Junyu Lu (卢俊宇)

Phone: (+86) 15841760236 · E-mail: dutljy@mail.dlut.edu.cn
Homepage: https://dut-lujunyu.github.io/

Personal Statement

My research interests lie in **Natural Language Processing**, **Text Mining**, and **AI Security**. Currently, I focus on detecting hate speech and hateful memes prevalent on social platforms by developing novel deep learning methods, specifically large language models, for semantic analysis. I have a particular interest in evaluating the model's ability to protect against toxic content in Chinese, and I have constructed two dataset resources, ***ToxiCN*** and ***ToxiCN MM***, from linguistic and psychological perspectives. These datasets have been widely adopted by over **600** researchers. Since 2022, I have published over **10** papers in leading conferences and journals in deep learning and natural language processing, including *NeurIPS*, *ACL*, *COLING*, and *TASLP*. Additionally, I have participated in multiple national and international technical evaluation tasks, achieving results within the **top 5%**.
My academic dream is to conduct interesting and meaningful research for a better world!

Education

Dalian University of Technology, School of Computer Science and Technology 2023.09 - Now
- Ph.D. Student, Research Direction: *AI Security*, Supervisor: *Hongfei Lin*

Dalian University of Technology, School of Computer Science and Technology 2021.09 - 2023.06
- M.Sc. Candidate, Research Direction: *AI Security*, Supervisor: *Hongfei Lin*

Dalian University of Technology, School of Computer Science and Technology 2017.09 - 2021.06
- B.Sc, Guaranteed Postgraduate Admission (***Top 15%***)

Main Publications

1. **Lu J**, Xu B, Zhang X. Min C, Yang L*, and Lin H. "Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks." *The 61st Annual Meeting of the Association for Computational Linguistics.* **ACL 2023**. [Paper] [Code]
   - We focus on the fine-grained detection of Chinese toxic language. We analyze toxic types and expressions of toxic language from a linguistic perspective, introduce a fine-grained dataset **ToxiCN**, and propose a **Toxic Knowledge Enhancement** method to incorporate lexical features for detection.
   - Our proposed **ToxiCN** has been downloaded **over 500 times** and has been selected as the **sole Chinese data source** for the international evaluation task, *"CLEF 2024: Multilingual Text Detoxification"*.

2. **Lu J**, Xu B, Zhang X, Wang H, Zhu H, Zhang D, Yang L, and Lin H*. "Towards Comprehensive Detection of Chinese Harmful Memes." *The 38th Annual Conference on Neural Information Processing Systems.* **NeurIPS 2024.** [Paper] [Code]
   - We focus on the comprehensive detection of Chinese harmful memes, presenting a definition tailored to the local online environment from a psychological perspective. We construct the first Chinese harmful meme dataset **ToxiCN MM** and propose a **Multimodal Knowledge Enhancement** detector for detection.

3. **Lu J**, Lin H, Zhang X, Li Z, Zhang T, Zong L, Ma F, Xu B*. "Hate Speech Detection via Dual Contrastive Learning." *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* **TASLP**. [Paper]
   - We propose a **Dual Contrastive Learning** framework to mitigate the interference of insulting words in hate speech detection, jointly optimizing self-supervised and supervised contrastive losses to leverage both textual semantics and label signals to capture context information.

4. **Lu J**, Xu B, Zhang X. Liu K, Zhang D, Yang L, and Lin H*. "Take its Essence, Discard its Dross! Debiasing for Toxic Language Detection via Counterfactual Causal Effect." *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.* **LREC-COLING 2024**. [Paper] [Code]
   - We examine the useful and misleading effects of lexical bias on model decision-making from **a causal perspective**. We present a **Counterfactual Causal Debiasing Framework** to retain its useful effects and mitigate misleading effects, improving fairness while maintaining accuracy.

***Note**: * indicates the corresponding author. The full list of my publications is shown in Google Scholar.*

Technical Evaluation Task

CCAC 2024 Task1 Track2: "LLM Security in Few-shot Model Response Detection" (**2/30+**) &
CCAC 2024 Task1 Track1: "LLM Security in Few-shot User Query Detection" (**1/30+**) 2024.07

- Propose a Human Adversarial-based Instruction Fine-tuning Framework to enhance the safety performance of the large language model across two tracks.

NLPCC 2022 Task5 Track1: "Multi-label Classification for Scientific Literature" (**2/40+**) 2022.05

- Propose a Multi-task Hierarchical Cross-Attention Network to capture the dependencies among hierarchical literature labels and the correlation between labels and text. (*Published in NLPCC 2022, First Author*)

Project

Smart Campus System based on LLM, Main Participant. 2024.05

- Design an LLM tailored for campus scenarios, which provides intelligent Q&A services. Propose a retrieval-augmented contextual learning method to ensure that the model's output aligns with students' values.
- The project has obtained China Software Copyright, ID: 2024SR0616080.

Hate Speech Dataset Catalogue, Main Participant. [Repo]. 2022.03

- The most extensive open-source directory of hate speech datasets, covering basic information on **125 datasets** in **25 languages**, with **300+ stars** on GitHub.

Internship Experience

AI Center of China Mobile Research Institute, Research Intern 2022.07 - 2022.12

- Responsible for enhancing the content moderation module of the **"Jiutian" AI platform.** Retrain the Chinese pre-trained model to improve their detection capabilities.

Teaching Experience

Large Language Model Technology and Applications, Teaching Assistant 2024.09 - Now

- Teach the sections of "*Fundamentals of Large Language Models*" and "*Safety Computation of Large Language Models*", assign course projects and handle grading.

Python Programming Language Design, Teaching Assistant 2022.03 - Now

- Held Q&A sessions, graded exams, assignments, and course projects.

Academic Service

- **Youth Working Committee of Chinese Information Processing Society**, Student Committee Member
- **Conference PC Member**: ACL2024, EMNLP2024, COLING2024, BIBM2024, ICASSP2025
- **Journal Reviewer**: TASLP, TALLIP, Journal of Chinese Information Processing (co-reviewer)

Languages and Skills

- **Standard Language Test**: CET-6
- **Programming Language**: Python / C / C++ / Java / LaTeX
- **Technology**: Pytorch / NumPy / Pandas / Matplotlib / Git / Linux OS / Markdown
- **Other Expertise**: Good communication skills, love for teamwork.

Awards

National Scholarship (2023), Huawei Scholarship (2022), Xianglian Scholarship (2024), First Prize Scholarship for Doctoral Candidate (2024), First Prize Scholarship for Postgraduate (2022-2023), Excellent Postgraduate (2022-2023).