# Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish

**Hannah Devinney**
Dept. Computing Sci.
Umeå Centre for Gender Studies
Umeå University
hannahd@cs.umu.se

**Jenny Björklund**
Centre for Gender Research
Uppsala University
jenny.bjorklund@gender.uu.se

**Henrik Björklund**
Dept. Computing Sci.
Umeå University
henrikb@cs.umu.se

## Abstract

Gender bias has been identified in many models for Natural Language Processing, stemming from implicit biases in the text corpora used to train the models. Such corpora are too large to closely analyze for biased or stereotypical content. Thus, we argue for a combination of quantitative and qualitative methods, where the quantitative part produces a view of the data of a size suitable for qualitative analysis. We investigate the usefulness of semi-supervised topic modeling for the detection and analysis of gender bias in three corpora (mainstream news articles in English and Swedish, and LGBTQ+ web content in English). We compare differences in topic models for three gender categories (masculine, feminine, and nonbinary or neutral) in each corpus. We find that in all corpora, genders are treated differently and that these differences tend to correspond to hegemonic ideas of gender.

## 1 Introduction

As Machine Learning (ML) models are increasingly applied in ways that affect our lives in significant ways, their fairness becomes a societal concern. Over the last few years, a number of highly publicized scandals have occurred. For example, Dastin (2018) reports on Amazon's problems with a recruiting tool that turned out to be biased against women, while Olson (2018) describes how Google Translate tended to translate gender neutral pronouns into e.g. masculine ones for engineers, but feminine ones for nurses. If we are to continue using ML models for decision making, it is crucial that we develop methods for ensuring their fairness.

When we say that we want a fair ML model, it is not always clear what we mean. From a gender-theoretical perspective, fairness is typically understood in relation to structural frameworks of power asymmetries, see, e.g., (Frye, 1983; Nussbaum, 1999). Various technical definitions of fairness exist in computer science, and which definition is appropriate may vary by application, complicating what it means to "not include" biased data; see, e.g., (Mehrabi et al., 2019). We believe that in the long run, methods and tools from the Humanities and Social sciences will be a necessary complement to mathematics and statistics in our quest for fair Natural Language Processing (NLP) systems. The current work is a small step in this direction.

ML models are trained using data produced by humans, such as medical diagnoses, image labels, and written text. As a natural consequence, these data generally reflect our society, including our biases and stereotypes (Caliskan et al., 2017). In fact, the data does not only reflect biases and stereotypes; it also contributes to shaping them (discussed in section 1.1).

There are two general approaches for analyzing and mitigating bias in the models: focusing on either the training data or the models themselves. (For a more fine-grained description of the approaches, see,





*Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92
Barcelona, Spain (Online), December 13, 2020.

e.g., (Shah et al., 2020).) Both approaches have their merits, but in this article we focus on the former as we believe understanding injustices in the data will help practitioners make more appropriate choices when training models. More specifically, we look at text corpora of the kind often used to train NLP models and explore the possibility of using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) Topic Modeling (TM) to investigate gender bias in such corpora.

A topic model is a statistical generative model that, during training, can be said to "discover" a set of topics implicitly underlying the documents in the corpus. It has previously been noted that, due to stereotypes and representational issues in the training data, some of the topics tend to be gendered, in the sense that they represent traditionally masculine or feminine aspects of life (Dahllöf and Berglund, 2019). Our aim is to further investigate this potential for discovering gendered topics.

To be able to more clearly find what words are associated with different genders, we make use of semi-supervised TM (see, e.g., Andrzejewski and Zhu (2009)). This means that some topics are seeded with gendered words, forcing the training procedure to treat these words as belonging to the same, explicitly gendered, topic. In addition, we use unsupervised TM to explore which topics are implicitly gendered.

After training the models, we manually inspect the results, looking first at the top 50 words of each topic and their respective weights, and then looking at the top 20 in more depth. This involves using a qualitative, rather than a purely quantitative approach. We argue that this is an advantage because bias and prejudice are complex, context-dependent concepts, and a purely quantitative approach does not lend itself to a complete understanding of the situation.

### 1.1 Theoretical Grounding

Bias is inherently human, and thus vague and fleeting. If we give a strict mathematical definition of what it means for a data set to be biased, we can only verify or falsify the presence of the particular features of our definition. As pointed out by Blodgett et al. (2020), the definitions in technical papers on bias in NLP are often inconsistent or implicit. The idea behind using TM is that, combined with qualitative analysis of the results, it has the potential to help discover ways in which representational bias is manifested in a corpus, rather than simply verifying that an expected bias exists. In other words, we expect to find differences given that we know we live in an inequitable world, but are also concerned with discovering *how* groups are treated differently in the data.

Under the taxonomy used in Blodgett et al. (2020), our work is concerned with discovering representational harms *within the training data* i.e. the potential for systems trained on such data to demean, misrepresent, or fail to represent particular groups. Such behavior is harmful in its own right, reinforcing the subordination of already-disadvantaged groups (Crawford, 2017). These biases may also contribute to "downstream" allocational harms when applied to systems concerned with distributing resources.

Language - in a broad sense - is the mechanism by which stereotypes are transmitted and maintained (see, e.g., Maass and Arcuri (1996)), and is more generally crucial for the construction of our worldviews. As scholars such as Hall (2013) have argued, the material world has no meaning in itself. Rather, meaning is created through language when we describe and represent the world, for instance in news articles, which often make up the corpora that ML models are trained on. Thus, language has material effects; how we describe or represent groups is intimately linked to power relations and affects the distribution of resources (Foucault, 1976).

We understand gender as socially and culturally constructed rather than as unchanging, innate characteristics of "women" and "men", tied to biological sex. Following Butler (1990) we see gender as constructed through performativity, i.e. acts that are repeated over time and produce our understanding of gendered categories. Hence, the words that are associated with women, men, and nonbinary[1] people in the corpora studied here do not necessarily reflect real-life experiences, but they contribute to (re)producing our ideas of femininity and masculinity.

We would like to treat gender not as a oppositional binary categorization, as in most of the existing literature on gender bias in NLP, but as much more flexible and fluid. As a first step in this direction, we

[1]Throughout this paper, we use 'nonbinary' as an umbrella term referring to all gender identities between or outside the 'binary' categories of men and women.

use three gender categories in this study: masculine, feminine, and nonbinary (which in practice is often mixed-gender or "neutral"). We investigate two corpora made up of mainstream news articles, one in English and one in Swedish. In order to make up for the fact that these corpora rarely mention nonbinary people, we also compare with a third, "queer" corpus, collected from sources that are explicitly oriented towards LGBTQ+ themes.

### 1.2 Related Work

Over the last few years, research interest in bias and fairness in ML models has increased, prompted in part by the highly publicized scandals referred to above. We mention some of the most immediately relevant work here. For a more comprehensive survey of the existing literature, see Mehrabi et al. (2019) for bias in ML generally, and Blodgett et al. (2020) for bias in NLP.

There is a growing body of work on measuring and mitigating bias in word embeddings; see, e.g., (Bolukbasi et al., 2016; Garga et al., 2018; Zhao et al., 2018b). As shown by Gonen and Goldberg (2019), however, the problem is hard to overcome, as the proposed methods leave substantial implicit bias in the embeddings.

Techniques for mitigating bias in other NLP applications have also been tried. For example, Zhao et al. (2018a) present methods for minimizing bias in coreference resolution, as do a number of articles resulting from the first Workshop on Gender Bias in Natural Language Processing (2019). Hoyle et al. (2019) use unsupervised latent variable modeling to investigate what words are used to describe men and women in texts. Their main conclusion is that positive adjectives referring to women are more often related to their bodies than is the case for men.

A few articles stress that there are different kinds of bias and that bias takes different forms over time, culture, genre, etc. For example, Hitti et al. (2019) propose a taxonomy of bias, where they identify four kinds of bias, two of which cannot be identified using today's quantitative methods. This points to the need for a mixture of qualitative and quantitative methods when studying bias and fairness in ML. There have been some efforts in this direction (Leavy, 2018; Dahllöf and Berglund, 2019; Hoyle et al., 2019), but they are few and most of the work remains to be done. Hovy and Spruit (2016) discuss in particular "demographic bias" in NLP datasets, where exclusion from or misrepresentation in the data leads to (or amplifies) social and material consequences for the "left out" groups.

## 2 Methods

We used semi-supervised TM to find explicitly-gendered topics in order to explore the differences in what words and concepts women, men, and nonbinary (or, in cases with low representation, "neutral") people are associated with. We trained these topic models using two different sets of seed words across three corpora, for 15 topics at sentence-level "documents." We also trained a baseline, unsupervised topic model for each corpus, which we use to explore implicitly-gendered topics. One key aspect of our approach was our use of qualitative analysis to interpret our topics.

### 2.1 Corpora

We used three corpora to make our comparisons across language and social context: Mainstream news corpora in both Swedish and English, and the English-only Queer corpus (news and web content by or relating to LGBTQ+ people and issues).

#### 2.1.1 Mainstream

The Mainstream corpora were made available to us by colleagues. They were produced using Scrapinghub[2] during 2019. Each corpus was collected from a relatively small number of news websites and contains 100 000 news and magazine articles, where each article is at least 1000 characters long. The Mainstream English (ME) corpus contains approximately 58 million words before preprossessing; Mainstream Swedish (MS), 44 million words.

[2]https://scrapinghub.com/

### 2.1.2 Queer (English-only)

The novel Queer English (QE) corpus was constructed using the corpus development tools provided by Sketch Engine.[3] (Kilgarriff et al., 2014) It contains 92 million words before preprocessing, over 66 thousand documents, collected over five weeks from January to early February 2020. Due to time constraints and the fact that there are relatively fewer sources for LGBTQ+ material in Swedish, a corresponding Swedish corpus was not constructed.

First, we applied Sketch Engine's web scraper tool to a list of LGBTQ+ publications' websites (including current newspapers and magazines, as well as archival material from print media) and the "LGBTQ+" pages from mainstream news websites such as the BBC. Approximately 28 million words of the corpus resulted from this step. The remaining two thirds of the corpus was built using the keyword search tool, which scrapes material from urls returned by Bing searches of 3 keywords at a time.

Our list of keywords, presented in Table 1, contains "definitional" LGBTQ+ words, such as acronyms for the community and names of orientations and gender identities;[4] "contextually" queer keywords and phrases, such as *coming out* and *drag*; pronouns; and general words for people and occupations, such as *woman* and *politician*. This last category was included as we found it to produce a wider variety of material.[5] To ensure the maximum number of unique permutations of search words, we shuffled the list of keywords and ran the searches in sets of 9. We repeated this procedure four times.

## 2.2 Preprocessing

While preprocessing the texts for use in training the topic models, we attempted to treat the corpora for both languages as equivalently as possible, given available resources. After reading in the corpus file, we made several standard replacements (newline and tab with a single space, etc.) and also merged any occurrences of the word "non-binary" with "nonbinary," before eliminating characters which were not alphanumeric, space, the ascii apostrophe, or a currency symbol. Texts were lemmatized and split into smaller *documents* for TM (see Section 2.3). For both languages, we employed a modified version of the NLTK stopword list, which did not include third person pronouns or negations such as "not."

### 2.2.1 Lemmatization

We used the NLTK[6] toolkit for tokenization, lemmatization, and POS tagging of the English corpora. Lemmas were concatenated with their POS tags in order to make disambiguation possible in analysis. We used the Penn Treebank tagset and ignored coordinating conjunctions, cardinal numbers, determiners, prepositions, possessive endings, particles, *to*, and wh-words. To better match the Swedish preprocessing and improve our ability to compare results across languages, we merged all sub-tags for nouns, proper nouns, adjectives, and verbs (e.g. *girl* `girl+NN` and *girls* `girl+NNS` are both included in the corpus as `girlNN`). After removing stopwords and unwanted parts of speech, we added our POS-tagged lemmas to the dictionary and new documents to a gensim (Řehůřek and Sojka, 2010) corpus, and stored both for use in training topic models.

For Swedish, we used the Stagger[7] (Östling, 2013) package for tokenization, lemmatization, and POS tagging. Again, we removed stopwords, concatenated lemmas and POS tags, and created a gensim dictionary and corpus.

## 2.3 Semi-Supervised Topic Modeling

We used both unsupervised and semi-supervised TM to explore the corpora. In short, semi-supervised TM lets us "force" certain words to be associated with certain topics. This can be used to make sure that the retrieved topics are more relevant to the user or to "guide the topic model towards the discovery of secondary or non-dominant statistical patterns in the data" (Andrzejewski and Zhu, 2009). We used

[3] http://www.sketchengine.eu

[4] Some of these terms may be considered outdated. We included them to get a better view of the community as a whole, as older members may continue to identify with and use them, and to capture a broader temporal slice of search results. Slurs were intentionally excluded from the list.

[5] i.e. stories about people who happen to be queer, in addition to stories *about* being queer.

[6] https://www.nltk.org

[7] https://www.ling.su.se/english/nlp/tools/stagger/stagger-the-stockholm-tagger-1.98986

| | | |
|---|---|---|
| ace | genderfluid | pansexual |
| actor | genderfluidity | pansexuality |
| actress | gender identity | performer |
| agender | genderqueer | person |
| aro | girl | politician |
| aromantic | he | queer |
| asexual | hetero | same-gender |
| asexuality | heterosexual | same-sex |
| bi | homosexual | sexuality |
| bigender | homosexuality | sexual orientation |
| bisexual | intersex | she |
| bisexuality | lesbian | spivak |
| boy | LGBT | straight |
| came out | LGBT+ | they |
| celebrity | LGBTQ | trans |
| child | LGBTQ+ | trans* |
| cis | LGBTQA | transgender |
| cisgender | LGBTQA+ | transsexual |
| closet | LGBTQI | transvestite |
| closeted | LGBTQIA | two dads |
| come out | LGBTQIA+ | two fathers |
| coming out | M2F | two moms |
| drag | man | two mothers |
| F2M | MTF | woman |
| FTM | neopronoun | xe |
| gay | nonbinary | ze |
| gender | non-binary | zie |

Table 1: **LGBTQ+ Keyword List:** Search terms used to build the QE corpus.

it to, in each topic model, create three "gendered" topics: one feminine, one masculine, and one neutral/nonbinary. This was achieved by "seeding" these topics with a number of gendered seed words; see Section 2.3.3.

For the topic inference, we used Parallel Semi-Supervised Latent Dirichlet Allocation (pSSLDA),[8] an implementation by Andrzejewski of the method described by Andrzejewski and Zhu (2009). This package makes it easy to seed topics by setting $z$-values (essentially weighted priors or feature labels, increasing the likelihood of a word to belong to a particular topic) for the relevant words. It implements LDA inference using Gibbs sampling, with relatively modest memory requirements. Another benefit is that it is a parallel implementation, which lets the user run the inference on many kernels simultaneously, saving time.

We piloted our experimental design with varying document sizes (paragraphs, sentences, and 25, 50, or 100 word chunks) and numbers of topics (5, 10, 15, and 20) to determine what was appropriate for our analysis of these corpora. The random seed (194582), number of samples (1000) and $z$-values (5.0) were kept constant throughout. Our final experimental suite uses sentence-level document size and 15 topics.

### 2.3.1 Number of Topics

We ran standard (unsupervised) TM with the same packages as our final experiments for all three corpora to determine the "natural" number of topics they split into, based on our subjective analysis. For all corpora, we found that using 15 topics produced the most coherent themes without blending themes

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

together (as in the cases of 5 or 10 topics) or producing too many topics with no discernible theme (as in the case of 20 topics). In retrospect, we might have also used a coherence measure to inform this decision, and will do so in future work.

#### 2.3.2 Document Size

To find the most appropriate document size (i.e. how much context to consider as "co-occurrence") we ran unsupervised TM for all three corpora, preprocessed using different methods to split the texts into documents. We found that, due to formatting differences across texts even within a particular corpus, paragraphs were too difficult to define and too varied in length to be an appropriate document size.

Sentences were split for the English corpora by naïve punctuation rules at full stops, exclamation points, and question marks; and for Swedish following the 'MAD' (major delimiter) tag produced by Stagger. For both corpora, word chunks of specified sizes were calculated within texts, meaning that a text containing 267 words would be split into three "100" word chunks: two of exactly 100 words, and one of the remaining 67 words.

In general across the different corpora, we found a sentence-level split to provide the "crispest" topics and it was therefore used in our final analysis. This somewhat matched our intuitions. As we were trying to find what words and concepts are associated with different genders by using explicitly gendered words as a proxy to discover implicitly gendered words, limiting context helped capture more closely-associated words.

#### 2.3.3 Seed Words

In addition to a fully unsupervised run for every experiment, we ran semi-supervised TM on two different sets of seed words, each with three lists serving as a proxy for social categories of gender (masculine, feminine, neutral/nonbinary). The division of lists into "base" and "relational" was based on the gendered terms used as a filter in (Hitti et al., 2019). In the base list, we included words we consider to be purely definitional, as opposed to "relational" words such as *mother-father-parent* or *wife-husband-spouse*. The reason for this was to ensure that such words did not skew the feminine category towards a false association with family. Related work e.g. (Lu et al., 2018; Hoyle et al., 2019), tends to include these relational words (as they are reliably gendered in English and other languages), so we constructed the relational list to ease comparison and see if there was any appreciable effect. Note that the relational list contains both base and relational words. The full lists are presented in Table 2. In addition to using these seed words to train our models, we counted the number of times each seed token appeared in the corpora.

### 2.4 Qualitative Analysis

Our final analysis is based on a total of nine topic models, keeping document size and the number of topics constant but varying the choice of corpus (QE, ME, MS) and seed word list (none, base, relational).

In order to answer our question of whether this method is appropriate for discovering potential gender bias in different corpora, we qualitatively analyzed our results by setting up a number of research questions. These questions reflect some of our expectations, as they were grounded in feminist and queer theories about gendered inequalities and stereotypes, as well as differences between, on the one hand, Sweden and English-speaking countries, and, on the other, queer and mainstream contexts, with regards to how gender and gender equality are conceptualized; see, e.g., (Beauvoir, 1949; Jagose, 1996; Martinsson et al., 2016). We conducted our initial analysis with respect to the following questions:

1. Are there gendered differences in the material?
    (a) Are women associated with the private sphere (family/relationships, the "home") and appearance?
    (b) Are men associated with the public sphere and allowed to "be" more things (i.e. represented in a more varied and neutral way)?
    (c) Is nonbinary representation scarce in the Mainstream corpora, and does this category therefore appear to be more "neutral" in mainstream news but more "nonbinary" in the QE corpus?

| | F-En | M-En | N-En | F-Sw | M-Sw | N-Sw |
|---|---|---|---|---|---|---|
| base | she | he | they<br>ze<br>xe | hon | han | hen |
| | woman | man | person | kvinna | man | person |
| | girl | boy | child | flicka | pojke | barn |
| | lady | guy | | tjej<br>dam | kille | |
| | female | male | neutral | kvinnlig | manlig | ickebinär |
| | feminine | masculine | nonbinary<br>enby<br>genderqueer | | | icke-binär<br>genderqueer |
| | Miss<br>Ms<br>Mrs<br>madam | Mr<br>sir | Mx | | | |
| rel | mother | father | parent | mamma<br>mor | pappa<br>far | förälder |
| | daughter | son | kid | dotter | son | barn |
| | niece | nephew | nibling | systerdotter<br>brorsdotter | systerson<br>brorson | syskonbarn |
| | grandmother | grandfather | grandparent | mormor<br>farmor | morfar<br>farfar | morförälder |
| | granddaughter | grandson | grandchild | dotterdotter<br>sondotter | dotterson<br>sonson | barnbarn |
| | aunt | uncle | | faster<br>moster | farbror<br>morbror | |
| | girlfriend<br>fiancee | boyfriend<br>fiance | partner | flickvän<br>fästmö | pojkvän<br>fästman | sambo |
| | stepmother | stepfather | stepparent | styvmor<br>bonusmamma | styvfar<br>bonuspappa | styvförälder<br>bonusförälder |
| | stepdaughter | stepson | stepchild | styvdotter<br>bonusdotter | styvson<br>bonusson | styvbarn<br>bonusbarn |
| | wife | husband | spouse | fru<br>hustru | (man) | partner |
| | sister | brother | sibling | syster | bror | syskon |

Table 2: Seed word lists. For each gender and language, corresponding words are horizontally aligned. The main differences between the English and Swedish lists are that titles are excluded from the Swedish lists, since they are very rarely used, and there are more relational words in the Swedish lists. This is because words such as *grandmother* have two versions in Swedish: the maternal and paternal grandmother. Recall that the base words are also included in the corresponding relational list.

2. Is there less gender bias in the MS corpus than the ME corpus?

3. Is there less (or different) gender bias in the QE corpus than the ME corpus?

4. Will women be associated with relationships when using the base wordlist (which does not contain relation information)? Will men also "become" associated with relationships when using the relational wordlist?

We performed our initial analysis as a group, looking at the top 50 words and their weights across the three corpora and three sets of seedwords. First we looked at the unsupervised topics, noting themes and anything we found striking. Then we compared the gendered topics: between each other within wordlists, and between the wordlists for each gendered topic. To examine gendered topics, we used a visual summary of the top 50 words and their weights (supplemented by the exact numbers), and similarly noted themes and anything striking.

We drew some initial conclusions but were also left with additional questions, which we set out to answer individually. In this layer of the analysis, we looked more closely at the top 20 words for each gendered topic. For each topic we grouped the words into categories such as 'relational verbs,' 'active verbs,' and 'other verbs,' and compared the different topics.

The full results of our experimental suite can be found at GitHub.[9] For each topic model, the provided file lists the top 50 words for each topic together with their weights. Relative weights are provided for the models used in the final analysis.

## 3 Results

Table 3 shows an example of our results, using the base seed word list and the ME corpus. Following Dahllöf and Berglund (2019) the words are listed in order of descending weight within each topic, and color coded according to how "exclusive" they are to the topic. In other words, for a topic $t$ and a word $w$, the ordering in the list is based on $p(w|t)$, while the color coding is based on $p(t|w)$ (**LemmaPOS** $\geq$ 90%, otherwise *LemmaPOS* $\geq$ 75%, otherwise *LemmaPOS* $\geq$ 50%, otherwise LemmaPOS $<$ 50%). Additionally, seed words are underlined.

### 3.1 Quantitative Results: Occurrence of Seed Words

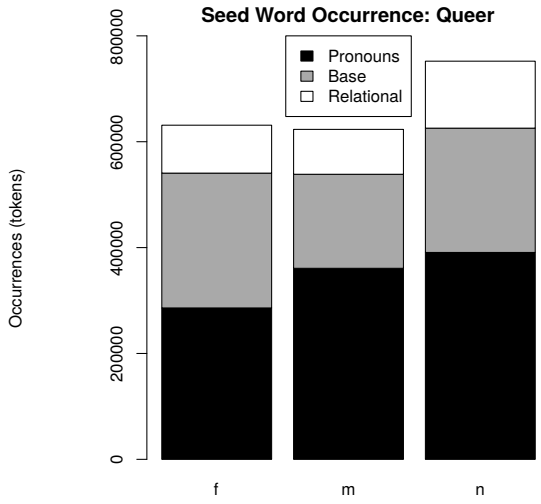


Figure 1: Number of occurrences for seed words in the QE corpus.

The bulk of tokens for each gender category in the seed word lists are common personal pronouns, although the QE corpus contains proportionally fewer than the Mainstream corpora. In both English

[9] https://github.com/TopicModelAnon/FullResults

F **herPRP$**, theirPRP$, **womanNN**, *familyNN*, *tellVB*, **mediumNN**, homeNN, askVB, herPRP, *friendNN*, *youngJJ*, showVB, alsoRB, *writeVB*, callVB, takeVB, timeNN, lifeNN, *socialJJ*, themPRP, *videoNN*, *questionNN*, **motherNN**, becomeVB, liveVB, *sendVB*, **wearVB**, leaveVB, *menNN*, *speakVB*, *postNN*, **readVB**, *hearVB*, *nameNN*, **messageNN**, **girlNN**, giveVB, nowRB, **daughterNN**, *parentNN*, *phoneNN*, *interviewNN*, findVB, useVB, ownJJ, **mrNN**, *shareVB*, *postVB*, *twitterNNP*, *sonNN*

M **hePRP**, **hisPRP$**, **himPRP**, **oldJJ**, timeNN, wouldMD, **manNN**, getVB, takeVB, goVB, *backRB*, tellVB, dayNN, justRB, startVB, tryVB, 'sVB leaveVB, **guyNN**, *agoRB*, *laterRB*, workVB, *awayRB*, giveVB, firstJJ, **himselfPRP**, stillRB, runVB, *spendVB*, fewJJ, *handNN*, *headNN*, neverRB, *'dMD*, dieVB, lookVB, keepVB, askVB, seeVB, *sawVB*, homeNN, turnVB, **boyNN**, believeVB, lifeNN, longJJ, injuryNN, sameJJ, moveVB, *walkVB*

N **theyPRP**, *theyPRP$*, notRB, *canMD*, *themPRP*, **asRB**, *wellRB*, soRB, willMD, **childNN**, wouldMD, *wayNN*, 're VB, *manyJJ*, moreRBR, takeVB, evenRB, needVB, lookVB, *mayMD*, wantVB, thereEX, giveVB, tooRB, onlyRB, seeVB, **personNN**, shouldMD, goVB, *mightMD*, veryRB, otherJJ, *farRB*, *keepVB*, *muchJJ*, timeNN, stillRB, uPRP, findVB, placeNN, tryVB, *ableJJ*, workVB, helpVB, moveVB, nowRB, believeVB, ownJJ, *possibleJJ*, feelVB

Table 3: Top 50 words (lemmas concatenated with merged Penn Treebank POS tags) in gendered topics for the ME corpus using the base wordlist. The ordering in the list is based on $p(w|t)$, while the color coding is based on $p(t|w)$ (**LemmaPOS** $\geq$ 90%, otherwise *LemmaPOS* $\geq$ 75%, otherwise *LemmaPOS* $\geq$ 50%, otherwise LemmaPOS $<$ 50%). Additionally, seed words are underlined.

corpora, the exception is the neo-pronouns *ze* and *xe*[10] which appear less than ten times each in the QE corpus and not at all in the ME corpus. In the MS corpus, the gender-neutral third person singular pronoun *hen* appears only 1128 times. *Hen* was added to the Swedish Academy Glossary in 2014, following public debate stemming from its inclusion in a 2012 children's book, and its reception is gradually becoming more positive (Gustafsson Sendén et al., 2015). This relative recency, initial unpopularity, and the fact that (unlike English *they*) it is exclusively singular may all contribute to the relative infrequence of *hen*. The number of occurrences of the different categories of seed words for the three corpora are depicted in Figures 1, 2, and 3.

Within both Mainstream corpora, words from our masculine seed lists occur more often than neutral seed words, and roughly twice as often as words from our feminine seed list. The vast majority of this difference is explainable by the personal pronouns *he, she, they* and *han, hon, hen* (all of which are only tracked as the subjective form). Notably, in ME the pronoun *he* occurs more often than all of the seed words combined for either other gender. Comparing only pronouns, the *he/she* ratio for the ME corpus is 2.53 and 1.26 for the QE corpus; *han/hon* for the MS corpus is 2.58.

The QE corpus by contrast is much better balanced than either Mainstream corpus, and contains explicit nonbinary representation. 3.75% of tokens within the neutral seed category are explicitly gendered (*ze, xe, nonbinary, enby, genderqueer*), compared to 0.05% in the ME corpus. We discovered after experiments were run that while *icke-binär* (nonbinary) does appear several dozen times within the MS corpus, it is tagged as a noun instead of an adjective, and therefore listed as occurring 0 times. The rate of occurrence is low enough that we do not believe its exclusion in the TM seriously impacts our results, but is worth mentioning as part of our overall observation that nonbinary people and issues are largely invisible in both the data and the tools used to process natural language.

[10]We did not include other neo-pronouns in our seed word lists. It is also possible that these pronouns do appear in the ME corpus but are improperly lemmatized.

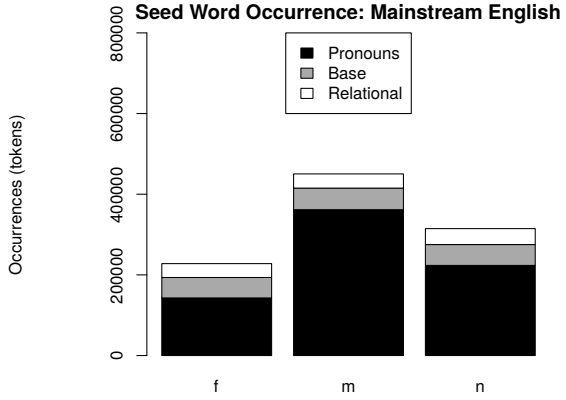


Figure 2: Number of occurrences for seed words in the ME corpus.

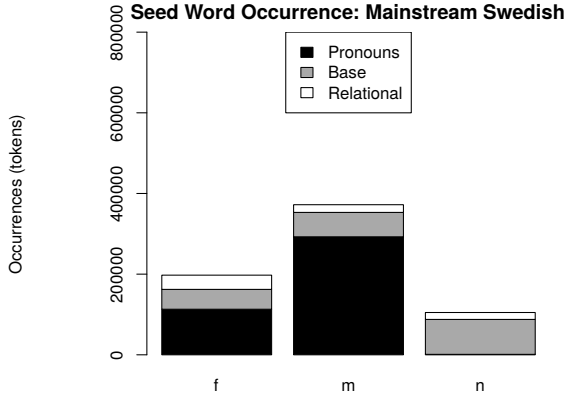


Figure 3: Number of occurrences for seed words in the MS corpus.

### 3.2 Qualitative Results

Our analysis reveals the presence of both explicitly- and implicitly-gendered topics, although these topics were not always aligned with the specific stereotypes we expected. We found gendered differences within and across our corpora, both with unsupervised and semi-supervised TM techniques.

#### What are the gendered differences?

Across all three corpora, the explicitly-gendered feminine topic is associated with the private sphere: family (*family, mother, father, parent, home*), relationships (*relationship, friend, love*), and communication (*tell, ask, write, call, see, meet, feel*). *She* in its subjective form is not present in the feminine ME topic's top 50 words, although it does appear more highly weighted in the QE corpus and the MS corpus (*hon*). Women also tend to be linked to time, in particular to youth in the ME corpus (where the masculine topic was more generally associated with time). Other than this association with youth, we did not find the link between women and appearance we expected.

We find that while men are associated with the public sphere, they are also "neutral" in the ME corpus: associated with general or generic terms similar to those in the neutral category. This suggests that material in this corpus implicitly treats men as the norm from which other genders deviate. 'People' are men unless otherwise specified, a sexist form of false generic (Mills, 1995). Although the masculine topic we obtain from this corpus using semi-supervised TM does not follow a particular theme, this does not mean that certain topics are not masculine. The "political" topic in unsupervised ME is dominated by masculine pronouns (*hePRP* 0.072 and *hisPRP$* 0.042) - the public sphere remains implicitly masculine. This was the only notable instance of strongly gendered associations within our unsupervised topic models.

We also note that the words in the feminine topics are more exclusive to those topics. If we look at the example (ME corpus, base seed word list) in Table 3, we see that 29 out of the 48 words that are not seed words are colored, indicating a relative weight ($p(t|w)$) of at least 0.5. For the masculine topic, this number is 14 out of 46 and for the neutral topic 13 out of 47. This indicates that the predominant themes in the feminine topic (family/relationships, communication/social media) are very strongly tied to femininity in the corpus, whereas the themes in the masculine and neutral topics do not have such strong connections to a gender.

Our experiments for the MS corpus and the QE corpus do not show this same generalization of men as neutral; the masculine topics are instead related to crime and death/Christianity, respectively.

Neither Mainstream corpus really contains enough nonbinary representation to produce a "coherent gender". Instead, we see that the third gender topic in these corpora are best termed "neutral", and are often not related to individuals, or even people as a category. In contrast, we do find that there is (more) adequate representation of people who do not fall neatly within the binary gender categories of "men" or "women" in the QE corpus, as expected. Although the third category for this corpus still contains primarily neutral or generic references to people, a coherent theme emerges relating to "acceptance" (both self-acceptance and the acceptance of others), with words such as *parent, question, love, feel, ask, share, accept, able, different, choose*.

**Is Swedish less gender-biased than English?**

There does not seem to be notably less gender difference in the MS corpus than in the corresponding ME corpus. Women are associated with family and relationships, as well as communication, in both corpora; although *hon* is more highly weighted in its subject form than *she* is. Perhaps the most interesting difference is in men: in English, men are neutral (the "norm") while in Swedish the masculine topic is best labelled "crime and punishment."

**Is the QE corpus less gender-biased than the ME corpus?**

Comparing between our two English corpora, we find that the QE corpus still strongly associates women with family/relationships (*family, father, friend, relationship, love*) and time (although here *old* is present in addition to *age, young, life*). The theme of the masculine category, however, is completely different: from a generic norm in the ME corpus to death and Christianity in the QE corpus. The exact reasons behind this difference is unclear; however, as the frequency of "feminine" and "masculine" tokens is more balanced in the QE corpus, it is unlikely that this is a case of misrepresentation caused by exclusion, as described in (Hovy and Spruit, 2016).

One key finding within the QE corpus is the presence of nonbinary people and the emergence of a coherent theme from the neutral/nonbinary topic. Within the ME corpus this topic is better described as "neutral" but in the QE corpus it can more honestly be termed "nonbinary." Where nonbinary representation is insufficient, such as in both Mainstream corpora, the neutral topic appears to refer to people in general, if it refers to "people" at all (compare the MS corpus, where this topic is dominated by local and international news). Only with sufficient representation does a coherent third gender category become evident.

**Does the relational seed word list "induce" an association between a gender and family/relationships?**

In general, we find that women are associated with family/relationships and communication regardless of whether relational seed words are used or not. We also find that men in the Mainstream corpora do not become more associated with these things when relational seed words are added. In fact, the seed words themselves fail to appear among the top 50 words. The ME neutral topic skews more towards a "real" theme with the addition of relational seed words: we find words such as *school* and *student*.

Interestingly, there seems to be a stronger effect of adding relational seed words when training on the QE corpus, although it does not really serve to alter the theme of any of the topics overall. The relational

version of the feminine topic adds *lesbianJJ, gayJJ,* and *gayNN*; and the relational seed words actually appear in the masculine topic. The nonbinary topic changes the least.

## 4 Discussion

Semi-supervised topic modeling seems to do a decent job of exposing the differences in treatment of gender in the text corpora we tested, suggesting it is indeed an appropriate method for discovering bias in data before it is used to train a biased model. We found evidence of gendered differences emblematic of structural power divides in all three corpora. Women tend to be strongly associated with the "home" (family, relationships) and communication; while men are more varied and nonbinary people are nearly invisible in "mainstream" contexts. Generally, this method constitutes a "middle ground" where we escape some limitations of purely quantitative metrics (e.g. understanding *how* representational harms manifest, rather than merely confirming the existence of expected biases) but still must reckon with others (e.g. the required subjective reading may overlook unexpected biases). We plan to expand this method, for example to include guidelines for qualitative analysis with an eye to structures of power borrowed from feminist research methods.

The models we trained require qualitative analysis in the form of human reading to interpret. This is a benefit, as it requires us to think through the how and why of these differences, but can also leave us with lingering questions. For example, we found a very strong theme of Christianity and death in the masculine topic for the QE corpus, but without further examination we cannot tell if this association with Christianity is positive (affirming ministry, messages of acceptance) or negative (condemnation, homophobia). Contrary to our expectations, we did not find a connection between women and appearance in any of our corpora - this may be due to genre (not many "lifestyle" articles) but again would require further examination to determine a cause.

Additionally, TM is not fully deterministic, so there can be some question of the reliability of the results across corpora. It might have been interesting to e.g. train one model for both the English corpora and then investigate them separately, and this may be an angle for future research. This behavior may also be an advantage for more involved investigations, as training multiple models on the same data with different random seeds could provide different "points of view" from which to investigate the corpus and allowing us to triangulate a more complete picture. This potential should also be investigated in future work.

More work is necessary to establish whether TM can help us "debias" corpora, e.g. by identifying and removing strongly-biased texts from the corpus. A natural next step in the line of research presented here is to use the semi-supervised topic models to classify documents and investigate how well this method does at identifying stereotypical writing. TM is relatively computationally cheap, making it an attractive first step in understanding the potential consequences of training a model on a given dataset.

Most work on bias in text so far deals only with gender and considers gender to be a binary category system. We want to contribute to more nuance by working with a nonbinary definition of gender and with a greater focus on intersectionality. This is important since research both in the humanities and in the sciences has shown that focus on only one category, such as gender, can hide prejudice against, for example, women of color; see, e.g., (Buolamwini and Gebru, 2018; Crenshaw, 1991). English and Swedish mark gender grammatically through third person pronouns and semantically in certain nouns (*mother, father, parent*), but there is no equivalent explicit marking for other aspects of identity such as race or class, meaning different strategies must be undertaken to discover intersectional associations. Our technique similarly may not generalize to languages which do not mark gender in this way (e.g. Finnish, which has no gendered third person pronouns), or which have noun cases with grammatical gender (e.g. French or German).

Although we make some progress towards better capturing fluid and multi-faceted understandings by expanding our fixed data categories of "gender" to include a third option, this remains an unsatisfactory solution as it fails both to separate nonbinary individuals from a group or generic (in the case of English *they*) and to provide an intersectional view of different experiences of gender within these three categories. As Bivens (2017) describes such a three-category practice, it "transgresses a rigid binary, yet

falls short of a fluid spectrum, positioning ... somewhere in-between". It remains an open question how to tackle these issues in practical NLP research.

## 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.

Simone de Beauvoir. 1949. *The Second Sex*. Alfred A. Knopf, New York. Translated by Constance Borde and Sheila Malovany-Chevallier, 2010.

Rena Bivens. 2017. The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society*, 19(6):880–898, jun.

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.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Fairness, Accountability and Transparency*, pages 77–91. PMLR.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York.

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

Kate Crawford. 2017. The trouble with bias. Keynote at NeurIPS.

Kimberlé Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

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*.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Accessed: 2020-05-06.

Michel Foucault. 1976. *The History of Sexuality. Vol 1, An Introduction*. Penguin. Translated by Robert Hurley, 1990.

Marilyn Frye. 1983. *The Politics of Reality: Essays in Feminist Theory. Berkeley*. Crossing Press.

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

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NACL: Human Language Technologies, 1*, pages 609–614.

Marie Gustafsson Sendén, Emma A. Bäck, and Anna Lindqvist. 2015. Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Frontiers in Psychology*, 6:893, jul.

Stuart Hall. 2013. The work of representation. In Stuart Hall, Jessica Evans, and Sean Nixon, editors, *Representation*, pages 1–59. Sage.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Workshop on Gender Bias in Natural Language Processing*, pages 8–17.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Annamarie Jagose. 1996. *Queer Theory: An Introduction*. New York University Press, New York.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Susan Leavy. 2018. Uncovering gender bias in newspaper coverage of irish politicians using machine learning. *Digital Scholarship in the Humanities*, 34(1):48–63.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender Bias in Neural Natural Language Processing. jul.

Anne Maass and Luciano Arcuri. 1996. Language and stereotyping. In C. Niel Macra, Charles Strangor, and Miles Hewstone, editors, *Stereotypes and Stereotyping*, chapter 6, pages 193–225. Guilford Press, New York, NY.

Lena Martinsson, Gabriele Griffin, and Katarina Giritli Nygren. 2016. Introduction: Challenging the myth of gender equality in Sweden. In *Challenging the Myth of Gender Equality in Sweden*. Policy Press, Bristol.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *ArXiv*, abs/1908.09635.

Sara Mills. 1995. *Feminist Stylistics*. Routledge, New York, New York, USA.

Martha C. Nussbaum. 1999. *Sex and Social Justice*. Oxford UP.

Parmy Olson. 2018. The algorithm that helped google translate become sexist. Forbes. Accessed: 2020-05-06.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for swedish. *Northern European Journal of Language Technology*, 3:1–18.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NACL: Human Language Technologies, 2*, pages 15–20, June.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Empirical Methods in Natural Language Processing*, page 4847–4853.