UMEÅ UNIVERSITET

Gender and Representation: Investigations of Bias in Natural Language Processing

Hannah Devinney

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i MIT.A.121, byggnad MIT-huset,
tordagen den 18 April, kl. 13:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Dr. Christian Hardmeier,
Computer Science, IT University of Copenhagen, Danmark.

Department of Computing Science

Organization
Umeå University
Department of Computing Science

Document type
Doctoral thesis

Date of publication
27 Mars 2024

Author
Hannah Devinney

Title
Gender and Representation: Investigations of Bias in Natural Language Processing.

Abstract
Natural Language Processing (NLP) technologies are a part of our every day realities. They come in forms we can easily see as 'language technologies' (auto-correct, translation services, search results) as well as those that fly under our radar (social media algorithms, `suggested reading' recommendations on news sites, spam filters). NLP fuels many other tools under the Artificial Intelligence umbrella – such as algorithms approving for loan applications – which can have major material effects on our lives. As large language models like ChatGPT have become popularized, we are also increasingly exposed to machine-generated texts.

Machine Learning (ML) methods, which most modern NLP tools rely on, replicate patterns in their training data. Typically, these language data are generated by humans, and contain both overt and underlying patterns that we consider socially undesirable, comprising stereotypes and other reflections of human prejudice. Such patterns (often termed 'bias') are picked up and repeated, or even made more extreme, by ML systems. Thus, NLP technologies become a part of the linguistic landscapes in which we humans transmit stereotypes and act on our prejudices. They may participate in this transmission by, for example, translating nurses as women (and doctors as men) or systematically preferring to suggest promoting men over women. These technologies are tools in the construction of power asymmetries not only through the reinforcement of hegemony, but also through the distribution of material resources when they are included in decision-making processes such as screening job applications.

This thesis explores gendered biases, trans and nonbinary inclusion, and queer representation within NLP through a feminist and intersectional lens. Three key areas are investigated: the ways in which "gender" is theorized and operationalized by researchers investigating gender bias in NLP; gendered associations within datasets used for training language technologies; and the representation of queer (particularly trans and nonbinary) identities in the output of both low-level NLP models and large language models (LLMs).

The findings indicate that nonbinary people/genders are *erased* by both bias in NLP tools/datasets, and by research/ers attempting to address gender biases. Men and women are also held to cisheteronormative standards (and stereotypes), which is particularly problematic when considering the intersection of gender and sexuality. Although it is possible to mitigate some of these issues in particular circumstances, such as addressing erasure by adding more examples of nonbinary language to training data, the complex nature of the socio-technical landscape which NLP technologies are a part of means that simple fixes may not always be sufficient. Additionally, it is important that ways of measuring and mitigating 'bias' remain flexible, as our understandings of social categories, stereotypes and other undesirable norms, and `bias' itself will shift across contexts such as time and linguistic setting.

Keywords
NLP, natural language processing, gender bias, social impact of AI, gendered pronouns, neopronouns, gender studies, topic modeling

Language
English

ISBN
print: 978-91-8070-336-9
PDF: 978-91-8070-337-6

ISSN
0348-0542

Number of pages
173