NLP for Counterspeech against Hate and Misinformation (CSHAM)

Daniel Russo, Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie and Marco Guerini
Invited speakers: Cathy Buerger and Simone Fontana

Website: https://sites.google.com/view/nlp4csham

This tutorial aims to bring together research from different fields such as computer science and the social sciences and policy to show how counterspeech is currently used to tackle abuse and misinformation by individuals, activists and organisations, how Natural Language Processing (NLP) and Generation (NLG) can be applied to automate its production, and the implications of using large language models for this task. It will also address, but not be limited to, the questions of how to evaluate and measure the impacts of counterspeech, the importance of expert knowledge from civil society in the development of counterspeech datasets and taxonomies, and how to ensure fairness and mitigate the biases present in language models when generating counterspeech.

The tutorial will bring diverse multidisciplinary perspectives to safety research by including case studies from industry and public policy to share insights on the impact of counterspeech and social correction and the implications of applying NLP to critical real-world problems. It will also go deeper into the challenging task of tackling hate and misinformation together, which represents an open research question yet to be addressed in NLP but gaining attention as a stand alone topic.

---

(1) Presenters:

Daniel Russo, PhD Student at University of Trento and Fondazione Bruno Kessler, Italy.
Email: drusso@fbk.eu
Website: https://drusso98.github.io/
Daniel Russo is undertaking a PhD in the field of natural language generation at the University of Trento, Italy, in collaboration with Fondazione Bruno Kessler, under the supervision of Marco Guerini. Here, he is a member of the Language and Dialogue Technologies Group. He holds an MSc in Cognitive Science and a BSc in Computer Science. He co-organised the PoliticIT shared task at the EVALITA 2023 Italian Evaluation Campaign. His principal research interest is the automatic countering of online misinformation.

Helena Bonaldi, PhD Student at University of Trento and Fondazione Bruno Kessler, Italy.
Email: hbonaldi@fbk.eu
Website: https://helenabon.github.io/
Helena Bonaldi is a PhD student in the LanD group at Fondazione Bruno Kessler, under the supervision of Marco Guerini. Her research mainly focuses on the automatic generation of counterspeech against hate. Recently, she has started investigating the intersection of hate and misinformation in the context of the Hatedemics project. She has co-organised the Counterspeech for Online abuse (CS4OA) and the Multilingual Counterspeech Generation workshops.

Yi-Ling Chung, Senior research scientist, Genaios
Email: yilingchung27@gmail.com
Website: https://yilingchung.github.io/
Yi-Ling Chung is a Senior research scientist at Genaios. Her work addresses misinformation and online harms through fact-checking, abuse detection, and response generation, and investigates the impact of new AI techniques on online safety. She co-organised the Workshop

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 9–10
July 27, 2025 ©2025 Association for Computational Linguistics

on CounterSpeech for Online Harms, and the Workshop on Online Abuse and Harms (WOAH 7 and 8).

Gavin Abercrombie, Assistant Professor, Heriot-Watt University, Edinburgh, Scotland.
Email: g.abercrombie@hw.ac.uk
Website: https://gavinabercrombie.github.io
Gavin Abercrombie is an Assistant Professor in the Interaction Lab at Heriot-Watt University. His research focuses on socio-technical issues and human aspects of NLP. He is Co-Investigator on the EPSRC project Equally Safe Online, and is a founding organiser of both the workshops on Counterspeech for Online abuse (CS4OA) and Perspectivist Approaches to NLP (NLPerspectives).

Marco Guerini, head of the Language and Dialogue Technologies group at Fondazione Bruno Kessler (FBK), Italy.
Email: m.guerini@fbk.eu
Website: https://www.marcoguerini.eu
Marco Guerini is the head of the Language and Dialogue Technologies group at FBK. He works on NLP for persuasive communication, sentiment analysis and social media. In recent years his research has focused on the development of generative AI technologies to fight online hate and misinformation. He is the coordinator of EU funded projects, author of scientific publications in top-level conferences and international journals and organiser of workshops and shared tasks.

(2) Invited Speakers:

Cathy Buerger, Director of Research at the Dangerous Speech Project.
Email: cathy@dangerousspeech.org
Website: https://cathybuerger.com/
Dr. Cathy Buerger is the Director of Research at the Dangerous Speech Project where her work is dedicated to understanding and mitigating harmful speech and its role in inciting violence. She is a Research Affiliate of the University of Connecticut's Economic and Social Rights Research Group and Managing Editor of the Journal of Human Rights. She holds a PhD in Anthropology from the University of Connecticut.

Simone Fontana, journalist, managing editor of Facta.
Email: s.fontana@facta.news
Website: https://muckrack.com/simone-fontana1
Simone Fontana is a journalist based in Italy and managing editor of Facta. He focuses on disinformation, politics, extremism and online communities, but he also covered social and economic issues related to the environmental and climate crisis. His work has been published in Italy and abroad in publications such as La Repubblica, L'Espresso, Domani, Wired, Rolling Stone, Green European Journal and The Daily Dot.