2

Hannah Devinney
Gender and Representation: Investigations of Bias in Natural Language Processing
Umeå Universitet, 2024

Until quite recently, competent use of language seemed to require a human brain, and language technology, or natural language processing, was restricted to specialist use cases. This has changed fundamentally. In online chatbots and other tools, language technology has reached the masses. People trust it enough to substitute its "judgment" for their own in high-stakes situations such as examinations or job interviews, and many users readily accept a veneer of eloquence as evidence of human-like cognitive capabilities. However, it is increasingly evident that this technology does not treat all humans equally. Trained on large amounts of text, these systems easily pick up stereotypes and misrepresentations and reproduce and reinforce these biases in their output.

Hannah Devinney's PhD thesis, *Gender and Representation: Investigations of Bias in Natural Language Processing,* builds a bridge between practical language technology and a well-founded and inclusive understanding of human gender by drawing from both computer science and gender studies. It promotes the creation of computer-based tools with awareness of non-binary gender and narrows the theoretical gap between gender studies and language technology research.

Devinney first surveys how gender is operationalized in over 200 academic papers addressing gender bias in natural language processing published before 2021. Many of them rely on vector space embeddings, which represent words or concepts as points in an abstract space. In such a space, relations between concepts can be characterized geometrically based on direction and distance, and from a technical perspective, it is tempting to view gender as a dichotomous axis, along which one moves in a more "masculine" or "feminine" direction. Non-binary gender can then, at best, be located at the seemingly "neutral" halfway point between the two binary gender extremes.

To the extent that the surveyed papers even have an identifiable concept of gender, Devinney finds that a binary conceptualisation reigns supreme, but many papers are based on vague and ill-defined gender concepts. Authors often mention non-binary gender as a "limitation" to be addressed by unspecified future methods, paying lip service to gender inclusivity without actually following through in modelling. Devinney's detailed discussion of the many pitfalls in the definition

Citation: *Tidskrift för genusvetenskap* 45 (2-3): 198–200
http://dx.doi.org//10.55870/tgv.v45i2-3.61725

and operationalisation of gender concepts is instructive for researchers more familiar with the technical aspects of language technology than its sociocultural constraints.

Concluding this part of the thesis, Devinney first recommends making the theoretical assumptions and social objectives of one's work explicit to allow readers and peer reviewers to engage with them. Further, they advocate for consistent, respectful, and accurate language, for instance, by preferring the more neutral terms *masculine* and *feminine* to *male* and *female* in linguistic contexts. This may seem straightforward, but if the main problem is indeed a lack of awareness, this recommendation will need additional education to have an effect. Finally, Devinney advocates for the use of feminist research methods, emphasizing the importance of researcher positionality, interdisciplinary collaboration, and the inclusion of stakeholders. Implementing these well-founded recommendations in language technology research also requires raising methodological awareness; Devinney's work lays a good foundation for that.

The second part of the thesis introduces methods for large-scale corpus analysis. Devinney advocates for a mixed-methods approach to identifying manifestations of gender bias in text. The proposed EQUITBL method begins with topic modelling, a statistical method to identify topically related clusters of words in large corpora. The outcome of this process is then analysed qualitatively with reference to seed word lists representing gender categories. This combination of automatic, statistical methods with qualitative work enables efficient analysis of large datasets while still emphasizing human judgement over rigid and overgeneralising categories. This approach is very plausible and adds welcome nuance to quantitative results.

In the third and last part of the thesis, Devinney studies how neopronouns and non-normative gender identities are handled by large language models. The models tested (Llama for English and GPT-SW3 for Swedish) do respect some explicitly stated pronoun choices, but choosing pronouns changes the character of the generated stories and makes characters more likely to be described as *the Other.* Another worrying result comes from the analysis of model refusals – that is, outputs where the model declares a prompt inappropriate or offensive instead of generating a meaningful response. In Devinney's intersectional analysis of gender, faith and ethnic identities, certain combinations of identity traits produce disproportionately many refusals. The models systematically discriminate against those identities. But even socially dominant identities attract higher numbers of refusals when made explicit: mentioning identity is treated as problematic in itself. In light of recent political developments, it is reasonable to expect this behaviour to increase further in commercial products.

At the core of many of these problems is a trade-off between generalisation and individualisation. Generalisation – that is, making inferences about individuals based on observations of others – is what enables machine learning, but it conflicts with individual identities that defy stereotypical expectations. For instance, language technology frequently generalizes by assuming that there is a small, fixed number of pronouns; it fails when an individual uses other pronouns. It also assumes everyone uses the same pronouns throughout their lifetime, and fails when a person's preferred pronouns change over time. In each case, rigid assumptions enable stronger generalisation – but only at the cost of misrepresenting parts of the population.

Intersectionality further complicates the picture. Those computational models that even consider more than one aspect of a person's identity often remain limited to a still highly reductive two-dimensional framework, such as binary gender and a small number of ethnicities. Each additional dimension or category multiplies the number of intersections to be considered, thereby reducing the possibility of making generalisations from the available data. Devinney's work does not directly address this issue, but it highlights how even in the reductive case of only two dimensions, many consequential problems remain unsolved.

Language technology is becoming increasingly widely adopted, but it still deals poorly with the diversity of its users. Devinney warns against the naïve acceptance of the "folk model" of gender – which treats gender as binary, physically determined, and immutable. This model not only limits the systems themselves but also the research intended to improve them.

Hannah Devinney's PhD thesis is one of the first and few works to offer a credible and competent approach to bridging the gap between gender theory and natural language processing. As such, it represents an important milestone.

Christian Hardmeier
Associate Professor, Data Science Section
IT University of Copenhagen