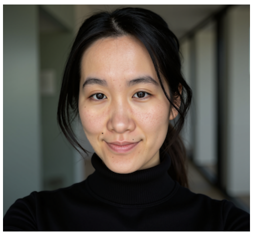
Xixuan Zhang



Berlin, Germany xixuan.zh@gmail.com +49-(0)15225127120
linkedin.com/in/xixuan-zhang xixuanzhang2022.github.io

Profile Summary

Data Scientist (NLP/LLMs) designs data-driven analyses and NLP pipelines—from collecting and curating large text corpora through feature engineering, modeling, and evaluation to clear delivery in dashboards and reports for diverse stakeholders.

Education

Freie Universität Berlin– Berlin, Germany

- Ph.D. Candidate, Computational Social Science (expected 2025) **2020 – Present**
- M.A., Media and Political Communication **2016 – 2019**
- B.A., Media and Communication Studies **2012 – 2016**

Technische Universität Berlin – Berlin, Germany

- M.Sc., Computer Science (coursework completed) **2022 – 2025**
- B.Sc., Media Informatics **2017 – 2022**

Professional Experience

Data Scientist & Research Associate **March 2022 – Oct 2025**

Freie Universität Berlin, Berlin, Germany

Built end-to-end data workflows for the research project *"NEOVEX,"* including data collection, curation, and analysis leveraging NLP methods and LLMs on conspiracy theory-related content; results published in top Q1 journals.

- Co-developed a graph-based dictionary expansion algorithm, boosting domain-specific keyword coverage for data collection by 30%.
- Built a 32M-text corpus from multiple platforms over 11 years through scraping/API; fine-tuned BERT models to detect conspiracy theory-related content, raising average F1 from 61 % to 77%.
- Developed a fine-grained narrative extraction pipeline, mapped cross-country convergence of 6,402 narratives from 123k Reddit posts with time-series analysis, and surfaced high-correlation conspiratorial belief signals—transferable to trend detection and brand/risk monitoring.

Research Associate **May 2019 – Sept 2022**

Weizenbaum Institute, Berlin, Germany

Conducted data-driven research using ML, NLP, and network analysis in the research group *"News, Campaigns, and the Rationality of Public Discourse"*; results published in top Q1 journals.

- Reconstructed dynamic diffusion graphs from 237k retweets (51.8k actors) plus 6.57M follow edges; segmented users with community detection (7 communities) for influence mapping.
- Built cascade analytics at scale—size, depth, max-breadth, structural virality and time-sliced growth —for 18,908 seed tweets, with integrated content classification for campaign monitoring.
- Implemented exposure-based triggers and Granger causality across communities to quantify cross-segment spillovers—actionable to product diffusion, referral funnels, and risk monitoring.

Other projects (selected)

Discourse Cohesion on Reddit | Python **Feb 2025 – Sept 2025**

- Built structure-aware embeddings for 231k comments/36k threads (r/climate) with GloVe + graph-weighted co-occurrence; integrated 7 discourse features; quantified cohesion (cosine, entropy) and modeled cross-positional engagement (logistic/OLS).
- Result: climate discourse is topic-driven and cohesive; politeness, continuity, on-topic exchange predict higher disagreement—actionable for moderation and conversation-quality KPIs.

Explainable AI: Unsupervised Concept Attribution of CNNs | Python **April 2024 – Sept 2024**

- Built unsupervised concept discovery from SimCLR last-layer activations; quantified importance (PCA, modified TF-IDF, Shapley) and benchmarked across linear/RF/XGBoost with a custom evaluation set.
- 81% concept-attribution fidelity in label-free settings; improved explainability of vision models.

Tracking Climate Change Revisions in Wikipedia | Python & R **Jan 2022 – July 2024**

- Parsed 930k sentence-level revisions from 891 climate articles; engineered relevance/timing/role features; built a 4-class taxonomy + labeling pipeline (13k labels via active learning); fine-tuned BERT models (+30% avg F1).
- Estimated revision hazards with a shared-frailty survival model + meta-analysis to surface drivers and optimize monitoring—analogous to churn/retention risk and incident triage.

Elastic Autoscaling for Real-Time Microservices | GKE/Kubernetes **April 2023 – Sept 2023**

- Decomposed a game mediator into role-based microservices, prototyped a serverless variant, staged K8s → GKE, and built an OpenAPI + k6 load-testing suite for six map/player scenarios.
- Tuning (autoscalers vs. EPMA) cut provisioning overhead −48%, scale-up 28→11 s, p99 610→390 ms, and cost/1k requests −29% under fluctuating load.

Skills

- Coding: Python (Pandas, NumPy, PyTorch, LangChain, SpaCy, Scikit-learn, Matplotlib), SQL, Java
- Data & Cloud: Google Cloud Platform (GCP), Kubernetes, Docker, HPC, Spark, Hadoop
- Analytics & BI: A/B testing, causal inference, forecasting, Tableau, Looker, Power BI
- Language: English (C1), German (C1), Chinese (Native)

Peer-reviewed journal publications (Selected)

- Zhang, X. (2025). Decoding revision mechanisms in Wikipedia: Collaboration, moderation, and collectivities. *New Media & Society*. https://doi.org/10.1177/14614448251336418
- Buehling, K., Zhang, X., & Heft, A. (2025). Veiled conspiracism. Particularities and convergence in styles and functions of conspiracy-related communication across digital platforms. *New Media & Society*. https://doi.org/10.1177/1461444825131575
- Schindler, J., Jha, S., Zhang, X., Buehling, K., Heft, A., & Barahona, M. (2025). LGDE: Local Graph-based Dictionary Expansion. *Computational Linguistics*, 1-32. https://doi.org/10.1162/coli_a_00562
- Zhang, X. (2023). Diffusion Dynamics and Digital Movement: the Emergence and Proliferation of the German-speaking #FridaysForFuture Network on Twitter. *Social Movement Studies*. http://doi.org/10.1080/14742837.2023.2211015