Agent-Based Simulations of Online Political Discussions: A Case Study on Elections in Germany*

Abdul Sittar[1,1,*,†], Alenka Guček[2,†] and Marko Grobelnik[3,†]

[1] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*

Abstract

User engagement on social media platforms is influenced by historical context, time constraints, and reward-driven interactions. This study presents an agent-based simulation approach that models user interactions, considering past conversation history, motivation, and resource constraints. Utilizing German Twitter data on political discourse, we fine-tune AI models to generate posts and replies, incorporating sentiment analysis, irony detection, and offensiveness classification. The simulation employs a myopic best-response model to govern agent behavior, accounting for decision-making based on expected rewards. Our results highlight the impact of historical context on AI-generated responses and demonstrate how engagement evolves under varying constraints.

Keywords

Conversational agents, Conversation history, Reward-driven mechanism, Sentiment analysis

1. Introduction

User engagement on social media platforms is influenced by a complex interplay of factors, including historical context, time constraints, and reward-driven interactions. Understanding how these elements shape online discourse is crucial for developing AI-driven models that can generate meaningful and coherent responses [1],[2]. However, existing research lacks a comprehensive framework to analyze how past conversation history, motivation, and resource constraints impact engagement dynamics and content generation [3],[4].

This study presents an agent-based simulation approach to model user interactions on social networks, focusing on political discourse on German Twitter data. The simulation is designed to reflect real-world engagement patterns by incorporating historical user interactions, motivation levels, and time budgets. We fine-tune AI models to generate posts and replies, using sentiment analysis, irony detection, and offensiveness classification to assess content quality. To model decision-making, we employ a myopic best-response model, where agents interact based on expected rewards, replicating the motivations of real social media users [5], [6], [7].

The research follows a structured implementation approach. We first collect German Twitter data containing posts from parliamentary delegates and replies from regular users. Two language models are then fine-tuned using a supervised fine-tuning (SFT) approach to generate contextually relevant tweets and replies. The third phase involves designing a simulation framework, including a database schema, user network structure, and ranking system. Finally, we conduct a series of experiments, toggling variables such as conversation history, time budget, motivation, and ranking mechanisms to analyze their impact on AI-driven engagement.

1.1. Problem Statement

There is limited understanding of how conversation history affects the quality and coherence of AI-generated responses. While context-aware models show improvements, their impact on sentiment,

Extended Semantic Web Conference, Jun 1–5, 2025, Portorož, SL

*Corresponding author: abdul.sittar@ijs.si

†These authors contributed equally.

abdul.sittar@ijs.si (A. Sittar); alenka.gucek@ijs.si (A. Guček); marko.grobelnik@ijs.si (M. Grobelnik)
0000-0003-0280-9594 (A. Sittar); 0000-0003-4453-1498 (A. Guček); 0000-0001-7373-5591 (M. Grobelnik)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

engagement, and user perception remains unclear. Additionally, social media engagement is reward-driven but constrained by users' limited time and motivation. Existing models often overlook these constraints and their effect on engagement over time. Evaluating AI responses also requires analyzing sentiment, irony, offensiveness, and relevance, yet current methods lack a comprehensive framework for tracking their evolution. Addressing these gaps is crucial for developing more effective and responsible AI-driven social media interactions.

1.2. Research Questions

- RQ1: How does the inclusion of previous conversation history impact AI-generated responses?
- RQ2: How does user engagement evolve under time and energy constraints in a reward-driven environment?

2. Related Work

The study of social influence dynamics in online networks has traditionally relied on models that assume uniform activity levels among users [8]. Classic models, such as Axelrod's cultural dissemination framework, posit that all users engage in communication and opinion updates with equal probability, leading to homogeneous interaction patterns [9]. For instance, at every time point a randomly picked agent is selected for update and can be influenced by a network neighbor. This homogeneity assumption is highly unrealistic for the context of online social networks, where a relatively small share of users are highly active while most users contribute little. However, empirical research suggests that social media engagement is highly skewed, with a minority of users driving most interactions while the majority remain relatively passive [10]. This discrepancy raises questions about whether existing models accurately capture real-world social media behavior.

Recent work has sought to address this limitation by incorporating heterogeneous user activity into agent-based simulations. For instance, studies leveraging success-driven user activity models suggest that individuals who receive positive reinforcement—such as likes, retweets, or other forms of approval—tend to increase their engagement levels over time [11]. This dynamic is rooted in social reinforcement learning, which theorizes that behavior reinforced by social feedback is more likely to be repeated. By extending Axelrod's model to account for such adaptive behaviors, researchers have explored whether these modifications influence key outcomes, such as the emergence of polarization in online discourse.

A central mechanism in these models is homophily, which describes the tendency of users to interact preferentially with those who share similar traits or opinions [12]. In Axelrod's framework, the probability of two agents interacting depends on their degree of cultural similarity. This principle has been widely applied in computational studies of social media, where ideological homophily has been shown to contribute to the formation of echo chambers and opinion clustering. However, the extent to which heterogeneous activity levels amplify or mitigate these effects remains an open research question.

Building on this foundation, our study integrates success-driven user activity and historical engagement patterns into an agent-based simulation of political discourse on German Twitter. By analyzing how these factors shape AI-generated interactions, we contribute to a more nuanced understanding of engagement dynamics in online social networks.

3. Implementation Approach

The research presented in this study is focused on the simulating user behaviors and engagement on social media. It relies on several assumptions about agent behaviors and interactions. Agents preferences are driven by rewards. Users are motivated to engage in activities, such as posting or interacting, when expected reward surpass the reward from not engaging. This reward can include social feedback or emotional satisfaction. The agents follow a myopic best-response approach, where they make decisions based on immediate or short-term rewards rather than long-term outcomes. Lastly, the model assumes that users face limitations in terms of time and energy, which are represented by a resource budget. In pursuit of this objective, we present implementations in four steps. In the initial phase, we collect German Twitter data focusing on political discourses. These datasets contain two distinct types of Twitter interactions such as posts and replies. Given these datasets, we define two separate tasks suitable for training a language model: post generation (creating tweets in line with political discourse) and reply generation (responding to posts with contextually relevant arguments).

The subsequent step involves training two model adapters on top of Llama-3.2-3B-Instruct using the supervised fine-tuning (SFT) paradigm. The models were fine-tuned using the transformers and PEFT Python packages and released on the Hugging Face Hub for accessibility. The training aimed to optimize two distinct tasks: 1) Given a list of topics and an ideology, generate a tweet similar to those written by parliamentary delegates, 2) Given a post and an ideology, generate a contextually relevant reply in the style observed in user interactions (see the detailed explanation in section 4).

In the third phase, we design the simulation mechanism where we define the database schema to store the conversation generated by agents, implement a network among the users in the database and ranking service to rank the content and finally we design the life cycle of agents for communication with each other. The simulations can be executed using alternative data structures like arrays or dictionaries instead of relying on a database design. Nevertheless, using a database allows for the possibility of integrating real-time interactions with actual users.

Subsequently, we run simulations in which we turn on and off multiple variables such as history, budget, motivation and ranking. The results of these simulations have been presented in Section 7. The source code of this approach is available at the GitHub Repository.

4. Data Preprocessing and Model Fine-Tuning

We collected two datasets containing German Twitter data focused on political discourse. The first dataset consists of posts made by delegates from the national parliament regarding political issues, primarily related to the energy transition and the rise of right-wing ideology. The second dataset includes replies from regular Twitter users in response to these posts. These two distinct types of data (posts and replies) allow for the development of two learnable tasks for a language model. Preprocessing these datasets involves content filtering to enhance the quality, with a focus on active users based on the number of posts or replies. Users contributing over 15 posts and 25 replies are selected, ensuring that both tasks can be modeled at a user-based level. Additionally, irrelevant external sources such as URLs are removed, and short content is excluded, aiming to retain only substantial arguments or opinions.

To further enrich the data, annotations are added to both posts and replies. These annotations include a classification of political leaning (left, neutral, right) and topic labels, which are derived using a prompt-based technique with the Llama3.1:70b-instruct-q6-K model. This step significantly influences the training of the language model, especially in content generation tasks. However, challenges emerge

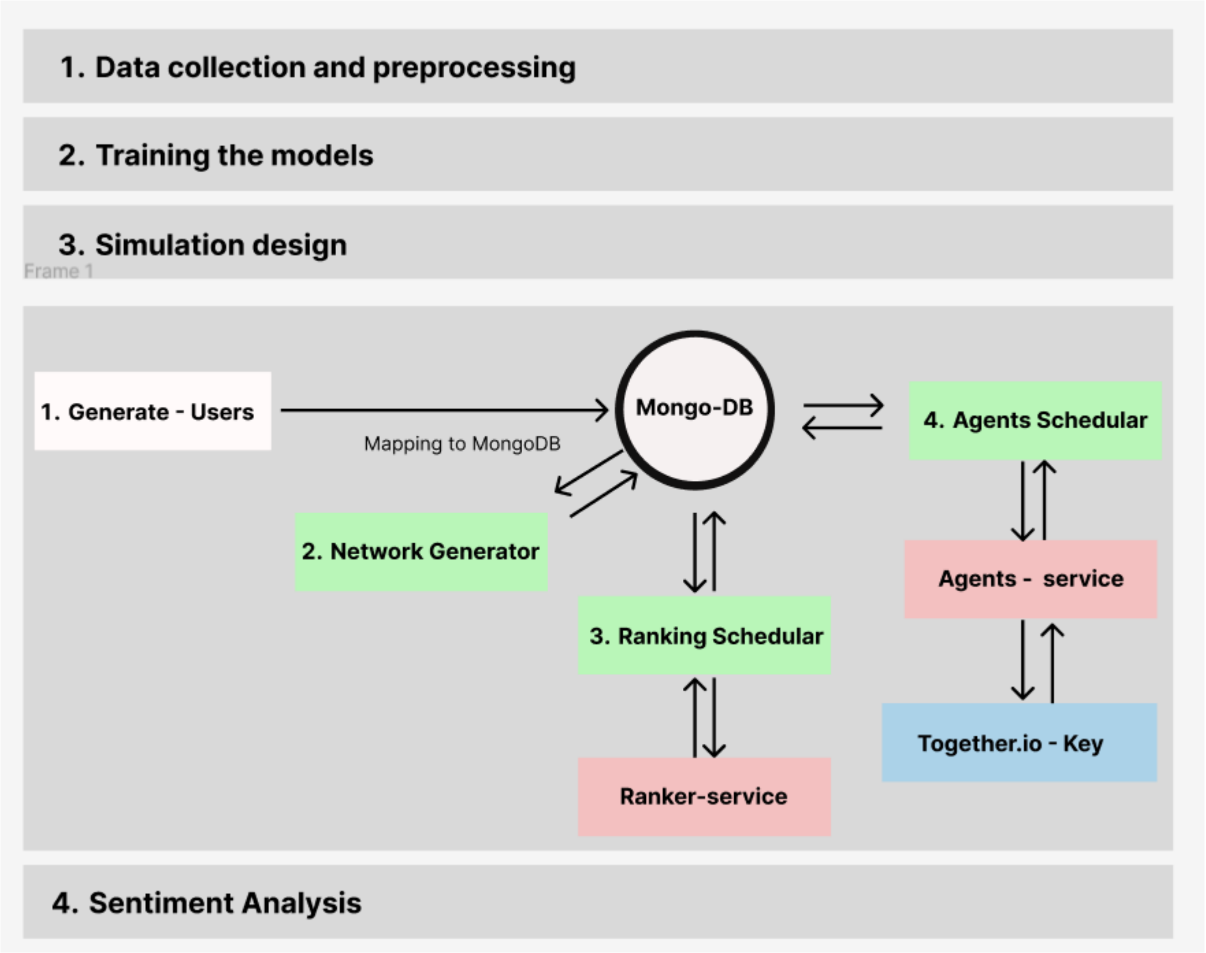

Figure 1: Implementation Approach: Covering data collection, model training, simulation design, and sentiment analysis of generated data.

due to the heterogeneous nature of topic naming, as the model uses both German and English terms and varying levels of detail in its descriptions. While refining the prompt-based approach can address some of these issues, human preprocessing is recommended for a more significant improvement in data quality. The results of this annotation step are critical, as the topics extracted serve as the context for the language model during training.

The models are trained using a supervised fine-tuning (SFT) approach, optimizing the Llama-3.2-3B-Instruct model for the two tasks: posting and replying. The SFT paradigm adapts the language model to these specific tasks by comparing the generated content with labeled data, which reflects the behavioral patterns of the most active users in the dataset. Despite this, there are limitations regarding the generalizability of these models across different social media contexts. The sampling bias in selecting the most active users raises concerns about whether these models can accurately reflect discourse across various user communities. The varying communication dynamics and argumentation styles present in different social media networks present both methodological and theoretical challenges, urging the need for more specialized models tailored to specific communities rather than universal frameworks.

5. Network Architecture and Modeling

The Myopic Best-Response Model, illustrated by a logistic function, helps explain user behavior in online networks. This model posits that users will engage in a particular activity, such as sharing content, when the expected reward from that activity surpasses the expected reward of not engaging in it. Central to this model are several assumptions about user behavior: preferences are driven by rewards, users receive feedback from other users (e.g., likes), and they apply a decision-making rule based on immediate rewards, considering constraints like limited time and energy. The model incorporates a "myopic best-response" approach, where users focus on the immediate expected rewards, and it assumes that users will select behaviors that maximize their short-term satisfaction.

To model user characteristics, the system introduces two types of resource budgets. The general resource budget, denoted as bgi,tbgi,t, represents the available energy and time for a user at a specific time. This budget decreases as users engage in activities, but it regenerates over time at a certain rate ss, creating variability in user resources. In addition to this general budget, a maximal round-resource budget bri,tbri,t determines how much energy and time a user can expend during a single activity session. These resource constraints ensure that users cannot engage in excessive activities within a limited timeframe. This two-tiered approach allows for personalized and dynamic representations of user behavior and resource allocation.

User activation within the model is governed by binary decisions of whether to log in or log off. The likelihood of logging in is modeled using a memoryless Poisson process, where the probability of a user logging in during a specific period depends on the expected reward of logging in versus staying offline. This probability is governed by a logistic function that accounts for various rewards associated with online activity, including time spent online, personal value (entertainment), social feedback, and notifications like the fear of missing out (FOMO). The system incorporates parameters that influence user activation based on both internal motivations and external social dynamics, providing a comprehensive picture of why users may choose to engage or disengage from the platform.

The model also addresses the expectation management of users by assuming they adjust their expectations based on past experiences. Users weigh previous online sessions using a weighted average that discounts older experiences, capturing the "shadow of the past." This approach considers time biases, where users might be overly optimistic about how much time they will spend online, based on their previous experiences. Additionally, the system incorporates a reward function that models both personal value and social feedback. Personal value is driven by the content consumed and the time spent engaging with it, while social feedback is shaped by the reactions of other users, reflecting both immediate and delayed responses. These factors combine to influence user behavior and decision-making within the system.

6. Experiment Setup

In the experimental setup to model user interactions on a social media platform, we implemented an agent-based simulation that generates posts, comments, likes and dislikes over multiple iterations. The simulation initializes a database with historical user interactions and sets key parameters, including 51 agents and 30 iterations. The agents are politically neutral, ensuring unbiased engagement.

Each iteration consists of two primary phases:

1. **Post Generation**
    - A subset (20%) of agents is selected to create posts.
    - Each agent's post is generated using the `"TWON-Agent-OSN-Post-en"` model, leveraging their previous interactions.
    - Posts are added to the database, and the agent's **time budget** is updated accordingly.
2. **Comment Generation**
    - A larger subset (80%) of agents engages by commenting on existing posts.
    - The system retrieves recent posts and assigns engagement probabilities using a classifier.
    - The agent-post pairs with the highest probability are selected for interaction.
    - Comments are generated via the `"TWON-Agent-OSN-Replies-en"` model, added to the database, and agent motivation is updated.

3. **Like/Dislike**
    - A subset (20%) of agents engages by liking on existing posts.
    - A subset (20%) of agents engages by disliking on existing posts.

6.1. Experimental Scenarios and Iterative Processes

Below are the various combinations of multiple variables used for running the simulation and generating synthetic data.

1. A history of real users engaging in discussions on the relevant topic using a classifier to select a user to reply to a specific post, without incorporating time budget and motivation features.
2. Without real user discussion history on the topic, use a classifier to determine a user's reply to a specific post, while considering time budget and motivation features.
3. Without real user discussion history on the topic, using a classifier to select a user to reply to a specific post, excluding time budget and motivation features.
4. A history of real users discussing the topic, using a classifier to choose a user to reply to a specific post, while incorporating time budget and motivation features.
5. A history of real users participating in discussions on the topic, integrating time budget and motivation features to assess user engagement in posting or replying.
6. Without a history of real users discussing the topic, incorporating time budget and motivation features to evaluate user participation in posting or replying.

6.2. Evaluation of Generated Social Media Conversations

To assess the nature and quality of the generated posts and comments, we conducted a multi-faceted analysis covering six key aspects: **topics, emotions, sentiment, irony, offensiveness, and hate speech**. Each post and comment was evaluated using pre-trained classification models, producing probabilistic scores for each category.

1. **Topic Classification:** Each sample was categorized into topics, such as *news & social concern.* The assigned probability indicates the relevance of a post/comment to the given category.
2. **Emotion Detection:** The system attempted to classify emotions in user-generated content (e.g., joy, anger, sadness). If no dominant emotion was detected, the field remained empty.
3. **Sentiment Analysis:** Posts and comments were labeled with sentiment scores for **neutral** and **positive** sentiment, helping to gauge the general tone of the conversation.
4. **Irony Detection:** A probability score was assigned to determine the likelihood that a given post or comment contained ironic elements.
5. **Offensive Language Detection:** Each sample was analyzed for offensive content, categorizing it as either *offensive* or *non-offensive*, with corresponding confidence scores.
6. **Hate Speech Detection:** The system evaluated whether the content contained hate speech, classifying it as either *HATE* or *NOT-HATE.*

A sample analysis output for a post is structured as follows:

> *Sample:* "The first bill to do this was the American Rescue Plan, passed in the Senate with support from all Democrats and 10 Republicans. Many of those Republicans who voted in support of the law were conservative Republicans who believed in the importance of fiscal stimulus."
>
> **Detected Features:**
>
> - **Topic:** *news & social concern* (0.97)
> - **Sentiment:** Neutral (0.74), Positive (0.84)

- **Irony:** Present (0.86)
- **Offensive:** Non-offensive (0.79)
- **Hate Speech:** NOT-HATE (0.97)

This classification enables us to systematically assess the content quality, tone, and engagement trends within the simulated social media conversations. By analyzing multiple iterations, we can observe how agent interactions evolve over time and how different factors influence engagement dynamics.

7. Results

The table 1 presents data from the agent-based simulation designed to model user interactions on a social media platform. It captures various experimental conditions, including whether historical interactions are considered, whether agents have constraints on budget and motivation, and the impact of different ranking mechanisms. The results are evaluated in terms of the number of posts, comments, likes, and dislikes generated under these conditions.

RQ1: One key observation is the effect of conversation history on engagement. When historical user interactions are included (rows 1-4), the number of posts and comments is generally higher. This suggests that users are more likely to contribute when they can build on past interactions. Conversely, when history is not considered (rows 5-8), engagement decreases, indicating that conversation history plays a crucial role in sustaining discussions (see Table 1).

RQ2 The influence of resource constraints, such as time and motivation, is also evident. In scenarios where both budget and motivation are restricted (rows 5-8), the number of posts and comments is lower compared to the history-only condition. However, likes and dislikes appear in these conditions, which were absent in earlier cases. This suggests that when agents face limitations, they may prefer passive engagement (liking/disliking) over active participation (posting/commenting).

Sentiment analysis is performed on the generated data using the set of political figures' social media interactions, including posts and comments. The sentiments are categorized into several emotional tones such as hate, not-hate, non-offensive, irony, neutral, positive, and negative. The scores represent the proportion of each sentiment relative to the total number of interactions (e.g., posts, comments) for each person.

In the context of these scores, hate and negative sentiment scores are relatively low for most individuals, with many scoring a 0, indicating little to no overt hostility. However, the negative sentiment does appear in varying degrees, highlighting some criticism or dissatisfaction expressed in the comments, such as with figures like Roderich Kiesewetter, Zoe Mayer, and Johannes Vogel in earlier sets. These scores suggest that while the conversations around these figures are not dominated by hate, there is some level of critique and discontent present.

On the other hand, figures like Michael Roth, Ralf Stegner, and Florian Hahn show a higher proportion of not-hate, indicating more neutral or positive reactions, where the overall sentiment isn't hostile. Figures with positive sentiment scores, such as Florian Hahn, Frank Schäffler, Maximilian Mordhorst, and Ralf Stegner, stand out as being generally well-received, with positive emotions dominating the conversation. This implies that their posts and comments tend to generate more favorable responses, with positive sentiment scores indicating approval, admiration, or general support from the audience.

The irony and neutral scores reflect the tone of comments that are neither overtly positive nor negative, showing a more detached or nuanced perspective. A high irony score, for example, can indicate that the post or comment may be sarcastic or not entirely sincere, while a neutral score suggests that people are responding without strong emotional engagement—neither supporting nor opposing the figure intensely.

Non-offensive sentiments, another key category, indicate how much of the content is deemed respectful or neutral, avoiding harmful language. The high levels of non-offensive sentiment across the board suggest that most users engage in discussions that avoid explicit offense, suggesting a preference for civil discourse, even when disagreements exist.

To summarize, the data shows that public figures with higher positive sentiment scores are generally more liked, while those with higher negative scores might be facing more criticism. Most figures, however, experience a mix of neutral or non-offensive responses, with only occasional spikes in irony or negative sentiment, highlighting the complexity of online political discussions.

no.	iteration	history	budget	motivation	ranking	posts	comments	likes	dislikes
1	-	y	n	n	-	83	169	-	-
2	1	y	n	n	-	123	334	-	-
3	2	y	n	n	-	212	354	-	-
4	3	y	n	n	-	113	264	-	-
5	-	n	y	y	-	77	27	6	4
6	3	n	y	y	-	80	122	12	8
7	-	y	y	y	ranked	71	111	18	20
8	1	y	y	y	ranked	50	98	14	23
9	2	y	y	y	ranked	53	89	15	21
10	-	y	y	y	chronological	63	107	22	27
11	1	y	y	y	chronological	59	151	20	21
12	2	y	y	y	chronological	55	117	27	16
13	-	y	y	y	random	56	140	23	31
14	1	y	y	y	random	70	142	31	27
15	2	y	y	y	random	61	140	49	36

8. Conclusion

This study demonstrates the significant impact of historical context, resource constraints, and reward-driven mechanisms on user engagement and AI-generated responses in political discourse on social media platforms. By employing an agent-based simulation approach and fine-tuning AI models, we were able to assess the quality and tone of content using sentiment analysis, irony detection, and offensiveness classification. Our results show that considering past conversation history positively influences user engagement, as interactions based on prior exchanges tend to foster more active participation. Conversely, when historical context is excluded, engagement levels drop, highlighting the importance of maintaining continuity in conversations.

Furthermore, the research reveals how time and motivation constraints can lead to passive forms of engagement, such as liking or disliking content, instead of more active contributions like posting or commenting. This finding underscores the challenge of designing AI-driven models that can effectively manage and balance limited resources while sustaining meaningful interactions. Sentiment analysis across the dataset further emphasizes that online political discussions often reflect a range of emotions, from positive to negative, with neutral and non-offensive sentiments prevailing in most cases. Although criticism and discontent are present, they are not overwhelmingly dominant, suggesting that civil discourse remains a significant aspect of social media interactions.

In conclusion, this study contributes to the understanding of how AI models can be better designed to simulate realistic social media interactions, accounting for historical context, resource constraints, and user motivations. The findings also highlight the need for more comprehensive frameworks to analyze and track the evolution of sentiment, irony, and engagement dynamics in digital spaces, which are critical for fostering more responsible and constructive online discourse.

9. Acknowledgments

I am grateful to the Trier University, with special thanks to Simon Münker and Prof. Dr. Achim Rettinger, for their insightful input and guidance on examining the role of previous conversation history in shaping AI-generated responses. I also extend my appreciation to the team at the Karlsruhe Institute of Technology—particularly Fabio Sartori, Andreas Reitenbach, and Prof. Dr. Michael Mäs—for their valuable support and contributions to the investigation of user engagement dynamics under time and energy limitations within reward-driven settings.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using this service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] L. Zhang, Y. Hu, W. Li, Q. Bai, P. Nand, Llm-aidsim: Llm-enhanced agent-based influence diffusion simulation in social networks, Systems 13 (2025) 29.

[2] T. Hu, D. Liakopoulos, X. Wei, R. Marculescu, N. J. Yadwadkar, Simulating rumor spreading in social networks using llm agents, arXiv preprint arXiv:2502.01450 (2025).

[3] L. Wang, J. Zhang, H. Yang, Z.-Y. Chen, J. Tang, Z. Zhang, X. Chen, Y. Lin, H. Sun, R. Song, et al., User behavior simulation with large language model-based agents, ACM Transactions on Information Systems 43 (2025) 1–37.

[4] X. Dai, Y. Xie, M. Liu, X. Wang, Z. Li, H. Wang, J. Lui, Multi-agent conversational online learning for adaptive llm response identification, arXiv preprint arXiv:2501.01849 (2025).

[5] W. Wang, A. Casella, Performant llm agentic framework for conversational ai, in: 2025 1st International Conference on Artificial Intelligence and Computing, 2025.

[6] M. Kumar, et al., Exploring hate speech detection: challenges, resources, current research and future directions, Multimedia Tools and Applications (2025) 1–37.

[7] A. Fiat, E. Koutsoupias, K. Ligett, Y. Mansour, S. Olonetsky, Beyond myopic best response (in cournot competition), Games and Economic Behavior 113 (2019) 38–57.

[8] S. Horn, S. Banisch, V. Batzdorfer, A. Reitenbach, F. Sartori, D. Schwabe, M. Maes, Success-driven user activity contributes to online polarization, Available at SSRN 5031685 (2020).

[9] R. Axelrod, The dissemination of culture: A model with local convergence and global polarization, Journal of conflict resolution 41 (1997) 203–226.

[10] F. Riquelme, P. González-Cantergiani, Measuring user influence on twitter: A survey, Information processing & management 52 (2016) 949–975.

[11] F. Wu, D. M. Wilkinson, B. A. Huberman, Feedback loops of attention in peer production, in: 2009 International Conference on Computational Science and Engineering, volume 4, IEEE, 2009, pp. 409–415.

[12] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks, Annual review of sociology 27 (2001) 415–444.

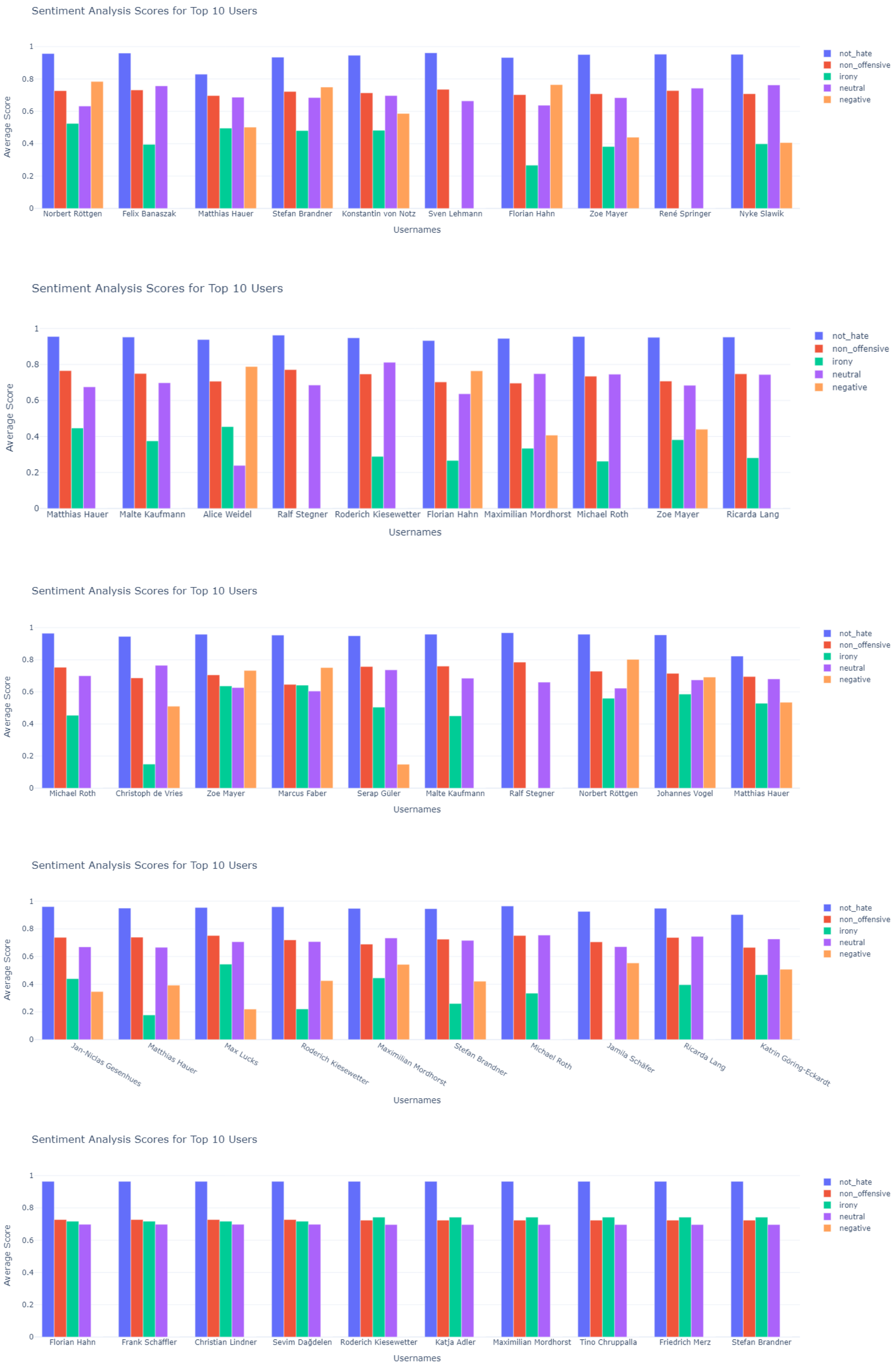

Figure 2: Sentiment analysis of posts from the top 10 users

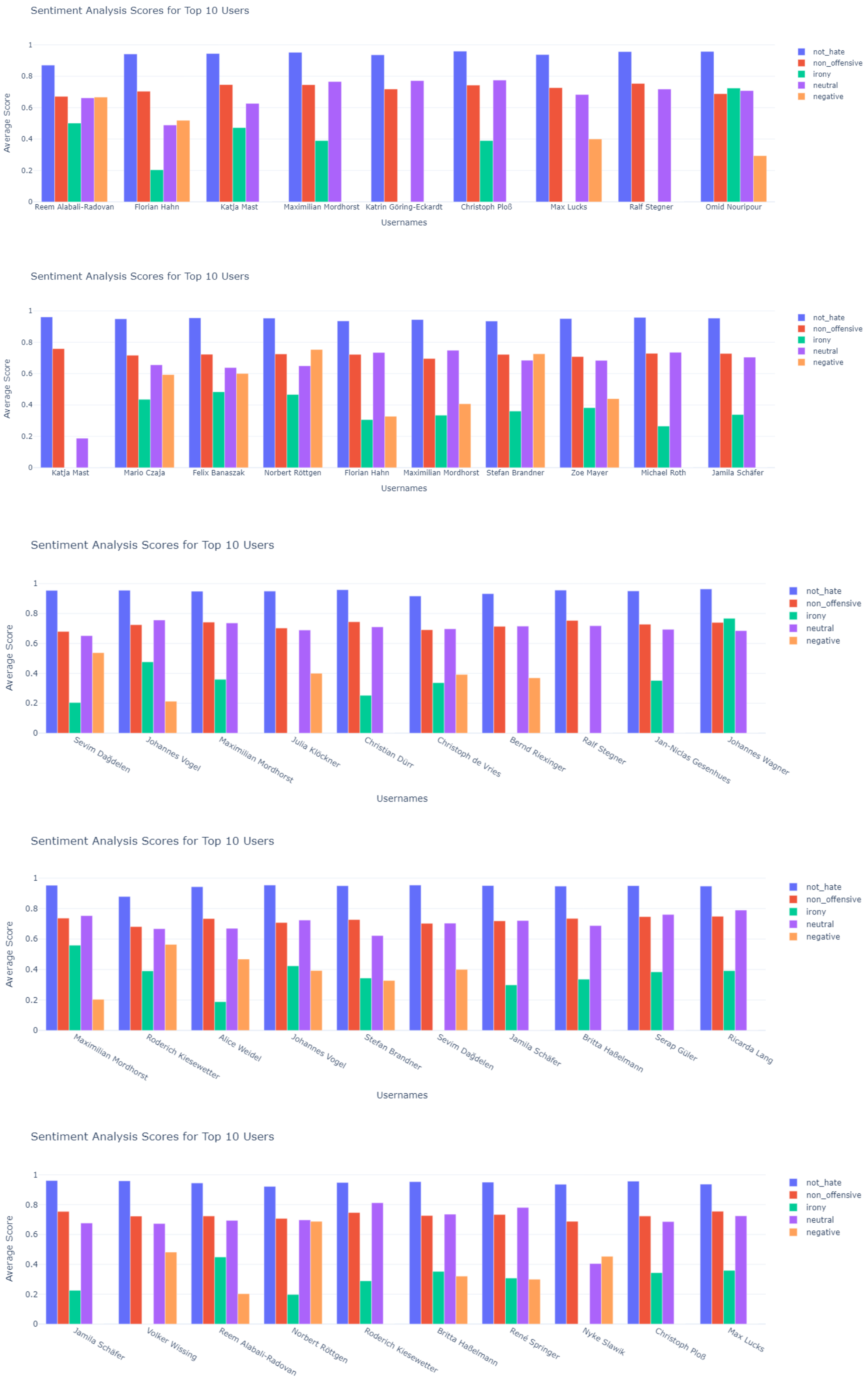

Figure 3: Sentiment analysis of comments from the top 10 users