# Using Gender- and Polarity-Informed Models to Investigate Bias

**Samia Touileb**
Department of Informatics
University of Oslo
samiat@uio.no

**Lilja Øvrelid**
Department of Informatics
University of Oslo
liljao@uio.no

**Erik Velldal**
Department of Informatics
University of Oslo
erikve@uio.no

## Abstract

In this work we explore the effect of incorporating demographic metadata in a text classifier trained on top of a pre-trained transformer language model. More specifically, we add information about the gender of critics and book authors when classifying the polarity of book reviews, and the polarity of the reviews when classifying the genders of authors and critics. We use an existing data set of Norwegian book reviews with ratings by professional critics, which has also been augmented with gender information, and train a document-level sentiment classifier on top of a recently released Norwegian BERT-model. We show that gender-informed models obtain substantially higher accuracy, and that polarity-informed models obtain higher accuracy when classifying the genders of book authors. For this particular data set, we take this result as a confirmation of the gender bias in the underlying label distribution, but in other settings we believe a similar approach can be used for mitigating bias in the model.

## 1 Introduction

As is well established, training data for NLP tasks may contain various types of bias that can be inherited by the models we train, and that may potentially lead to unintended and undesired effects when deployed (Bolukbasi et al., 2016). The bias can stem from the unlabeled texts used for pre-training of language models (LMs), or from the language or the label distribution used for tuning a downstream classifier. Typically, when a classifier is fitted on top of a pre-trained LM for a given task, only textual data is considered by the learned representations.

In this work we investigate the effect of adding metadata information about demographic variables that are known to be associated with bias in the training data. Specifically, we focus on the task of binary sentiment classification based on data where gender has previously been shown to be correlated with the label distribution. The data we use are Norwegian book reviews, where the gender of both critics and book authors have previously been annotated (Touileb et al., 2020). When considering all pairs of male/female critics/authors, Touileb et al. (2020) showed that female critics tended to assign lower ratings to female authors, relative to other gender pairs. In this work we explore the effect of adding information about gender to a document-level polarity classifier trained on top of a pre-trained BERT model for Norwegian, showing that the model is able to take this metadata into account when making predictions. Through experiments with gender classification on the same data set, we also demonstrate that the language of the reviews is itself indeed gendered.

We believe that adding this type of metadata about *e.g.,* demographic information when available can in many cases be used to mitigate bias in models. Consider the case of a model for toxic language classification; it seems intuitively plausible that incorporating information about users could help reducing the risk of false positives for self-referential mentions by marginalized groups. However, we have a different focus for the particular experiments reported here: we show how adding information about gender in a polarity classifier confirms gender bias, by showing how a gender-informed model obtains substantially higher accuracy when evaluated on a biased label distribution.

In what follows, we start in Section 3 with an overview of related work, after providing a brief bias statement in Section 2. In Section 4 we present our dataset, and give a detailed description of our experiments in Section 5. We present and analyse our results in Section 6, followed by an error analysis in Section 7. Finally, we summarize our findings and discuss future works in Section 8.

*Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 66–74
August 5, 2021. ©2021 Association for Computational Linguistics

## 2 Bias statement

This work focuses on gender bias, which we identify as the differences in language use between persons, on the unique basis of their genders. The concrete task that we deal with in the current paper is that of polarity classification of book reviews, using labels derived from the numerical ratings assigned by professional critics. We use an existing dataset of book reviews dubbed NoReC$_{gender}$ (Touileb et al., 2020), which is a subset of the Norwegian Review Corpus (Velldal et al., 2018), a dataset primarily used for document-level sentiment analysis. The subset NoReC$_{gender}$ has previously been augmented with information about the gender of both critics and book authors. Through experiments with gender predictions of both critics and book authors, we demonstrate the presence of gendered language in these reviews. Previous work has also shown that the distribution of ratings in the dataset to some degree is correlated with the gender of the critics and the authors. Consequently, work on sentiment classification on the basis of the dataset could risk inheriting aspects of gender bias unknowingly, either in the model predictions themselves or in how these are evaluated, or both. One of our motivations in this work is exactly to assess whether the predictions of sentiment classifiers trained on review data may to some degree depend on gender, by explicitly incorporating this as a variable in the model.

Note that there are also issues of what could be argued to be representational harm (Blodgett et al., 2020) associated with the underlying encoding of gender itself, since only the binary gender categories of male/female are present in the data. While the dataset we use only reflects binary gender categories, we acknowledge the fact that gender as an identity spans a wider spectrum than this.

## 3 Related work

State-of-the-art results for various NLP tasks nowadays typically build on some pre-trained transformer language models like BERT (Devlin et al., 2019). Despite their great achievements, these models have been shown to include various types of bias (Zhao et al., 2020; Bartl et al., 2020; Basta et al., 2019; Kaneko and Bollegala, 2019; Friedman et al., 2019; Kurita et al., 2019).

Recent works have shown the advantage of adding extra information to pre-trained language models for numerous tasks, *e.g.,* dialog systems (Madotto et al., 2018), natural language inference (Chen et al., 2018), and machine translation (Zaremoodi et al., 2018). Knowledge graphs have also been used to enrich embedding information. Zhang et al. (2019) use entries from Wikidata, as well as their relation to each others, to represent and inject structural knowledge aggregates to a collection of large-scale corpora. They show that their approach reduces noisy data and improves BERT fine-tuning on limited datasets. Bourgonje and Stede (2020) enrich a German BERT model with linguistic knowledge represented as a lexicon as well as manually generated syntactic features. Peinelt et al. (2020) enrich a BERT with LDA topics, and show that this combination improves performance of semantic similarity. Ostendorff et al. (2019) use a combination of metadata about books to enrich a BERT-based multi-class classification model. They train a BERT model on the title and the texts of each book, and concatenate the output with metadata information and author embeddings from Wikipedia, and feed them into a Multilayer Perceptron (MLP).

When it comes to gender and gender bias, previous research has been devoted to the identification of bias in textual content and models (Garimella and Mihalcea, 2016; Schofield and Mehr, 2016; Kiritchenko and Mohammad, 2018), and in input representations as static and contextualised embeddings (Takeshita et al., 2020; Bartl et al., 2020; Zhao et al., 2020; Basta et al., 2019; Kaneko and Bollegala, 2019; Friedman et al., 2019; Bolukbasi et al., 2016). A considerable amount of previous work has also gone into either mitigating existing bias in embeddings (Takeshita et al., 2020; Maudslay et al., 2019; Zmigrod et al., 2019; Garg et al., 2018), making them gender neutral (Zhao et al., 2018), or using debiased embeddings (Escudé Font and Costa-jussà, 2019). Instead of debiasing and mitigating bias in embeddings, some work has focused on creating gender balanced corpora (Costa-jussà et al., 2020; Costa-jussà and de Jorge, 2020).

Several previous studies have focused on gender and gender bias in sentiment analysis, both from data and model perspectives. To name a few: Kiritchenko and Mohammad (2018) propose an evaluation corpus (Equity Evaluation Corpus) that can be used to mitigate biases towards a selection of genders and races. Occupational gender stereotypes exist in sentiment analysis models (Bhaskaran and Bhallamudi, 2019), both in training data and in pre-trained contextualized models.

Models have also been proposed to uncover gender biases (Hoyle et al., 2019). Incorporating extra demographic information into sentiment classification models have also been successful. Hovy (2015) has shown that incorporation gender information (as embeddings) in models can improve sentiment classification. They show that such an approach can reduce the bias towards minorities, as for example females, who tend to communicate differently from the norm.

In this paper, we do not focus on biases present in existing systems , nor do we try to mitigate them in a traditional way. We use a dataset of Norwegian book reviews for which a previous study has indicated some degree of gender bias in the label distribution of review ratings (Touileb et al., 2020). Here, we investigate whether this bias is reflected in the text, as measured by classification scores on two tasks, namely binary sentiment and gender classification, and whether adding metadata information explicitly providing the gender of the authors and critics of the reviews, or the sentiment score of the review increases classification performance. Similarly to (Ostendorff et al., 2019), we explore the effects of adding this metadata information to document classification tasks using a BERT-based model, in this case the Norwegian NorBERT (Kutuzov et al., 2021).

## 4 Dataset

In this work, we focus on gender effects in reviews written by male or female critics, which in turn rates the works of male and female authors. The dataset we use is the NoReC$_{gender}$[1] (Touileb et al., 2020) subset of the Norwegian Review Corpus (NoReC (Velldal et al., 2018)). NoReC$_{gender}$ is a corpus of 4,313 professional book reviews from several of the major Norwegian news sources. Each review is rated with a numerical score on a scale from 1 to 6 (represented by the number of dots on a die), assigned by a professional critic. The reviews also contain additional metadata information like the name of the critics, name of the book authors, and their respective genders.

The numerical ratings and name of the critics were already provided in the metadata data of NoReC (Velldal et al., 2018), while the name of the authors and the information about the genders were manually annotated with the release of NoReC$_{gender}$ (Touileb et al., 2020).

| | M | F | Total |
|---|---|---|---|
| Unique critics | 125 | 74 | 199 |
| Unique authors | 1,435 | 882 | 2,317 |

Table 1: Total number of unique male and female critics and authors in NoReC$_{gender}$.

| | Train | Dev. | Test | Total |
|---|---|---|---|---|
| pos | 568 | 69 | 71 | 708 |
| neg | 568 | 60 | 55 | 683 |

Table 2: Total number of positive and negative reviews in the data splits of NoReC$_{gender}$.

As pointed out by Touileb et al. (2020), some of the reviews were written by children, unknown authors/critics, or by editors, these were not assigned genders and were therefore not included in our work. This results in a set of 4,083 documents. Table 1 shows an overview of the NoReC$_{gender}$ dataset in terms of total number of critics and authors, and their distribution across genders.

Each review in NoReC$_{gender}$ comes with a numerical dice score from 1 to 6. Similarly to Touileb et al. (2020), we choose to focus on clear positive and negative reviews and therefore only use reviews with negative ratings representing dice scores 1, 2, and 3, and reviews with positive ratings representing scores 5 and 6. However, in order to control for the distribution of positive and negative labels, we have selected a subset of reviews with rating 5 to have a balanced distribution of positive and negative reviews in the train set. This results in a subset of 683 negative and 708 positive reviews for NoReC$_{gender}$. A distribution of these across the train, dev, and test splits can be seen in Table 2.

The dataset NoReC$_{gender}$ also contains a bias in the distribution of labels, based on the gender of the critics and the authors (Touileb et al., 2020). Figure 1 shows the total number of ratings in our dataset, where the first letter (M/F) indicates the gender of the critic and the second letter indicates that of the author. For example, *MF* represents reviews written by male critics reviewing the works of female authors. Here we observe a clear difference in the ratings given by female critics to female authors (*FF*). While most reviews seem to have a certain amount of balance between positive and negative polarities with slightly more positive than negative

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

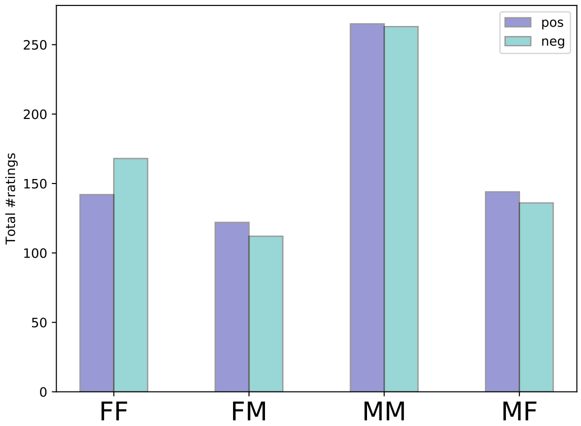

Figure 1: Distribution of ratings given by critics to works of authors. The first letter (M/F) indicates the gender of the critic and the second that of the author.

reviews, for *FF* it is the opposite. This, in addition to the unbalance between the total number of reviews based on gender, represent the bias present in NoReC$_{gender}$'s label distribution.

## 5 Experiments

We use the Norwegian BERT model NorBERT[2] (Kutuzov et al., 2021). The model uses the same architecture as BERT base cased (Devlin et al., 2019), and uses a 28,600 entry Norwegian-specific sentence piece vocabulary. It was jointly trained on both official Norwegian written forms Bokmål and Nynorsk, on 200M sentences (around 2 billion tokens) from Wikipedia articles and news articles from the Norwegian News Corpus.[3]

We use a similar architecture to Ostendorff et al. (2019) as shown in Figure 2. We feed our review texts to a NorBERT architecture of 12 hidden layers consisting of 768 units each. These representations and the metadata are subsequently concatenated and passed to a two-layer Multilayer Perceptron (MLP), using ReLu as activation function. The output layer (SoftMax) gives for each task its binary output, *i.e.,* either binary sentiment classification labels, or binary gender classification labels. We set the learning rate for AdamW (Loshchilov and Hutter, 2019) to $5e-5$, and batch size to 32. We train the model for 5 epochs, and keep the best model on the dev set with regards to $\mathrm{F}_1$.

We have experimented with various input sizes (first 300 tokens, first 512 tokens, and first 128 + last 383 tokens) both with tokenized and untokenized texts. The best results were achieved using untokenized texts, and using the first 128 and last 383 tokens, as pointed out by Sun et al. (2020). These are the input sizes used in the models we report in this work.

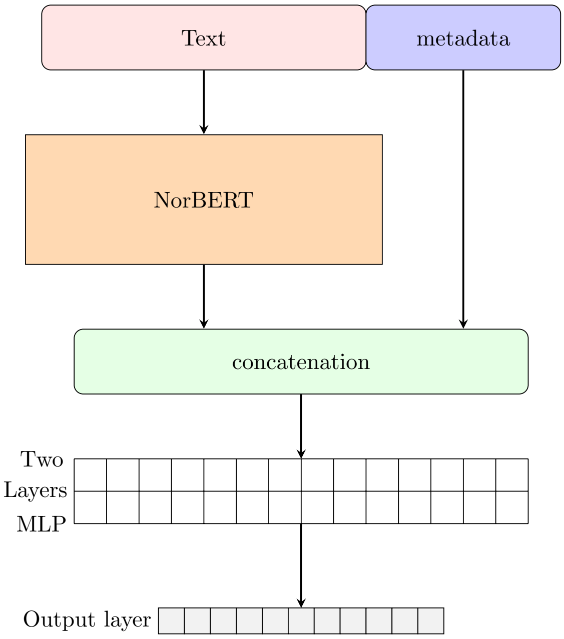

Figure 2: Architecture of our metadata-enriched classification model. Our baseline model has the same architecture except for the metadata input and the concatenation step.

Our metadata is one-hot encoded, and has a dimension of two for gender (female and male), and two for polarity (positive and negative). In the case where we combine information about the genders of both authors and critics, the dimension is four (*i.e.,* two gender dimensions each).

For the task of binary gender classification, we perform a set of four experiments:

- *NorBERT–none*: without any metadata.
- *NorBERT–ga*: adding information about the gender of authors.
- *NorBERT–gc*: adding information about the gender of critics.
- *NorBERT–gac*: adding information about the gender of both the authors and the critics.

For each of the binary classification of genders of authors or critics, we perform the following two experiments:

- *NorBERT–none*: classifying the gender of authors or critics without any metadata.

[2] https://huggingface.co/ltgoslo/norbert

[3] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

| Model | dev | test |
|---|---|---|
| *NorBERT–none* | 82.45 | 80.66 |
| *NorBERT–ga* | 84.51 | **84.21** |
| *NorBERT–gc* | 84.92 | 82.33 |
| *NorBERT–gac* | **85.25** | 82.92 |

Table 3: Model performance on dev and test for binary sentiment classification. *NorBERT–none* is the baseline model. All models report mean $F_1$.

| | Model | dev | test |
|---|---|---|---|
| Author | *NorBERT–none* | 89.57 | 90.12 |
| | *NorBERT–polarity* | **94.93** | **94.60** |
| Critic | *NorBERT–none* | **70.40** | **63.84** |
| | *NorBERT–polarity* | 64.99 | 57.76 |

Table 4: Model performance of binary gender classification on dev and test for authors and critics. Models report mean $F_1$.

- *NorBERT–polarity*: classifying the gender of authors or critics by adding information about the polarity (positive and negative) of the review.

In all of our experiments, we use the task specific *NorBERT–none* as baselines.

## 6 Results

Table 3 shows $F_1$ scores of our binary sentiment classification models on both dev and test splits of NoReC$_{gender}$. The baseline model *NorBERT–none* that only uses NorBERT without metadata performs quite well on both dev and test splits with $F_1$ scores of 82.45 and 80.66 respectively. But as can be seen, the model is the least accurate in our set of experiments.

We observe that the *NorBERT–ga* model, which incorporate information about the gender of the authors is the most accurate model on the test set, with an $F_1$ score of 84.21, while it is the third most accurate on the dev split with an $F_1$ score of 84.51. *NorBERT–gc*, which adds information about the gender of the critics, also yields better results than the baseline with an $F_1$ score of 84.92 on dev, and 82.33 on test. The best performing model on the dev set is *NorBERT-gac*, with added information about the genders of both authors and critics. This model is also the second best model on test with a $F_1$ score of 82.92.

The results presented in Table 3 show that gender-informed models with metadata information improve the task of binary sentiment classification with respectively 2.06, 2.47, and 2.8 $F_1$ points on the dev set, and 3.55, 1.67, and 2.26 $F_1$ points on test for the three models *NorBERT-ga*, *NorBERT-gc*, and *NorBERT-gac*. This suggests that for a binary classification task on NoReC$_{gender}$, knowing the gender of the authors and critics clearly influences the performance of the model.

The scores of our gender classification tasks are presented in Table 4. As previously mentioned, for the gender classification, we have two tasks: classification of the gender of the authors, and classification of the gender of the critics.

For the classification of the authors' genders, the baseline classifier *NorBERT–none* performs quite good with a $F_1$ score of 89.57 and 90.12 on dev and test respectively. However, adding the metadata about the polarity of the review (if it's positive or negative) influences the classification task by 5.36 and 4.48 points on dev and test respectively.

Interestingly, we observe the opposite situation for the classification of the gender of critics. Here, the baseline model *NorBERT–none* outperforms the *NorBERT-polarity* model by 5.41 and 6.08 $F_1$ score points on respectively dev and test splits.

For the task of author gender classification, knowing the polarity of the review clearly influences the classification. Again, this indicates that gender and polarity are correlated in our data. The results also point to a difference between the gender of authors and critics. However, additional information about the polarity of the review, seems to hurt the classification of the genders of critics.

## 7 Error analysis

In order to gain further insight into the differences between the models we are comparing and in particular, the classification differences caused by the addition of information on gender/polarity, we perform an error analysis by comparing, for each task, how our models perform compared to the task-specific baselines.

Figure 3 shows how the three models *NorBERT–ga*, *NorBERT-gc*, and *NorBERT-gac* have different predictions than their baseline *NorBERT–none* for binary sentiment classification. We show the relative differences of true positives as a heatmap. These are made on the test predictions of each model over all five runs. Positive numbers (dark purple) specify that the model made more correct predictions than the baseline *NorBERT–none*, while negative numbers (white) indicate it made fewer correct predictions. The abbreviations *FF*, *FM*, *MF*, and *MM* represent the gender of the critic reviewing the work of an author of a given gender. *FF* refers to female critic and female author, *FM* female critic and male author, *MF* male critic and female author, and *MM* for male author and male critic.

It is clear that all three gender-informed models become more accurate in the classification of reviews written by female critics and reviewing the works of female authors (*FF*). As previously mentioned, and as pointed out by Touileb et al. (2020), female critics tend to be more negative towards female authors, and therefore there are few reviews that fall within this category with positive polarity. Adding information about the gender of the authors and the critics, seems to help the model identify some of the *FF* reviews that *NorBERT–none* was not able to classify correctly. This information seems to be particularly important for *NorBERT–ga*, which was the best model on the test set achieving 12 $F_1$ points more than the baseline on *FF*. This model also seems slightly better at identifying reviews for *MM*. A closer analysis differentiating the positive and negative polarities also shows that the three models are more accurate precisely in identifying the positive reviews in the *FF* subset.

The same applies to a lesser degree for *FM*. Knowing the gender of the authors and the critics, separately, enables the models to correctly classify more reviews than *NorBERT–none*. In contrary, for *MF*, only knowing the gender of both the critics and authors seems to slightly improve classification. For the *MM* reviews, the *NorBERT–ga* model is better at identifying the positive reviews, while *NorBERT–gac* is better at identifying the negative reviews.

Figure 4 shows the breakdown of the relative differences of true positives. Here again, the relative differences are made on the test predictions of each model over all five runs. Positive numbers (dark blue) represent the cases where the model made more correct predictions than the baseline *NorBERT–none*, while negative numbers (white) indicates the opposite. For clarity, we add a prefix to each model in the figure to specify the task. *GA-NorBERT–pn* represent the model *NorBERT–pn* for the task of author gender classification, while *GC-NorBERT–pn* represents the task of critic gender classification.

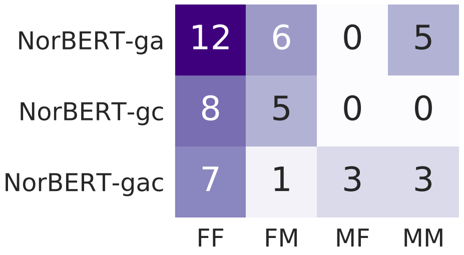

Figure 3: Relative differences of true positives for binary sentiment classification on test compared to their baseline *NorBERT–none*. Darker colors represent more correct predictions than the baseline.

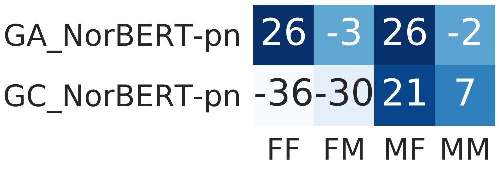

Figure 4: Relative differences of true positives for binary authors and critic gender classification on test compared to their relative baselines *NorBERT–none*.

For the author gender classification task, as can be seen in Figure 4, having extra information about the polarity of the review helps the model *NorBERT–pn* (*GA_NorBERT–pn*) to better predict the gender of the author if she's a female. This again is compared to the task specific baseline *NorBERT–none*. It also seems that this model makes a few more mistakes than the baseline when it comes to the author being a male. For gender classification of the critics, adding metadata information seems to negatively affect the model's ability to identify female critics. The model *NorBERT–pn* (*GC_NorBERT–pn*) is more accurate when it comes to identifying the gender of male critics compared to the baseline, achieving 21 and 7 $F_1$ points more than the baseline on respectively *MF* and *MM*.

This corroborates our previous observations, that adding metadata information about the polarity of reviews aids the identification of female authors for author gender classifiers. While for critic gender classification it fails at identifying female critics, but is accurate in identifying males.

## 8 Conclusion

In this work, we have investigated the effect of adding information about the gender of critics and book authors when classifying the polarity of book reviews, and the polarity of the reviews when classifying the genders of authors and critics. Using a document-level classifier on top of a recently released Norwegian BERT-model, we have shown that gender-informed models obtain substantially higher accuracy, and that polarity-informed models obtain higher accuracy when classifying the gender of the book authors. In further analysis, we have observed clear differences in the classification results for male/female authors/critics. Specifically, we demonstrated that adding to NorBERT information about the genders of critics and book authors influences a binary sentiment classification task by being more accurate in predicting positive reviews for female authors.We have also shown that using polarity information helps the identification of female authors, but seems to greatly hurt the identification of female critics. Some directions for future work include quantifying the bias in the original NorBERT model. As our experiments showed, using the baseline model with only NorBERT and no metadata achieves good results, and we therefore plan to evaluate the existing biases in NorBERT.

## Acknowledgments

All of our experiments were run on resources provided by UNINETT Sigma2 of the Norwegian National Infrastructure for High Performance Computing and Data Storage, under the NeIC-NLPL umbrella.

## References

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. 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. 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. 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.

Peter Bourgonje and Manfred Stede. 2020. Exploiting a lexical resource for discourse connective disambiguation in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5737–5748, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). 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. 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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. 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. 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. The COLING 2016 Organizing Committee.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. 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. 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. 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. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification. *arXiv preprint arXiv:1909.08402*.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55, Barcelona, Spain (Online). Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2020. Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 125–138, Barcelona, Spain (Online). 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.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. 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. Association for Computational Linguistics.

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