Fatma Elsafoury

Berlin
Germany
✉ *e.fatma.e@gmail.com*
efatmae.github.io

I'm currently a Post-doctoral at the Weizenbaum institute in Berlin. I'm funded by the Fraunhofer-FOKUS research institute. I submitted my PhD thesis at the university of the West of Scotland in July 2023. My research interests are in social science computing and ethics in AI. In particular, I focus in my research on hate speech detection, social bias, and fairness in natural language processing (NLP). I also organize the women_in_NLP talk series.

Research Experience

Sept 2023 **Post-Doctoral Researcher**, *Weizenbaum Institute*, Berlin, Germany.
Responsibilities: I joined the Data, Algorithmic Systems and Ethics research group. I'm working on investigating bias in multilingual and low-resourced-languages large language models as well as studying bias against under-represented identity groups from the Middle-East.

June 2022 – Sept 2022 **Research Intern**, *IBM research*, New York, US.
Responsibilities: I worked with the multi-language NLP team on measuring bias in and fairness in language models, and the effectiveness of different debiasing methods.

Sep 2019 – Feb 2022 **Research associate**, *Knowledge Transfer Partnership (KTP)*, Glasgow, UK.
Responsibilities: I built an online platform to detect hateful textual content.

Publications

2024 **Fatma Elsafoury**. "Systematic Offensive Stereotyping (SOS) Bias in Language Models". A long paper **arXiv preprint:https://arxiv.org/abs/2308.10684v1 .**

2024 **Fatma Elsafoury**, and Stamos Katsigiannis. "On Bias and Fairness in NLP: Investigating the Impact of Bias and Debiasing in Language Models on the Fairness of Toxicity Detection". A long paper **under-submission at the Computational Linguistics journal.**

2023 **Fatma Elsafoury**. "Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection". A long paper **Published at the Big Picture workshop at EMNLP 2023.**

2023 **Fatma Elsafoury**, Gavin Abercrombie. "On the Origins of Bias in NLP through the Lens of the Jim Code". A long paper **arXiv preprint arXiv:2305.09281, 2023.**

2022 **Fatma Elsafoury**, Steve Wilson, Stamos Katsigiannis, and Naeem Ramzan. "SOS: Systematic Offensive Stereotyping Bias in Word Embeddings". A long paper **Published at COLLING 2022.**

2022 **Fatma Elsafoury**, Steve Wilson, and Naeem Ramzan. "A Comparative Study on Word Embeddings in Social NLP Tasks". A long paper **published at the SocialNLP workshop at NAACL 2022.**

2022 **Fatma Elsafoury**. "Darkness can not drive out darkness: Investigating Bias in Hate Speech and Abuse Detection Models". A long paper **published at the Student Research Workshop (SRW) workshop at ACL 2022.**

2021 **Fatma Elsafoury**, Stamos Katsigiannis, Steve Wilson, and Naeem Ramzan. "Does BERT Pay Attention To Cyberbullying?". A short paper at **SIGIR 2021.**

2021 **Fatma Elsafoury**, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. "When the timeline meets the pipeline: A survey on automated cyberbullying detection". **Published in the IEEE-ACCESS Journal 2021.**

2020 **Fatma Elsafoury**. "Teargas, Water Cannons and Twitter: A case study on detecting protest repression events in Turkey 2013". Published in the **Text2Story** Workshop at **ECIR 2020.**

2017 Sarah Birch, and **Fatma Elsafoury**. "Fraud, Plot, or Collective Delusion? Social Media and Perceptions of Electoral Misconduct in the 2014 Scottish Independence Referendum". **Published in the Election Law Journal 2017**.

Public Outreach Articles

2024 **Fatma Elsafoury**. "AI discriminiation and Deepfakes". Published at the Federal Agency for Civic Education (German) https://www.bpb.de/lernen/bewegtbild-und-politische-bildung/556762/diskriminierung/

Talks

January 2025 **Guest Lecture**: *"Introduction to Bias and Fairness in AI, Natural Language Processing and Content Moderation"* at the **Introduction to research methods and computer science concepts course, Technische Universität, Berlin**

November 2024 **Invited Talk**: *"Sufis are Buddhists, and Amazighs are Native South Venezuelans: LLMs and the Arab world"* at the **Middle Eastern Studies Association**

October 2024 **Guest Lecture**: *"Data Ethics, NLP and toxicity detection"* at the **Data Ethics course, Technische Universität**

July 2024 **Invited Talk**: *"Systematic offensive stereotyping (SOS) bias in static and contextual word embeddings"* at the **MilanNLP research group, Bocconi University.**

March 2024 **conference Talk**: *"Darkness Cannot Drive Out Darkness: Investigating the Impact of Bias in NLP Models on Hate Speech Detection"* at the **NLP as a Method workshop, Weizenbaum Research Institute.**

March 2023 **Invited Talk**: *"Systematic offensive stereotyping (SOS) bias in static and contextual word embeddings"* at the **Bias in AI network meeting, Durham University.**

Dec 2022 **Invited Talk**: *"Bias and fairness in hate speech detection"* at the **SMASH group, University of Edinburgh.**

Nov 2022 **Invited Talk**: *"Bias and fairness in hate speech detection"* at the **Interaction Lab, Heriot-Watt University.**

Oct 2022 Conference Talk: *"SOS: Systematic Offenisve Stereotyping Bias in Word Embeddings"* at the **COLING main conference.**

Aug 2022 Seminar Talk: *"On bias and fairness in large language models"*, the **Exit talk at IBM Research.**

July 2022 Workshop Talk: *"Comparative study on word embeddings and social NLP tasks"* at the **SocialNLP workshop at NAACL**.

June 2022 Seminar Talk: *"Different biases in word embeddings"* at the **Language Models seminars at IBM Research**.

May 2022 Workshop Talk: *"Darkness can not drive out darkness: Investigating Bias in Hate Speech and Abuse Detection Models"* at the **SRW at ACL**.

April 2022 **Invited Talk**: *"Bias in NLP"* at the **DAAI seminar** at Birmingham City University.

July 2021 Conference Talk: *"Does BERT Pay Attention To Cyberbullying?"* at the **SIGIR 2021** conference.

June 2021 Seminar Talk: *"A true cyberbullying type lies in the eye of the beholder: A comparative study on detecting different types of cyberbullying using different word embeddings"* at the 73rd **Language Lunch at the University of Edinburgh**.

Nov 2020 Seminar Talk: *"Does BERT pay attention to attribution?"* at the 72nd **Language Lunch at the University of Edinburgh**.

Apr 2020 Workshop Talk: *"Teargas, Water Cannons and Twitter: A case study on detecting protest repression events in Turkey 2013"* at the **Text2Story** Workshop at **ECIR 2020**.

Teaching Experience

April 2019 – Aug 2019 **Lecturer (Part-time)**, *School of Computing*, Dundee University.
Responsibilities: I worked with Data Science MSc students where I helped students with their Programming and Machine Learning assignment and marking assignments.

Sept 2017 – Jan 2019 **Lab Assistant**, *School of Computing*, Glasgow University.
Responsibilities: I worked as Python and Alice lab tutor for undergraduate students and Java lab tutor for MSc students. I was involved in exam invigilation and Marking. I also supervised 2 MSc students.

Oct 2018 – Dec 2018 **Web development Tutor (Volunteer)**, *Code First Girls*, Glasgow University.
Responsibilities: I worked as a web development tutor for female students and employees at the University of Glasgow. I was responsible for teaching HTML, CSS, JavaScript, BootStrap, JQuery and GitHub.

July 2018 – July 2018 **Python Tutor (Volunteer)**, *PWS Africa*, Ibadaan University, Nigeria.
Responsibilities: I worked as a Python tutor for undergraduate and MSc students at the School of Math at Ibadan University. I was responsible for preparing course material, and teaching Python (Basics) and Data Science toolkits (Pandas, Numpy and MatplotLib).

Awards

2022 **Enrichment scheme (Community award)**, *Alan Turing Institute*, London, UK.
I received this award to work with the *Online Hate* project team and the causal inference study group to measure the causal inference on bias in word embeddings on hate speech detection models. I'm the only PhD student from the University of the West of Scotland who won this award.

2020 **Studentship award**, *ECIR*, Lisbon, Portugal.
I received this award for free registration to attend the conference and present my work.

2014 **Lord Kelvin Adam Smith (LKAS)**, *Glasgow University*, Glasgow, UK.
I was one of 11 who goth is scholarship in 2011. I was the only student from Egypt, Africa, and the Middle East to win this award to study in the UK.

2011 **ERASMUS MUNDS**, *Twente University*, Twente, The Netherlands.
I was one of only 3 Egyptian women who won this scholarship to study in the Netherlands.

Education

Nov 2019 – July 2023 **PhD student**, *Computer Science*, The University of the West of Scotland.
Working on social biases and their influence on Toxicity and Hate Speech Detection.

Dec 2019 **MSc by Research**, *Computer Science*, University of Glasgow.
Thesis Title: Detecting Protest Repression Incidents from Tweets.

Mar 2013 **MSc**, *GeoInformatics*, Twente University, Enschede, Netherlands.
Thesis Title: Monitoring Urban Traffic Status Using Twitter Messages.

May 2008 **BSc**, *Computer and information sciences*, Helwan University, Cairo, Egypt.
Thesis Title: Personal identification through iris recognition system.

Community Service

2025 Affinity Group co-officer at the EquiCL intiative at ACL research community.

2024 Organizing member of the Network of Arab Women in AI.

2024 Served on Served on the reviewing committee for the Workshop on Online Hate (WOAH) and the Workshop on Safety in Conversational AI.

2023 Served on the organizational committee for the international network on digital labour (INDL) conference 2023.

2023 Served on the reviewing committee for ACL 2023 for the Ethics in NLP track.

2022 Served on the reviewing committee for EMNLP 2022 for the Ethics in NLP track.

2021 – Present Organizer of the **Women_in_NLP** Talk Series. My role is to invite female researchers or practitioners in NLP, organize the event, announce it, and host it. Since starting, I have hosted speakers from Allen AI, Google, Microsoft Research, and others.

2021 – Present Volunteer at the **ACL Year Round Mentorship** which is a network that helps and supports junior researchers in natural language processing (NLP).

2020 – Present Volunteer at the Scottish Informatics and Computer Science Alliance **(SICSA) PhD Peer Support Group** where mental support is provided to PhD students in Scottish universities by fellow PhD students. My duties include co-hosting two sessions a month to allow the students to express themselves in a safe and friendly environment.

Jan 2022 – May 2022 Member of the **students' mental health advisory board at the University of the West of Scotland** where we meet monthly to discuss strategies to support the students' mental health.

2021 Volunteer at the **WiML workshop at NeurIPS conference** My role was to help participants in Gather Town virtual environment and to micro-blog about the posters and the talks presented.

2018 Volunteer at **Code First Girls** which is an initiative that aims at empowering women by teaching them basic web technologies.

2018 Volunteer at the **Programming Workshop For Scientists In Africa (PWS Africa)** which is an initiative to teach programming to students with fewer opportunities in African countries. The activities included teaching Python to undergraduate students at Ibadan University, Nigeria.

Professional Experience

Sep 2019 – Feb 2022 **Data Scientists (NLP)**, *Seric Systems*, Glasgow, UK.
Responsibilities: Responsible for developing a safeguard platform that uses machine learning to detect cyberbullying incidences in school collaborative platforms.

Dec 2013 – Aug 2014 **Software Developer**, *ESRI-NEA*, Cairo, Egypt.
Responsibilities: Responsible for developing web-mapping applications using Javascript and ArcGIS web.

May 2009 – Aug 2011 **Software Developer**, *ICON Technologies*, Cairo, Egypt.
Responsibilities: Responsible of developing web and desktop mapping applications using C#, ASP.NET and ArcGIS (desktop and web).

Languages

Arabic, *Native Speaker*.
English, *Fluent*.
German, *B1-level(Intermediate)*.