CASS ZHIXUE ZHAO

44 7516862694
zhixue.zhao@sheffield.ac.uk
S1 4BP Sheffield
Personal Site
Google Scholar

I am a lecturer in Natural Language Processing (NLP) at the Computer Science Department of the University of Sheffield. My long-term research goal is to enable **trustworthy, responsible, and efficient NLP models.** These days, I am interested in anything related to interpretability and large language models (LLMs). My recent research projects focus on model compression, model editing, and text-to-image models (with Toshiba EU).

Previously, I worked as a Postdoc researcher on explainable AI and responsible AI. The overarching aim is to demystify predictions made by black-box LLMs, making them easier to understand and trustworthy. The work also addresses model hallucination to ensure the reliability of LLMs, alongside exploring model compression techniques that mitigate compute demands and thus foster inclusivity within NLP research. Back in 2020, I worked as a research assistant within the same department, working on NIHR-funded NLP projects for systematic reviews of public health research. My Ph.D. research, which was funded by the University of Sheffield, looked at transfer learning and mitigating model bias for hate speech detection.

During my postdoctoral tenure, I published three first-author papers in top-tier NLP conferences: EMNLP2022, ACL2023, and NAACL2024. Two of the papers are oral presentations. Furthermore, I have a first-author paper accepted in the top NLP journal, TACL2024. I also published actively during my Ph.D., with publications in prestigious journals and conferences such as WWW and Online Social Networks and Media. Moreover, I have actively contributed to the academic community by serving on the Program Committee for major conferences including ACL and EMNLP, and by participating as a reviewer for conferences like NAACL, ACL ARR, and AAAI.

Education

01/2019~06/2023	**The University of Sheffield**, PhD, Information School (funded by the Uni)
Research Topic:	Transfer Learning for Hate Speech Classification, Bias in pre-trained language models
09/2017~09/2018	**The University of Sheffield**, M.Sc. Data Science, Distinction
Core Modules:	Data Analysis, Database Design, Data Mining and Visualization.
09/2008~07/2012	**Shanghai Institute of Technology**, B.S. Food Engineering
Core Module:	Advanced Mathematics, Linear Algebra, Microbiology, Chemistry

Publications

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing Explanation Faithfulness between Multilingual and Monolingual Fine-tuned Language Models. 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL 2024 Main (oral presentation)

George Chrysostomou, **Zhixue Zhao**, Miles Williams, and Nikolaos Aletras. 2024. Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization. 2024 Transactions of the Association for Computational Linguistics. TACL 2024 (accepted)

Paul Youssef, **Zhixue Zhao**, Jörg Schlötterer, and Christin Seifert. 2024. Detecting Edited Knowledge in Language Models. 2024 Transactions of the Association for Computational Linguistics. TACL 2024 (under review).

Zhixue Zhao and Nikolaos Aletras. 2023. Incorporating Attribution Importance for Improving Faithfulness Metrics. The 61st Annual Meeting of the Association for Computational Linguistics. ACL 2023 Main (oral presentation).

Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022. On the Impact of Temporal Concept Drift on Model Explanations. In Findings of the Association for Computational Linguistics: EMNLP 2022 Findings.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2022. Utilizing Subjectivity Level to Mitigate Identity Term Bias in Toxic Comments Classification. Online Social Networks and Media, 29, 100205.

Mark Clowes., Stansfield Claire, Thomas James, Shemilt Ian, Paisley, S., Mark Stevenson, **Zhixue Zhao**, Marshall Iain, Gregory Kell, June 2022. All is FAIR in health inequalities research: using machine learning to build a new database of health equity studies. European Association for Health Information and Libraries 2022.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification. In Companion Proceedings of the Web Conference 2021 (pp. 500-507)

Research & Teaching Experience

11/2023~Now Department of Computer Science, The University of Sheffield

Lecturer in Natural Language Processing

Delivering natural language processing modules at the undergraduate and graduate levels. Curriculum development, course design, and assessment methods. Supervising and mentoring undergraduate, graduate, and PhD students in research projects, theses, and career development. Research and grant applications.

Teaching modules:

- COM MSc Dissertation Project
- COM3110 Text Processing
- COM6911 Team Project

01/2022~11/2023 Department of Computer Science, The University of Sheffield
Postdoc Researcher Model Explanation
This research is part of a large interdisciplinary project, “Social Explainable Artificial Intelligence”, funded by the EPSRC. The overall goal of this project is to explore a wide range of research directions on responsible AI and model interpretability. I work independently on programming and, with supervision, writing papers. The main output is research publications.

07/2022~08/2022 English and American Studies, The University of Manchester
Research Assistant Nineteenth-Century Nature Writing in English and Twenty-First-Century Environmentalism
This pilot project aims to open out research questions for a larger project on women’s nature writing in the long nineteenth century. My duty focuses on pre-processing related text and exploratory research on the related corpus, such as temporal shift and geographic drift.

07/2021~12/2021 Department of Computer Science, The University of Sheffield
Research Assistant Automatically mapping and assessing inequalities in public health research
This was a multi-discipline project, collaborated with UCL and King's College, funded by the National Institute for Health Research. This project aims to develop automated methods to find, organize, and describe scientific literature relevant to public health and understand its findings concerning inequalities in health. Particularly, NLP techniques will be used to automatically organize these documents into topics, identify the research method used, and identify whether it contains information about factors related to health inequalities at scale. My role was for programming and analysis.

11/2020~03/2021 Department of Computer Science, The University of Sheffield
Research Assistant Text Processing for Health Technology Assessment
The TePHTA is a cross-discipline study with researchers from the Computer Science Department and the School of Health and Related Research, University of Sheffield. This project aims to improve healthcare decision-making through NLP methods. I participated in a part of this project, the medical system review analysis, which utilizes NLP models to assist healthcare experts in gaining insights from a massive volume of medical literature regarding employees' health situations and working conditions. I was in charge of programming and analysis.

12/2018~03/2021 Information School, The University of Sheffield, Sheffield
Teaching Assistant (for below modules)

- INF4002 Introduction to programming
- INF6027 Introduction to Data Science
- INF6060 Information retrieval
- INF6028 Data mining and visualization
- INF6032 Big data analytics
- INF6050 Database design
- INF6024 Researching social media
- INF109 Digital media and society

Professional Activities

2024~Now **Well-being & Neurodiversity Rep** COM ED&I Committee, The University of Sheffield
2023~2024 **Communication officer** Staff Race Equality Network, The University of Sheffield
2021~2021 **Organizer** NLP Reading Group, The University of Sheffield