Fairness-aware Natural Language Processing

Keywords

Fairness, Machine Learning, Natural Language Processing

Location

Laboratoire I3S, Université Côte d'Azur, Sophia Antipolis

Supervisors

- Amaya Nogales Gómez, Associate Professor, 3IA Côte d'Azur, Université Côte d'Azur, nogales@i3s.unice.fr
- Serena Villata, Tenured Researcher, CNRS, villata@i3s.unice.fr

Duration

4 months

Target

Level: M1

To apply

Send CV, master and bachelor transcripts to email addresses above.

Description

In Natural Language Processing (NLP), a large variety of fairness-aware solutions have been proposed to handle the ethical and social risks of harm from Language Models [2]. A fundamental part for advancing the state-of-the-art is the empirical evaluation of new approaches in benchmark datasets that represent realistic bias and fairness settings [1]. The objective of this internship is to investigate real-world and synthetic datasets used for fairness-aware NLP. More specifically, it will involve the following tasks:

- Categorize the different real-life and synthetic datasets used in the NLP literature by their risk of harm.
- Analyse the relationships between the different attributes w.r.t the protected attribute.
- Provide a comparison based on the fairness metrics used in the ML literature using several benchmark approaches in NLP.

References

[1] T. Le Quy, A. Roy, V. Iosifidis, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning, 2021. arXiv 2110.00530.

[2] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L.-A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from language models, 2021. arXiv 2112.04359.