researchers to systematically study user behavior, information exposure, and interaction dynamics under controlled conditions.

One of the key strengths of the platform is its flexibility. The modular architecture allows researchers to configure different ranking algorithms, interaction mechanisms, and experimental interventions. This flexibility enabled the platform to support mul- tiple case studies across diverse research domains, including misinformation analysis, scientific communication, and toxicity mitigation in online discussions. These deploy- ments demonstrate the adaptability of the platform for interdisciplinary research collaborations.

Another important contribution of the platform is its support for both human- centered experimental deployments and synthetic agent-based simulations. The inte- gration of automated agents and configurable ranking strategies allows researchers to explore how algorithmic mediation influences conversational structure, influence con- centration, and information diffusion patterns. Such simulations provide a valuable complement to empirical studies by enabling large-scale experimentation that would otherwise be difficult to conduct with human participants.

Despite these advantages, several limitations remain. First, while experimen- tal platforms such as TWON allow greater control over the research environment, they cannot fully replicate the scale, complexity, and social context of large com- mercial social media platforms. User behavior in experimental settings may differ from real-world interactions due to differences in network size, incentives, and plat- form familiarity. Second, maintaining realistic interaction dynamics requires careful experimental design and calibration of ranking algorithms and agent behaviors.

Future work will focus on expanding the platform’s capabilities in several direc- tions. These include integrating more advanced recommendation and moderation algorithms, improving support for large-scale simulations, and enabling richer forms of multimodal interaction. Additionally, further work could explore the integration of large language models and automated moderation systems to study emerging challenges in online communication ecosystems.

Overall, TWON provides a valuable experimental infrastructure that bridges the gap between simulation-based studies and real-world social media environments. By enabling controlled, reproducible, and scalable experiments, the platform contributes to advancing methodological approaches in computational social science and the study of online information ecosystems.

7 Conclusions and future work

This paper introduced TWON, a modular social media experimentation platform designed to support controlled studies of online behavior, information exposure, and algorithmic interventions. Built on a MERN-stack architecture, the platform integrates user interaction modules, experimental intervention engines, detailed data logging, and AI-driven analytical capabilities within a unified research environment. TWON enables researchers to simulate realistic social media environments while maintaining full experimental control over feed ranking, content exposure, and user interaction dynamics.

The platform demonstrates strong flexibility and scalability through multiple empirical deployments across different research contexts. These deployments include studies on disinformation engagement, scientific communication interventions, toxic- ity reduction through AI-assisted moderation, and large-scale simulations involving autonomous agents. Across these studies, TWON successfully captured fine-grained behavioral data, enabling the analysis of user engagement patterns, content interac- tion dynamics, and the effects of different experimental manipulations. The results illustrate that controlled social media simulations can generate valuable behav- ioral datasets while maintaining ecological realism comparable to real online social networks.

Beyond human-participant experiments, the platform also supports hybrid exper- imental designs that integrate AI-based agents and language models. This capability allows researchers to simulate large-scale information ecosystems, explore algorithmic amplification dynamics, and test intervention strategies in environments that combine both synthetic and human actors.

Future work will focus on several directions to further enhance the platforms capa- bilities. First, the integration of more advanced recommendation algorithms, including reinforcement learning and adaptive ranking systems, would allow the simulation of more complex algorithmic feed dynamics. Second, expanding human-agent interaction frameworks will enable richer hybrid experiments where LLM-driven agent interact directly with human participants. Finally, additional work will focus on improving scal- ability, real-time analytics, and reproducibility tools, facilitating large-scale behavioral experiments and longitudinal studies of online information ecosystems.

Declarations

Funding:

This work is supported by TWON (project number 101095095), a research project funded by the European Union under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, Topic 07).

• Conflict of interest/Competing interests: The authors declare that they have no

known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

• Ethics approval and consent to participate: This studies was reviewed and approved

by the appropriate ethics review board at the University of Amsterdam. All research procedures were conducted in accordance with established institutional and ethical guidelines.

• Consent for publication: All authors have reviewed and approved the final version

of the manuscript and consent to its publication.

• Data availability: All data generated or analyzed during this study are included in the manuscript.

• Materials availability: All materials necessary to reproduce the study are described

in the manuscript.

Author contribution:

Conceptualization: Abdul Sittar, Mateja Smiljani´c, Michael Heseltine, Fran¸cois t'Serstevens, Natan Vite˘znik, Corinna Oschatz. **Software and implementation**: Abdul Sittar, Mateja Smiljani´c. **Case study design and experimental configu- ration**: Michael Heseltine, Fran¸cois t'Serstevens, Natan Vite˘znik, Corinna Oschatz. **Review and validation**: Alenka Gu˘cek. **Funding acquisition and supervision**: Damian Trilling, Marko Grobelnik. **Writing – original draft and revisions**: All authors.

References

1. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600
2. Guess AM, Malhotra N, Pan J, Barber´a P, Allcott H, Brown T, Crespo- Tenorio A, Dimmery D, Freelon D, Gentzkow M et al (2023) How do social media feed algorithms affect attitudes and behavior in an election. campaign? Sci 381(6656):398–404
3. Tigas P, Annadani Y, Jesson A, Sch¨olkopf B, Gal Y, Bauer S (2022) Interven- tions, where and how? experimental design for causal models at scale. Adv Neural Inf Process Syst 35:24130–24143
4. Zheleva E, Arbour D (2021) Causal inference from network data. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 4096–4097
5. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Gen- erative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology, pp. 1–22

6. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. science 359(6380), 1146–1151
7. Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on twitter during the 2016 us presidential election. Science 363(6425):374–378
8. Brum P, Cˆandido Teixeira M, Vimieiro R, Arau´jo E, Meira W Jr, Lobo Pappa G (2022) Political polarization on twitter during the covid-19 pandemic: a case study in brazil. Social Netw Anal Min 12(1):140
9. Zhang OX (2025) Enhancing moderation in decentralized social media: A comprehen- sive approach to community-level blocklist management and customization on mastodon. Master's thesis, Princeton University
10. Bono CA, La Cava L, Luceri L, Pierri F (2024) An exploration of decentral- ized moderation on mastodon. In: Proceedings of the 16th ACM Web Science Conference, pp. 53–58
11. La Cava L, Mandaglio D, Tagarelli A (2024) Polarization in decentralized online social networks. In: Proceedings of the 16th ACM Web Science Conference, pp. 48–52
12. Biswas A, Talukdar W (2025) Building Agentic AI Systems: Create Intelligent, Autonomous AI Agents that Can Reason, Plan, and Adapt. Packt Publishing Ltd, ???
13. Guo T, Chen X, Wang Y, Chang R, Pei S, Chawla NV, Wiest O, Zhang X (2024) Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680
14. Takata R, Masumori A, Ikegami T (2024) Spontaneous emergence of agent indi- viduality through social interactions in large language model-based communities. Entropy 26(12):1092
15. Sittar A, Gu˘cek A, Grobelnik M, Rettinger A, Smiljani´c M (2026) Simulating multi-agent social media ecosystem with llms: Modeling, evaluation, and struc- tural effects of ranking mechanisms. Evaluation, and Structural Effects of Ranking Mechanisms03, (February (2026)
16. Sittar A, Cesnovar M, Gucek A, Grobelnik M (2026) Constructing a dataset to support agent-based modeling of online interactions: Users, topics, and interaction networks. arXiv preprint arXiv :260112628

Figures

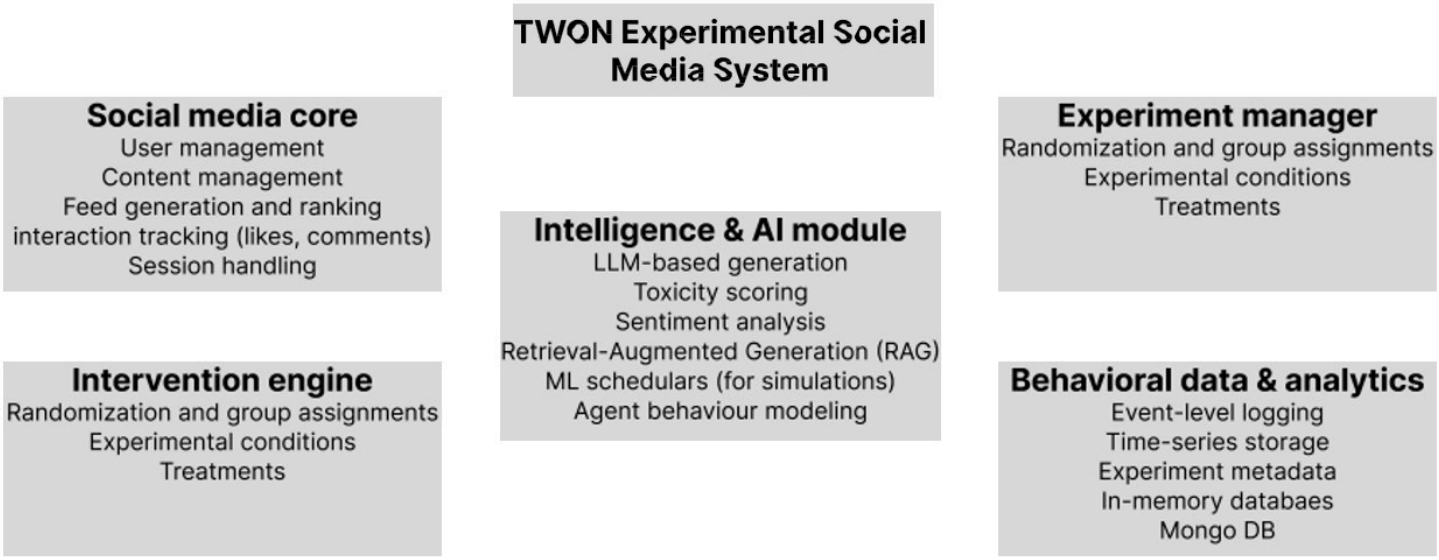


Figure 1

TWON social media: A component-based experimental social mediasystem.

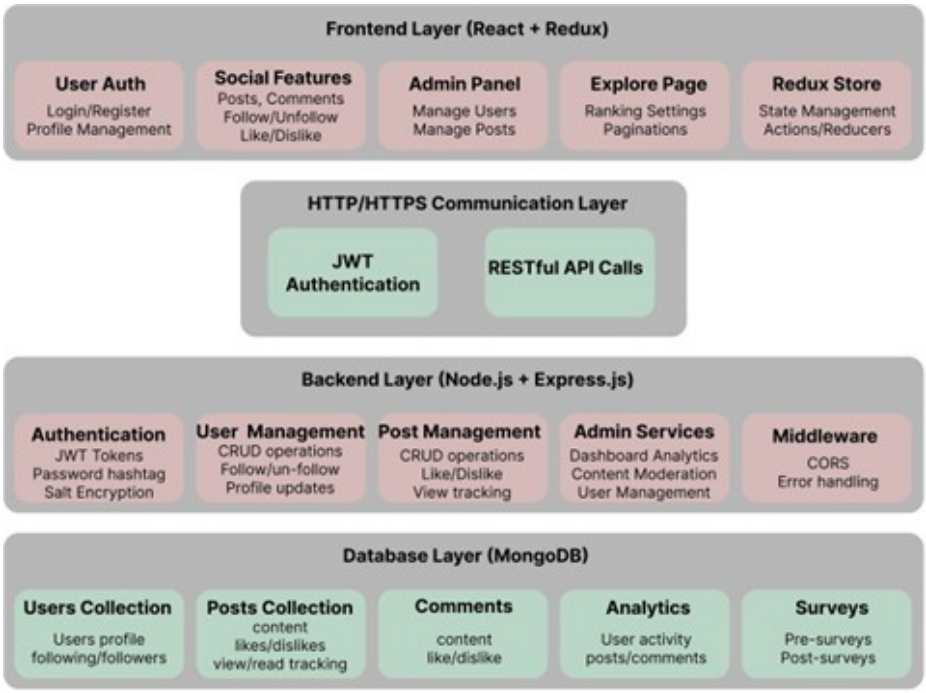


Figure 2

How TWON Social Media Works: A visual breakdown of the application's architecture from user interface to database storage.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx