Ewelina Gajewska

NLP Researcher | Data Scientist | AI Fairness
Warsaw, Poland | LinkedIn | GoggleScholar | Hugging Face

SUMMARY

NLP Researcher and Data Scientist specializing in **AI fairness, hate speech detection, and computational social science**. Currently pursuing a PhD in Information and Communication Technology at the Warsaw University of Technology. Expertise in building identity-aware AI and multi-agent LLM systems to tackle implicit toxicity, mitigate algorithmic bias, and foster trustworthy digital ecosystems. Proven track record of developing human-centric and identity-aware language models to identify and counter harmful discourse while maintaining equitable performance across marginalized communities. Passionate about leveraging computational social science and ethical ML to build safer, fairer, and more inclusive AI products.

PROFESSIONAL EXPERIENCE

Warsaw University of Technology | Warsaw, Poland *Proje*ct *Leader* | 2026 – Present

- **Decoding Persuasion Project (2026–, Polish National Science Centre):** The aim of this project is to systematically decode how artificial persuasion operates relative to human persuasive tactics. It integrates corpus studies, computational modelling, and behavioural evaluation to establish reliable methods for identifying and assessing persuasive intent in emerging technologies.

Warsaw University of Technology | Warsaw, Poland *Doctoral Researcher* | 2023 – Present

- **AI Fairness & Hate Speech Research:** Pioneering research on algorithmic fairness by developing identity-aware LLMs to improve human-centric hate speech detection. Successfully reduced out-group under-detection biases against marginalized communities.
- **Multi-Agent Moderation Systems:** Designed multi-agent LLM architectures to detect implicit hate speech and minimize algorithmic hallucination by grounding AI context in cultural realities.

Warsaw University of Technology | Warsaw, Poland *Research Assistant: NLP and Data Science* | 2022 – 2026

- **iTRUST Project (2023–2026, Polish National Science Centre & CHIST-ERA):** Leading ML interventions against societal polarization. Developing debiasing techniques for transformer models to diagnose and mitigate harmful online discourse.

- **DeLab (Deliberation Laboratory) Project (2022–2023, VolkswagenStiftung):** Developed NLP pipelines to analyze online public discourse, employing sentiment analysis and emotion detection algorithms to map factual disagreement.
- **Con2Con (From Controversy to Consensus) Project (2022, CyberiADa-2):** Researched online conflict dynamics, leveraging transformer-based models to extract and analyze argumentative structures in text.

Adam Mickiewicz University | Poznan, Poland *Research Assistant: NLP* | 2021 – 2023

- **ComPathos Project (2021-2023, Polish National Science Centre):** Contributed to the computational modeling of pathos in natural language. Developed tools for extracting emotion-eliciting words and mapping argument schemes in large textual datasets.

EDUCATION

Warsaw University of Technology | Warsaw, Poland *Ph.D. in Information and Communication Technology* | 2023 – Present
- **Research Area:** Algorithmic Fairness in NLP, Implicit Hate Speech Detection, and Mitigating Algorithmic Bias.

Adam Mickiewicz University | Poznan, Poland *Master's Degree in Cognitive Science* | 2021 – 2023
- **Research Area:** Developing Human-Centred Models for Improved Recognition of Emotions in Text.
- *Awards*: Headmaster's Award for Scientific Activity.

Adam Mickiewicz University | Poznan, Poland *Bachelor's Degree in Cognitive Science* | 2018 – 2021
- *Awards:* Headmaster's Scholarship.

RECENT PUBLICATIONS

- **Gajewska, E.**, et al. (2026). *Algorithmic Fairness in NLP: Persona-Infused LLMs for Human-Centric Hate Speech Detection.* 59th Hawaii International Conference on System Sciences (HICSS-59). https://hdl.handle.net/10125/112188
- **Gajewska, E.**, et al. (2026). *Improving Implicit Hate Speech Detection via a Community-Driven Multi-Agent Framework.* Proceedings of the 18th International Conference on Agents and Artificial Intelligence. http://doi.org/10.5220/0014434500004052
- Malvicini, S., **Gajewska**, **E**., Derbent, A., Budzynska, K., Chudziak, J., & Martinez-Echevarria, M. (2026). *A Natural Language Agentic Approach to Study Affective Polarization*. Proceedings of the 18th International Conference on Agents and Artificial

Intelligence. https://doi.org/10.5220/0014309200004052

- **Gajewska**, **E**. (2026). *An interpretable method of political bias detection in news media through entity framing analysis*. Journal of Computational Social Science 9, 43. https://doi.org/10.1007/s42001-026-00474-3
- **Gajewska, E.**, et al. (2025). *Leveraging a Multi-Agent LLM-Based System to Educate Teachers in Hate Incidents Management.* 26th Conference on Artificial Intelligence in Education (AIED). https://doi.org/10.1007/978-3-031-98462-4_42
- Konat, B., **Gajewska, E.,** & Rossa, W. (2024). *Pathos in natural language argumentation: Emotional appeals and reactions*. Argumentation, 38(3), 369-403**.** http://dx.doi.org/10.1007/s10503-024-09631-2
- **Gajewska, E**. (2023). *eevvgg at SemEval-2023 Task 11: Offensive Language Classification with Rater-based Information.* 17th International Workshop on Semantic Evaluation (SemEval-2023). https://aclanthology.org/2023.semeval-1.24/

OPEN SOURCE & AI CONTRIBUTIONS

Hugging Face (eevvgg)

- **LLaMA-2 Fine-tuning:** Deployed specialized text-generation models (eevvgg/llama-2-7b-hvv) for computational social science applications.
- **Custom BERT/RoBERTa Models:** Engineered light-weight models for argument type classification, including ethos and pathos recognition, and multiple ad-hominem attack classifiers (eevvgg/AdHominem-ElecDeb60To16, eevvgg/Ad-attacks-online-hate-bert).
- **Political & Social NLP:** Developed bert-polish-sentiment-politics for analysis of emotions in Polish political discourse (eevvgg/bert-polish-sentiment-politics).

TECHNICAL SKILLS

- **NLP & Large Language Models:** Multi-Agent LLM Frameworks, Toxicity Moderation, Argument Mining, Fine-tuning (LLaMA-2, BERT, RoBERTa), Prompt Engineering.
- **Programming & Frameworks:** Python, PyTorch, Hugging Face Transformers, Scikit-learn, Pandas, NumPy.
- **Research & Analytics:** Computational Social Science, Disaggregated Data Analytics, Statistical Analysis, Data Visualization.