

GeBNLP 2024

**The 5th Workshop on Gender Bias in Natural Language  
Processing**

**Proceedings of the Workshop**

August 16, 2024

©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
317 Sidney Baker St. S  
Suite 400 - 134  
Kerrville, TX 78028  
USA  
Tel: +1-855-225-1962  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 979-8-89176-137-7

## Message from the Organisation Committee

This volume contains the proceedings of the Fifth Workshop on Gender Bias in Natural Language Processing held in conjunction with the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). This year, the organizing committee underwent changes in membership, with Christine Basta and Marta R. Costa-jussà extending a warm welcome to Agnieszka Faleńska, Seraphina Goldfarb-Tarrant, and Debora Nozza as new co-organizers. We greatly appreciate the invaluable insights and expertise they contribute to our team.

This year, the workshop received 36 submissions of technical papers, of which 26 were accepted (20 long, 5 short, and one non-archival), for an acceptance rate of 72%. We are pleased to report a slight increase in submissions compared to the previous editions over the last three years. This year, we received 36 papers, compared to 33 in the last edition and around 19 in the three years before. Once more, we thank the Programme Committee members, who provided extremely valuable reviews in terms of technical content and bias statements, for the high-quality selection of research works. We want to extend our deep gratitude to the individuals who played pivotal roles in assisting us in conducting a highly successful workshop in person: Jasmijn Bastings, Agostina Calabrese, and Amanda Cercas Curry.

The accepted papers represent a broad spectrum of Natural Language Processing (NLP) research areas. They explore key NLP tasks, including language modeling and generation, machine translation, relation extraction, hate speech detection, fake news identification, sentiment analysis, and authorship profiling. Novel approaches to bias analysis and debiasing methods are introduced. Additionally, compelling studies are presented on underrepresented languages such as Turkish, Bangla, Hindi, and Norwegian. Several research studies have been conducted to study gender inclusivity in NLP, showing important developments in this area.

This year, the workshop featured a Shared Task on Machine Translation Gender Bias Evaluation with Multilingual Holistic Bias. This task allows for investigating the quality of Machine Translation systems in the specific cases of gender specification, gender robustness and unambiguous gender.

Finally, the workshop will feature two distinguished keynote speakers: Isabelle Augenstein, University of Copenhagen, and Hal Daumé III, University of Maryland and Microsoft Research NYC.

We are very pleased to keep the high interest that this workshop has generated over the last four editions, and we look forward to an enriching discussion on how to address gender bias in NLP when we meet in a hybrid event on August 16th, 2024!

August 2024

*Christine Basta, Marta R. Costa-jussà, Agnieszka Faleńska, Seraphina Goldfarb-Tarrant,  
Debora Nozza*

# Organizing Committee

## Program Chairs

Christine Basta, Alexandria University, Egypt

Marta Costa-jussà, FAIR, Meta

Agnieszka Faleńska, University of Stuttgart, Germany

Seraphina Goldfarb-Tarrant, Cohere

Debora Nozza, Bocconi University, Italy

# Program Committee

## Program Committee

Gavin Abercrombie, Heriot-Watt University  
Rupam Acharyya, State University of New York, Buffalo  
Bashar Alhafni, New York University  
Jasmijn Bastings, Google DeepMind  
Jenny Björklund, Uppsala University  
Ankani Chatteraj, NVIDIA  
Hongyu Chen, University of Stuttgart  
Amanda Cercas Curry, Bocconi University  
Hannah Devinney, Umea University  
Marco Gaido, Fondazione Bruno Kessler  
Matthias Gallé, Cohere  
Nizar Habash, New York University Abu Dhabi  
Lucy Havens, University of Edinburgh  
Wael Khreich, American University of Beirut  
Svetlana Kiritchenko, National Research Council Canada  
Tomasz Limisiewicz, Charles University Prague  
Ziqian Luo, Oracle  
Mercedes García Martínez, Uniphore  
Carla Perez-Almendros, Cardiff University  
Michael Roth, University of Stuttgart  
Rafal Rzepka, Hokkaido University  
Gerasimos Spanakis, Maastricht University  
Karolina Stanczak, Mila - Quebec Artificial Intelligence Institute and McGill University  
Masashi Takeshita, Hokkaido University, Tokyo Institute of Technology  
Samia Touileb, University of Bergen  
Sorouh Vosoughi, Dartmouth College  
Azmine Toushik Wasi, Shahjalal University of Science and Technology

# Keynote Talk

## Gender, Stereotypes, and Harms

Hal Daumé III

University of Maryland and Microsoft Research NYC

**Abstract:** Gender is expressed and performed in a plethora of ways in the world, and reflected in complex, interconnected ways in language. I'll discuss recent and ongoing work measuring how modern NLP models encode (some of) these expressions of gender, how those encoding reflect cultural stereotypes (and whose cultural stereotypes), and how that impacts people using these models. This will reflect joint work with a number of collaborators including students Haozhe An, Connor Baumler, Yang Trista Cao, Eve Fleisig, Amanda Liu, and Anna Sotnikova.

**Bio:** Hal Daumé is a Volpi-Cupal endowed Professor of Computer Science and Language Science at the University of Maryland, where he leads TRAILS, an NSF & NIST-funded institute on Trustworthy AI; he is also a Senior Principal Researcher at Microsoft Research NYC. His research focus is on developing natural language processing systems that interact naturally with people, promote their self-efficacy, while mitigating societal harms. Together with his students and colleagues, he has received several awards, including best paper at AACL 2022, ACL 2018, NAACL 2016, CEAS 2011 and ECML 2009, test of time award at ACL 2022 (and nomination at ACL 2017), and best demo at NeurIPS 2015. He has been program chair for ICML 2020 (together with Aarti Singh) and for NAACL 2013 (together with Katrin Kirchhoff), and he was an inaugural diversity and inclusion co-chair at NeurIPS 2018 (with Katherine Heller). When not sciencing and teaching, he spends most of his time climbing, yogaing, cooking, backpacking, skiing, and biking.

# Keynote Talk

## Quantifying societal biases towards entities

**Isabelle Augenstein**  
University of Copenhagen

**Abstract:** Language is known to be influenced by the gender of the speaker and the referent, a phenomenon that has received much attention in sociolinguistics. This can lead to harmful societal biases, such as gender bias, the tendency to make assumptions based on gender rather than objective factors. Moreover, these biases are then picked up on by language models and perpetuated to models for downstream NLP tasks. Most research on quantifying these biases emerging in text and in language models has used artificial probing templates imposing fixed sentence constructions, been conducted for English, and has ignored biases beyond gender including inter-sectional aspects ones. In our work, we by contrast focus on detecting biases towards specific entities, and adopt a cross-lingual inter-sectional approach. This allows for studying more complex interdependencies, such as the relationship between a politician’s origin and language of the analysed text, or relationships between gender and racial bias.

**Bio:** Isabelle Augenstein is a Professor at the University of Copenhagen, Department of Computer Science, where she heads the Copenhagen Natural Language Understanding research group as well as the Natural Language Processing section. Her main research interests are fair and accountable NLP, including challenges such as explainability, factuality and bias detection. Prior to starting a faculty position, she was a postdoctoral researcher at University College London, and before that a PhD student at the University of Sheffield. In October 2022, Isabelle Augenstein became Denmark’s youngest ever female full professor. She currently holds a prestigious ERC Starting Grant on ‘Explainable and Robust Automatic Fact Checking’, as well as the Danish equivalent of that, a DFF Sapere Aude Research Leader fellowship on ‘Learning to Explain Attitudes on Social Media’. She is a member of the Royal Danish Academy of Sciences and Letters, and President of SIGDAT, which organises the EMNLP conference series.

## Table of Contents

<i>A Parameter-Efficient Multi-Objective Approach to Mitigate Stereotypical Bias in Language Models</i> Yifan Wang and Vera Demberg .....	1
<i>Do PLMs and Annotators Share the Same Gender Bias? Definition, Dataset, and Framework of Contextualized Gender Bias</i> Shucheng Zhu, Bingjie Du, Jishun Zhao, Ying Liu and Pengyuan Liu .....	20
<i>We Don't Talk About That: Case Studies on Intersectional Analysis of Social Bias in Large Language Models</i> Hannah Devinney, Jenny Björklund and Henrik Björklund .....	33
<i>An Explainable Approach to Understanding Gender Stereotype Text</i> Manuela Nayantara Jeyaraj and Sarah Jane Delany .....	45
<i>A Fairness Analysis of Human and AI-Generated Student Reflection Summaries</i> Bhiman Kumar Baghel, Arun Balajee Lekshmi Narayanan and Michael Miller Yoder .....	60
<i>On Shortcuts and Biases: How Finetuned Language Models Distinguish Audience-Specific Instructions in Italian and English</i> Nicola Fanton and Michael Roth .....	78
<i>The power of Prompts: Evaluating and Mitigating Gender Bias in MT with LLMs</i> Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari and Maite Melero .	94
<i>Detecting Gender Discrimination on Actor Level Using Linguistic Discourse Analysis</i> Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann and Stephanie Thiemichen .....	140
<i>What Can Go Wrong in Authorship Profiling: Cross-Domain Analysis of Gender and Age Prediction</i> Hongyu Chen, Michael Roth and Agnieszka Falenska .....	150
<i>Towards Fairer NLP Models: Handling Gender Bias In Classification Tasks</i> Nasim Sobhani and Sarah Jane Delany .....	167
<i>Investigating Gender Bias in STEM Job Advertisements</i> Malika Dikshit, Houda Bouamor and Nizar Habash .....	179
<i>Dissecting Biases in Relation Extraction: A Cross-Dataset Analysis on People's Gender and Origin</i> Marco Antonio Stranisci, Pere-Lluís Huguet Cabot, Elisa Bassignana and Roberto Navigli . .	190
<i>Gender Bias in Turkish Word Embeddings: A Comprehensive Study of Syntax, Semantics and Morphology Across Domains</i> Duygu Altinok .....	203
<i>Disagreeable, Slovenly, Honest and Un-named Women? Investigating Gender Bias in English Educational Resources by Extending Existing Gender Bias Taxonomies</i> Haotian Zhu, Kexin Gao, Fei Xia and Mari Ostendorf .....	219
<i>Generating Gender Alternatives in Machine Translation</i> Sarthak Garg, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao and Matthias Paulik .....	237



<i>Beyond Binary Gender Labels: Revealing Gender Bias in LLMs through Gender-Neutral Name Predictions</i>	
Zhiwen You, HaeJin Lee, Shubhanshu Mishra, Sullam Jeoung, Apratim Mishra, Jinseok Kim and Jana Diesner . . . . .	255
<i>Is there Gender Bias in Dependency Parsing? Revisiting Women’s Syntactic Resilience"</i>	
Paul Stanley Go and Agnieszka Falenska . . . . .	269
<i>From ‘Showgirls’ to ‘Performers’: Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs</i>	
Marion Bartl and Susan Leavy . . . . .	280
<i>Sociodemographic Bias in Language Models: A Survey and Forward Path</i>	
Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson and Rebecca J. Passonneau . . . . .	295
<i>Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP</i>	
Vagrant Gautam, Arjun Subramonian, Anne Lauscher and Os Keyes . . . . .	323
<i>Evaluating Gender Bias in Multilingual Multimodal AI Models: Insights from an Indian Context</i>	
Kshitish Ghate, Arjun Choudhry and Vanya Bannihatti Kumar . . . . .	338
<i>Detecting and Mitigating LGBTQIA+ Bias in Large Norwegian Language Models</i>	
Selma Kristine Bergstrand and Björn Gambäck . . . . .	351
<i>Whose wife is it anyway? Assessing bias against same-gender relationships in machine translation</i>	
Ian Stewart and Rada Mihalcea . . . . .	365
<i>Analysis of Annotator Demographics in Sexism Detection</i>	
Narjes Tahaei and Sabine Bergler . . . . .	376
<i>An Empirical Study of Gendered Stereotypes in Emotional Attributes for Bangla in Multilingual Large Language Models</i>	
Jayanta Sadhu, Maneesha Rani Saha and Rifat Shahriyar . . . . .	384
<i>Overview of the Shared Task on Machine Translation Gender Bias Evaluation with Multilingual Holistic Bias</i>	
Marta R. Costa-jussà, Pierre Andrews, Christine Basta, Juan Ciro, Agnieszka Falenska, Seraphina Goldfarb-Tarrant, Rafael Mosquera, Debora Nozza and Eduardo Sánchez . . . . .	399