Gender and sentiment, critics and authors: a dataset of Norwegian book reviews

Samia Touileb
Language Technology Group
Department of Informatics
University of Oslo
samiat@ifi.uio.no

Lilja Øvrelid
Language Technology Group
Department of Informatics
University of Oslo
liljao@ifi.uio.no

Erik Velldal
Language Technology Group
Department of Informatics
University of Oslo
erikve@ifi.uio.no

Abstract

Gender bias in models and datasets is widely studied in NLP. The focus has usually been on analysing how females and males express themselves, or how females and males are described. However, a less studied aspect is the combination of these two perspectives, how female and male describe the same or opposite gender. In this paper, we present a new gender annotated sentiment dataset of critics reviewing the works of female and male authors. We investigate if this newly annotated dataset contains differences in how the works of male and female authors are critiqued, in particular in terms of positive and negative sentiment. We also explore the differences in how this is done by male and female critics. We show that there are differences in how critics assess the works of authors of the same or opposite gender. For example, male critics rate crime novels written by females, and romantic and sentimental works written by males, more negatively.

1 Introduction

Gender is a widely studied source of bias in textual content (Garimella and Mihalcea, 2016; Schofield and Mehr, 2016; Kiritchenko and Mohammad, 2018). There has been considerable previous work analyzing gender bias in NLP models and in particular, in input representations such as static and contextualized word embeddings (Kaneko and Bollegala, 2019; Friedman et al., 2019; Bolukbasi et al., 2016; Zhao et al., 2020; Basta et al., 2019).

Gender-annotated datasets largely focus on the gender of the author of a specific piece of text, such as a blog (Mukherjee and Liu, 2010; Liu and Mihalcea, 2007) or a tweet (Burger et al., 2011) and has given rise to considerable research focused on author gender identification (Mukherjee and Liu, 2010; Rangel and Rosso, 2019). Datasets which enable the study of response to gender in text, however, are considerably fewer (Voigt et al., 2018). With a few noteworthy exceptions (Zhao et al., 2020; Sahlgren and Olsson, 2019), a majority of previous work has focused on gender modeling and the study of gender bias in English.

Social psychological research on gender bias in language has shown that there are sociocultural stereotypes inherent in the language used to describe females and males (Menegatti and Rubini, 2017). While the descriptions of females tend to focus on their communal traits, males are described for their agentic traits (Menegatti and Rubini, 2017). Madera et al. (2009) show that the gender of the writer can also influence how females and males are described. They show that gender stereotypes can discriminate female applicants in an academic setting, due to their recommendation letters which tend to contain more communal-related words, in contrary to letters written for males which focus more on their agentic abilities. Also, males in their recommendation letters, tend to describe the agentic traits of females more often than females do (Madera et al., 2009). This makes explicit the need to investigate the gender of both sides: the writer, and the person being written about.

This paper introduces a dataset of Norwegian book reviews with information about the gender of both the (professional) critic and the book author. Each review comes with a rating on a scale of 1–6, which can be used as a supervision signal for overall positive/negative sentiment of the text. As a part

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 125–138
Barcelona, Spain (Online), December 13, 2020.

of describing the provided dataset, we include an exploratory analysis of the data through a series of empirical experiments on gender- and sentiment classification. The combination of gender information on two sides in addition to ratings allows for investigating several interesting research questions. Mainly, we here seek to address the following two closely related questions, mostly differing with respect to perspective:

- (R1.1) Are there differences in how the works of male and female authors are critiqued, and in particular in terms of positive and negative aspects? Moreover, (R1.2) are there differences in how this is done by male and female critics?
- Conversely: (R2.1): How do male and female critics choose to word positive and negative criticism? Moreover, (R2.2) are there differences with respect to how they choose to do this with respect to the works of male and female authors?

As a simplifying assumption, we only consider gender as a binary category (male and female) in this work. We acknowledge the fact that gender as an identity spans a wider spectrum than this, but this simplification was here deemed necessary to enable our annotation of the reviews. It is worth noting that during our manual annotation we did not come across any mention of known (to us) non-binary or transgender authors or critics. However, it is quite possible that some of the gender labels assigned would have been different had we been able to rely on self-identification.

Bias statement: This work mainly attempts to shed light on whether there are latent biases inherent in the data directly, focusing on book reviews. As part of this, we investigate whether polarities associated with certain words to some degree are correlated with the gender of either the critic or the author (or both, in combination). One of our motivations is to assess whether the predictions of sentiment classifiers trained on review data – as is commonly the case – may to some degree depend on the gender of either the critics, the creator of the work being reviewed (the author), or both. By extension, and in terms of possible harms, the dataset we present here suffers from representational harms (Blodgett et al., 2020). The book reviews present in the NoReC corpus, which are written by professional Norwegian critics, seem to contain gender stereotypes when describing the works of female and male authors. The societal lexical asymmetries in how females and males are portrayed is present in the language use. For example, words related to emotions and feelings are used negatively to describe the works of female authors, but positively when the works described are written by males. Also, words related to achievements with regards to literary genre or the process of publishing in general are positively used when describing the work of males, and negatively for the works of females. These observations seem to maintain the existing social hierarchies that tend to focus on emotional traits when describing females, while focusing on competence traits when describing males (Menegatti and Rubini, 2017).

2 Related work

Much of the previous work on bias in ML models within NLP has focused on identifying biases in word embeddings and how to mitigate them (Maudslay et al., 2019; Kaneko and Bollegala, 2019; Zmigrod et al., 2019; Friedman et al., 2019; Garg et al., 2018; Bolukbasi et al., 2016), or even make them gender neutral (Zhao et al., 2018b). However, such efforts have received criticism by Gonen and Goldberg (2019) who argue that the biases have not been removed, but only "hidden" and kept at a deeper level in the embedding space. Bias has also been investigated in several other settings, like multilingual embeddings (Zhao et al., 2020), deep contextual representations (Basta et al., 2019; May et al., 2019), language models (Qian et al., 2019), coreference resolution (Cao and Daumé III, 2020; Zhao et al., 2018a; Rudinger et al., 2018), and machine translation (Escudé Font and Costa-jussà, 2019), just to name a few of the more recent efforts.

Another line of work has focused on investigating gender representations in corpora and models, and release gender-neutral corpora (corpora in which either the distribution of genders is balanced, or where gender stereotypes and gendered words are removed). Schofield and Mehr (2016) use film scripts to analyse the linguistic and structure variations in dialogues and how these differ based on gender. Garimella

and Mihalcea (2016) investigate gender biases and discrimination in blogposts. They use a metric to compute the salience of word classes combined with semantic and psycholinguistic resources to identify dominant word classes. These latter are used to uncover the underlying differences in the choice of word classes and concept usage between man and women. They show that the gender of a blog author can be identified using gender-based word disambiguation techniques, and that changes in word frequencies and contexts contribute to the differences between genders. Costa-jussà et al. (2020) present a tool *GeBioToolkit* that automatically extracts multilingual parallel sentences using Wikipedia biographies from several languages, which also relies on gender information to create a gender-balanced corpus. They also introduce the multilingual, parallel and gender-balanced corpus *GeBioCorpus*, a corpus for machine translation applications covering the three languages Catalan, English, and Spanish.

Several studies have also focused on gender and gender bias in sentiment analysis. Kiritchenko and Mohammad (2018) present the Equity Evaluation Corpus (EEC) that contains a set of manually crafted English sentences, with the sole purpose to mitigate biases towards certain races and genders. They also show that using the EEC corpus helped uncover the existing biases in over two hundred sentiment analysis systems, which seemed to give higher sentiment predictions for sentences associated with one given race or gender.

Hoyle et al. (2019) use a generative latent-variable model to represent collocations of positive and negative adjective and verb choices, given a gendered head noun. Their analyses goes beyond qualitative analysis, and shed light on the differences on how men and women are described differently. They use a corpus of books spanning various genres, and show for example that positive adjectives used to describe women are related to their bodies more often than is the case for men. Bhaskaran and Bhallamudi (2019) analyse the existence of occupational gender stereotypes in sentiment analysis models. They show that all their tested models (BOW+logistic regression, BiLSTM, BERT (Devlin et al., 2019)) contain occupational gender stereotypes to some extent. They also show that simple models seem to show biases in training data, while contextual models might reflect biases introduced while pretraining.

Voigt et al. (2018) present an annotated corpus for the gender of the addressee and the sentiment and relevance of comments. The corpus comprised comments from responses to Facebook and Reddit comments, TED talks, and posts on Fitocracy. This work has similarities to ours, since they look at the responses to gender e.g. how the content can differ based on the gender of the person being addressed. However, in our work we also add the dimension of the gender of the critic, and how it can positively or negatively affect the description of the book authors' gender.

In this paper, we do not focus on the differences in gender representations and biases present in existing systems, nor do we try to mitigate them. We rather investigate the differences in gender descriptions in Norwegian book reviews, and if this affects the ratings of the reviews. To this end we introduce a new dataset of rated reviews with meta-information about the gender of both critics and authors of the work under review. We focus on how positive and negative words can be informative for a simple machine learning model, and how these differ between genders.

3 Gender-coded book reviews

The underlying source data in this study is the Norwegian Review Corpus (NoReC) comprising professional reviews across a wide variety of domains, collected from several of the major Norwegian news sources (Velldal et al., 2018). Each review is rated with a numerical dice score on a scale from 1 to 6. In the current work, we only deal with the subset of 4,313 *book reviews*, for which we have extended the meta-information for each review to include manually coded information about gender – both of the critics and book authors. In what follows, we give an overview of our manual and semi-automatic annotation efforts, and provide the resulting corpus statistics.

3.1 Annotation process

We use two simple approaches to annotate the genders of critics and authors: (*i*) a semi-automated approach; use a list of male and female names and match them with the critics, the title, and the excerpt of each review, followed by manual correction, and (*ii*) a manual approach; examine titles, excerpts, and

	Author in title	Author in excerpt	Main text	Total
Semi-automatic	1,324	367	–	1,691
Manual	151	368	1,898	2,417
Non-identifiable	–	–	69	69
Child critic	–	–	31	31
Mixed authors	–	–	105	105
Total	1,475	735	2,103	4,313

Table 1: Annotation process summary.

reviews to manually identify the authors being reviewed.

Authors For the identification of the gender of the book authors, we use a list of predefined male and female names[1] and perform a simple string matching against the title and excerpt of each review. We thereafter manually examine the identified authors and their genders. The list of names contains overlaps between genders, and some names can be both male and female names. In our data, we identified 95 critics and 178 authors that were automatically assigned both genders, these were manually adjudicated and corrected. Table 1 presents the total number of correctly identified authors and their genders using our semi-automatic approach, as well as the number of manually annotated names and their respective genders.

An extensive manual analysis showed that our naive semi-automated approach correctly identified 1,324 authors and their genders in the review titles, and 367 in the excerpts. However, we had to manually correct 151 and 368 authors and genders respectively. Most of these corrections were due to mentions of book characters in titles and excerpts. For example, most of the reviews of Harry Potter books mention Harry Potter in the title or excerpt and not the author, J.K. Rowling.

In addition to the above, we manually annotated 2,103 reviews, either by manually examining the titles and excerpts for names that do not exist in our lists, or by reading the reviews. During this annotation we identified 31 reviews written by children[2], and 105 reviews reviewing books written by both male and female authors. Furthermore, we were not able to identify who the authors are in 69 reviews. These three categories (written by children, reviewing both male and female, and unknown authors) are not included in our investigations, as our main focus is to investigate the differences in reviews written about the works of male and female authors by professional male and female critics.

Critics The names of the critics were already provided by the metadata of the NoReC corpus, and we use the semi-automated approach described above to identify their gender. We also performed a manual check of the whole corpus, and corrected the gender of 39 critics. During this process, some of the critics were identified as *redaksjonen* 'the editors'. A total of 343 of these were manually corrected after inspecting the online published version of the reviews. Still, there are 23 reviews written by unknown critics labeled as "redaksjonen" that could not be identified.

Summarizing statistics In all of the following counts, we disregarded all reviews written by both a male and a female critics, written about both a male and a female author, written by children, and unknown critics. However, this information will be present in the released gender annotations of NoReC[3]. The final dataset comprises reviews written by 199 unique reviewers: 125 male and 74 female critics. These reviews rate the works of 2,317 unique book authors, from which 1,435 were written by males, and 882 by females.

[1] http://clarino.uib.no/iness/page?page-id=Resources

[2] Some sources in NoReC have books reviewed by children and teenagers who have the appropriate age levels for the books.

[3] https://github.com/ltgoslo/norec_gender

		Author		
		M	F	Total
Critic	M	1,748 (73.7%)	623 (26.3%)	2,371
	F	825 (48.2%)	887 (51.8%)	1,712

Table 2: Total counts of reviews by gender.

	Authors		Critics	
	pos	neg	pos	neg
Acc	0.83	0.95	0.86	0.69
MC	0.57	0.55	0.60	0.44
$F1_M$	0.86	0.95	0.89	0.70
$F1_F$	0.79	0.94	0.82	0.68

Table 3: Accuracy of gender classification of authors and critics. Here, pos and neg represent to which polarity the test set belongs. $F1_M$ and $F1_F$ represent class-level F1 scores. *MC* represents majority class values.

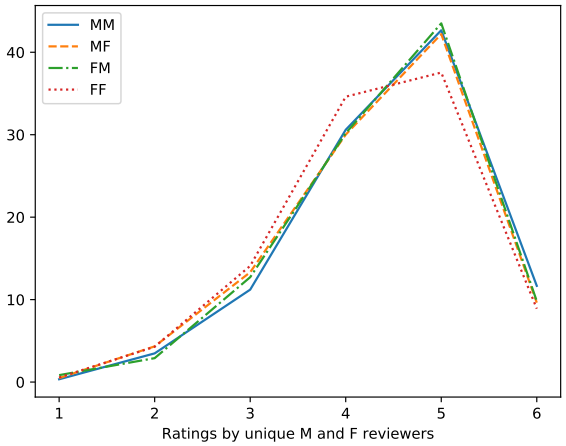

Figure 1: Distribution of ratings given by unique male and female critics to works of male and female authors. The first letter (M/F) indicates the gender of the critic and the second letter that of the author; e.g., *FM* plots ratings by female critics for male authors. The y axis represents normalized percentages of each rating.

3.2 Initial data analysis

Table 2 shows the total document counts broken down along the gender of both the critics and the book author. We see that while the majority of reviews written by male critics targets the work of male book authors (73.73%), female critics tend to have a more balanced review distribution, with a small majority in reviewing female authors (51.81%).

Another interesting aspect of the dataset is the distribution of ratings given by male and female critics. Figure 1 shows the normalized percentages of each rating, where the first letter (M/F) indicates the gender of the critic and the second letter indicates the gender of the author; e.g., *MF* corresponds to reviews by male critics of works by female authors. Here we observe a clear difference in the ratings given by female critics to female authors (*FF*). In general, female critics tend to give works by women lower ratings. Ratings 2, 3, and 4 were given by female to female on 4.28%, 14.09%, and 34.61% of the time. Compared to *MM*, *MF*, *FM* where rating 2 respectively represents 3.48%, 4.33%, and 2.90% of the total ratings. For rating 3 the trend is similar with 11.21% for *MM*, 13.32% for *MF*, and 12.72% for *FM*. Similarly, *MM*, *MF*, and *FM* gave rating 4 to respectively 30.60%, 30.01%, 30.18% of their reviews. On the upper range of the scale the trend is the opposite, with *FF* giving ratings 5 and 6 to respectively 37.54% and 8.90% of their reviews, compared to 42.67% and 11.67% for *MM*, 42.21% and 9.63% for *MF*, and 43.51% and 9.81% for *FM*.

4 Gender Classification

Regardless of sentiment, our first experiments were based on the language use in gender classification. We used our corpus for binary gender classification using Logistic Regression and cross validation, from both the authors' and the critics' perspectives. We opted for this simple classifier because it allows us to easily access the most informative features (in our case words) that guide the classification during training. We manually analysed the top 200 most informative words for each gender, and we were able to see that there were differences in the use of language in relation to gender, regardless of sentiment. However, in an effort to see if there are indeed differences in the language with regard to sentiment, we tested our gender classifiers on two different subsets: (i) a positive test set containing reviews with ratings 5 and 6, (ii) a negative test set comprising reviews with ratings 1, 2, and 3.

Our experiments as presented in Table 3 show that gender classification of authors yields higher accuracy for the negative reviews, while gender classification of critics show the opposite effect (higher accuracy for the positive reviews). More interestingly, looking at F1 values of both genders female (F) and male (M) of book authors, it is clear that our model is able to classify reviews about works of male authors in both positive and negative context, with a slightly higher accuracy for the negative test set. For reviews about the works of female authors our model is much better at identifying the gender in the negative test set. Conversely, in the classification of critics' genders, our model is better at classifying both female and male critics in the positive test set. Moreover, and as can be seen in table 3, our models perform better than the majority class baseline (classify all reviews as M) for both authors and critics.

5 Sentiment Classification

Based on the observations presented in Section 4, our main interest in the current work is therefore to investigate if there are any differences in how the literary works of female and male authors are described in positive or negative reviews by female and male critics. In particular, we want to investigate whether the gender of the author or the critic affects the language use of the review and its rating. In order to do so, we train a number of models on differing critic–author combinations to predict the rating of a review, e.g. male critics reviewing female authors, male critics reviewing male authors, etc. We then go on to analyze the most informative words of these models using clustering over their word embeddings.

For transparency and ease of interpretability, we first make use of machine learning models based on traditional approaches (i.e. with discrete and count-based features) that allows for straightforward extraction of the most informative features. We thereafter use word embeddings to cluster these features and identify representations of the content. To this end, we focus solely on the use of content words, i.e. adjectives, nouns, and verbs. These are already available in NoReC which is annotated with PoS tags using UDPipe (Velldal et al., 2018).

To train our models, we create two subsets of our gender-annotated dataset: (i) a subset containing reviews reviewing female authors (R_F), and (ii) a subset of reviews reviewing male authors (R_M). Instead of looking at the full range of ratings as given in NoReC, we focus on the lowest and highest values of the rating range. The reason for this is that we want to analyze cases of clear positive or negative sentiments. We select all reviews with rating 1, 2, 3, and 6, and randomly select reviews with rating 5 to balance the distribution between the lower and higher ranges. Ratings 1, 2, and 3 represent negative reviews, while 5 and 6 are positive. These categories are consecutively used for binary sentiment classification using Logistic Regression and cross validation.

In order to obtain a richer picture of the important features for classification, and how these differ between genders, we investigate the results of two different strategies for training using our gender-annotated data:

Authors: We combine the train and dev splits within each of R_F and R_M for cross validation. Also, as previously mentioned, we balance the data within the splits such that the positive and negative classes are equally distributed. We thereafter analyse the results of four testing strategies: (1) train on R_F train+dev, test on R_F test, (2) train on R_F train+dev, test on R_M test, (3) train on R_M train+dev, test on R_M test, and (4) train on R_M train+dev, test on R_F test.

Critics: We follow the same steps as above, but run different models based on the gender of the critic. We add an additional dimension to the previous analysis by comparing the author and critic aspects. More concretely, we analyse results of four training combinations: (1) R_{FF}: female critics, female authors, (2) R_{MF}: male critics, female authors, (3) R_{FM}: female critics, male authors, and (4) R_{MM}: male critics, male authors.

For each of these strategies, we have manually analysed the 200 most informative words, and looked at the overlap between them. We provide additional details in Section 5.1 and Section 5.2.

5.1 Authors

As previously mentioned, we separate the reviews about female and male authors and create two subsets R_F and R_M. Then, we balance the number of positive and negative reviews, such that all reviews with ratings 1, 2, 3, and 6 are selected, and we select a random sample of the reviews with rating 5 to make the distribution of positive and negative labels balanced. Thereafter, for each of these subsets, we train three separate Logistic Regression models with cross validation for binary sentiment classification using as features the word counts of adjectives, verbs, and nouns of each review. Balancing the distribution of positive and negative labels in each of the subsets R_F and R_M considerably decreases the size of the data, which is already small to start with. We therefore run a 10-fold cross validation approach on the combined train and dev splits as identified in NoReC (for each of the subsets), and use the test split for final evaluation.

For each of the subsets R_F and R_M we first test on the test splits of the same subset (R_F test and R_M test respectively), and then on the test split of the opposite gender. Here, we do not focus on achieving the best accuracy, but rather on understanding what guided the model during classification. However, for the record, we did a simple grid search to identify the most suitable parameters (we focused on the parameters penalty, solver, and max-iter). We found that there are no obvious differences in the accuracy of models tested on data describing the same gender versus data describing the opposite gender. We therefore focus on the actual word usage.

For each subset, we identify the 200 most informative words during training. We believe that these words give insights into the classification process and can help us identify the differences between important words for the identifications of positive and negative reviews about female and male authors. Moreover, we cluster these 200 most informative words of each subset using pre-trained word embeddings[4] to identify the different clusters of words, which adds a second level of analysis to our investigation. We used the Silhouette method (Rousseeuw, 1987) to determine the optimal number of clusters, which was 25 clusters. We manually analysed these 25 cluster of words and labeled them as shown in Figure 2. Using the information from the clusters, we analysed which adjectives, nouns, and verbs were positive in the R_F but negative in R_M, and vice versa. We also looked at which cluster of content words were positive and negative in both subsets R_F and R_M. These are presented in Figure 2.

Most positive adjectives used to describe females, which also are negative when describing male authors are uplifting words (*morsom* 'funny', *sjelden* 'rare', *utmerket* 'excellent') and adjectives relating to quality characteristics (*tydelig* 'clear', *rett* 'right'). Also, adjectives describing emotions (*rørende* 'touching', *treffende* 'aptly'), and socially critical and beliefs (*filosofisk* 'philosophical', *historisk* 'historical',), seems to be positive when describing female and negative for male. Most adjectives negatively used to describe the works of female authors and positively used for works by males are derogatory adjectives (*kaotisk* 'chaotic', *mislykket* 'unsuccessful'). These are in themselves negative words, but seem to be present in positive descriptions of work by male authors, which might either reflect that even if a book is unsuccessful, the male author might still be positively reviewed, or that unsuccessful events happening in a book might still be a positive aspect of the content of the book.

Some quality characteristics (*uventet* 'unexpected') seem to also be negative for female but positive for male works descriptions. Description of literary genre (*selvbiografisk* 'autobiographical', *skjønnlitterær* 'fiction'), pain inducing (*farlig* 'dangerous', *tragisk* 'tragic') and emotional (*dyster* 'gloomy', *vittig* 'witty') are also negatively used to describe the works of female authors, while positive for works by

[4]Model 2 from `http://vectors.nlpl.eu/repository/`

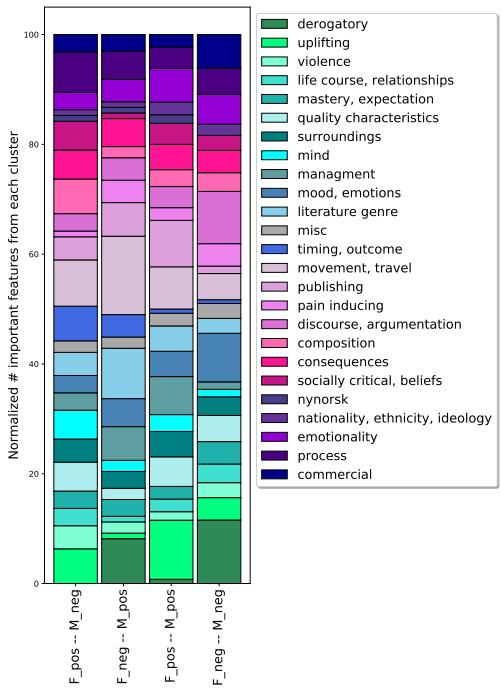

Figure 2: Distribution of clusters of most informative words for sentiment classification in R_F and R_M in NoReC.

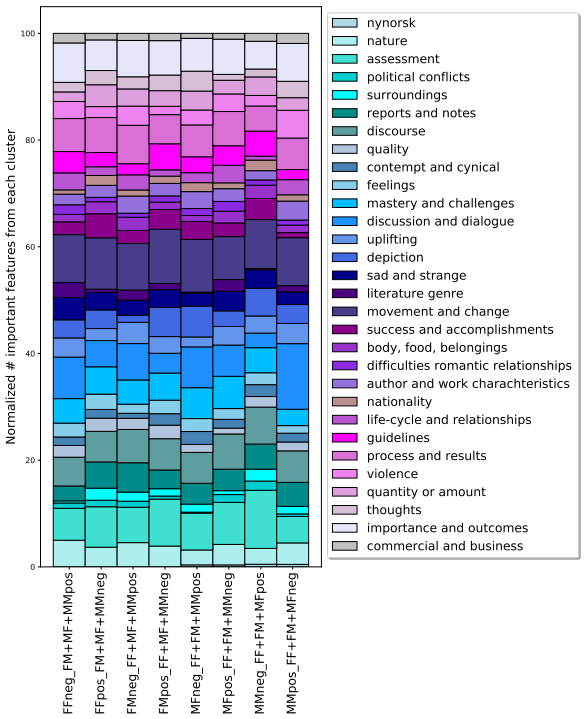

Figure 3: Distribution of clusters of most informative words for sentiment classification in R_{FF}, R_{FM}, R_{MM}, and R_{MF} in the gender-annotated NoReC Corpus.

males. Another interesting observation, is that adjectives related to violence are used to positively describe male, while they are negative for females (e.g. *død* 'death').

Nouns related to literary genre (*klassiker* 'a classic'), life-cycles and relationships (*kone* 'wife', *søster* 'sister'), and violence (*offer* 'victim') tend to be positive when describing books written by female authors. In contrast, nouns related to consequences (*reaksjon* 'reaction'), pain inducing descriptions (*smerte* 'pain', *skyldfølelse* 'guilt'), and commercial (*penger* 'money'), but also literary genre (*essay* 'essay', *dikting* 'poetry') are negative for females and positive for males.

When it comes to verbs, the four clusters movements and travel, discourse and argumentation, consequences, and process seem to be used both positively and negatively when describing female and male authors. Verbs pertaining to mood and emotions (*angre* 'regret'), discourse and argumentation (*avdekke* 'uncover', *snakke* 'to talk'), reflect females positively and males negatively. Moreover, verbs associated with mind (*evne* 'ability', *reflektere* 'reflect') and violence (*drepe* 'kill', *kidnappe* 'kidnap') tend to reflect females' work as negative and males' work as positive. This might indicate that critics dislike crime fictions written by female authors.

5.2 Critics

We take the same splits R_F and R_M as in Section 5.1 and split them further based on the gender of the critic. This results in the four splits introduced in Section 5: R_{FF}, R_{FM}, R_{MM}, and R_{MF}. We once again balance the distribution of positive and negative reviews, by selecting all reviews with ratings 1, 2, and 3 as negative, all reviews with rating 6 as positive, and a random sample of reviews with rating 5 to make the distribution of positive and negative balanced.

We follow the same steps introduced in Section 5.1, and train a simple Logistic Regression model with a 10-fold cross validation. We use the combination of train and dev splits for training, and keep the

test split for final evaluation. We used different testing strategies to investigate whether the gender of the critic has a say on both the sentiment and the words used to describe the works of an author. These are: (1) train on R_{FF}, test on: R_{FF}, R_{FM}, R_{MM}, R_{MF}, (2) train on R_{FM}, test on: R_{FM}, R_{FF}, R_{MM}, R_{MF}, (3) train on R_{MF}, test on: R_{MF}, R_{FF}, R_{MM}, R_{FM}, (4) train on R_{MM}, test on: R_{MM}, R_{FF}, R_{MF}, R_{FM}. We analysed the accuracy, macro F1, and class-level F1 of each of our testing strategies. However, as in the case of authors, we could not identify any considerable differences between the values. There were small nuances in the values, but making sense out of them was not trivial. We therefore rather focus on the differences in language use, and how this is reflected in the most informative words during training.

After training, we identify the 200 most informative words for each of the subsets R_{FF}, R_{FM}, R_{MM}, and R_{MF}. We use the same pre-trained word embeddings as in Section 5.1, and cluster the most informative adjectives, nouns, and verbs. We identified 30 clusters. Each of the clusters represent the theme or topic of the set of words it comprises. The themes were manually attributed after careful analysis of the clusters. These clusters are shown in Figure 3. To analyse the differences in each of the subsets R_{FF}, R_{FM}, R_{MM}, and R_{MF}, we focus on what is positive (negative) for each subset, but negative (positive) in the other subsets. This allows us to see the distinctive word differences, and which clusters seems to dominate these differences. As can be seen in Figure 3, the overall distribution of clusters seems to have small variations, but in what follows we show examples of the actual words that were used, and how these differ.

Most words used negatively by female critics to describe female authors in short R_{FF}, and which are positive in R_{FM}, R_{MM}, and R_{MF} are related to assessment; where words like *flink* 'clever' and *bra* 'good' which in themselves are positive words are negatively used in R_{FF}. Words like *dramatikk* 'dramatic', *absurd* 'absurd', *trist* 'sad', and *håpløs* 'hopeless' are representative of the cluster sad and strange, which female critics negatively employ to describe the works of female authors. Other interesting negatively representative words of the subset R_{FF} are the words of the life-cycle and relationships cluster. The words *barn, jente, gutt* 'children, girl, boy', *dame, mann* 'woman, man', *far* 'dad', *forelske* 'fall in love', *gift* and *gifte* 'married' and 'get married' respectively, as well as *føde* 'give birth' are positively used in the other subsets, but when female critics review female works, these seem to be negatively perceived. Moreover, some words with generally more positive connotations are representative of negativity in R_{FF}, as e.g. *stil* 'style' and *presis* 'precise' (cluster quality), *fascinere* 'fascinate', *fin* 'nice', *solid* 'solid' (cluster uplifting), and *elske* 'love', and *glad* 'happy' (cluster feelings). Conversely, negative words that are also negatively used in R_{FF} are related to violence as *død* 'dead', *drepe* 'kill', *mord* 'murder', and *morder* 'murderer'; and words related to contempt and cynical as *selvopptatt* 'selfish' and *ulykkelig* 'unhappy'.

On the other hand, words that are positively used by R_{FF} while negatively used in the remaining subsets are mainly related to the clusters movement and change, and process and results. These exhibit some differences in how simple words can be mostly used to write about a given gender, and not another. The positive words from the life-cycle and relationships cluster are *bror, søster* 'brother, sister', *venn, vennine* 'friend(male and female)', and *forelskelse* 'infatuation'. While this cluster is also negatively used in (R_{FF}), the words are different. The words negatively used seem to be about advanced relationships, either with family members or love relationships (where getting married and having children seem to be negative), while in the positive R_{FF} these words seems to be more about friendships, brothers and sisters, and early or short-term love interests. The same applies to the sad and strange clusters, which in positive (R_{FF}) comprises *dramatisk* 'dramatic', *mystisk* 'mysterious', and *dyster* 'gloomy'. Some of the positive qualities in R_{FF} that are negative in the remaining subsets are *humoristisk* 'humorous', *klasisk* 'classical', *poetisk* 'poetic', and *sentimental* 'sentimental'. On the contrary, Some negative words from the cluster violence are also used positively *blodig* 'bloody', and *dø* 'dead'. Another interesting set of words that are positive in R_{FF} but not elsewhere, are the words *familieliv* 'family life' and *kjærlighet* 'love', which reflect the content of the cluster difficulties in romantic relationships.

In subset R_{FM}, there is an interesting difference in positively and negatively words used from the cluster contempt and cynical. For example, words like *desperat, hevn, løgn* 'desperate, vengeance, lie' are negatively used, while the more feeling oriented words are positive: *ensom, hjertekjærende, ulykkelig*

'lonely, heartbreaking, unhappy'. When it comes to the feeling cluster, female critics use the words *ambisjon, drømme* 'ambition, dream' to negatively describe the work of male authors, while they use *elske, glede* 'love, joy' to positively describe works. As in the previous subset, words from life-cycle and relationships referring to marriage and wives are negatively used, while words referring to love and giving birth are positive. Coming from a female critic, this might be an indication for not adhering to the "traditional" views of relationships. This however, goes in contrast with the words from the political conflicts cluster, where words related to power and Christianity are positively used, while words related to rebellion and religious people are negative.

Some of the same observations can be found in R_{MF}. When it comes to difficulties in romantic relationships, discussing family life, sex, and physical relationships is seen as negative by male critics when discussing the work of females, while using words about love and gender are positive. However, in contrast to the two previous subsets, discussing marriage and fatherhood is positive, while talking about love and giving birth is negative. This difference is particularly interesting, because we can see the effect of having a male or female critic. Another compelling difference, is that male critics perceive female authors who write crime fictions to be negative, while works of female authors are positively described if they are from other genres (e.g. biographies and autobiographical books, novels and novel collections).

When male critics assess the work of male authors, they positively view works that are literary and poetic, but negatively describe biographies and prose. When it comes to difficulties in relationships, erotic works are positively seen, while those triggering anxieties are negative. Moreover, mentions of love, marriage, and children are actually perceived both positively and negatively, which is in contrast to the previous subsets where there was a clear difference in the polarity of early romance and stable relationships. Concerning political conflicts, works about Islam, Christianity, rebellions, and politics are negative, while those covering wars and power are positive. Another fascinating difference in this subset compared to the others, is that male critics who assess the work of male authors seems to be negative to romantic and sentimental books (*romantisk, sentimental*), while classics, witty and entertaining books, or books about music are deemed positive.

6 Non-professional reviews – the Bokelskere corpus

Our gender-annotated literature subset of the NoReC corpus reflects how works by female and male authors are positively and negatively described. However, since these reviews are written by professionals, the language can be expected to be of a more formal and possibly less affective style. In order to investigate to what extent this proves to be correct, we carried out the same analysis done on NoReC on a corpus of non-professional reviews.

We use a corpus comprising user-generated book reviews from `bokelskere.no` compiled by the National Library of Norway (Språkbanken)[5]. We will refer to this corpus as the Bokelskere (*book lovers*) corpus in what follows. The Bokelskere corpus contains the raw texts from discussions and book reviews written by users of the `bokelskere.no` web community. The ratings follow the same scheme as in NoReC, with numerical scores ranging from 1 to 6. The reviews are structured as both reviews and comments on reviews. The corpus is in JSON format and contains a total of 219,000 review comments. For each of these, the corpus provides (amongst others) information about the book being reviewed (title and author), and the rating.

We annotated the Bokelskere corpus with PoS tags using the same version of UDPipe that was used to annotate NoReC (Velldal et al., 2018). Moreover, neither the gender of the users (i.e. critics), nor the gender of the book authors are provided in the Bokelskere corpus. We therefore used our annotations from the NoReC corpus to automatically annotate Bokelskere. We only annotate the authors from our gender-annotated corpus for whom we know the gender. We were able to identify the gender of 9,833 female authors, and 15,544 male authors. However, 1,691 and 2,815 reviews of female and male authors respectively did not contain ratings and were therefore disregarded.

Figure 4 gives an overview of the rating distributions of the remaining 8,142 female and 12,729 male

[5]The corpus can be found here: `https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-53/`

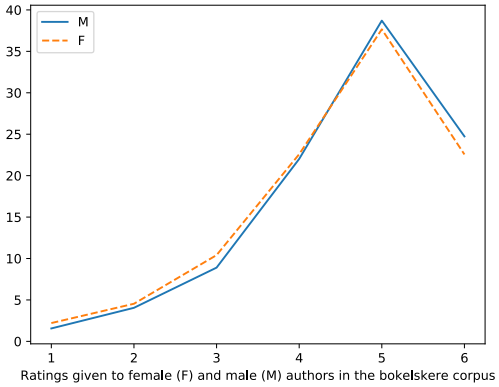

Figure 4: Distribution of ratings given to Female (F) and male (M) authors in the Bokelskere corpus. The y axis represents normalized percentages of each rating.

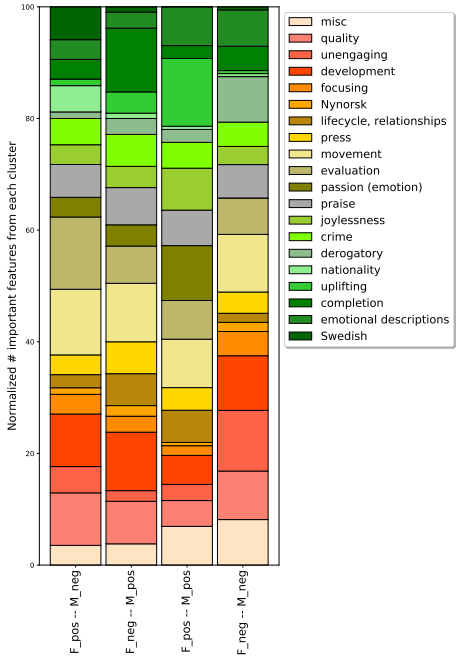

Figure 5: Distribution of clusters of most informative words for sentiment classification in F_{subset} and M_{subset} in Bokelskere Corpus.

reviews. The trend is similar to the distribution of ratings in the gender-annotated NoReC. Female authors tend to get more ratings on the lower range than male authors, while it is the opposite on the higher range. For ratings 1, 2, 3, and 4 female authors are given respectively 2.2%, 4.5%, 10.4% , and 22.6% of the total reviews, and male authors are given 1.5%, 4%, 8.9% , and 22% respectively. On the contrary, for ratings 5 and 6 female authors are respectively given 37.6%, and 22.5% of total ratings, while male are given 38.7%, and 24.7%. Since the gender of the users reviewing books is not available for the Bokelskere corpus, we focus our analysis on the gender of the reviewed authors and follow the same methodology described in Section 5.1.

The Bookelskere corpus do not have predefined train, dev, and test splits. We therefore follow the same strategy as for the splits in NoReC by first sorting the reviews by date and then reserving the first 80% for training, the following 10% for dev split, and the remaining 10% for testing. However, during this work, we combine the train and dev splits, and train a Logistic regression with 10-fold cross validation, and do a final evaluation on the test split. We also balance the distribution of positive and negative reviews for each of the genders.

Clustering the 200 most informative words for binary sentiment classification on Bokelskere, enabled us to identify 20 clusters. These are shown in Figure 5. The distribution of overlap of these clusters based on how often they are used to positively or negatively describe books written by female or male authors also offers an interesting overview (see Figure 5).

When it comes to adjectives, words that are positively used for female authors but negatively used to describe the works of male authors are mostly related to quality and nationality (*mild* 'mild', *solid* 'solid', *engelsk* 'English', *fransk* 'French'). But also words related to the expression of emotions as passion (*inderlig* 'sincerely'), praise (*begeistre* 'exciting'), and general descriptions (*komisk* 'comical'). Adjectives that are negatively used to describe the works of females, but positively used for male, tend also to relate to quality (*åpenbar* 'obvious', *personlig* 'personal', *realistisk* 'realistic'), but also development (*paralell* 'parallel', *tilgjengelig* 'available'), praise (*imponere* 'impress', *positiv* 'positive'), derogatory (*håpløs* 'hopeless', *vond* 'bad') and uplifting words (*gøy* 'fun', *inspirere* 'inspire').

The nouns *venn* 'friend' (related to life-cycle and relationships) is used to positively describe female

works and negatively describe male works, while the words *datter* 'daughter' and *søster* 'sister' are negative in female description but positive for male descriptions. Nouns related to crimes seems also to be positive for male, while negative for females (e.g. *gjerningsmann* 'perpetrator'). Most verbs used to positively describe females' while negatively describe males' works are development (*sammenligne* 'compare', *presentere* 'present'), and evaluation (*forestille* 'imagine', *oppfatte* 'perceive'). Conversely, verbs used negatively when describing works of females but positively for males seem to be related to completion (*ende* 'end', *gjennomføre* 'conduct'), development (*lage* 'make'), evaluation (*forstå* 'understand'), life-cycle and relationships (*føde* 'give birth', *gifte* 'marry'), and passion (*forelske* 'fall in love').

7 Conclusion and limitations

We present a gender-annotated dataset of professional book reviews, where both the gender of critics and the book authors are annotated. We also present a corpus of user reviews annotated for the gender of the book authors. We make all annotations and reviews publicly available. We have shown that there are differences in how female and male book authors are positively or negatively described, and that the gender of the critics influences the differences. For example, male critics deem female crime novels and male romantic and sentimental books as negative. This shows that book reviews contain the social hierarchies that tend to focus on emotional traits to describe females as in Menegatti and Rubini (2017).

There are several ways in which the preliminary analysis of the current work can be improved and extended. First, the annotations of the book authors are based on which book is being reviewed, and not if the author is being mentioned in the review. This can lead to issues during classification, since it might be possible that the review in itself contains references to the characters of the book, which might or might not be of the same gender as the author. Therefore, the word usage might not actually reflect the book author, but rather the fictional characters of the book. Secondly, we were able to identify differences in how female and male critics describe the works of female and male authors, but we did not quantify to which degree this is true. The distribution of ratings gives an indication of this, but more extensive analysis is necessary. In future works, we aim to explore how to quantify the amount of bias, but also identify if a review is discussing the quality of the book (as in the work of the author), or if it only focuses on the characters and the storyline.

Acknowledgements

This work has been carried out as part of the SANT project (Sentiment Analysis for Norwegian Text), funded by the Research Council of Norway (grant number 270908).

References

Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August. Association for Computational Linguistics.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy, August. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

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*, pages 4349–4357.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July. Association for Computational Linguistics.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France, May. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, page 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August. Association for Computational Linguistics.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy, August. Association for Computational Linguistics.

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

Aparna Garimella and Rada Mihalcea. 2016. Zooming in on gender differences in social media. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 1–10, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June. 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, Florence, Italy, July. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy, July. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hugo Liu and Rada Mihalcea. 2007. Of men, women, and computers: Data-driven gender modeling for improved user interfaces. *ICWSM*, 7:26–28.

Juan M Madera, Michelle R Hebl, and Randi C Martin. 2009. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5267–5275, Hong Kong, China, November. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. pages 223–228, July.

Francisco Rangel and Paolo Rosso. 2019. Overview of the 7th author profiling task at pan 2019: Bots and gender profiling in twitter. In *Proceedings of the CEUR Workshop, Lugano, Switzerland*, pages 1–36.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June. Association for Computational Linguistics.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender bias in pretrained Swedish embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland, September–October. Linköping University Electronic Press.

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, San Diego, California, USA, June. Association for Computational Linguistics.

Erik Velldal, Lilja Øvrelid, Cathrine Stadsnes Eivind Alexander Bergem, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July. Association for Computational Linguistics.