Crime and Relationship: Exploring Gender Bias in NLP Corpora

Hannah Devinney
Umeå University

Jenny Björklund
Uppsala University

Henrik Björklund
Umeå University

Abstract

Gender bias in natural language processing (NLP) tools, deriving from implicit human bias embedded in language data, is an important and complicated problem on the road to fair algorithms. We leverage topic modeling to retrieve documents associated with particular gendered categories, and discuss how exploring these documents can inform our understanding of the corpora we may use to train NLP tools. This is a starting point for challenging the systemic power structures and producing a justice-focused approach to NLP.

1 Introduction

Machine learning models for Natural Language Processing (NLP) have been shown to reflect implicit human bias (Caliskan et al., 2017; Bolukbasi et al., 2016; Garg et al., 2018). By replicating and even amplifying these biases, NLP systems risk causing a variety of harms to groups and individuals based on their identities (Crawford, 2017).

Although definitions and measures of "bias" and "fairness" vary as Blodgett et al. (2020) discuss, an important source of these behaviors is the text data used to train such NLP models. Due to the size of these corpora, it is difficult to know what goes into a model: they are too large for humans to analyze in detail in order to discover potential patterns of misrepresentation and under-representation. Purely computational measures of bias are capable of processing this data; however, they are likely to miss context and nuance a human reader would not.

This work is a step towards tools which would allow us to combine the advantages of both computers and humans. By using computational methods to reveal words and ideas associated with different social groups and leaving the results to human interpretation, we can critically examine large amounts of text data with respect to power structures. Related work in this field includes Hoyle et al. (2019), who investigate gendered differences in descriptive words using unsupervised latent variable modeling and Dahllöf and Berglund (2019), who perform a gendered analysis of topic models trained on Swedish literary corpora, demonstrating how certain topics relate to gender.

We explore one option for exploring large text data sets by training semi-supervised topic models on three corpora representing different social contexts to investigate differences in how men, women, and nonbinary people are represented in the corpora. In section 2, we summarize our initial work and findings from (Devinney et al., 2020). In section 3, we use the trained topic models to retrieve documents from the corpora which are highly weighted with respect to a particular topic. We read these documents to verify our interpretations of the theme(s) associated with each topic and to demonstrate some of the advantages and potential pitfalls of topic modeling as an exploratory method. Section 5 discusses potential future directions and applications of this work.

2 Topic Model Experiments

Topic Modeling (TM) using Latent Dirichlet Allocation (Blei et al., 2003) is a statistical method for creating a generative model from a corpus of documents. The model has a number of *topics*, a mix of which is assumed to underly the corpus. Each topic is a probability distribution over the vocabulary of the corpus. Using semi-supervised TM allows us to seed certain topics with a number of words before training, essentially forcing them to be prominent within the topic. In this project, we used pSSLDA,[1] an implementation of the metod developed by Andrzejewski and Zhu (2009).

Our models each have 15 topics, three of which we made "gendered" by seeding them with "gen-

[1] https://github.com/davidandrzej/pSSLDA

dered words". (Masculine such as *man*, *he*, *male*; feminine such as *woman*, *she*, *female*; neutral/nonbinary such as *person*, *they*, *nonbinary*.) We used two versions of these lists, one containing only words that we consider to be purely definitional and one where we also include "relational" words such as *father*, *wife*, *partner*. We tried to make the English and Swedish lists as similar as possible given the differences in the languages, exemplified by the presence of the exclusively singular neutral pronoun *hen* in Swedish.

We trained models on three different corpora: mainstream news articles in English (ME) and in Swedish (MS), and one collected from LGBTQ+ publications, forums, and LGBTQ+ sections of mainstream publications (Queer English, QE). The last corpus was used because of the scarcity of representation of nonbinary people and themes in the mainstream corpora. We also trained unseeded (i.e. unsupervised) models for each corpus, which did not provide many conclusions relating to gender.

In our qualitative analysis of the seeded models, we looked at the 50 most heavily weighted words in each of the "gendered" topics. We found systematic differences in which words showed up across genders. The differences vary by context, but still broadly correspond to hegemonic ideas about gender roles.

In the feminine topics, we found a prevalence of words linked to home, family, and relationships. In general, they were also "narrower", i.e. more focused on specific themes.

The masculine topics, by contrast, were more "open", i.e. more varied in theme, seemingly reflecting the perception of masculinity as the norm, and as such neutral. These topics were also more directed towards the public sphere. In the MS corpus, there was also a connection to crime and punishment, while in the QE corpus, there was a theme linked to Christianity and death.

For the mainstream corpora, the nonbinary topics do not appear as coherently gendered, presumably due to lack of representation. Instead, like the masculine topics, these were very neutral in character. The corresponding topic for the QE corpus, however, can more honestly be called nonbinary. A coherent theme relating to "acceptance" emerges.

This leaves us with further questions: What are the "typically gendered" documents like? What are some of the advantages and drawbacks of topic modeling as an exploration method?

3 A Closer Reading of "Highly Gendered" Documents

In order to investigate what kinds of documents the gendered topics primarily correspond to, we calculated, for each document in each corpus, how likely it is to be generated by each of the three gendered topics for that corpus, i.e. $p(d|t)$. We then extracted the 50 top scoring documents for each topic and corpus.

We read the top 50 articles[2] for the gendered topics in each corpus, keeping notes on any patterns. For the masculine topics in the MS and QE corpora, we in particular looked at whether these articles could explain the associations we found to "crime and punishment" and "death and Christianity", respectively. Whenever percentages are mentioned they may not sum to 100%, as a single document can belong to more than one category.

3.1 Mainstream English

We initially found that both the masculine and neutral topics were similarly "neutral," but a closer reading of the actual articles reveals a difference in this neutrality. Only 6% of the articles linked to the neutral topic, are about a specific person. These articles tend to be advice columns or horoscopes (genres intended to apply to as many people as possible), although there are also articles discussing protests and other community-oriented events. Most of the articles in the masculine topic (86%) are about specific people.[3] It is possibly the most varied in terms of distinct themes, with stories about sports, crime, violence, winning the lottery, and travel. Violence and injury is the most common of these themes, with about half of the articles featuring stories of various men surviving stabbing attacks, being tried for murder and abuse, or dramatically rescuing women from precarious situations (fires[4] or domestic abuse).

Within the feminine topic, a very different kind of violence is represented. 38% of articles discuss violence in the form of abuse, usually in the context of relationships and often sexual or sexualized. In these articles, women are universally victimized, although many feature a story of "overcoming" this

[2]Where articles were not in the target language or were otherwise "broken" due to the scraper, they were left out. The English documents were reviewed by the first author and the Swedish ones by the last (native speakers).

[3]All human men, except one woman and one cat.

[4]Interestingly, the woman mentioned in the previous footnote saved *herself* and her child from a house fire.

adversity. 22% of the articles cover celebrity gossip, and 10% are advice columns. Most of the articles feature relationships prominently. One interesting exception to this pattern is a story covered by four articles: a woman lost and injured in the wilderness in Hawai'i survives until her rescue.

3.2 Mainstream Swedish

The top articles for the masculine topic in the Mainstream Swedish corpus verify the theme of "crime and punishment", present in 40% of the articles. Two thirds are about more or less well-known persons.[5] There are, however, also a substantial number of articles about relationships (26%), in particular grief for a lost family member (20%). Finally, stories about injury and illness are prevalent (20%).

For the feminine topic, 100% of the articles are about "celebrities", all female. Most of them (all but two) can be categorized as celebrity gossip. The main themes are relationships (73%) and injury/illness (16%). Again, among the relationship-related articles, grief is a prominent theme (23%). There are certain persons and families that feature in more than one article. For instance, 16% of the articles feature the daughters of Swedish singer Lill-Babs, 13% Swedish influencer Bianca Ingrosso, and 10% the Swedish royal family.

The articles for the neutral topic are indeed neutral. None of them are explicitly about a particular person. Rather, they report on events, property sales, and finance.

3.3 Queer English

The articles in the feminine topic in the QE corpus support our reading of this topic as centered around family and relationships (80%). These articles are primarily about people who have come out, usually to family, and many describe parent's process of coming to terms with their child's sexual orientation or gender. Many are first-person records of someone's own experience or are forum discussions. The other main themes are articles about fiction (18%, film recommendations and character descriptions) and "official" resources from organizations about LGBTQ+ inclusion (16%). A notable difference from the feminine topic in the Mainstream corpora is the comparative lack of violence and celebrity gossip, the former of which contains much less intimate partner violence.

Although both men and women in the QE corpus are subjected to homophobic violence, this violence is much more clearly associated with the masculine topic: only about 9% of the documents for this topic are *not* about violence and/or homophobia. Gay and bisexual men are blackmailed, assaulted, and murdered, often in pre-meditated acts involving the perpetrator seeking out victims specifically because of their sexual orientation. 26% of the articles involve death,[6] suggesting that the "death" theme we originally identified in this topic is more linked to homophobic violence (8 articles) and less to e.g. AIDS (1 article). The other theme we found, "Christianity", is present in 17% of articles, varying in sentiment. Some offer affirmation; others quote pastors equating queerness with sin.

Unlike the neutrality of the Mainstream corpora, the nonbinary topic in this corpus contains actual nonbinary representation, with about a third of the documents being specifically about nonbinary people. We use "nonbinary" as an umbrella term referring to anyone whose gender identity doesn't fit into a "binary" state of being a man or a woman, but it is important to note that these documents express a wide range of gender diversity, with people describing themselves as nonbinary, genderfluid, agender, genderqueer, bigender, and more. The dominant theme in this topic is coming out (72%), which supports our initial reading of the topic as one of "acceptance". Most of the remaining articles are about figuring out your own identity,[7] and seeking advice or offering personal experiences about navigating various situations (dating, jobs, friendship) as a queer person. Most of the "neutral" (as opposed to explicitly nonbinary) documents in this topic are about coming out, and many are about coming out as asexual or bisexual: orientations not attracted to one-and-only-one gender, which may be less likely to use gendered language. Of the documents about coming out, most are either individual posts offering advice about how to come out or forum threads of people seeking advice and support before coming out. Some offer advice for supporting newly-out friends and family.

4 Discussion

Most of our findings in these closer readings support our initial impressions of the topic "themes" based on highly-weighted tokens, but provide a

[5] All human men, except one woman, a UCF fighter, and a kangaroo named Ripped Roger, who boxes

[6] Two of these are about historical figures.

[7] Mostly from sites about asexuality and aromanticism.

much richer picture of the types of stories represented in each corpus and the patterns in language used to discuss people of different genders.

Starting documents may influence results. Topic modeling clusters documents. Documents "assigned" to a topic early in this process are likely to influence what other documents are included in that topic. For example, the horoscopes we observed in the ME neutral topic all list every astrological sign, meaning that these documents are certain to contain otherwise-rare words such as *capricorn* or *sagitarius* and therefore they will cluster together. That these documents tend to use very neutral language means they become over-represented in this topic. This non-deterministic behavior may actually be an advantage in exploration: although in our experiments we did not use a random seed for training, doing so and training several models could reveal other patterns.

Articles may not "belong" in the corpus. Our corpora are intended to be monolingual, but the scraping methods used to produce them may capture extraneous documents. We observe 15 of the top 50 articles in the ME masculine topic are written in Italian; and several of the QE feminine documents are in Spanish. This draws our attention to an issue we were unaware of in these corpora, giving us a starting point to find other, similar, documents which do not belong and weed them out.

Over-representation of people/stories. For the Mainstream corpora we found many articles containing celebrity gossip. The same persons tended to appear in multiple stories. The same stories also appeared multiple times, from different outlets, but often with very similar content. This can be due to plagiarism or just the fact that the writers quote from celebrities' social media. The risk here is that a relatively small number of people's activities and views have a disproportional effect on the tendencies of models trained on the corpora.

Variations in genre may also be informative. The QE corpus is composed of a variety of web content instead of news articles specifically, and we find that the genre of document varies between gendered topics as well. The masculine topic contains almost entirely "news" or other articles describing people in third person; the feminine topic is mostly "news" with some forum discussions and first-person articles about own experiences; and the nonbinary topic is predominantly first-person discussions and sharing of experiences. Similarly, advice columns show up in the ME feminine and neutral topics, but not the masculine one. It is impossible to tell whether this is due to the themes associated with each gendered topic or vice versa, but there is some interesting analysis to be done here with regard to which voices are heard. "More privileged" groups get their stories told, while less privileged groups must carve out their own spaces and tell their own stories.

5 Future Work

Having demonstrated topic modeling as an aid in corpus exploration, our goal is next to develop a tool to help NLP practitioners better understand the potential risks to their models by uncovering bias in the training data. Such a tool should does not necessarily have to use LDA and should be applicable to more aspects of identity than gender, and indeed cover more than one aspect at a time to enable intersectional analysis. As not all of these categories are so explicitly semantically encoded as gender is in English and Swedish (with relatively-equivalent words such as *person-woman-man*), it would be useful to be able to automatically generate seed words from exemplar texts - which may themselves be informative as to the characterization of groups.

We also intend to experiment with this technique for reducing stereotypes in corpora. That is, we would like to see what happens if we train the models, remove some documents based on this type of analysis of what is strongly gendered and stereotypical, and then retrain the models. It would also be informative to test if e.g. language models trained on the two copora (original and edited) behave differently on bias benchmark tasks.

Our goal is to produce corpora with better representation, as a harm reduction tactic distinct from conventional "debiasing" where a system is manipulated to improve its score with respect to some well-defined "bias" measure. We note that different applications may have different conditions for being "bias free" (e.g. one system may seek to reflect "world knowledge" of labor statistics, while others should focus on making sure an occupation is equally likely to be held by a person of any gender), and that being "unbiased" may not be the same as being equitable or just. We hope that our approach maintains enough flexibility to allow researchers to pursue the most appropriate definition.

References

David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet Allocation with topic-in-set knowledge. In *Semi-supervised Learning for Natural Language Processing*, pages 43–48.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *ArXiv*, abs/2005.14050.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kate Crawford. 2017. The trouble with bias. `https://www.youtube.com/watch?v=fMym_BKWQzk`. Keynote at NeurIPS.

Mats Dahllöf and Karl Berglund. 2019. Faces, Fights, and Families: topic modeling and gendered themes in two corpora of swedish prose fiction. In *Proceedings of the 4th Conference of The Association of Digital Humanities in the Nordic Countries*.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-supervised topic modeling for gender bias discovery in English and Swedish. Technical Report UMINF 20.12, Umeå University, Dept. Computing Sci. To appear in Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16):E3635–E3644.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716. Association for Computational Linguistics.