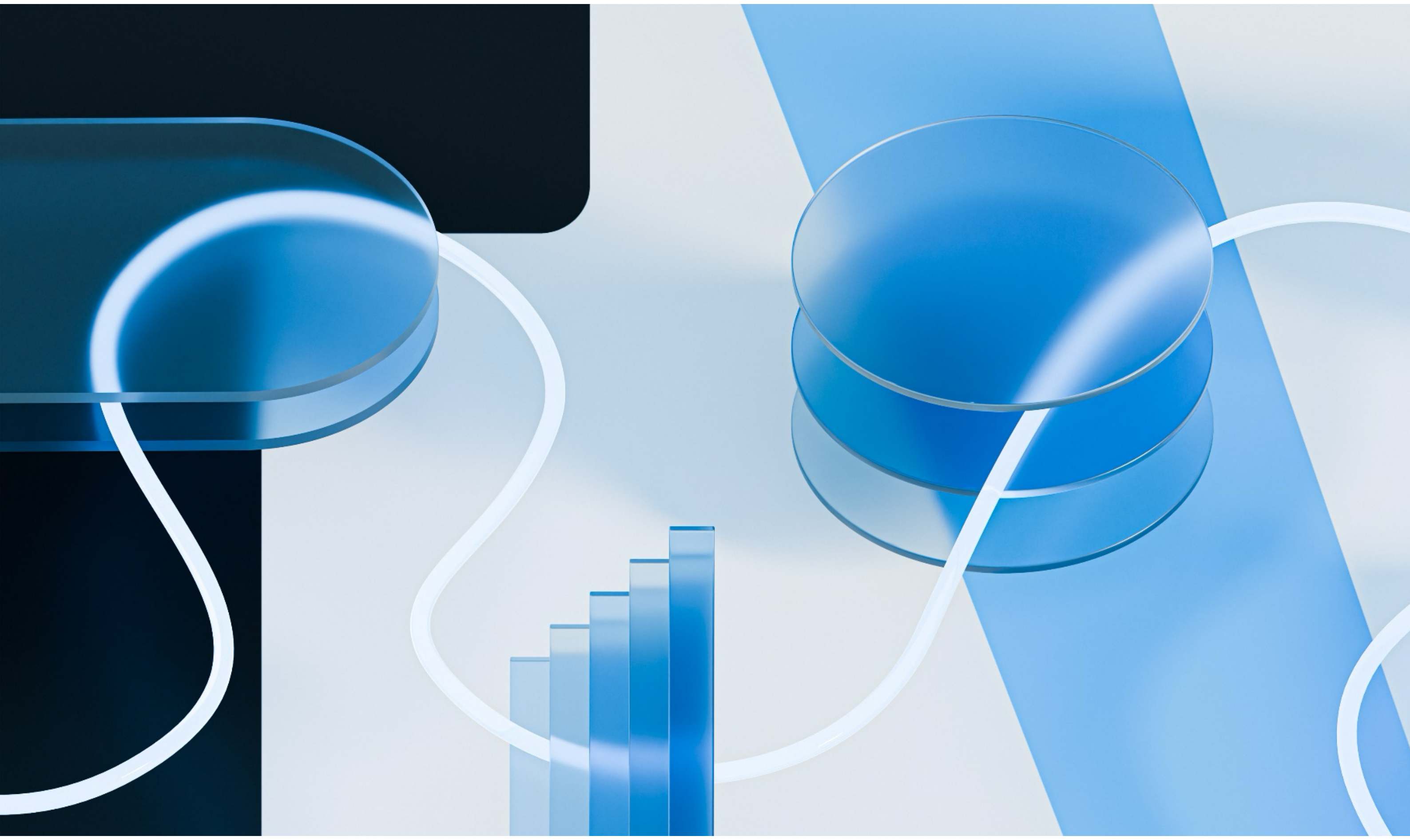

TWin of Online Social Networks

Deliverable D3.2

ML Dataset

Main Authors: Simon Münker & Achim Rettinger

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) and Slovenska Tiskovna Agencija (Slovenia), 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	D3.2
Deliverable Name	ML Dataset
Due Date	31.03.2025
Submission Date	XX.03.2025
Type	DATA
Dissemination Level	PU - Public
Work Package	WP 3
Lead beneficiary	2-UT
Contributing beneficiaries and associated partners	Forschungszentrum Informatik (FZI), Karlsruher Institut für Technologie (KIT), Institut Jožef Stefan (JSI)

Executive Summary

This report details the development of the Machine Learning Dataset (Deliverable D3.2) for the TWON project, a key component of Work Package 3 focused on data-driven estimation of digital twins for online social networks. The ML Dataset is derived from raw data consisting of posts from 𝕏 in German and English, categorized into content from influential public figures and replies from regular users. The dataset development followed a three-phase approach:

Preliminary Analysis We conducted comprehensive distribution and frequency analyses of the raw data containing over 22 million entries across German and English content. This analysis revealed skewed distribution patterns, with most users contributing only a single reply, a significant presence of retweets, and varied text quality metrics. Topic classification identified news and social concerns as dominant themes across both languages.

Preprocessing and Enrichment Based on the preliminary analysis, we applied strategic filtering to enhance data quality, including removal of external URLs, retweets, and extremely short posts. We focused on the most active users to enable user-level modeling and enriched the data with political party affiliations and topic labeling using LLM-based classification.

Task-specific Formatting The dataset was structured into dialogue formats for specific machine learning tasks, including post generation, reply generation, and reply decision modeling. This formatting supports the instruction learning approach detailed in Deliverable D3.1.

The final outcome consists of two enriched datasets containing 7,088 (English) and 5,219 (German) samples with comprehensive metadata including author information, political affiliations, and topic classifications. These datasets will serve as the foundation for training machine learning models that can accurately replicate the dynamics of real-world online social networks.

Contents

List of Tables

List of Figures

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
OSN	Online Social Network
TWON	TWin of Online Social Networks
TWON	Twin of Online Social Networks
WP	Work Package

ML Dataset

Simon Münker & Achim Rettinger [*]

March 26, 2025

1 Introduction

This report, Deliverable D3.2 (ML Dataset), is a crucial component of WP3, which focuses on the data-driven estimation of TWONs. The primary objective of WP3 is to develop machine learning models that can accurately replicate the dynamics of real-world OSNs based on empirical data. The ML Dataset described in this report is derived from the raw data (D4.1.1 Initial Curated Data Set) comprising an unannotated collection of posts from $\mathbb{X}$ (formerly Twitter) in German and English. The dataset is categorized into two main groups: content produced by influential public figures and content generated by regular users in the form of replies. This dichotomy is fundamental to our data transformation and enrichment strategy, which involves three main phases: preliminary analysis, preprocessing and enrichment, and task-specific formatting. The preliminary analysis phase involves understanding the distribution and frequency of interactions within the dataset, identifying potential biases, and determining optimal sampling and filtering criteria. This phase is critical for ensuring the quality of the data used to train our machine learning models. The preprocessing and enrichment phase focuses on removing noise, such as external URLs and retweets, and enriching the data with semantic information, such as topic classifications and political affiliations. Finally, the task-specific formatting phase aligns the dataset with the machine learning objectives outlined in Deliverable D3.1 (Deep Learning Report), ensuring that the data is suitable for training generative models and other machine learning tasks.

The ML Dataset is closely connected to other work packages within the TWON project. In WP2 (Computational Modelling of Complex Social Networks), the dataset is used to inform the development of computational models that simulate the opinion dynamics and network mechanics of OSNs. In WP4 (TWON Implementation and Computation), the dataset is integrated into the TWON software platform,

[*] This report draws strongly on collaborative research projects. We would like to acknowledge the co-authors of these papers: Nils Schwager (UT) and Kai Kugler (UT). We would like to thank Alenka Guček (JSI) and Michael Heseltine (UvA) for reviewing the internal draft of the paper. The report has benefited considerably from the comments.

where it is used to run large-scale simulations and generate insights into the effects of various platform mechanics. Finally, in WP5 (Field Studies), the dataset is used to calibrate and evaluate the digital twin, ensuring that it accurately replicates real-world OSN dynamics.

All implementation code and associated pipelines are publicly available in an open-access repository (`https://github.com/cl-trier/TWON-Agents`) under the Apache 2.0 license, ensuring reproducibility and transparency of our methods. While the repository contains all necessary computational steps and procedures, the raw and processed datasets derived from $\mathbb{X}$ are currently excluded pending resolution of copyright and data usage restrictions. Once these legal considerations are addressed, we plan to integrate the complete dataset into the repository. All data processing steps are conducted using the Python packages Pandas (pandas development team, 2020) and Seaborn (Waskom, 2021).

2 Preliminary Analysis

Initially, we conduct a comprehensive distribution and frequency analysis to determine optimal sampling and filtering criteria for our machine learning tasks. Our approach prioritizes data quality over quantity to ensure the development of high-performing generative models. The raw subsets contain $154,834$ (German | Posts), $3,226,277$ (German | Replies), $932,755$, (English | Posts) and $17,803,216$ (English | Replies). The linked notebook (`https://github.com/cl-trier/TWON-Agents/blob/master/pipeline/analysis/00--Preliminary-Analysis.ipynb`) contains the below described analysis.

Frequency Distribution Analysis of the replying dataset reveals a highly skewed distribution of responses per user across both languages. The majority of users contributed only a single reply, making them unsuitable for our modeling process due to insufficient data points per individual.

lang	subset	mean	std	min	25%	50%	75%	max
English	Posts	97.98	23.90	3.00	94.00	104.00	112.00	138.00
	Replies	2.70	2.61	1.00	1.00	2.00	3.00	31.00
German	Posts	145.56	236.57	1.00	18.50	72.00	163.00	2111.00
	Replies	2.25	3.18	1.00	1.00	1.00	2.00	147.00

Table 1: User interaction (post, reply) frequency

Textual Quality Analysis of word distribution and sample lengths reveals a significant proportion of retweeted content and extremely short responses that lack discourse-relevant information.

lang	subset	mean	std	min	25%	50%	75%	max
English	Posts	193.43	84.13	1.00	127.00	208.00	271.00	684.00
	Replies	140.25	96.00	1.00	59.00	116.00	221.00	982.00
German	Posts	164.37	78.84	1.00	125.00	140.00	235.00	885.00
	Replies	135.27	86.78	4.00	62.00	112.00	199.00	616.00

Table 2: User Tweet length (after removing urls and mentions)

External References or Images Furthermore, a substantial portion of the samples contained links to external sources and embedded images. Although these elements provided discourse-relevant information, we decided to focus exclusively on text-based discourse and postpone the retrieval of external knowledge for the initial iteration of TWON.

lang	English		German	
subset	Posts	Replies	Posts	Replies
contains url	0.70	0.56	0.34	0.18

Table 3: Proportion of samples that contain external urls

Topic Classification To ensure comprehensive analysis of political discourses, we utilize a BERT-based fixed-class prediction framework (Barbieri et al., 2020). Our findings reveal that news and social concerns consistently demonstrate the highest predicted values by a significant margin. Nevertheless, each dataset contains minor proportions of alternative classifications, with similar distributions observed across both data subsets and language categories.

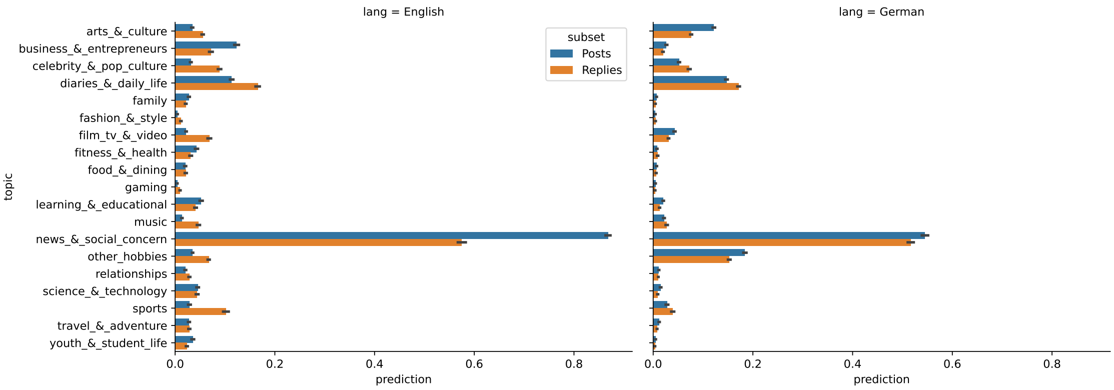

Figure 1: Predicted topics using the TweetEval (Barbieri et al., 2020) classification system

3 Preprocessing

Based on the insights gained during our preliminary anaylsis, we apply significant content filtering upfront to enhance the quality of the selected data. We acknowledge that the following steps reduce ”realism” of the data and alter the original proportions. However, since we lack information on the representativeness of the raw samples, we assume these steps, while potentially increasing skew, ultimately improve performance for our defined tasks.

Remove external source We remove samples that include URLs to external sources. While this significantly reduces the number of samples, our models cannot access this information during the training and inference process, which would marginally lessen the quality of the generated content. We propose addressing this issue with content retrieval-based systems; however, this lies outside the current TWON scope.

Remove retweets Tweets starting with *”RT”* mark a shared posting. The author does not add new content but replicates the original text and shares it with their followers. We exclude these samples as we cannot easily verify if the shared content is in line with the retweeting author’s opinion. This decreases the quality in later stages when we use the dataset to model the behavior of individuals rather than ideology groups.

Remove reaction (public figures) Similar to Retweets, we exclude reactions inside the post dataset as we miss the link between the delegate’s reaction to the original post reacted to. Our goal is to have a post-dataset that initializes conversation without external or preceding measurable input.

Filter by content length We aim to focus on content that contains arguments or opinions, particularly given the later applied discourse metrics. Thus, we remove samples with less than 32 characters based on our heuristic assumption that a specified length is necessary for conveying an opinion.

Sample most active users After the initial preprocessing, we further reduce both datasets to retrieve only the most active users measured by the number of posts and replies. This allows modeling both tasks on a user-based level by providing user examples (few-shot) during prompting and modeling them based on their interaction timeline. We restrict the selection of posts to users who wrote more than 15 posts and the selection of comments to users who reacted more than 25 times.

Joining discourse items Finally, we join both dataset groups (posts and replies) into a conversation format, where every reply is matched with its corresponding post. This yields a discourse structure and ensures that our replying agents have access to the original stimulus.

4 Enrichment

Matching Politicians with Partys We retrieve data from the Deutscher Bundestag and US Congress to incorporate information about politicians' party affiliations. This enhancement enriched our post generation task at an individual level. The dataset also enabled us to analyze Twitter metrics — likes, retweets, and replies — by political party (for German politicians only). Our findings support existing research (Serrano et al., 2019) suggesting that the Alternative für Deutschland (AfD) achieves significant success on Twitter based on engagement metrics. However, our data revealed a high standard deviation, indicating that our sample may not be fully representative of the broader political landscape.

	likes				retweets				replies			
	min	max	mean	std	min	max	mean	std	min	max	mean	std
AfD	7	12771	1442.0	1715.4	0	1856	199.9	256.2	0	2026	158.7	202.1
CDU/CSU	0	16133	619.7	1470.4	0	7692	152.1	639.0	0	4018	133.0	346.5
SPD	1	9519	472.0	1153.9	0	5951	131.5	557.8	0	4314	122.3	319.6
FDP	3	8035	355.3	633.2	0	761	33.7	58.7	0	4423	202.5	420.5
Linke	0	7295	318.9	824.9	0	1172	40.5	120.1	0	2227	46.6	143.5
Grüne	0	15699	318.2	952.6	0	1663	30.3	94.2	0	6349	161.9	481.3
BSW	2	6074	210.5	489.1	0	823	21.4	71.3	0	1304	10.7	68.6

Table 4: Distribution of engagement metrics grouped by political party (Germany)

Topic labeling We perform unrestricted topic labeling to both replies and posts on a prompt-based level (Zheng et al., 2023) utilizing `Llama3.1:70b-instruct-q6_K` (Dubey et al., 2024). These annotations serve during the instruction training as the context provided to the language model.

5 Task-specific Formatting

We implement the following formatting steps in memory during training to a) create a dataset that follows an instruction-learnable format and b) augment enrichment for perspectives that may be missing during data collection.

Posting/Replying The post structuring component transforms the preprocessed and sampled dataset into a structured dialogue format. Each entry is contextualized with relevant metadata, including author characteristics and political affiliation. The interaction sequence generation component creates training sequences that capture the dynamics of social media interactions. By incorporating multiple examples of exchanges between users, this approach facilitates the study of response patterns and conversation dynamics.

Reply Decision As we only collected pairs of posts and replies where a user wrote an actual response, we lack information about content users might have seen but chosen not to respond to. Therefore, for this task, we assume that every non-matching sample in our dataset indicates a post the user has seen but decided not to engage with. Consequently, we augment our dataset with these negative samples (posts not replied to) and generate a balanced dataset to learn how users decide which content to respond to among the posts they've seen.

6 Outcome

The ML Dataset described in this report is a cornerstone of the TWON project, providing the empirical data needed to understand, measure, and simulate the impact of OSN mechanics on democratic debates. By delivering a high-quality, enriched, and task-specific dataset, this report supports the project's objectives and contributes to the development of a robust digital twin of OSNs. The outcomes of this work are two enriched datasets containing 7088 (EN) and 5219 (DE) rows including the columns: id_{post, reply, author_post, author_reply}, author_post_{first_name, last_name, party}, text_{post, reply}, topics_{post, reply}. We utilize the processed dataset to align the posting and replying behavior of our agents through supervised fine-tuning. Thus, the datasets build the foundation for a data-driven alignment procedure for replicating OSN user behavior. We discuss training procedures and results in the Deliverable D3.1 (Deep Learning Report).

7 Possible Improvements

News and External Sources Our analysis revealed that a significant proportion of the samples include links to external sources. During this iteration, we omitted external content; however, we recognize the need to incorporate these sources as they reflect a fundamental aspect of social media communication. For future iterations, we plan to provide our agents with tools that allow the retrieval of external content or augment the generation process with a predefined vector database including scraped news articles selected based on the links in our samples. We plan to utilize news data from D4.1.1 (Initial Curated Data Set) that was collected through Event Registry and matches references in the posting and replying dataset.

Topic Labeling The performance heavily depends on the topics extracted in during the unrestricted prompt-based topic labeling task. Based on a superficial analysis of the results, we see the main problem in the heterogeneous naming of topics. The model does alter between German and English terms *(Rechtsextremismus v. Right-Wing Extremism)* and utilizes different levels of granularity as descriptions *(Politik, Vertrauen, Deutschland v. Frauen in Führungspositionen, LKW Maut Steuererhöhung, Kritik an älterer Generation)*. While improving the prompt technique can address these issues, we assume that human preprocessing is necessary to improve the quality significantly.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020. 6, 10

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 12

The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`. 9

Juan Carlos Medina Serrano, Morteza Shahrezaye, Orestis Papakyriakopoulos, and Simon Hegelich. The rise of germany's afd: A social media analysis. In *Proceedings of the 10th international conference on social media and society*, pages 214–223, 2019. 12

Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL `https://doi.org/10.21105/joss.03021`. 9

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 12

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