Recognizing and Reducing Bias in NLP Applications

Dirk Hovy Università Bocconi, Italia
dirk.hovy@unibocconi.it

Abstract

As NLP technology becomes used in ever more settings, it has ever more impact on the lives of people all around the world. As NLP practitioners, we have become increasingly aware that we have the responsibility to evaluate the effects of our research and prevent or at least mitigate harmful outcomes. This is true for academic researchers, government labs, and industry developers. However, without experience of how to recognize and engage with the many ethical conundrums in NLP, it is easy to become overwhelmed and remain inactive. One of the most central ethical issues in NLP is the impact of hidden biases that affect performance unevenly, and thereby disadvantage certain user groups.

This tutorial aims to empower NLP practitioners with the tools spot these biases, and a number of other common ethical pitfalls of our practice. We will cover both high-level strategies, as well as go through specific case sample exercises. This is a highly interactive workshop with room for debate and questions from the attendees. The workshop will cover the following broad topics:

- Biases: Understanding the different ways in which biases affect NLP data, models, and input representations, including including strategies to test for and reduce bias in all of them.
- Dual Use: Learning to anticipate how a system could be repurposed for harmful or negative purposes, rather than its intended goal.
- Privacy: Protecting the privacy of users both in corpus construction and model building.