Fatma Elsafoury

Glasgow
United Kingdom
✉ *e.fatma.e@gmail.com*
efatmae.github.io

My research interests are in social science computing and AI ethics. In particular, I focus in my research on hate speech detection, social bias, and fairness in natural language processing.

Research Experience

Oct 2022 - Oct 2023 **Enrichment scheme (Community award)**, *Alan Turing Institute*, London, UK.
Responsibilities: I'll be working remotely 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.

June 2022 - Sept 2022 **Research Intern**, *IBM research*, New York, US.
Responsibilities: I worked on measuring bias,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: My KTP project was a collaboration between a business partner (Seric systems) and the University of the West of Scotland to build a platform for detecting hate speech in social media. My role as a research associate was to bridge the gap between academia and business by building a tool that meets the business needs based on the latest research in hate speech and machine learning. My tasks included a literature review, getting relevant datasets, developing a machine learning model, and building an online platform to easily use the trained models. The research project was successfully executed with all partners happy with the outcome.

Publications

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

2022 **Fatma Elsafoury**, Steve Wilson, and Naeem Ramzan. "A Comparative Study on Word Embeddings in Social NLP Tasks". A full 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 full 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**.

Talks

October 2022 *"Bias and fairness in hate speech detection"* at the **Interaction Lab, Heriot-Watt university.**

August 2022 *"On bias and fairness in large language models"* at the **Exit talk at IBM Research.**

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

June 2022 *"Different baises in word embeddings"* at the **Language Models seminars at IBM Research**.

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

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

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

June 2021 *"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 *"Does BERT pay attention to attribution?"* at the 72nd **Language Lunch at the University of Edinburgh**.

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

Education

Nov 2019 - Nov 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

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 **t**he WiML workshop at NeurIPS confernce 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 through 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 less opportunities in African countries. The activities included teaching Python to undergraduate students in Ibadan University, Nigeria.

Professional Experience

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

May 2009 – Aug 2011 **Software Developer**, *ICON Technologies*, Cairo, Egypt.
Responsibilities: Responsible of developing web and desktop mapping applications.

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 web development tutor for female students and employees in 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 Python tutor for undergraduate and MSc students at the school of Math in Ibadan university. I was responsible for preparing course material,teaching Python (Basics) and Data Science toolkit (Pandas, Numpy and MatplotLib).

Languages

Arabic, *Native Speaker*.

English, *Fluent*.

German, *A2-level(post-Beginner)*.