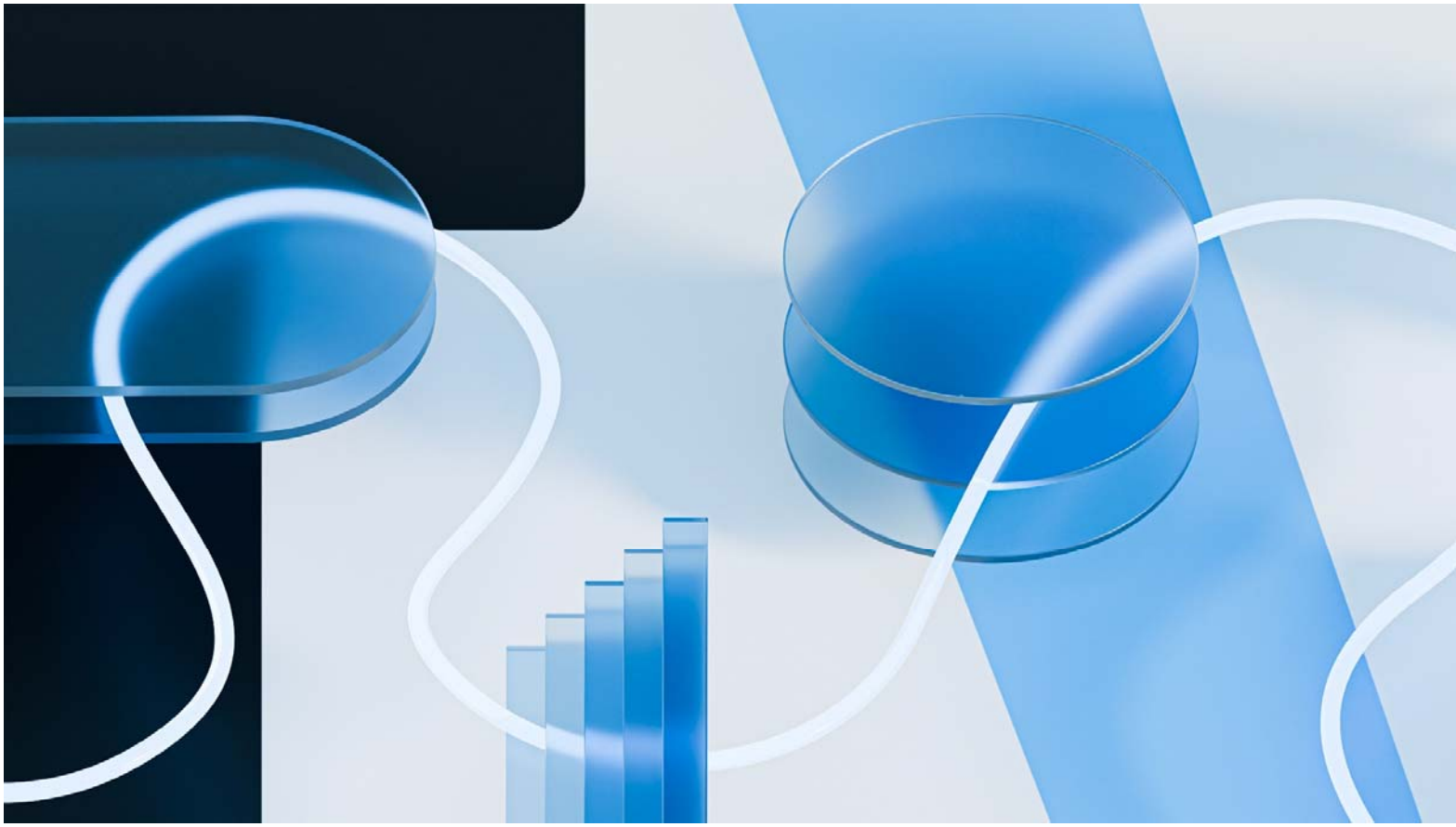
twon

twin of online social networks



TWin of Online Social Networks

Deliverable D4.2

Final data set DS1-DS4

Main Authors: Alenka Guček, Abdul Sittar, Michael Heseltine

Funded by the European Union

About TWON

TWON (project number 101095095) is a research project, fully funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). TWON started on 1 April 2023 and will run until 31 March 2026. The project is coordinated by the Universiteit van Amsterdam (the Netherlands) and implemented together with partners from Universität Trier (Germany), Institut Jozef Stefan (Slovenia), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia), Slovenska Tiskovna Agencija (Slovenia) and Dialogue Perspectives e.V (Germany).

Funded by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

twon
twin of online social networks

Project Name	Twin of Online Social Networks
Project Acronym	TWON
Project Number	101095095
Deliverable Number	D4.2
Deliverable Name	TWON Final Data Set DS1-DS4
Due Date	30.06.2025
Submission Date	26.06.2025
Type	DEM — Demonstrator, pilot, prototype
Dissemination Level	PU - Public
Work Package	WP 4
Lead beneficiary	6-JIS
Contributing beneficiaries and associated partners	Institut Jozef Stefan (JSI), Universiteit van Amsterdam (UvA)

Executive Summary

To provide scientifically grounded evidence on the impact of online social networks (OSNs) on democratic debate, the project develops simulations of OSNs through digital replicas known as TWONs. Determining appropriate parameters for the underlying mathematical models—such as the weights used in large language models—requires robust empirical data. For this purpose, we have collected data from four key sources: online news monitoring (DS1), social media monitoring (DS2), individual data donations (DS3), and debate simulations (DS4).

Datasets from news (DS1) and social media (DS2), previously described in Deliverable D4.1, are presented together with their corresponding summary statistics. In line with the project's work plan, the final dataset presented in this deliverable draws on data from individual donations (DS3) and debate simulations (DS4).

This deliverable introduces and briefly describes the final datasets and provides a set of descriptive statistics and preliminary findings that will support subsequent phases of the project. Importantly, we are now making these datasets (DS1–DS4) publicly available, and include the corresponding access links and summary statistics.

twon
twin of online social networks

Contents

List of Tables

List of Figures

List of Abbreviations

DS	Data Source
LLM	Large Language Model
OSN	Online Social Network
SFT	Supervised Fine Tuning
TWON	Twin of an Online Social Network

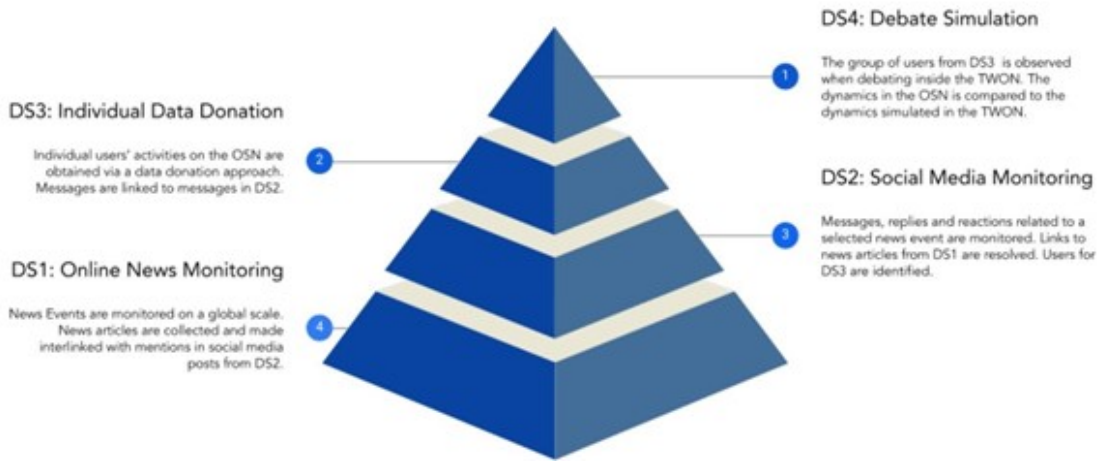


Figure 1: Data sources considered in the TWON project

1 Introduction

The project aims to simulate online social networks in order to provide scientifically-founded evidence about the effect of online social networks on democratic debates. Classic simulations are based on simplistic models and heavily rely on assumptions (e.g., about human behavior), the correctness of which may turn out to be a bottleneck and compromise the accuracy of the results. In contrast, in order to simulate online social networks as realistic as possible, we follow a data-driven approach and use state-of-the-art models (such as large language models, or LLMs). The number of parameters of such models ranges from millions to hundreds of billions. Finding appropriate parameter values of such complex mathematical models (e.g., weights of the aforementioned LLMs) is a challenging task, for which we use real-world data. In particular, we use data from four different sources: online news monitoring (DS1), social media monitoring (DS2), individual data donation (DS3) and debate simulation (DS4), see also Figure 1. The final dataset presented in this deliverable integrates data from all four sources. For DS1 and DS2, we now provide newly added descriptive statistics. The data from individual donations (DS3) and debate simulations (DS4) are presented up to the current phase of the project, with additional simulated data to be produced in the final stages.

The following pages provide a brief description of the final dataset and the process by which it was created.

2 DS1: Data from Online News Monitoring

For DS1, no new data was collected. We however analysed the existing data set described in D4.1, that was previously published as a sensitive dataset. 70508 news articles, mentioned in the DS2 from beforementioned D4.1 were analysed and the analysis is available at `https://github.com/TheBigSM/News_Dataset_Analysis`. DS1 data (metadata), without the news text (which is protected by copyright) has been

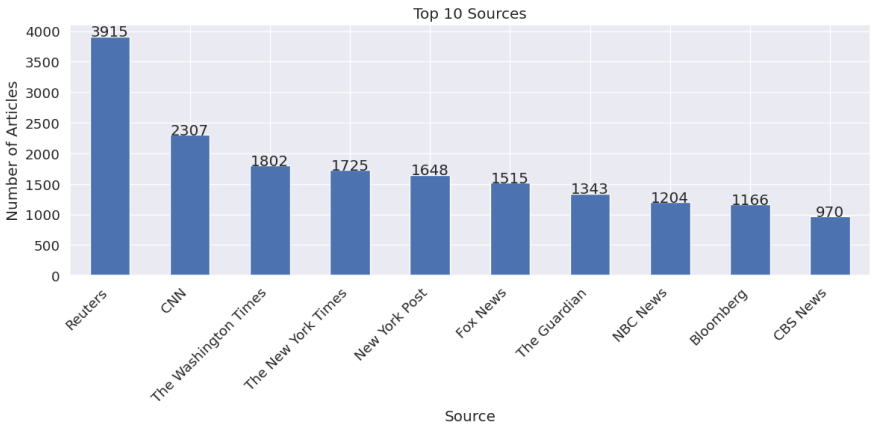


Figure 2: Top 10 sources of news providers

published as a public dataset available at https://zenodo.org/records/15577602. The news dataset consists of articles collected from a diverse set of online sources, with the top 10 sources contributing a significant portion of the overall content (see Figure 2). The most represented source is Reuters with 3915 articles, followed by CNN (2307 articles) and The Washington Times (1802 articles), among others. Together, the top 10 sources account for approximately 25% of the dataset. The average article length is approximately 528 words, with a median of 359 words, indicating a moderate variance in content length across sources. This distribution reflects a mix of short-format updates and longer analytical pieces.

A word frequency analysis was conducted to identify the most commonly used terms across the news dataset. As visualized in the word cloud in Figure 3, the most prominent words reflect core themes and topics covered in the articles, including Trump, president, and house. Sentiment analysis of the news dataset shows that overall sentiment tends slightly positive, with a mean polarity of 0.0638 and a median of 0.0527. The standard deviation of 0.0856 indicates relatively low variation in sentiment across articles. The average subjectivity score of 0.3287 suggests that most content leans toward a neutral or fact-based tone, with limited use of subjective or emotionally charged language. Topic modeling with LDA revealed eight main themes, including U.S. politics, the Ukraine war, public health, and economics. Political and international topics (e.g., Topics 0–2) were most prevalent across articles. Topic distribution over time showed shifting coverage aligned with major events. Full topic details and temporal visualizations are available in the accompanying GitHub repository `https://github.com/TheBigSM/News_Dataset_Analysis`.

3 DS2: Data from Social Media Monitoring

3.1 USA Test User Summary Statistics

We use the USA user sample to establish some baseline account, content, and exposure for Twitter users, generally. The dataset is publicly available at https://zenodo.org/records/15577602. To do this we first

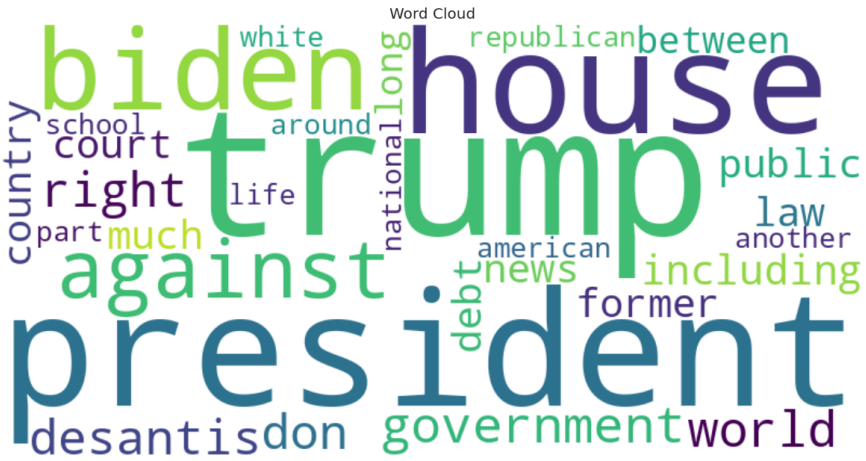


Figure 3: Word cloud of the most frequent words in the analyzed news

present basic aggregated metrics of account activity. We then develop a series of machine learning classifiers to classify both the content that these accounts post and like as well as the content that is posted by the accounts that they follow (and, thus, the type of content that out users are likely to be exposed to in their feeds). Specifically, we use BERTweet, a RoBERTa-based transformer model developed for classifying social media data, to create four classifiers: 1) A binary classifier of political content, 2) a binary classifier of sport-related content, 3) a binary classifier of entertainment content, and 4) a three-way classifier of content ideology (left, center, right). We note here in advance that our sampling method likely oversamples both active accounts (in terms of posting and engaging with other user's content) as well as political accounts and political content. We therefore see these statistics as more so representative of politically engaged users, rather than necessarily all users.

Account Metrics

The median account-level metrics of interest for our accounts were:

- 257 followers
- 697 following
- 11,438 tweets (includes retweets)

The followers to following ratios align with expectations, as the average user is likely to follow more accounts than have other accounts following them. The absolute numbers are also relatively low, reflecting that most users have contained social media networks, following only a manageable number of accounts and having their content only followed by a limited number of users. The number of tweets, however, is higher than might be expected – this likely reflects the sampling process, wherein a user, by definition, had

to be active to be included in the sample. We therefore see a high volume of content being seen by a low number of followers.

Topics

From a total of 394,353 tweets across time from test users

- 55.7% Political
- 4.3% Sport
- 2.9% Entertainment
- 0.00 average ideology (22% left, 22% right, 55% center)
- 0.01 average ideology among political posts (36% left, 37% right, 26% center)

From 363,200 likes across time from our users

- 55.7% Political
- 5.2% Sport
- 5% Entertainment
- -0.01 average ideology (25% left, 23% right, 51% center)
- 0.00 average ideology among political likes (39% left, 40% right, 21% center)

Our users primarily post about and like political content. This is not necessarily surprising given the sampling strategy. However, the percentage of content from other major categories such as sport and entertainment is low, reflecting the tendency for these users to engage narrowly in political spaces and rarely deviate to other topical discussions. The ideology of content is very much balanced, meaning that all sides of the ideological debate are present and being represented in a relatively balanced manner.

Total tweets “seen” between 2023-5-25 and 2023-5-31

- 2,605,082 combined
- 551,649 unique
- From 4,482 accounts

In just a 6 day window, our users potentially encountered a large amount of content. Each post was seen by an average of 5 users, suggesting a 5% crossover of content among our users. Of course, we cannot verify, at this time, which of these were actually seen by users, but the pool of content actively circulating within even this small network is very large – more than any one individual could realistically read.

Engagement

- 0.01869423% liked
- 0.005258952% replied
- 0.008867283% retweeted
- 0.001957712% quoted

Cross-referencing this content with the engagement behaviors of our users during the same time period, we see very low levels of engagement. This suggests two things: 1) most content is not actually viewed by individuals in the follower network, and 2) even when viewed, users very rarely actually engage in any meaningful way. This means that while it is possible to easily collect and analyze engagement data at the aggregate level, at the individual level the signal level is incredibly low. Therefore, by definition, when we see engagement activity captured in the data, we are very much seeing an exceptional response from any given user and not the norm.

3.2 Germany Conversations Summary

Having established some baseline metrics for account types and content exposure using US data, we then turn to a similar pool of German data and follow a similar process to classify the content of German conversations on Twitter. The dataset is publicly available at https://zenodo.org/records/15577602. First, we collect all replies sent to German Members of the Bundestag on Twitter between January and May of the year 2023. This resulted in a dataset of 3,226,277 tweets. This full dataset can be used for further training of language models and for comparisons of simulated and real-life German social media content at the individual message and complete conversation levels. For now, we analyze a sample of the data.

We randomly sample 250 users from the users who replied and then cross-reference all of their replies in the dataset. We then cross-reference the top-level post that they were originally replying to to create a three-tier dataset: Original post <- Reply <- Replying User Information.

We then classify the content of these replies and top level posts. In this instance, rather than hand-coding training data and fine-tuning transformer models, we take a Large Language Model approach to classifying the text. Specifically, we use Llama3 70b, the largest, most up to date, and most accurate open-source LLM available at the time of coding, to classify the original posts for the same four content criteria (political, sport, entertainment, ideology), plus message negativity. We also classify the content of the replies to these posts for 1) political content, 2) message ideology, 3) negative content tone, 4) a binary classification of whether the reply is substantive in nature (e.g. contributes meaningfully to the conversation and is not simply limited to an emoji, and short statement of agreement or disagreement, or is just a link).

Account Metrics

The median account-level metrics of interest for our accounts were:

- 861 followers
- 2,359 following
- 11,774 tweets (includes retweets)

Here again we see a skewed ratio of account followed and the number of followers. In the German case, however, the absolute numbers are much higher than in the US. This may reflect the exact types of accounts captured in the sample. Despite verified accounts being excluded, it is still possible that more institutional accounts are actively mentioned or involved in these political conversations. Still, the number of total tweets matches the US sample and is, again, relatively high.

Original Post Metrics

In total, we take a random sample of 6594 top-level posts from 487 individual accounts. These posts were classified as the following:

- 86.5% political
- 7.3% entertainment
- 4% sports
- 0.01 average ideology (11% left, 12% right, 77% center)
- 0.01 average ideology among political posts (13% left, 14% right, 73% center)
- 36% have a negative sentiment

In the German case, the posts are overwhelmingly political and significantly more political than the posts seen in the US data. The percentages of sport and entertainment content are similar to the US and are, again, relatively small. The ideology of posts is again balanced perfectly around the ideological center – however, in the German case we actually see significantly less ideological content, with the vast majority being ideologically centrist.

The general engagement with these posts also provides important information into the type of content that social media users find most engaging.

On average non-political posts generate 15 replies and 80 likes, while political posts receive 78 replies and 315 likes (a more than 400% increase). Left-wing posts receive 103 replies and 357 likes, neutral posts

55 replies and 232 likes, right-wing posts 121 replies and 536 likes. Negative posts receive, on average, 91 replies and 451 likes, compared to 56 replies and 188 likes for non-negative posts

Political content with a clear ideological lean, especially a right-wing lean, therefore receives the most engagement from users. Negative content also receives increased engagement. The combination of ideology and negativity is especially conducive for engagement – negative left-wing posts receive 127 replies and 497 likes, while negative right-wing content receives 139 replies and 678 likes, on average.

Reply Metrics

In total, these posts generated 295,146 replies from 79,070 unique accounts. These replies were classified as the following:

- 92% political
- 0.1 average ideology (8% left, 17% right, 75% center)
- 74.3% negative tone
- 61% substantive content

First, the percentage of political content is very high. This, in part reflects the nature of the conversations, coming in response to overtly political accounts. However, there may also be some technical considerations which influence these numbers – specifically, the content of the reply also tags/mentions to author of the original top-level post that they are replying to. As these posts come from political accounts, it is likely that Llama3 is able to identify the tagged accounts as being political, therefore skewing the numbers upward. The ideology of these replies is more right-wing skewed than the original posts. Almost twice as many replies are classified as right leaning than left leaning. However, the vast majority of replies are classified as centrist, suggesting that replies (at least based on the raw text), contain less overt ideological cues. Notably, almost three quarters of all replies are classified as negative. This suggests that negative engagement is commonplace, especially in the political space being examined. Finally, almost two-thirds of replies are classified as substantive. This, from a normative standpoint, is potentially a positive finding, suggesting that despite the tendency for responses to be negative, users are still responding in a way that meaningfully contributes to the debate and at least engages substantively with the original content in some way, as opposed to merely posting one-word negative comments.

Takeaways

With these two datasets we have offered broad analysis of the types of content that users follow, see, and engage with. Based on these analyses, there are some general takeaways that can be used to inform our expectations about user behavior, both in our online case studies and in our virtual simulations.

1. Following patterns: most users are following significantly more accounts than they are being followed by. Thus, while they are exposed to a relatively large amount of content, the potential exposure for the content that they create is quite limited. This suggests that most users have little influence within the network, while a small number of influential accounts drive the majority of engagement. 2. Politics is prevalent: The majority of content, both original posts and replies are political in nature. These political posts in turn generate significantly more engagement than non-political posts. As our sample is drawn from political spaces, the broader implications are limited, but this does suggest a siloing of users and topics, where the politically engaged stay engaged within the field of politics and rarely venture to non-political topics. 3. Ideology: The overall ideology of messages is generally quite balanced. Right-wing content generates more engagement and is especially prevalent in the replies to political posts. This suggests that while all viewpoints are represented, right-wing users may be the most engaged (within the general user base and not necessarily within the more elite high-profile user tier). 4. Tone: Negativity is popular. Not only are a majority of comments negative, but negative posts also generate more engagement. When combined with ideological posts, these engagement effects are amplified even further. This has implications for the tenor and quality of public debate – on the one hand, negativity spurs debate and engagement, but in turn produces highly negative forms of engagement, creating something of a reinforcing ecosystem of negativity on the platform.

4 DS3: Individual Data Donations

Case studies generated two datasets: German case study focuses on pre-bunking of misinformation and communicating statistical uncertainty, Serbian case study focuses on disinformation in the Ukraine war.

4.1 German case study

To generate real data from real users in a controlled environment, we conducted an experimental study in Germany, in conjunction with our partners at the Robert Koch Institute. The focus of this study was on two facets of scientific communication:

1. Pre-bunking of misinformation
2. Communicating statistical uncertainty in scientific information

To test these areas, we designed an experiment around social media posts that had previously been designed and used by the Robert Koch Institute in their social media messaging. In particular, these posts focused on the topic of vaccines, with the pre-bunking posts focusing on MRNA vaccines, generally, while the uncertainty posts focused on the Monkeypox vaccine.

The study was conducted between March 24 and March 28 2025. Participants were recruited through Prolific, with pre-participation screens set for Germany location, fluency in the German language, desktop participation only, and a 90%+ prior task approval rating. In total, 1,295 participants began the study and 1,200 successfully finished.

Upon entering the study, these respondents were randomized into five groups: a control group for our pre-bunking and uncertainty messaging groups, a pre-bunking treatment, a pre-bunking treatment plus additional "truth sandwich" content, and an uncertainty treatment.

Pre-survey

Each respondent received a URL link taking them to our realistic online social media platform (TWON). Before entering the TWON platform, participants were asked to complete an initial survey which included questions for: self-rated ideology (1–11, left to right), age, gender, education level, and level of political interest (1–5). After completing these questions, to enhance the realism of the user experience, respondents were asked to select one of four pre-made avatar and username combinations to represent them on the platform.

Experimental Platform Design

After completing the pre-survey, respondents were then taken to the platform main page, with content displayed specific to the randomly assigned treatment group. Each user saw eight social media posts with text and images (see Figure4), displayed in a random order. For all treatment groups, seven of these posts were identical, taken from real-life posts from major institutional accounts (newspapers and brands), covering a range of topics including politics, sports and entertainment. The eighth post contained content specific to the randomized treatment group (pre-bunking control, pre-bunking treatment, truth sandwich treatment, uncertainty control, and uncertainty treatment - see Figure 5). Each of these experimental posts was identical in style and contained only minor manipulations of the text and images.

Figure 4: Sample Treatment Post

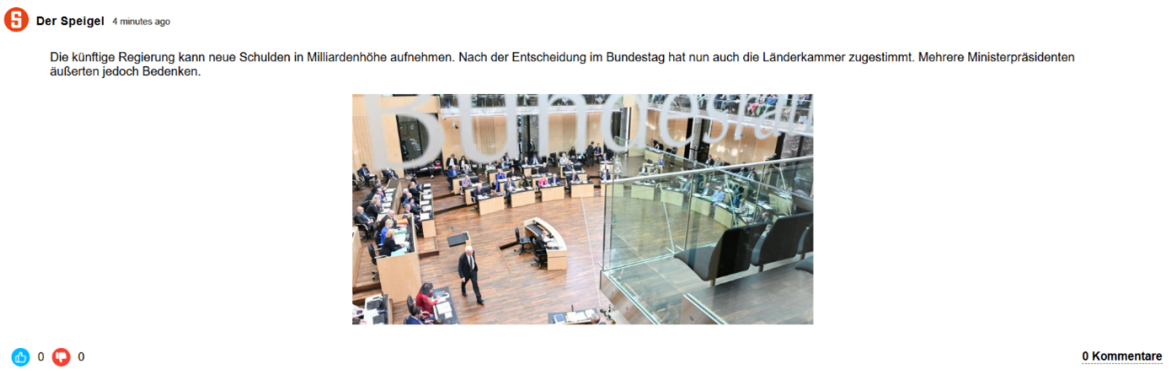


Figure 5: Non-Treatment Post

Once on the platform, a notification appeared on screen asking users to scroll and read all posts and encouraged them to interact with the content just as they would on a normal social media site. To ensure that users interacted with the content and did not simply scroll through without engaging, we then asked users to like at least two of the eight posts and to comment on at least three of the posts.

Post-Survey

After completing the assigned tasks, participants were directed to a post-platform survey which asked key outcome questions relating to:

1. The content seen on the platform
2. Attitudes towards scientific institutions

3. Attitudes towards vaccines

These questions are used to specifically test for potential attitudinal effects across treatments.

Results – Descriptive

As the primary focus of this study is the data generated from user interactions on the platforms, we first provide a broad descriptive overview of the likes, dislikes, and user comments collected.

Table 1: Post-Level Engagement Metrics

Post	Avg. Comment Sentiment	Comments	Likes	Dislikes
Monkeypox Control Post	1.5778	90	116	13
DFB Team Football Post	1.4277	484	338	110
Monkeypox Uncertainty Treatment Post	1.3918	97	109	25
MRNA Vaccine Treatment Post	1.2280	193	218	30
MRNA Vaccine Control Post	1.2165	97	119	23
Assassin's Creed Shadows Review	1.1200	525	413	97
Istanbul Mayor Arrest	0.7910	311	395	65
Bundestag Debt Discussion	0.7716	359	302	143
Chancellor's Last Council	0.7647	340	231	101
Too Hot to Handle (Netflix)	0.7512	414	130	394
Trump and Musk Diversity Debate	0.4280	528	149	352

Table 1 shows the breakdown of comments on each post, based on average comment sentiment and comment frequency. The sentiment of each comment was classified using a simple zero-shot prompt with Phi3.5, on a 3-point negative, neutral, positive scale (0,1,2). These scores were then aggregated to give an average sentiment score for comments to each post.

As can be seen, the RKI healthcare posts consistently have the most positive (or close to the most positive) sentiment. Notably, the uncertainty treatment post did have a slightly less positive comment sentiment than the control post.

Outside of the treatment posts, the sports post generated the most positive responses, followed by the video game review post. All other posts had more negative comments than positive. Political posts showed similar levels of negative sentiment, with the US politics post having the lowest sentiment overall. The Netflix post also had low sentiment due to negative reception of its programming.

Likes and Dislikes

Healthcare posts generally received fewer likes and dislikes. The video game post received the most likes, followed by the Turkish protests post. The US politics and Netflix posts received the least likes. In contrast, the Netflix post had the most dislikes, followed by the US politics post. Healthcare posts received the fewest dislikes.

Engagement by User Type

We also present a set of sub-category trends in engagement, based on the gender of the user.

Table 2: Engagement by Gender

Post	Likes Women %	Dislikes Women %	Comments Women %	Women Sentiment Difference
Too Hot to Handle	59.8	33.0	40.1	-0.52
Monkeypox Control	49.2	23.0	53.3	0.26
Monkeypox Uncertainty	48.6	32.0	50.5	0.05
MRNA Vaccine Treatment	46.0	43.3	46.6	-0.43
MRNA Vaccine Control	42.9	21.7	46.4	-0.17
European Council	41.7	39.6	44.4	-0.01
Turkey Protests	37.0	29.0	38.3	-0.36
Video Game	33.8	37.1	34.0	-0.10
Bundestag Debt	32.4	39.3	34.5	0.22
Trump and Musk	32.2	47.8	46.8	0.11
German Football Team	27.7	43.1	34.2	-0.01

Women made up 40% of our sample. Female respondents showed distinct patterns:

- More likes on the Too Hot to Handle post (+20% above baseline)
- Fewer likes on football and general politics posts
- More likes and comments on healthcare content
- Less engagement with video game content

Dislike patterns mirrored likes. Women disliked Too Hot to Handle less, disliked politics posts at or above baseline, and showed fewer dislikes on healthcare posts.

Sentiment analysis showed:

- Women had significantly lower sentiment for Too Hot to Handle despite high liking.
- Turkey protests post received lower sentiment from women than men.
- Sentiment differences for other posts were small.

Overall, the data generated from this case study serves as useful information for simulation and language model training. It provides insight into:

- Which types of posts generate the most engagement
- What types of engagement are used in different contexts
- How different types of users (e.g., by gender) engage with content

4.2 Serbia Ukraine Disinformation Study

Our third study focused on Serbian language data generation in the context of news and information engagement around the war in Ukraine. To do this, we conducted a multi-round study on our TWON platform, in conjunction with our partners at The Slovenian Press Agency (STA). Together, we designed an experiment using manipulated Serbian language news articles, alongside general Serbian language social media posts. The topic of these articles and posts was either specifically about the war in Ukraine or about general topics such as sports or entertainment. Centrally, the Ukraine news articles, contained varying degrees of disinformation, allowing for a test of 1) who clicks on these disinformation articles, and 2) how users respond when exposed to disinformation.

The study was limited to Serbian participants and fielded on May 28, 2025. Participants were recruited through Latenta, a Serbian user recruitment specialist, with participation limited to desktop users. In total 272 participants successfully finished the study.

Experimental Platform Design

For this study we utilized a multi-round system, wherein we simulated a user refreshing their feed and thus reading new content over time through the inclusion of a refresh button at the top of TWON feed. Through this feature, we observe user behavior over multiple rounds on the platform, accounting for their behavior in the previous round.

In this design, upon entering the study, respondents were randomized into three groups: a control group, a “reinforcing” group, and an “opposing” group. All three groups saw the same randomized feed

once initially entering the platform. This state represents “Round 0”. After refreshing, the control group saw another randomized feed, the reinforcing group saw a feed which upranked content which aligned with the type of content that they had previously engaged with in round 0 (meaning this content was shower higher in the feed), while the opposing group had the type of content they engaged with from round 0 downranked. This logic was then repeated again for another round after a further refresh.

The criteria used for ranking was engagement with Ukraine content and engagement with disinformation. Each post shown in the feed was given an underlying score for both of these criteria. With likes counting positively and dislikes having an inverted score (multiplied by -1). For news articles, clicking to read counted as positive interaction. Each feed therefore had the following posts and scores:

Table 3: Post Ukraine and Disinformation Scores

Post	Ukraine Score	Disinformation Score
neutral sports (news)	-1	-0.1
neutral (general) A	-1	-0.1
disinfo ukraine (news)	1	1
pro russia (general)	1	-0.1
neutral (general) B	-1	-0.1
neutral ukraine (news)	1	-0.1
neutral general (news)	-1	-0.1
pro ukraine (general)	1	-0.1
neutral (general) C	-1	-0.1
neutral entertainment (news)	-1	-0.1
mixed (general)	1	-0.1
light ukraine (news)	1	0.5

Collectively, this design resulted in 19 different feed pathways repeated over three rounds in the following logic (see Figure 6).

Descriptive Results

In total, 3,663 likes, 2,055 news article clicks, 1,860 dislikes, and 464 comments were collected. The two main behaviors that users could perform (and where instructed to perform) on the platform were likes (dislikes) and news articles views, meaning that these behaviors of most interest in this dataset.

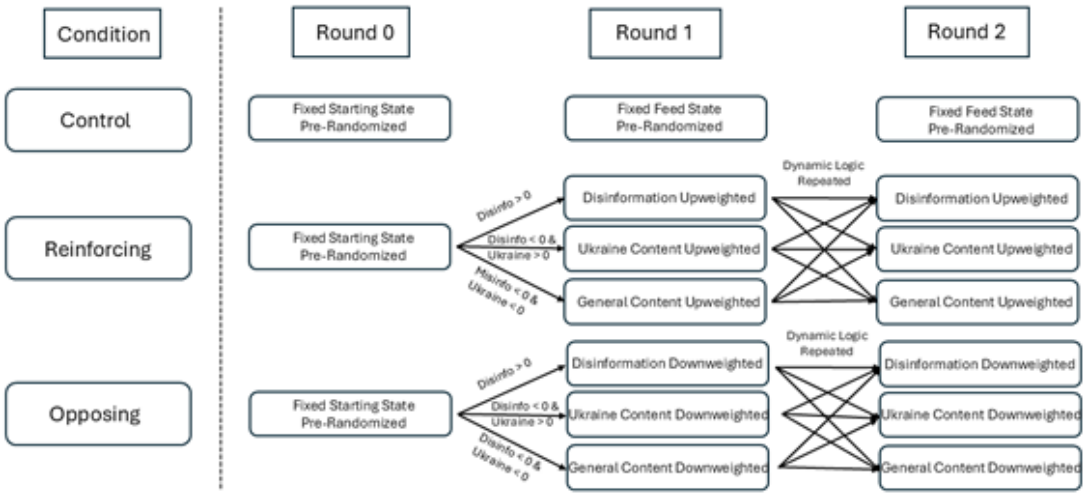


Figure 6: Visual representation of conditions and three rounds of the case study

In terms of article views, articles about non-Ukraine content generated more clicks compared to Ukraine content in every round (each round had 3 Ukraine stories and 3 non-Ukraine stories, providing an expected baseline of an equal split). In terms of disinformation, looking at just those posts flagged as Ukraine content, there is an approximate balance, relative to baseline expectations, in terms of the ratio of disinformation and non-disinformation link clicks (2 out of 3 links contain some disinformation, meaning that disinformation clicks should be approximately twice as common as non-disinformation. In the initial round (round 0), the proportion of disinformation clicks is higher than the baseline, while in round 1 the level is much lower.

Table 4: News Article Link Clicks Engagement Counts

Round	Ukraine Post Status	Link Click Count	Disinformation Post Status	Link Click Count
Round 0	Non-Ukraine	533	Non-Disinfo	109
Round 0	Ukraine	371	Disinfo	262
Round 1	Non-Ukraine	323	Non-Disinfo	135
Round 1	Ukraine	305	Disinfo	170
Round 2	Non-Ukraine	281	Non-Disinfo	80
Round 2	Ukraine	242	Disinfo	162

Note: Disinformation Post Status includes only those posts relating to Ukraine. All 3 non-Ukraine news posts are automat-

ically Non-Disinformation. 2 out of 3 Ukraine news links contain disinformation.

In terms of likes, Non-Ukraine content clearly and consistently garnered more engagement than the Ukraine posts. For the disinformation engagement, the ratios generally match expectations (with 2 out of 6 posts containing disinformation). In the first round, similar to the link clicks, the proportion of disinformation likes exceeds baseline expectations, while the following rounds are much more balanced. This suggests that disinformation is potentially engaging at first, but loses its appeal after repeated exposure.

Table 5: Post Likes Engagement Counts

Round	Ukraine Post Status	Like Count	Disinformation Post Status	Like Count
Round 0	Non-Ukraine	1010	Non-Disinfo	415
Round 0	Ukraine	666	Disinfo	251
Round 1	Non-Ukraine	566	Non-Disinfo	311
Round 1	Ukraine	465	Disinfo	154
Round 2	Non-Ukraine	557	Non-Disinfo	259
Round 2	Ukraine	399	Disinfo	140

Note: Disinformation Post Status includes only those posts relating to Ukraine. All 6 non-Ukraine posts are automatically Non-Disinformation. 2 out of 6 Ukraine posts contain disinformation.

Engagement by User Type

Finally, as with the German study, we can assess some basic descriptive differences between types of respondents, again using gender as a primary example. For reference, the makeup of our final sample was 32% women. In general, the link clicks are relatively balanced between the genders and across content types. In terms of disinformation clicks, differences do occur in terms of women being more or less likely to click disinformation, although these differences alternate in direction between rounds. In terms of likes, the gender ratios for Ukraine and non-Ukraine content are relatively balanced. For disinformation though women were more likely to like disinformation in the starting round, although this difference does not persist into subsequent rounds.

Table 6: News Article Link Clicks Engagement Percentages Based on User Gender

Round	Ukraine Post Status	Link Click Women Pct.	Disinformation Post Status	Link Click Women Pct.
Round 0	Non-Ukraine	35.3	Non-Disinfo	26.5
Round 0	Ukraine	35.1	Disinfo	38.4
Round 1	Non-Ukraine	34.9	Non-Disinfo	36.6
Round 1	Ukraine	32.1	Disinfo	28.7
Round 2	Non-Ukraine	24.0	Non-Disinfo	25.0
Round 2	Ukraine	31.7	Disinfo	35.1

Note: Disinformation Post Status includes only those posts relating to Ukraine. All 3 non-Ukraine news posts are automatically Non-Disinformation. 2 out of 3 Ukraine news links contain disinformation.

Table 7: Post Likes Engagement Percentages Based on User Gender

Round	Ukraine Post Status	Like Count	Disinformation Post Status	Like Count
Round 0	Non-Ukraine	32.2	Non-Disinfo	31.8
Round 0	Ukraine	35.7	Disinfo	41.3
Round 1	Non-Ukraine	28.0	Non-Disinfo	25.1
Round 1	Ukraine	25.1	Disinfo	25.0
Round 2	Non-Ukraine	29.1	Non-Disinfo	27.7
Round 2	Ukraine	26.1	Disinfo	23.0

Note: Disinformation Post Status includes only those posts relating to Ukraine. All 6 non-Ukraine posts are automatically Non-Disinformation. 2 out of 6 Ukraine posts contain disinformation.

In general then the descriptives suggest that 1) the Ukraine content was less engaging compared to the non-Ukraine content, 2) that disinformation content is not necessarily more engaging than non-disinformation content, although users may be more likely to initially engage with disinformation, although this does not persist over time, and 3) that men and women engage at similar levels with Ukraine/non-Ukraine and disinformation/non-disinformation content. These insights, among others, highlight how the data can be incorporated into wider content differences and user decision-making design within the final simulation studies.

5 DS4: Debate Simulation

This section presents two datasets developed to support simulation-based studies of online discourse dynamics. The first dataset is an agent-based simulation focused on political discussions on German Twitter, capturing how historical context, time constraints, and reward-driven interactions influence user engagement. The dataset includes generated posts from parliamentary delegates (agents) and replies from regular users (agents), annotated with sentiment, irony, and offensiveness to enable fine-grained content evaluation. The second dataset explores interaction capabilities of large language models (LLMs) across multiple languages. It is designed to test how LLMs respond in structured, multilingual dialogue settings, offering insights into model generalization, context retention, and discourse coherence beyond a single linguistic domain to understand which LLMs work best for different languages.

5.1 Mid-scale simulation

Part of this deliverable is the dataset developed to support mid-scale simulation of user engagement on social media, specifically within the domain of German political discourse on Twitter. The dataset is publicly available at https://zenodo.org/records/15577602. The dataset includes simulated agent posts from parliamentary delegates and corresponding replies from agent regular users, capturing conversational structure, temporal context, and interaction dynamics. By modeling agents with bounded rationality—using myopic best-response strategies—and incorporating variables such as conversation history, motivation, and time budget, the simulation replicates core features of real-world social media behavior.To enrich this data, annotations for sentiment, irony, and offensiveness are included, supporting fine-grained analysis of engagement patterns and discourse tone. The dataset was constructed as part of a larger simulation study exploring how historical context, motivation, and resource constraints influence AI-driven interactions. Two language models were fine-tuned using a supervised fine-tuning (SFT) approach on this dataset to enable generation of contextually relevant posts and replies using the Llama-3.2-3B-Instruct model for the two tasks: posting and replying, developed in the scope of WP3. Models used to generate the dataset can be found here.

The dataset is designed to support experiments in agent-based simulation, where agents generate and respond to content based on expected rewards under limited time and motivational constraints. It includes the inputs and outputs of a series of simulation runs where key parameters—such as the presence of conversation history, available time budget, and ranking mechanisms—were systematically varied. Details on how this variations were conducted and be found in Figure 7.

Data contains users, posts, comments, actions, intacts, motivation analytics, reply likelihoods and time budget analytics for different iterations and conditions explained in Figure 7. Both json and csv formats are available in the dataset. As an example, we show an excerpt of our data, in particular, one user:

```
{
    "_id": "67cd8c02b27a6ff3fc51e84e",
    "username": "Roderich Kiesewetter",
    "loggedIn": false,
    "name": "CDU/CSU",
    "email": "Roderich.Kiesewetter@example.com",
    "motivation": 37,
    "notificationEffect": 28,
    "engagement": 70,
    "success": 10,
    "timeSpent": 25,
    "feedbackScore": 4,
    "entertainmentScore": 4,
    "timeBudget": "67cd8c01b27a6ff3fc51e7b5",
    "followers": [],
    "followings": [],
    "isPublic": true,
    "opinionModel": {
        "opinion": {
            "politics": 0,
            "sports": 0,
            "technology": 0
        }
    },
    "logger": {},
    "actor": {},
    "frustration": 33,
    "biases": {
        "politics": 0.1064047799566532
    },
    "timeBudgetRemaining": 100,
    "createdAt": "2025-03-09 12:39:30.630000",
    "updatedAt": "2025-03-09 12:39:30.630000",
    "__v": 0
}
```


Listing 1: Example JSON object for one agent

The simulation framework is detailed in the workshop paper https://ceur-ws.org/Vol-3977/SemGenAge-6.pdf. This dataset supports the first simulation run, linking conversational context with sentiment and engagement data. Additional simulations will be conducted in the last months of the project.

no.	iteration	history	budget	motivation	ranking
1	-	y	n	n	-
2	1	y	n	n	-
3	2	y	n	n	-
4	3	y	n	n	-
5	-	n	y	y	-
6	1	n	y	y	-
7	2	n	y	y	-
8	3	n	y	y	-
9	-	n	n	n	-
10	1	n	n	n	-
11	2	n	n	n	-
12	3	n	n	n	-
14	-	y	y	y	-
15	1	y	y	y	-
16	2	y	y	y	-
17	3	y	y	y	-
18	-	y	y	y	ranked
19	1	y	y	y	ranked
20	2	y	y	y	ranked
21	3	y	y	y	chronological
22	4	y	y	y	chronological
23	5	y	y	y	chronological
24	6	y	y	y	random
25	7	y	y	y	random
26	8	y	y	y	random
27	-	n	y	y	ranked
28	1	n	y	y	ranked
29	2	n	y	y	ranked
30	3	n	y	y	chronological
31	4	n	y	y	chronological
32	5	n	y	y	chronological
33	6	n	y	y	random
34	7	n	y	y	random
35	8	n	y	y	random

Figure 7: Iterations for the midscale simulation with corresponding details on history, budget, motivation and ranking type

5.2 LLM efficiency study

This dataset was created to evaluate the efficiency and output quality of large language models (LLMs) in the generation of social media content. It supports comparative analysis of model performance across dimensions such as language, topic, persona, and platform. The dataset includes content generated by two LLMs: GPT-3.5 (175B parameters) and Mistral-7B-Instruct (7B parameters), selected for their contrast in scale and accessibility. Both models were prompted using a zero-shot, text-to-text approach to produce posts in three languages—English, German, and Dutch—reflecting varying levels of language model training exposure. The dataset and corresponding analysis can be found at https://github.com/cl-trier/TWON-Agents/tree/legacy-API/experiments/Etmaal-2024.

A total of 1000 samples were generated, equally distributed across combinations of language, platform (Twitter and Reddit), topic (e.g., economy, healthcare, Ukraine war), persona (liberal, conservative, alt-right), and length category ("few-word" to "long"). A subset of 600 samples was manually annotated by native or C2-level speakers, with a stratified selection of 100 samples per language used for preliminary analysis. Each annotated sample includes ratings on three five-point scales: topic alignment, persona alignment, and overall authenticity. Results show that while both models achieve comparable topic alignment, differences appear in persona fidelity and authenticity—particularly for underrepresented languages like Dutch, where generated content exhibited lower realism and cultural specificity.

6 Conclusions

The main output of this deliverable is the public release of four datasets (DS1–DS4), each collected (DS1, DS2, DS3) or generated (DS4) through carefully designed experimental studies. These datasets provide a foundational resource for downstream case study analysis and further simulations within the project. Use case study analysis which will be explained in detail in D5.3, due M36 and simulation computation report results in D4.4, due M36.

twon

twin of online social networks

Contact us

Damian Trilling
Project Coordinator

+31 62 782 7904

d.c.trilling@uva.nl

University of Amsterdam
Postbus 15791
1001 NG Amsterdam

Funded by
the European Union