Analyzing gender and sentiment in Norwegian book reviews

Tellef Seierstad

Thesis submitted for the degree of
Master in Language Technology
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2023

Analyzing gender and sentiment in Norwegian book reviews

Tellef Seierstad

Analyzing gender and sentiment in Norwegian book reviews

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

This thesis investigates gender and sentiment in Norwegian book reviews and their impact on machine learning models trained on the book review data. Our analysis reveals that female critics and authors give and receive significantly lower ratings than males. Using methods from interpretable machine learning, we go on to show that these statistical differences make models trained on the data associate features related to female gender with a lower sentiment than features related to male gender. We also explore the effects of gender normalization on the models' predictions and the impact of supplying models with gender knowledge during training. Our findings demonstrate the potential of interpretation methods for transformer models on Norwegian text and highlight the strengths and weaknesses of different methods for interpreting machine learning models.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Erik Velldal and Samia Touileb, whose input and guidance has been highly appreciated and helped completing this thesis.

All the experiments using transformer models have been run on the *Machine learning infrastructure (ML Nodes)* of the University Centre for Information Technology at the University of Oslo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is bias everywhere, and likely even more so with the increasingly high rate new content is created on the internet. In a general sense, bias simply means 'inclination', and can be perfectly neutral (Friedman & Nissenbaum, 1996). A grocery shopper buying apples rather than oranges displays a bias. However, the term bias can also carry moral meaning. We will use Friedman and Nissenbaum (1996)'s definition of bias as systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. Further on, gender bias can then be defined as systematic unequal treatment based on one's gender.

In this thesis, we perform an exploratory analysis of gender in Norwegian book reviews, with a particular emphasis on its relation to sentiment. We further investigate how this relation affects machine learning models trained on the data. The thesis builds on previous work on sentiment analysis, bias in machine learning and also more specifically on gender bias.

This thesis focuses on $\text{NoReC}_{\text{gender}}$, a data set of Norwegian book reviews that includes information about the gender of the authors and critics (Touileb et al., 2020). The *Norwegian Review Corpus (NoReC)* is a sentiment data set (Velldal et al., 2018), of which $\text{NoReC}_{\text{gender}}$ is a subset. This means that we mostly examine gender effects in sentiment analysis, but we also train machine learning models to predict the gender of the critics and authors of the reviewed books. We do this in order to see how different aspects of gender interact, and we examine the most informative features for all three classification tasks.

1.1 Motivation

As machine learning models have increasingly higher impacts on people's lives, it also gets more important to be aware of the adverse consequences such models can have when they contain bias. This thesis is motivated by addressing the problem of bias in machine learning. By investigating the relation between gender and sentiment in Norwegian book reviews, we can shed light on how using machine learning models might lead to unintended and unfair consequences when the training data is not distributed equally

across genders. The machine learning *algorithms* may contain no bias against specific groups by themselves, but they can perpetuate bias present in the data they are trained on. Another motivating factor is to work on interpretation methods for machine learning. With the current proliferation of machine learning, it is also important to understand the models to some degree, and not treat them as unknowable black boxes.

1.2 Research questions

We are interested in investigating how gender affects sentiment in *NoReC*. In order to do this, we hypothesize that normalizing the gender within a text can affect model predictions. By normalizing gender we mean to somehow remove the gender information contained in the text, e.g. by masking gendered pronouns like 'she'. Investigating *how* the model predictions are affected by the normalization can give insight into the relation between gender and sentiment in $\text{NoReC}_{\text{gender}}$. Conversely, we also hypothesize that *adding* gender as metadata to the model might affect its predictions. Finally, we want to understand why the models are affected the way they are, and consequently we propose the following research questions:

RQ1 Can a model still predict the gender categories and sentiment if we first normalize gendered words in the texts?

- a How could such gender normalizing pre-processing best be carried out?
- c Does gender normalization affect prediction of author gender?
- c Does gender normalization affect prediction of critic gender?
- d Does gender normalization affect prediction of sentiment?

RQ2 What is the effect of supplying knowledge of the gender during training of the models?

- a What is the effect of supplying knowledge of the author as a variable when attempting to predict the gender of the critic, and vice versa?
- b What is the effect of supplying knowledge of the author and/or critic as a variable when attempting to predict the sentiment?

RQ3 Is it possible to use methodology from interpretable machine learning to shed more light on what information is used by the models when predicting gender and/or sentiment?

- a To what extent does using linear interpretable models satisfy both predictive performance and explainability of the models?
- b To what extent do methods for interpreting deep neural networks give insight into why the models make certain predictions?

To answer these research questions, we first investigate whether there actually is gender bias in $\text{NoReC}_{\text{gender}}$. Then we use different combinations of gender normalization and supplying metadata on a variety of models, in order to examine and compare their effects. Finally, we explore several methods to interpret the models. The next section shows an outline of how the remaining chapters, where we describe these experiments, are structured.

1.3 Thesis overview

Chapter 2 - Background introduces key definitions and resources that this thesis builds upon.

Chapter 3 - Data presents the Norwegian Review Corpus (NoReC), whose subset $\text{NoReC}_{\text{gender}}$ is the data set used for the experimentation in this thesis.

Chapter 4 - Models and experiments provides an overview of the models that were used for the experiments, the computational environment and how the models were trained. Furthermore, it contains evaluations of the models given different constraints on the data.

Chapter 5 - Interpretability first introduces explainable artificial intelligence in general and some of its methods, before using those methods to interpret first the bag-of-words models and then the transformer models.

Finally, **Chapter 6 - Conclusion** concludes the thesis, providing a summary of our work and possible future work.

Chapter 2

Background

In this chapter we introduce key concepts necessary for this thesis. There has been done much research in sentiment analysis, bias in NLP and classification of long documents, some of which will be mentioned in this chapter. The most relevant resources that can be used for this research are also introduced.

2.1 Sentiment analysis

Sentiment can be characterized as a positive or negative evaluation expressed through language. Sentiment analysis is a field that attempts to automatically determine the sentiment of a text, and is for example often used to determine whether a review is positive or negative towards the item being reviewed. There exist two main approaches to the problem, machine learning and lexicon based (Taboada, 2016). In later years, deep machine learning models have become most popular (Yadav & Vishwakarma, 2020).

In lexicon-based sentiment analysis, sentiment values of a text are derived from the sentiment orientation of the individual words in the text, using an existing dictionary that contains words and their polarity, and then aggregating their values using an appropriate algorithm. (Taboada, 2016). The algorithms used need to be more complex than a simple sum of all the sentiment values, in order to deal with negation and other sentiment modifiers.

In practice, this thesis will only consider machine learning based sentiment analysis. The reason for that is that a lot of words do not have only one polarity but may be either positive or negative depending on the context in which they appear. These effects may be either contextual, such as ambiguous words getting their concrete meaning from the context, or compositional, like negation. Lexicon-based approaches need to use somewhat naive heuristics to handle these effects. The advantage to using lexical approaches is that it leads to more transparent models, which makes it simple to see why the model ended up with a specific output for a given input. Neither do the lexicon approaches need to be trained using an annotated data set, like machine learning approaches do, but they would still need annotated data for evaluation.

Sentiment analysis can be used at the document level, sentence level, and token level. Sentiment analysis at the document level takes in a document as input and in some way aggregates the whole document, before returning a sentiment score for the document. At the sentence level, sentiment analysis models would do this for each sentence in a text. Simple document level sentiment analysis models could simply use a sum of the sentence or token level sentiment as the aggregated sentiment for the whole document. In practice, more advanced methods are usually used at the document level, like transformer models, e.g. Bidirectional Encoder Representations from Transformers (BERT), introduced by (Devlin et al., 2019), or Recurrent Neural Networks (RNNs). A disadvantage of document level sentiment analysis is that it is hard to extract sentiment about distinct entities contained in the text separately. Sentence level classification makes it easier to distinguish sentiment towards different entities (Yadav & Vishwakarma, 2020).

In addition to these two levels, which are simply text classification for some unit of text, there is fine-grained sentiment analysis, which classifies on a token level what is the polar expression, who the holder of the sentiment is, and what is the target of the sentiment. In this way, it can be seen as a kind of entity recognition, where the entities are holder, polar expression, and target of the sentiment. Feldman (2013) and Yadav and Vishwakarma (2020) also mention aspect-based sentiment analysis, which is an extension to fine-grained sentiment analysis where aspect is added as a conceptual category on top of the target. Its task is to decompose the sentiment for each aspect of an item being reviewed. An example could be 'The main dish was great, but it took a long time coming'. In this case, both 'main dish' and 'it' (referencing the 'main dish') are targets of sentiment, but here they also represent two aspects of the restaurant being reviewed; the food, or taste, and the service. These aspects are quite domain specific – while they may for example be service, price, location and taste for restaurants, they may be something totally different for other domains, like literature or video games. For a review which includes several evaluations of different aspects, such as the example sentence above, Feldman (2013) argues that just classifying the whole as either positive or negative would miss valuable information about different aspects.

Sentiment analysis at document and sentence level is a subset of the more general text classification, which is the process of categorizing texts into organized groups (Minaee et al., 2021). Apart from sentiment analysis, other text classification tasks are news categorization and topic analysis. With deep learning-based classifiers, extractive question answering and natural language inference may also be cast as text classification problems (Minaee et al., 2021).

2.1.1 Previous work and state of the art

This machine learning performed in this thesis is document-level text classification, and thus the previous work discussed will also mainly be document-level text classification.

Since 2018, transformer-based pretrained language models, like BERT, have created a new state of the art in many NLP tasks, like text classification, and using these pretrained models lead to significant improvements across all popular text classification tasks (Minaee et al., 2021). While using these models by themselves often lead to better results than simpler models, Lyu et al. (2020) improve the sentiment analysis performance on the IMDB, Yelp-13, and Yelp-14 data sets by incorporating text from all the reviews belonging to a single user or product in the inputs to their model. This is to alleviate the challenge of different critics meaning different things by the words they use, and different products being evaluated using different words. Noting that computing representations of all reviews of a user for each training sample would be too expensive, they propose an incremental approach where they first obtain the review text representation and then use the current user and product vectors to compute biased document representations which are then used to get the sentiment. Then, they update the user and product vectors with the biased document representations. The IMDB, Yelp-13, and Yelp-14 data sets that Lyu et al. (2020) used, contain ratings made by users, and so does $\text{NoReC}_{\text{gender}}$, the data set used for this thesis, which will be introduced in Section 2.3. Using such user ratings as labels, instead of manually annotating the data, has some implications, which will be discussed in the next section.

2.1.2 Using ratings as labels

Traditionally, labels for text classification are manually annotated by specifically appointed annotators. However, for sentiment analysis, a common approach has been to use ratings made by users or professional critics that are already present in the data set, which alleviates some of the cumbersome manual annotation process. Nonetheless, using such ratings means that the researchers have less control over the labeling process, and the rating scale can be used inconsistently between different critics. Another challenge is that such user ratings are seldom a binary negative or positive evaluation, but often on a scale from e.g. 1 to 5. Sentiment analysis can be made to handle such scaled label input, but as soon as the task goes from binary to multiclass classification, the complexity of the task goes up and consequently the performance suffers.

Instead of using a numeric scale, another usual method when labelling sentiment data is to include a third, neutral category in addition to the positive and negative categories (Taboada, 2016). This could be useful if the model is used to classify texts that may not contain a specific sentiment, as opposed to classifying reviews, which one can expect to always contain some kind of sentiment.

Scaled sentiment output can carry a lot more information than just a binary positive/negative, but it also brings some new challenges, like different reviewers using the scale differently, and also that a given review score can mean different things for different items (Pang & Lee, 2005). Furthermore, if the model predicts 4 when the real score was 5, it is more correct than a prediction of 2, even if it is technically incorrect. If one

wants to take this difference into account, one would thus need a way to weight these two errors differently when training the model, e.g. by using regression instead of discrete classification (Feldman, 2013). If one treats the task as a standard multi-class problem, these class similarities will not be taken into account. However, the standard approach is to use multi-class evaluation, which in practice gives good results; Bergem (2018) did not get increased score when using a similarity-aware model compared to logistic regression on the *NoReC* data set, which is the same data set we will use for this thesis, except that we will only look at a small part of it, called $\text{NoReC}_{\text{gender}}$.

Previous work

When it comes to previous work for this problem, Pang and Lee (2005) have done research on different ways to model the problem of rating scales, and found that utilizing the similarity between texts with the same label improved the results of the models, often to a significant degree, compared to multi-class approaches that do not use any such similarity information. They collected internet movie reviews from four authors, which they used as their data set, and noticed that these four authors diverged significantly when it comes to what they mean by a given rating, and also by how the text relates to the rating. Because of that, they made a separate data set for each author, facilitating analysis of the results by not having to calibrate the authors' scales (Pang & Lee, 2005).

Mukherjee et al. (2019) have done similar research on the Amazon Reviews Dataset, which also contains ratings on a scale from 1 to 5. They found that due to the label imbalance, with high ratings comprising most of the reviews, oversampling the smaller classes during training leads to improved accuracy. Pang and Lee (2005) also had to handle label imbalance, but instead of oversampling the minority class, they folded it into the adjacent class, arriving at a four-class problem with more equal label distribution.

2.1.3 Evaluation

As mentioned above, to evaluate performance of models trained on scaled data, whose labels are not independent of each other, one need a metric that takes similarity between labels into account. Since the label distribution in these data sets is often skewed, one also needs to consider the label distribution by e.g. using a macro-averaged score. Bergem (2018) discusses these challenges in his Master's thesis and decided to go with macro-average mean absolute error (MAE_{M}), which is also used in Task 4 of SemEval-2017 for Subtask C: *Topic-based Classification on a 5-point Scale* (Rosenthal et al., 2017). While standard mean absolute error is enough to take the order of the classes into account, MAE_{M} has the advantage of also being robust to class imbalance. MAE_{M} is an error measure, and therefore lower values are better, while the standard multi-class metrics give scores in the range $[0, 1]$, where higher is better.

The previous work that this thesis builds on converted the ratings to binary positive/negative sentiment, removing the reviews that did not fit in either category in the process (Touileb et al., 2020). In this thesis we will do something similar, but keeping the class that they removed as a third class between positive and negative. This third class should not be seen as a neutral class, since there is no such thing as a neutral review, but rather as *fair*. Using three sentiment classes means to go from binary to multiclass classification, and will from here on be called ternary sentiment analysis. Instead of just training the models using ternary sentiment as labels, one could also train on the original ratings from 1–6 and map those predictions back to ternary sentiment. In this way the model could learn the difference between ratings that are merged together into one ternary sentiment class.

The corpus we will use for this thesis is The Norwegian Review Corpus (NoReC), and more specifically its subset, $\text{NoReC}_{\text{gender}}$, both of which will be presented in more detail in Section 2.2.1 below. Like the research discussed in this subsection, the data set used in this thesis, $\text{NoReC}_{\text{gender}}$, contains ratings that will be used as a weak label. Since the ratings are already there, we do not need to annotate the data manually. Unlike Amazon Reviews, $\text{NoReC}_{\text{gender}}$ does not include reviews from users, but only from professional reviewers.

2.2 Resources for Norwegian sentiment analysis

Most resources for sentiment analysis are available for English, and not as many exist for lower-resourced languages like Norwegian. Earlier, machine translation to English and then performing sentiment analysis on the translation has been done (Feldman, 2013; Taboada, 2016). This thesis will use new resources for sentiment analysis in Norwegian, described in detail in this section.

2.2.1 NoReC

The Norwegian Review Corpus (NoReC) is a corpus of Norwegian product reviews across a range of categories, e.g. music, restaurants and games (Velldal et al., 2018). This work uses resources that are part of the SANT project (Sentiment Analysis for Norwegian Text), aiming to provide training and evaluation data for the task of sentiment analysis for Norwegian, which was not available before this corpus was published. The corpus comprises more than 43 000 reviews collected from several Norwegian news sources, each of them with a score on a scale from 1–6, mainly intended to be used for evaluating models for document-level sentiment analysis. These scores, as mentioned above, are not manually annotated, but rather ratings given by professional reviewers (Velldal et al., 2018).

2.3 NoReC$_{\text{gender}}$

This thesis will focus on a gender-annotated subset of NoReC, only dealing with 4 313 book reviews of the original 43 000 total reviews in NoReC from diverse categories. The names of the critics and authors were already provided in the NoReC corpus, and Touileb et al. (2020) used a semi-automated approach followed by a manual correction as well as a fully manual approach to annotate the gender of the critics and authors. Their semi-automated approach uses a list of male and female names, matching them with the title and the excerpt for each review. Where the author's name is only in the main text of the review, not in the title or excerpt, the fully manual approach was used.

NoReC, and therefore NoReC$_{\text{gender}}$ as well, contains sentiment annotations on a scale from 1 to 6, representing the dots of a die (Velldal et al., 2018). However, Touileb et al. (2020) converts this classification problem to a binary one by selecting reviews with ratings 1, 2 and 3 as negative and rating 6 as positive, randomly sampling in ratings of 5 as positive to balance the distribution of sentiment. This is a similar approach to Pang and Lee (2005), except that Pang and Lee made four classes out of five, instead of two classes out of six.

's ()touilebGenderSentimentCritics2020's paper introducing NoReC$_{\text{gender}}$ is part of the research on bias in textual content, specifically gender bias. Noting that gender bias has been widely studied in NLP, they studied a combination of two aspects which had mostly been studied in isolation; how female and male reviewers express themselves and how works of female and male authors are described. They conclude 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 (Touileb et al., 2020). Their further work on the NoReC$_{\text{gender}}$ data set has shown that gender-informed models obtain higher accuracy than models without gender information (Touileb et al., 2021).

A limitation that Touileb et al. (2020) note about their research is that document-level sentiment analysis, as we have described above, fails to take into account the different aspects of the review, but returns only a thresholded aggregate. The reviewer could praise some specific aspects of the book while giving a tepid review overall, or they could write about characters in the book, not the book itself. These aspects can not be distinguished when classifying sentiment at the document level.

Identifying what the gender bias in NoReC$_{\text{gender}}$ consists of and how it is expressed in the texts will be a goal of this thesis. In the next section we will describe bias in general and gender bias specifically, as well as what implications bias has for NLP.

2.4 Classification of long documents

Document-level sentiment analysis can be seen as a subset of the more general text classification, and thus this section will deal with some general

aspects of text classification that are relevant for the thesis. The average length of documents in NoReC is 463 tokens (Barnes et al., 2020), which means that quite a few documents will be longer than BERT's limit of 512 tokens. As a consequence of this, some care will have to be taken when dealing with those long texts. In this section a few of the possible directions will be outlined.

Barnes et al. (2020) present two promising complementary directions for document-classification: transfer learning and hierarchical modeling. Transfer learning is transformer models like BERT which creates contextualized representations of text using large amounts of unlabeled text, whereas hierarchical models uses the document structure by first building sentence representations, before aggregating them into document representations (Barnes et al., 2020).

Using hierarchical models is one way of dealing with long documents, and Barnes et al. (2020) also found that hierarchical models outperformed those that did not incorporate the document structure, with best results from a hierarchical attention network (HAN). In addition, their results show that while the transfer learning model mBERT performs better on short documents, HAN performs better on longer documents, an improvement over mBERT that seems to increase with the document length.

At a lower level, the way text classification is traditionally done with BERT is using the CLS-token, which is the first token of every sequence when using the BERT architecture (Devlin et al., 2019). This token is a special classification token whose final hidden state can be used as the aggregate sequence representation for classification tasks (Devlin et al., 2019). Another way to do text classification is to start from the token embeddings and aggregate them to a document representation that is the input to a final classification layer, as done in the HAN mentioned by Barnes et al. (2020) and presented by Yang et al. (2016). This second approach can also be used with transformer models, but the most common way to do text classification with transformer models is to use the embeddings of the CLS-token.

There are ways to make transformer models more robust to long documents as well, with research being done both into making transformers more efficient and dealing with document length (Beltagy et al., 2020; Kitaev et al., 2020; Zaheer et al., 2020; Park et al., 2022). The main issue with 'vanilla transformers', which is mentioned in all of the articles above, is that the quadratic complexity of the transformer self-attention does not scale to long sequence lengths. Park et al. (2022) identify four standard approaches to long document classification: truncating long documents, using an efficient (linear attention) transformer model, chunking documents and selecting key sentences most central to the classification.

Tay et al. (2021) propose a benchmark for evaluating models in long-context scenarios and test 10 proposed models against a vanilla transformer and local attention baseline. They find that the tested models are trained faster and with lower memory usage than the vanilla transformer, while getting mostly equal or better results. For text classification, *Performers* (Choromanski et al., 2020) and *Linear Transformers* (Katharopoulos et al.,

2020) performed best (Tay et al., 2021). Both of those models use a kernel-based attention mechanism that scales linearly with the sequence length. On the other hand, Park et al. (2022) found that models made specifically for long documents require more time, do not perform consistently across data sets and are often outperformed by the BERT baseline. It should be noted however, that they did not test all the same models as (Tay et al., 2021), only the *Longformer* model was used in both articles.

In practice, if one does not do anything to address the problem of documents being longer than 512 tokens, the BERT tokenizer will simply truncate the input at 512 tokens. This might give sufficient results for one's needs. Sun et al. (2019) find that keeping the first and last part of the text, ignoring the middle, achieves better results than keeping only the first part for texts longer than 510 tokens. They call this the **head+tail** truncation method, which uses the first 128 and last 382 tokens of the text, in addition to the *[CLS]* token that begins the text and the *[SEP]* token that ends it. Sun et al.'s (2019) results also shows that the results for the **head+tail** are better than the results for hierarchical methods.

2.5 Other resources for Norwegian NLP

In order to do the research for this thesis, several resources will be used. In an NLP context, Norwegian is not a high-resource language, but nonetheless several important resources for Norwegian NLP exist, some of which will be outlined below.

2.5.1 BERT

In the earlier research on $\text{NoReC}_{\text{gender}}$, Touileb et al. (2021) used the NorBERT model (Kutuzov et al., 2021). At the time of writing, the NorBERT2 model[1] has been released, which is reported to have better scores on binary sentiment analysis than the earlier NorBERT. NorBERT is trained for Norwegian from scratch on Norsk Aviskorpus, a collection of Norwegian news texts[2] with 1.7 billion words, as well as Norwegian Wikipedia dumps in 'bokmål' and 'nynorsk' with respectively 160 million words and 40 million words (Kutuzov et al., 2021). NorBERT2 is also trained from scratch for Norwegian, but on a bigger corpus: the Norwegian part of the C4 web-crawled corpus, containing 9.5 billion words (Xue et al., 2021), as well as the non-copyrighted part of the Norwegian Colossal Corpus (NCC), containing 5 billion words. The NCC contains scanned books, newspapers and reports from 1814 and onward, as well as the Wikipedia dumps and Norsk Aviskorpus that was also used for training of NorBERT.

There are other relevant models for Norwegian using the BERT architecture: NB-BERT (Kummervold et al., 2021), m-BERT, released by Devlin et al. (2019) and XLM-Roberta, presented by Conneau and Lample (2019). m-BERT is multilingual BERT, trained on 104 different languages,

[1] https://huggingface.co/ltgoslo/norbert2

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

including Norwegian, and found to create multilingual representations that work well for cross-lingual model transfer (Pires et al., 2019). NB-BERT, like the standard BERT, has both a base and large version available, and improved evaluation results for several Norwegian NLP tasks compared to NorBERT when it was released (Kummervold et al., 2021). NB-BERT was initiated using the weights and tokenizer from the pretrained m-BERT, which means that it has a vocabulary of around 120 000 words, four times larger than the vocabulary of 30 000 for NorBERT, while NorBERT2 has a vocabulary of 50 000. NB-BERT was trained on the complete NCC corpus mentioned in Section 2.5.1, i.e. also the copyrighted parts that are not used for pre-training of NorBERT2, in total 18.4 billion words. (Kutuzov et al., 2021). After further pre-training on Norwegian, NB-BERT was shown to be better than m-BERT for named entity recognition (NER) in Norwegian, Danish, Swedish and English, using micro F1 score, while regressing compared to m-BERT for the more dissimilar languages Spanish and Finnish (Kummervold et al., 2021).

Kummervold et al. (2021) do not conclude whether the improved results for Swedish and Danish comes from close linguistic similarities to Norwegian or the possibility that some Swedish or Danish texts were present in the corpus. However it is not a far stretch to believe that the effect is caused by the language similarities. The improvement was also notably higher for Danish than for Swedish (1.7 pp against 0.6 pp), which could be caused by the fact that written Norwegian is closer to Danish than to Swedish, and that the written language in Norway was Danish until the beginning of the 20th century (Papazian, 2012). Some books and other documents in the corpus are from as far back as 1814. Kummervold et al. (2021) themselves assess that up to 1% of the corpus are texts written in Sami, Danish, Swedish and traces from other languages, while around 4% is written in English. Thus it is even harder to know whether the improvement over mBERT for English comes from training on some more English data or the linguistic similarities between Norwegian and English.

2.5.2 XLM-Roberta

XLM is an acronym for cross lingual language model (Conneau & Lample, 2019). XLM-Roberta (XLM-R), presented in 2019 by Conneau et al., was shown to outperform m-BERT on cross-lingual classification, especially for low-resource languages. Conneau et al. (2020) also show that XLM-R performs on par with monolingual models for several tasks.

All of the mentioned models have some inherent bias, so it should be interesting to investigate how they compare to each other in that regard.

2.5.3 Other resources

NorSENTLEX is a sentiment lexicon for Norwegian that was presented by Barnes et al. (2019) where they also show that incorporating information from sentiment lexicons using multi-task learning can improve model

performance. The sentiment lexicon was created by machine-translating an English lexicon and curating the results.

Another sentiment analysis resource for Norwegian based on NoReC is $\text{NoReC}_{\text{fine}}$, which is annotated with respect to polar expressions, targets and holders of opinion (Øvrelid et al., 2020). Like $\text{NoReC}_{\text{gender}}$, it is a subset of NoReC, but $\text{NoReC}_{\text{fine}}$ is annotated on entity-level, not document-level.

In addition to the $\text{NoReC}_{\text{gender}}$ data set, there is the Talk of Norway (ToN) data set, which also includes gender as a variable (Lapponi et al., 2018). ToN is a collection of Norwegian Parliament speeches from 1998 to 2016, annotated with metadata and augmented with several feature annotations. The thesis will mainly use the $\text{NoReC}_{\text{gender}}$ data set, but it may be of value to investigate if similar result may be achieved on different data sets as well.

2.6 Bias

Bias will be an integral focus of this thesis, since it undertakes to understand the gender bias present in $\text{NoReC}_{\text{gender}}$. This section will start by defining bias in general and the several different meanings and interpretations of bias, before gender bias in particular will be discussed.

Merriam-Webster (n.d.) gives several definitions of bias:

a: an inclination of temperament or outlook especially : a personal and sometimes unreasoned judgment : prejudice

b: an instance of such prejudice

c: bent, tendency

d (1): deviation of the expected value of a statistical estimate from the quantity it estimates

(2): systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others

Among these definitions, the first, second and third are probably the ones meant in colloquial use of the word, whereas for statistics and machine learning, the last two definitions are most relevant. For machine learning, *predictive bias* is a relevant subset of bias, which Shah et al. (2020) define as occurring when the label distribution of a predictive model reflects a human attribute in a way that diverges from a theoretically defined "ideal distribution.", a definition similar to *d(1)* above. It is important to note the different possible meanings of bias, but as mentioned in the introduction, in this thesis we use Friedman and Nissenbaum's (1996) definition of bias as systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others.

Even in machine learning, bias can have several different meanings. One much used phrase is the bias-variance trade-off for supervised learning or the bias-variance dilemma (Luxburg & Schölkopf, 2011). If we only use linear models, every functional dependency one could discover would be linear, they illustrate. This would, however, not result from the data, but

would be a *bias* imposed by us. On the other hand, an extremely close fit to the data tends to generalize poorly to future data, because such a fit entails fitting random aspects of the sample in addition to the replicable trends (Briscoe & Feldman, 2011), making it suffer from large *variance*. Too much bias can lead to *underfitting*, whereas too much variance can lead to *overfitting*.

A standard visualization of this problem is to have a data set with a curved decision boundary, where a linear classifier will not fit the data well. Such a classifier would have high bias, because it makes the assumption that the data is linearly separable. This kind of bias is also called inductive bias. On the other hand, one could imagine a classifier with a winding decision boundary that curves around every data point, which would be a classifier with high variance.

Apart from bias in the algorithms, there can also be bias in the data. Some models are pretrained on very large data sets, many of them coming from the internet without any manual corrections. That means that any bias present in text that people write on the internet would also be present in these large models that are trained on all of that data.

What this shows, is that bias is an inherent property of any NLP system, both when it comes to the models used and the data it is trained on. This bias is not a problem per se, but it could be, depending on the application. On one side, bias can be used to explain the world around us, and in a Bayesian framework the prior probability distribution serves as a bias (Hovy & Prabhumoye, 2021). There can be a lot of bias in a model without any problem, as long as the researcher is aware of it. Bias first becomes a problem when it is used in a non-intended way with an adverse outcome, by e.g. making a group of people systematically treated differently, like when a machine learning algorithm prioritizes men over women.

An issue with the word bias is that it has negative connotations, like its connection to the word 'prejudice', defined as a synonym of bias by Merriam-Webster (n.d.), so bias is almost negative per definition. In reality bias is neutral, and all models have some kind of bias. Bias might carry useful information, so it may not be a good idea to remove the bias of a model, even if possible (Hovy & Prabhumoye, 2021). Bias is also intrinsic to human language. As a car with many breakdowns is more prone to accidents, a patient with a chronic disease could have higher probability of worsening (Garrido-Muñoz et al., 2021).

With that in mind, it is clear that NLP can have negative consequences. Hovy and Prabhumoye (2021) list several of these, such as unequal performance for different user groups and the proliferation of harmful stereotypes, and they stress that NLP has a real impact on people's lives. Since the focus has moved from using models as a tool for understanding to be used as predictive models, they have become hard to analyse. While they solve their intended task, they also pick up secondary aspects of language that may be exploited to fulfill their function, and these aspects of language may carry a lot of subtle information about the speaker (Hovy & Prabhumoye, 2021). With the new large neural networks, one cannot simply look at a model weight and say how it impacts the result, and for the large

pretrained models one cannot even control the data that has been used for pretraining.

2.6.1 Gender bias

This thesis will not focus on applications, but rather seek to understand a predictive model. Hopefully it will manage to discern some of the 'secondary aspects of language' that the model uses to make its predictions. In the $\text{NoReC}_{\text{gender}}$ data set, we know that there is bias, and the goal of this thesis is to investigate where the bias is. While there could be a lot more to write about bias in general, the focus from here on will be gender bias, since that is important to the investigations of this thesis.

Stanczak and Augenstein (2021) define gender bias as 'the systematic, unequal treatment based on one's gender'. This means that they define gender bias in another way than Touileb et al. (2021), who define it as: 'the differences in language use between persons, on the unique basis of their genders'. Where Touileb et al. (2021) define it as the differences in the way people of different genders *act*, Stanczak and Augenstein (2021) define it as the differences in the way they are *treated*. In this thesis both these aspects of gender bias will be relevant, given that male authors and critics express themselves differently than female authors and critics (*how they act*), but male and female authors are also portrayed differently by others (*how they are treated*). In an earlier article on the same topic, Touileb et al. (2020) explicitly state that combination of these two aspects is their focus of study.

Stanczak and Augenstein (2021) also provide a definition of causal (gender) bias as 'the disparity in the output when model is feeded with different genders'. Using this way of measuring bias, there is clearly bias in the NoReC data set as found by Touileb et al. (Touileb et al., 2021).

According to Hovy and Prabhumoye (2021), the five sources where bias can occur in NLP systems are: '(1) the data, (2) the annotation process, (3) the input representations, (4) the models, and finally (5) the research design (or how we conceptualize our research)'. This thesis will mostly consider bias in the textual data and try to find out what the difference between the genders consists of. However, it must be kept in mind that there may be considerable bias from the input representations and the models as well. The difference could also come from some statistical bias in the data, rather than from the textual content. Models might use spurious correlations and statistical irregularities to increase it accuracy, and may thus give correct answers for the wrong reasons (Hovy & Prabhumoye, 2021).

In this particular case, binary gender annotation is a part of the research design that could be critiqued, because they base it on binary gender categories annotated by the researchers, and not by the people involved themselves. However, it is hard to do much about it from a researcher's point of view, given that it is how the data has been collected. You could also ask whether the annotation itself is part of the second source of bias or the first, since the gender was annotated by the researchers from the data (Touileb et al., 2020) without the possibility for asking the authors and critics about self-identified gender.

It is also important to note that Hovy and Prabhumoye's categories are not normative, but a framework to describe different kinds of sources for bias. It is not that important whether one specific source of bias is from one category or another, but to be aware that bias can be introduced in every step of the research process, from the data to the research design.

For example, we know that there is gender bias in the society at large, and this will be reflected in the data we use for NLP (Natural language processing) tasks. Several articles suggest that women use a less vague tone when they describe sentiment (De Amicis et al., 2021; Thelwall, 2018), but they also emphasize that this difference does not necessarily reflect actual difference of content, but simply a difference of style. Nonetheless, this might be a feature that distinguishes female sentiment from male sentiment, and may make a model fit easier on female sentiment (Thelwall, 2018).

On the other hand, these results are only valid in the context in which they were observed, and may not generalize to other contexts. In the first of these articles, De Amicis et al. (2021) analyse the gender difference of gender sentiment in earning conference calls and found a difference both between how women and men convey sentiment, and how they in turn are portrayed by others. As mentioned above, women use a less vague tone than men, but the financial analysts in this case were less positive when portraying a woman than a man (De Amicis et al., 2021).

Stanczak and Augenstein (2021) describe two of the core limitations of research on gender bias as treating gender as a binary value, neglecting it fluidity, and that most of the work is done on high-resource languages like English. This thesis will not do anything to counteract the first issue, given that the data set we use treats gender as a binary variable. On the other hand, it will deal with gender bias in Norwegian, which is not one of the highest-resource languages.

According to Stanczak and Augenstein (2021), gender bias has been confirmed to be prevalent in literature, news, and media, and in communication about and directed towards people of different genders. All these sources of gender bias are relevant to this thesis, especially the part about 'communication directed towards people of different genders'.

Language can be used as a substantial means of expressing gender bias (Stanczak & Augenstein, 2021). These biases present in the language data will then be transported and possibly augmented by machine learning algorithms, which, as mentioned above, will use any correlations they can find to achieve their objective. This means that detecting, analysing and mitigating bias is a pressing topic for NLP (Stanczak & Augenstein, 2021). On the other hand Garrido-Muñoz et al. (2021) asks whether the main task to be solved by the NLP systems could be damaged by the intervention to mitigate bias. There is probably a balance to be found here; trying to remove all bias could remove a lot of valuable information and significantly damage the performance, whereas not doing anything to address the bias could lead to highly unfair predictive systems.

Chapter 3

Data

The focus of this chapter will be a deeper analysis and exploration of the data that will be used in this thesis, i.e. the $\text{NoReC}_{\text{gender}}$ corpus. The data in this corpus are book reviews published between 2002 and 2019 in 6 Norwegian newspapers: *Aftenposten*, *Bergens Tidende*, *Dagbladet*, *Fædrelandsvennen*, *Stavanger Aftenblad*, and *Verdens Gang*. Figure 3.1 shows how many reviews in the data set were published during each year between 2002 and 2019, and Table 3.1 shows which sources the reviews come from.

As mentioned in Section 2.1.2 on using ratings as labels, interpreting ratings can be a challenge since different critics may not use the scale in the same way, and the scale may also mean something different for ratings of books by well-established authors vs. debut authors. The rating 4 out of 6 for example, may mean 'mediocre' when a specific critic reviews a book by some author, but it can also mean 'pretty good' when another critic reviews a book by another author, or even when reviewing the same book. With this rating challenge mentioned, it is however believed that the different usages of the rating scale averages out across the groups we are interested in, and that the general trends in the corpus are still valid. In this thesis we will interpret the rating 4 as *fair* and use it as a third sentiment class separate from the positive and negative class.

		rating		length	
source	count	mean	std	mean	std
ap	13	4.31	0.86	59.69	16.72
bt	186	4.41	0.90	578.11	142.16
dagbladet	1040	4.45	0.92	508.33	197.16
fvn	563	4.56	0.99	347.54	97.71
sa	454	4.40	0.99	414.91	143.12
vg	1822	4.35	1.03	334.01	127.53

Table 3.1: Summary of the different sources with their rating and review length distribution.

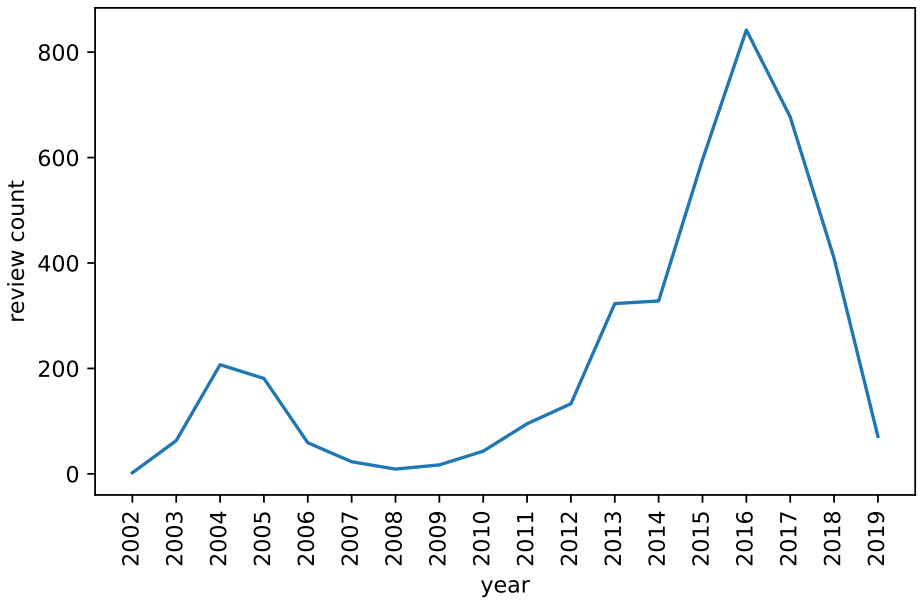

Figure 3.1: Number of reviews per year.

3.1 Data distribution

As mentioned in the earlier Chapter 2, $\text{NoReC}_{\text{gender}}$ comprises 4 313 book reviews. Among these there are in total 212 critics reviewing 2 320 authors. The overall rating distribution is shown below in Table 3.2. $\text{NoReC}_{\text{gender}}$ is split into a train, dev and test set (Touileb et al., 2021), but for the statistical analysis in this chapter the data from all three splits is included. The split follows a standard 80-10-10 ratio and is sorted by date, in such a way that the training data is older than the dev and test data. Velldal et al. (2018) write that this split by time reduces the risk of having the same item in different splits, and presents a more realistic test scenario where models trained on existing data are applied to

	mean	std	min	25%	50%	75%	max
rating	4.40	1.00	1	4	5	5	6
word count	395.40	172.46	18	287	360	481	1525

Table 3.2: Overall distribution of ratings and review lengths across the corpus.

We can see in Table 3.2 that the mean is 4.4 and that at least 75% of the reviews have a rating of 4 or higher, and at least 50% of them have a rating of 5 or higher. The number of words used in each review is also shown here. The numbers of words correlates positively to some extent with the rating, with a Pearson correlation coefficient of 0.15 and p-value 2.8e-22. This correlation is also shown in a scatter plot with color-graded density in Figure 3.2. The blue line for median review length and red

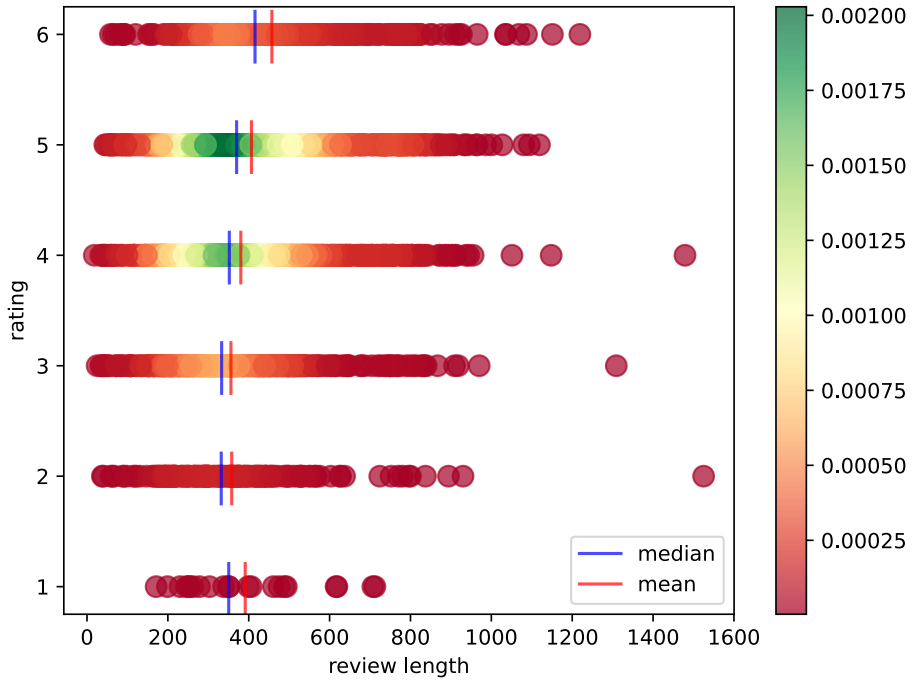

Figure 3.2: Scatter plot showing review length by rating.

line for mean length show that rating 3 receives the shortest reviews and rating 6 receives the longest ones. We can also see from the colored density gradient in Figure 3.2 that the most common review has rating 5 and is fairly short, i.e. containing 350 words or less, followed by reviews with rating 4 and approximately the same length. There seems to be a general trend that reviews with higher ratings use more words, as indicated by the positive correlation. Another possible complementary interpretation could be that ratings at the ends of the scale, i.e. 1 and 6, require more justification than the ratings in the middle, causing those reviews to be longer. These two explanations together could explain why reviews with the rating 1 are longer than reviews with the rating 2 and 3, but still shorter than reviews with rating 4, 5 and 6 on average.

If the ratings were normally distributed between the 6 different ratings, one would expect a mean of 3.5, with as many 3s as 4s, as many 2s as 5s and as many 1s as 6s. In reality the mean is 4.4 and the median is 5, showing a skewed distribution with a tendency of high ratings. This is another one of the challenges with using ratings as labels: critics are more likely to choose books they actually want to read, meaning that reviews also can be seen as recommendations by the critics, not only unbiased reviews. Another part of this selection bias could be that books that publishers believe would receive very low ratings may not be published at all, and books that are not published will not be reviewed, a kind of survivorship bias.

3.1.1 Gender

This thesis will deal with gender, and thus it is important that the gender of critics and authors is correct. Some of the books that are reviewed has more than one author, and some reviews are written by several critics. Any review where either the authors or the critics are of both genders will be discarded, in accordance with the approach used by Touileb et al. (2020). The overall analyses above used all the data, but from here on only the 4 078 reviews of the original 4 313 reviews with well defined genders will be used. Among these data there are in total 198 critics reviewing 2 208 authors. By gender they are 124 male critics (63%), 74 female critics (37%), 1 359 male authors (62%) and 849 female authors (38%).

When it comes to the distribution of gender of critics and authors for each review in the data set, there are somewhat more men in both categories. There are 2 366 reviews (58%) by male critics and 1 712 reviews (42%) by female critics. When it comes to the authors the gender discrepancy is a bit larger, 2 570 (63%) of the reviews were about books written by male authors, while 1 508 (37%) of them review books written by female authors. This is the same as the ratio of female authors to male authors reviewed, which indicates that each female author in the data set on average gets as many reviews as each male author, but more male authors are reviewed.

gender			
critic	author	count	ratio (%)
F	F	887	22
	M	825	20
M	F	621	15
	M	1745	43

Table 3.3: Total number of reviews grouped by gender of critic and author.

The review counts grouped by both critic gender and author gender is shown in Table 3.3. This table shows that female critics review more or less the same amount of books by male and female authors, just around 10% more for female authors. On the other hand, male critics review almost three books by male authors for each book by a female author they review and 43% of the data consists of men reviewing men. Male critics' preference to review male authors will later be shown to have some impact on whose gender, i.e. the critic's or the author's, is the most important factor for predicting the rating.

3.2 Distributions grouped by gender

This section will investigate how different aspects of the data are distributed across the genders. In some of the figures and tables, we will use a shorthand notation with *F* for the women and *M* for the men, placing the gender of the critic in front of the author gender. *FM* would thus mean the

group of female critics reviewing male authors, and *MF* would conversely mean the group of male critics reviewing female authors.

In Table 3.4 and 3.5 we can see that male critics on average give 0.08 higher ratings and male authors receive on average 0.11 higher ratings than their female counterparts. This may suggest that the gender of the author is more important for the rating than the gender of the critic. The five number summary shows the same tendency for both groups, mostly equal except that female critics and authors have a rating median of 4, not 5, like the male groups have. Otherwise the five number summary is the same as in Table 3.2 for all groups. Table 3.4 also shows that female critics write somewhat longer reviews, 10 words longer by average.

gender		rating			word count	
critic	count	mean	std	median	mean	std
F	1712	4.36	0.99	4	405.29	167.05
M	2366	4.44	0.99	5	395.49	171.73

Table 3.4: Summary of ratings and review length grouped by critic gender.

gender		rating			word count	
author	count	mean	std	median	mean	std
F	1508	4.34	1.00	4	400.69	165.98
M	2570	4.45	0.98	5	398.96	172.08

Table 3.5: Summary of ratings and review length grouped by author gender.

gender			rating			num words	
critic	author	count	mean	std	median	mean	std
F	F	887	4.31	1.00	4	404.66	162.71
	M	825	4.42	0.99	5	405.96	171.70
M	F	621	4.38	1.01	5	395.02	170.51
	M	1745	4.47	0.98	5	395.66	172.20

Table 3.6: Summary of ratings and review length grouped by critic gender and author gender.

In Table 3.6, the results are grouped by both critic gender and author gender. The same tendency can be seen here; the female gender is associated with lower ratings, no matter if it is the critic or author that is female, i.e. that women both give and receive lower ratings than men. The group with lowest ratings is female critics reviewing female authors, with a mean rating of 4.31. On the other hand, male critics reviewing male authors give the highest ratings, 4.47, 0.16 higher than women reviewing women.

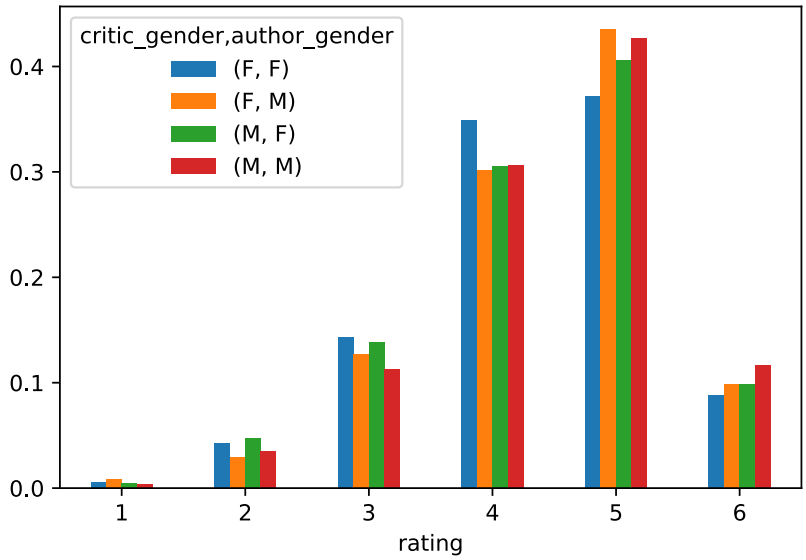

Figure 3.3: Bar plot of the normalized ratings for each group.

The standard deviation is more or less the same for all 4 groups, and so is the five number summary, not shown in this table but equal to the one in Table 3.2 except for the women reviewing women group, which again has a lower median rating of 4 as opposed to 5 for the other three groups.

In Figure 3.3 one can clearly see that the FF group, representing women reviewing women, has fewer of the high ratings 5 and 6, more of the low ratings 1, 2 and 3, and especially more of the *fair* rating 4.

3.2.1 Sources

The ratings by gender also vary among the different sources of the data set. Since AP only has 13 reviews with well defined gender, as shown in Table 3.1, it is excluded from the analysis grouped by the sources.

In Figure 3.4 and 3.5, we can see that mean rating by gender differs across the gender combinations, especially for *Fædrelandsvennen (fvn)* where the mean rating for MM is 0.64 higher than for FF. The review lengths, on the other hand, seem to differ more by source than by the gender combination.

3.2.2 T-tests

In previous work on gender bias, Lassen et al. (2022) have used literary reviews 'to identify systematic components that can be attributed to bias'. Their work is relevant for this thesis since they use use book reviews from Denmark, which is another Scandinavian country that is similar to Norway. Lassen et al. (2022) did not have access to the text of the reviews, but their analysis of the metadata shows the same tendencies as the work of Touileb et al. (2020) on $\text{NoReC}_{\text{gender}}$. Lassen et al. (2022) used t-tests to

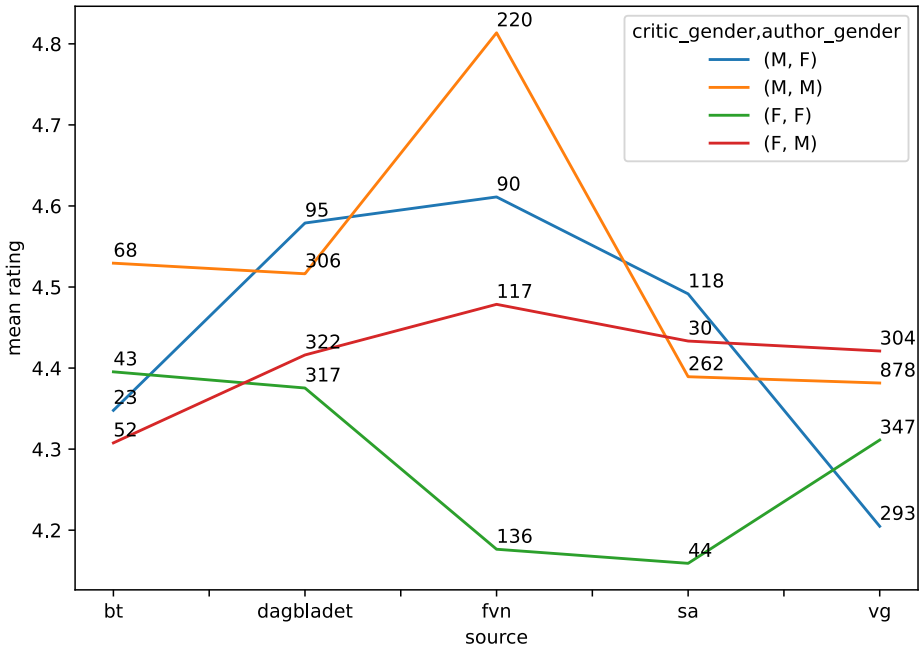

Figure 3.4: Line plot showing mean ratings grouped by source and genders. The numbers in the plot is the support for each data point, i.e. the number of ratings from which the mean is calculated.

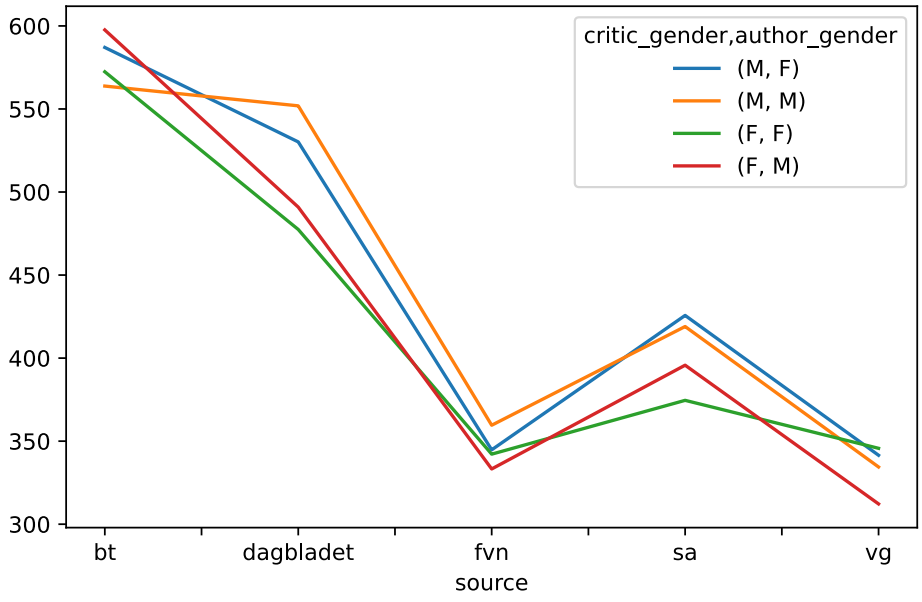

Figure 3.5: Line plot showing mean review length grouped by source and genders.

investigate whether there are statistically significant differences between the four gender combinations. Like they found in their data set, the t-tests performed in this section shows some significant differences in ratings given and received between the genders at a significance level of 0.05. In order not to make too many assumptions about the distributions, a Welch t-test was used, which does not assume the standard deviation is equal for the different distributions. The tests were chosen to be one-sided, testing whether female critics and authors are associated with lower ratings than male critics and authors, not only whether their ratings are different.

The t-tests on the ratings grouped by critics and the one grouped by authors had p-values of 5.3e-3 and 2.4e-4 respectively, clearly significant results at a significance level of 0.05 and thus the zero hypothesis of equal means for both groups was rejected. We conclude that there is sufficient evidence to say that female critics give lower ratings than male critics and that female authors receive lower ratings than male authors. For the ratings grouped by both critic and author gender at once, there were 6 different tests, the results of which are shown in Table 3.7.

alternative hypothesis	p-value
FM < MM	0.14
MF < FM	0.22
MF < MM	0.032
FF < MM	6.9e-5
FF < FM	0.011
FF < MF	0.091

Table 3.7: Results of Welch t-tests on the six combinations of gender groups. The first letter of the two-letter combination is the critic gender and the second is the author gender. Thus FM < MM is the hypothesis that female critics give male authors lower ratings than male critics give male authors.

While both the t-tests when just grouping by one gender variable at a time were significant, when grouping on both gender variables, Table 3.7 shows that when keeping the author gender the same and just changing the gender of the critic, the t-tests do not give significant results and the null hypothesis cannot be discarded. This means that female critics reviewing male authors do not give statistically significantly lower ratings than male critics reviewing male authors do (p-value 0.14), nor do female critics reviewing female authors give statistically significantly lower ratings than male critics reviewing female authors do (p-value 0.091). However, all the tests where the critic gender is held the same and the author gender is changed show significant results, as can be read from Table 3.7.

3.3 Outliers

The number of ratings for each critic is a long-tailed distribution. While most critics have reviewed quite few of the books in the corpus, some critics

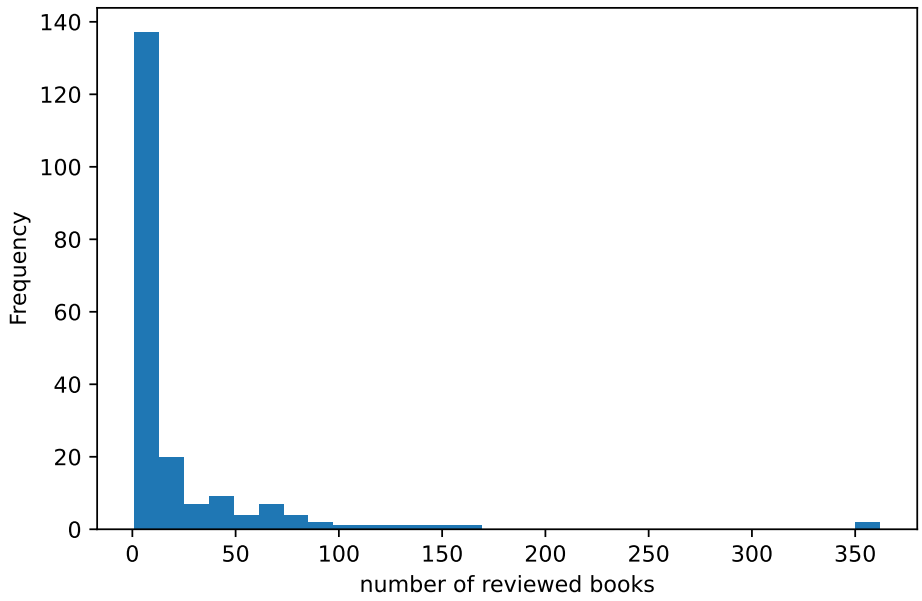

Figure 3.6: Histogram showing how many books each critic has reviewed.

have written a lot of reviews, as shown in Figure 3.6 and Table 3.8.

count	mean	std	min	25%	50%	75%	max
198	20.61	45.21	1	1	3	20	362

Table 3.8: Statistical summary of how many books each critic in the data set has reviewed.

The patterns that are of interest to the research of this thesis are general and systematic differences at group level, and not individual differences between the critics. If some of the critics who write lots of reviews are biased, that could impact the data a lot, giving an illusion of general validity for some trends, even though they may only describe how one critic writes reviews and give little insight into literary reviews in general. For this reason we did some additional analysis of the critics with at least 50 reviews. They are 25 people, 16 male and 9 female critics who account for 2743 reviews, i.e. 67% of the total number of reviews. 1075 of them are written by the 9 female critics and 1668 of them by the 16 male critics.

Sindre Hovdenakk, who has written 362 reviews for VG with a mean rating for female authors of 3.64 is close to being an outlier, as indicated by the box plot in Figure 3.7. He is also the critic with most total reviews and the male critic with most reviews of female authors, even at only 56 such reviews versus 306 reviews of male authors. The reviews he has written comprise more than 8% of $\text{NoReC}_{\text{gender}}$. His mean rating of male authors is 4.36, 0.72 higher than for female authors, but still lower than the overall average rating of 4.41. Among these 25 critics Hovdenakk is also the one who gives female authors the lowest ratings. Morten Abrahamsen is also

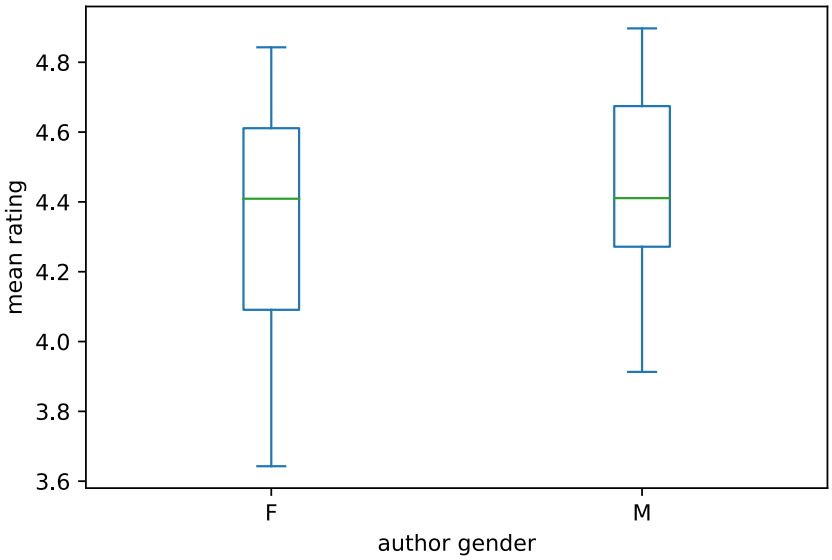

Figure 3.7: Box plot showing the mean ratings for the 25 most prolific critics by author gender.

almost such an outlier, with a mean rating of 3.69 for female authors, against 4.12 for male authors. However, he has only reviewed female authors 13 times, and male authors 59 times.

To better understand the possible impact of the most prolific critics, the next section compares the three critics with most reviews.

3.3.1 Rating comparison of the three most prolific critics

In order to see the difference between the three critics with most reviews visually, the normalized rating distribution is put into a line plot in Figure 3.8.

We can see in Figure 3.8 that Fredrik Wandrup, in blue and orange, gives comparatively high ratings, where e.g. almost 70% of the books of male authors that he reviews receive a 5 and the same goes for more than half of his reviews of books by women. Marie L. Kleve, in green and red, gives almost the same ratings to both male and female authors, whereas Hovdenakk's ratings of female authors in violet are clearly to the left of the other two critics. When it comes to the rating 6, Hovdenakk has the highest ratio of the three for male authors, and the lowest ratio for female authors. While Hovdenakk gives rating 4 at more or less an equal rate to both genders, Figure 3.8 shows that Hovdenakk gives male authors rating 5 at almost double frequency than for female authors, and male authors receive the highest rating 6 several times more often from him than female authors do. Of course the picture is the opposite for the lower ratings. Considering how Hovdenakk has written such a large part of the reviews in the corpus and with such a large difference for male and female authors, we

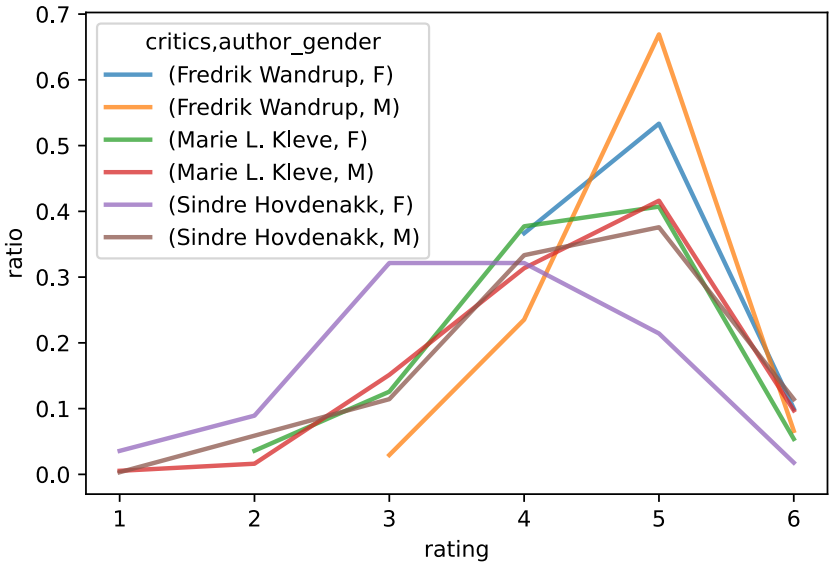

Figure 3.8: Line plot showing normalized ratings for the three most prolific critics grouped by author gender.

do some analysis of the data where we exclude his reviews in Section 3.3.2 below.

3.3.2 T-tests without Hovdenakk

In order not to allow one person to skew the whole corpus, the same analyses have been done on the data excluding reviews made by Hovdenakk. The results grouped by critic gender and author gender independently are shown in Table 3.9 and 3.10. The t-tests still show significant differences for these groups, with p-values of 4.6e-4 and 1.4e-4, respectively.

critic	count	mean	change	std	median
F	1784	4.36	0.0	0.99	4
M	2102	4.48	0.035	0.97	5

Table 3.9: Statistic summary of ratings grouped by critic gender with change of mean without reviews by Sindre Hovdenakk.

For all the four groups, however, shown in Table 3.11, there are some changes, indicating that removing Sindre Hovdenakk changes which groups have significant t-tests. After removing Hovdenakk's reviews from the data, the alternative hypothesis $MF < MM$, i.e. the hypothesis that male critics give female authors lower ratings than they give to male authors is no longer significant at significance level 0.05, with a p-value of 0.23, shown in Table 3.12.

author	count	mean	change	std	median
F	1554	4.37	0.027	0.99	4
M	2291	4.46	0.012	0.97	5

Table 3.10: Statistic summary of ratings grouped by author gender with change of mean without reviews by Sindre Hovdenakk.

gender			rating			
critic	author	count	mean	change	std	median
F	F	887	4.31	0.0	1.00	4
	M	825	4.42	0.0	0.99	5
M	F	565	4.45	0.073	0.99	5
	M	1439	4.49	0.022	0.96	5

Table 3.11: Statistic summary of ratings grouped by critic gender and author gender with change of mean without reviews by Sindre Hovdenakk.

alternative hypothesis	p-value
FF < MM	1.17e-5
FF < FM	0.0107
FF < MF	3.40e-3
FM < MM	0.0559
FM < MF	0.271
MF < MM	0.230

Table 3.12: Results of t-tests without Hovdenakk's reviews.

The hypothesis FF < MF, that female critics give female authors lower ratings than male critics do, is now statistically significant, with a p-value of 3.40e-3. This means that while the overall gender effects where female critics and authors correlate to lower ratings are still the same, after removing Hovdenakk the gender of the critic seem to have relatively more importance than the gender of the author. On the other hand, the hypothesis FM < MM, that female critics give male authors lower ratings than male critics do, is still not significant, with a p-value of 0.056. What this means is that, were Hovdenakk not part of the data set, male critics would not statistically significantly discriminate against female authors, whereas on the other hand, female critics would give female authors statistically significantly lower ratings than male critics do. These new results show how large impact one person can have on the corpus, and might put to doubt any generalizations of the effects we see here. But despite Hovdenakk giving female authors lower ratings on average, his reviews are still part of the corpus, and excluding them because he is an outlier in this regard might be seen as cherry-picking the data to get the results we want. For this reason, none of his reviews will be excluded in the further analysis, and this section may simply be regarded as a warning of the impact one outlier can have on the overall data statistics in a small data set like $\text{NoReC}_{\text{gender}}$.

3.4 Reviews of the same book by critics of different gender

An issue with the previous analyses in Section 3.1.1 and Section 3.3 is that the critics could be reviewing only different books altogether, and that could in large part be why the ratings are different for each gender. By assuming that reviews of a given author within the same year regard the same book, one can find multiple reviews of a single book by several critics and compare the results by gender. We also set the constraint that the books chosen for this must be reviewed by critics of both gender. Using this method, 836 reviews of 307 books have been found. Of the reviews, 403 were made by female critics and 433 by male critics, whereas 108 of the books were written by female authors and 199 by male authors, nearly the double, and male authors also get almost the double amount of total reviews in this case; 550 for them vs. just 286 for the female authors.

3.4.1 T-tests

As in earlier cases in Section 3.2.2 and Section 3.3.2, Welch t-tests were performed on the groups. For the critic gender and author gender groups by themselves, the p-values for alternative hypothesis F < M, i.e. that women give and receive lower ratings, are 6.0e-4 and 3.5e-2 respectively, both significant at significance level 0.05. Table 3.13 and Table 3.14 show that the means for the critic gender groups are 0.20 lower for female critics and 0.13 lower for female authors, so that unlike earlier, the gender of the critic seems to have the larger impact on the rating.

		rating						
critic	count	mean	std	min	25%	50%	75%	max
F	407	4.41	0.87	2	4	4.5	5	6
M	438	4.61	0.94	1	4	5	5	6

Table 3.13: Rating summary grouped by critic gender.

		rating						
author	count	mean	std	min	25%	50%	75%	max
F	286	4.43	0.92	2	4	4.5	5	6
M	550	4.56	0.91	1	4	5	5	6

Table 3.14: Rating summary grouped by author gender.

gender			rating						
critic	author	count	mean	std	min	25%	50%	75%	max
F	F	146	4.32	0.87	2	4	4	5	6
	M	257	4.47	0.88	2	4	5	5	6
M	F	140	4.56	0.96	2	4	5	5	6
	M	293	4.63	0.93	1	4	5	5	6

Table 3.15: Rating summary grouped by critic gender and author gender.

test	p-value
FF < MF	0.013
FF < FM	0.047
FF < MM	2.5e-4
MF < MM	0.22
FM < MM	0.016

Table 3.16: Welch t-tests for the data isolated on books reviewed by critics of both genders.

The statistical summary for all the 4 groups are shown in Table 3.15 and the results of the t-tests on the same data in Table 3.16. Unlike for the other t-tests in Section 3.2.2 and 3.3.2, here all tests are significant except one, the $MF < MM$ test, on the hypothesis that male critics rate books by female authors lower than books by male authors. This result is interesting, because having isolated only reviews of books which has at least one review by a critic of another gender, we can be quite sure that the difference in ratings between genders does not only occur since male and female critics review different books, but are actually giving the same books different ratings. This isolated analysis also differs from the other analysis by the fact that the gender of the critic has a much larger impact here on the rating than the gender of the author. Female critics give 0.195 lower ratings than male critics, while female authors receive 0.12 lower ratings than male authors. Looking at all the data together in Table 3.4 and 3.5, the numbers were 4.36 and 4.44 for different critic genders, just 0.08 difference, while they were 4.34 and 4.45 for different author genders, a difference of 0.11. it is not immediately clear why there is such a difference between what gender has more impact for the data subsets. It could be that books have to be deemed of at least a certain quality in order to be reviewed by several critics, thus possibly lending the gender of the author smaller impact after this 'screening' has already been passed.

3.5 Discussion

Our results of the data analysis in this chapter clearly show that female gender is associated with lower ratings, both for female critics and female authors. Despite being statistically significant differences, they are still only correlations, and we cannot conclude that there are causal relations between the gender of either critic or author and the ratings they give or receive. This means that we cannot say for certain if there is gender bias in $\text{NoReC}_{\text{gender}}$. Saying there is gender bias would be to argue that, were the genders of the authors switched and the text of the books the same, the critics would on average still give higher ratings to the 'male' authors, which would have indicated that gender itself matters for the judgement of ratings. Lassen et al. (2022), having found the same correlations in their data, add the same caveats, writing that it could be that the perceived gender bias is confounded with expertise bias, i.e. that specific literary language leads to higher appreciation among the critics. So if women in general write more genre literature, then the observed difference may stem from a difference in complexity of linguistic features (Lassen et al., 2022). On the other hand, one could argue that this by itself also would constitute a bias against women, if genres that are predominantly written by women are associated with lower literary value or lower linguistic complexity.

When it comes to the female critics giving lower rating than men, one could also ask if they simply review other books than the male critics do, books that the men also would have given the same, lower rating on average. This is an explanation that differs from the gender bias explanation, where

female critics would give lower ratings than men simply because they are women. We tried to control for this aspect in the previous section, Section 3.4, making sure we only analysed books reviewed by at least one male and one female critic, but any of those books could still be reviewed by ten men and one woman. $\text{NoReC}_{\text{gender}}$is also not a large corpus, and we cannot be sure if the book reviews it contains are representative of Norwegian book reviews in general. Nevertheless, gender bias could perfectly well be the explanation for some of the observed differences, we just cannot know for certain.

Chapter 4

Models and experiments

In addition to statistics on the data, we want to do machine learning experiments to investigate gender and sentiment further. In order to do that, we require some appropriate models. In Chapter 2, we introduced NorBERT and NorBERT2, as well as XLM-Roberta. NorBERT2 will be the main transformer model used withing the thesis, but we also include results from XLM-Roberta$_{\text{Large}}$ for comparison. Despite XLM-Roberta showing good results, it is a much larger model with a vocabulary of 250 000 words (Conneau et al., 2020), compared to 50 000 words in NorBERT2, and it also has a large number of weights. This makes XLM-Roberta less viable for some of the experiments where models are trained several times in a row and for interpretability methods.

We also establish baselines for our classification tasks: author gender classification, critic gender classification, and sentiment classification by using three different linear BoW models. These are two support vector machines: a support vector classifier (SVC) and a support vector regressor (SVR), and lastly a Ridge regressor. Support vector machines have been shown to give consistently good results on text classification (Barry, 2017), and are therefore included as a baseline. Ridge regression is included since it is similar to standard linear regression, but uses the l2-norm as a regularization penalty. This penalty makes the model's coefficients more robust to collinearity (McDonald, 2009), which is important since the TF-IDF features from the review texts used as input to the models are not independent, as they represent the words used in the texts.

This chapter will begin with describing the computational environment, the implementation details and the performance of the models, before doing some experiments on training models with different variations of the input data.

4.1 Computational environment

Training of deep neural networks is not possible without the appropriate hardware. In particular one needs accelerators like GPUs to make training of transformer models tractable. We used University of Oslo's *Machine Learning Infrastructure (ML Nodes)*, provided by University Centre for

Information Technology (n.d.). These ML nodes are part of University of Oslo's (2020) *AI hub-node project*, which aims at building up IT resources and competence in using these to support machine learning, deep learning and data science in research and education at the University of Oslo. There are eight available ML nodes with slightly different specifications. For the training of our transformer models we have used the seventh node[1], which contains 32 AMD EPYC 7282 16-Core Processors, 128 GiB RAM and 8 NVIDIA GeForce RTX 2080 Ti GPUs (University of Oslo, 2023), each with 11 GiB of memory. We have been able to restrict our resource usage to one of the available eight GPUs and we trained our BoW models on a personal computer with an Intel Core i5-8250U processor and 8 GB RAM.

4.2 Implementation details

We trained the baseline models using *Scikit-learn* (Pedregosa et al., 2011), and the transformer models using *PyTorch*, introduced by Paszke et al. (2019). Before going into details on each specific model, we describe implementation details and some factors that are shared between the baseline and transformer models.

The first of these factors is the pre-defined train/dev/test split. Even though it could be useful to do cross-validation on the BoW models in particular, to ensure soundness of the results, we chose to strictly follow the pre-defined splits. The data was originally split in this way to ensure replicability, and as we write in Section 3.1, citing Velldal et al. (2018), having the data split by timeline makes for a more realistic test scenario. The second shared factor also relates to the data, but now more specifically to its imbalanced nature, which compels us to use *class-balanced loss*, described in Section 4.2.1 below.

4.2.1 Class-balanced loss

As we show in Chapter 3, the data for all three classification tasks is imbalanced, especially the ratings on the scale from 1 to 6. In order not to lose a lot of performance due to this, one needs to find a way to balance the classes during training. Possible directions for this are downsampling the large classes, upsampling the small ones or weighting the classes differently, using e.g. the inverse class frequencies as weights. Cui et al. (2019), however, argue that such re-samplng or inverse weighting gives poor results and formulate the *Class-Balanced Loss* to address these problems. They introduce a weighting factor inversely proportional to what they call the *effective number of samples*. The idea of the effective number of samples is to capture the diminishing marginal benefits of using more data points in a class, since as the number of samples grow, it gets more likely that newly added samples are near-duplicates of existing samples (Cui et al., 2019). This means that their loss weighting would penalize large classes

[1] ml7.hpc.uio.no

less heavily than weighting by inverse class frequency does. In practice, Cui et al. (2019) write the class balance as:

$$CB(\mathbf{p}, y) = \frac{1-\beta}{1-\beta^{n_y}}\mathcal{L}(\mathbf{p}, y),$$

where n_y is the number of samples in the ground-truth class y, $\mathbf{p}$ is the predictions for class y, y is the true labels for class y, and β is a hyperparameter. Since the class-balanced loss simply consists of a weighting factor for the original loss $\mathcal{L}$, it is model independent. Setting $\beta = 0$ is the same as using no re-weighting and as β approaches 1, the re-weighting approaches re-weighting by inverse class frequency. Cui et al. (2019) find that $\beta = 0.999$ and $\beta = 0.9999$ are reasonable values for β.

We use mostly $\beta = 0.9999$ when re-weighting the imbalanced classes. Using class weighting like this is of course only possible for the classification models, and not for the regression models.

4.2.2 BoW models

The BoW models we used were, as listed in the introduction to this chapter: two support vector machines, i.e. an SVC and an SVR, and a Ridge regressor from *Scikit-learn* (Pedregosa et al., 2011). To get the results we show below in Section 4.3, we chose to use linear kernels in both of the support vector machines. Using linear kernels ensures interpretability and efficient training. For the SVR, all the other parameters were set to the *Scikit-learn* default. Apart from the linear kernel, the SVC parameters were also set to the default, except for using the *class-balanced loss* described in Section 4.2.1 as class weights. For gender classification, we set $\beta = 0.9999$, for ternary sentiment classification, $\beta = 0.9993$, and for classification of the rating from 1 to 6, we set $\beta = 0.9$. The ratings are so imbalanced and the number of training samples so few, that setting $\beta > 0.9$ did not lead to good results. For the Ridge regressor we set the regularization parameter $\alpha = 0.1$ instead of the default $\alpha = 1$ and otherwise left the default parameters unchanged.

To create the BoW representations of the reviews, we use a TF-IDF vectorizer from *Scikit-learn*, to which we pass an argument to remove accents and perform character normalization, and we convert all the input to lowercase. For author gender classification, we use only unigrams, while we use both unigrams and bigrams for critic gender classification and sentiment analysis. When adding bigrams as well, we set the minimum document frequency to three, in order not to fill the BoW representation with bigrams that only occur once or twice in the training data.

4.2.3 Transformer models

The current BERT models are limited to 512 input tokens, as we describe in Chapter 2. In order to address this problem, we have chosen two different truncation methods, forgoing more complex hierarchical methods. We use either **head-only** tokenization, keeping the first 510 tokens or **head+tail** tokenization, keeping the first 128 and the last 382 tokens, depending on

the classification task. For author gender classification, we use **head-only** tokenization, while we use **head+tail** tokenization for the other tasks. The reason for difference will me made more apparent in Section 4.4.1 and Section 5.2.5.

We also use several methods that reduce the memory footprint of the training of the transformer models and increase training speed, in order to fit all the experiments on one GPU. The techniques used were gradient accumulation, gradient checkpointing, mixed precision training, and using an 8-bit Adam optimizer, introduced by Dettmers et al. (2022), insread of the standard 32-bit AdamW optimizer. These techniques were used for all the experiments we performed using transformer models.

Gradient accumulation

Gradient accumulation means to sequentially send smaller subsets of each minibatch (called *microbatches*) through the network, accumulating the gradients until the whole minibatch has been processed. This reduces the memory used during training, allowing us to increase the overall batch size to numbers that would not fit into memory, while increasing the training time slightly (Sohoni et al., 2022).

Gradient checkpointing

Gradient checkpointing also reduces the amount of activation memory, by only storing a subset of the network activations instead of all of the intermediate outputs. This saves memory, but also increases computation, since the activations that are not stored must be computed during the backward pass, write Sohoni et al. (2022). Using this technique, they managed to reduce the memory required for the activations by a factor of 5.8, while increasing the computation required by 30%.

Mixed precision training

Using mixed precision training is primarily to increase the speed of the training, but it can also reduce memory requirements (Micikevicius et al., 2018). The weights of a model are usually stored as 32-bit floating point numbers (fp32). Mixed precision training requires an additional copy of the weights as 16-bit floating point numbers (fp16), increasing the memory requirements for the weights themselves. This increase notwithstanding, Micikevicius et al. (2018) argue that the memory consumption during training is dominated by the activations of each layer, not the model weights, and demonstrate that overall memory consumption using mixed precision training is roughly halved.

In our case, the NorBERT2 weights takes up just below 500 MiB of memory. During training the memory consumption can easily exceed 10 GiB of the 11 GiB available on the GPU, showing that the weights themselves do not consume the most memory.

8-bit Adam optimizer

Stateful optimizers, like Adam, can accelerate optimization compared to plain stochastic gradient descent, but use more memory, which could otherwise be used for model parameters, write Dettmers et al. (2022). They introduce optimizers that use 8-bit statistics while still maintaining the performance of using 32-bit optimizers without requiring any change in hyperparameters. Dettmers et al. (2022) note that 8-bit optimizers only reduce the memory consumption compared to other optimizers, proportional to the number of model parameters. This means that the 8-bit optimizer we use will have a larger effect on XLM-Roberta$_{\text{Large}}$, containing around 2.1 GiB of model parameters, than it does on NorBERT2, with 500 MiB of parameters.

Hyperparameters

Since NorBERT2 and XLM-Roberta are different models and gender classification and sentiment classification are different tasks, we used different hyperparameters. In general, for NorBERT2 we used a batch size of 16 and 8 gradient accumulation steps for an effective batch size of 128. For XLM-Roberta, we used batch size 4 and 8 gradient accumulation steps, for an effective batch size of 32. We used a weight decay of 0.003, learning rate 2e-5 and a cosine learning rate scheduler with a warm-up ratio of 0.3 for all experiments. For our gender classification tasks, we used binary classification with binary cross-entropy loss, allowing us to set the weight of one class. Using the *class-balanced loss* described in Section 4.2.1, we set $\beta = 0.9999$ to re-weight the male class in the loss function.

For sentiment analysis we used NorBERT2 to train models with both classification and regression heads, while we used only regression for XLM-Roberta. In order to re-weight the classes for ternary sentiment classification we set $\beta = 0.999$, and we did not re-weight the classes for rating classification. For the regression models, re-weighting can not be done in this manner.

4.3 Model performance

In large, the prediction tasks that we perform in this thesis are binary gender classification and multiclass sentiment classification. The gender classification is separated into author gender classification and critic gender classification, while for sentiment we predict the ternary sentiment that we introduce in Section 2.1.3. From the original ratings of 1–6, we categorize ratings 1–3 as *negative*, rating 4 as *fair* and rating 5–6 as *positive.*. Despite the negative class comprising the largest range, it contains the least amount of reviews, only 17% of the total, followed by the fair class at 31% and the positive class, which contains 52% of the reviews. For sentiment classification, we also show results for the classification of the rating, to demonstrate the impact of changing back from 3 to 6 ordinal classes. Finally

we include *converted sentiment*, which is the predictions for ratings converted to ternary sentiment predictions.

Getting the highest predictive performance was not the focus of this thesis, but rather interpreting the models, so we did not do any rigid hyperparameter optimization like grid search or more advanced methods for transformer models. However, we did use our knowledge of text classification to set reasonable hyperparameters that have given decent results. In order to interpret the models, it is also necessary to have a certain performance above random chance, otherwise there is no signal on which to explain the predictions.

We first describe the performance results for the BoW baselines and then the results for the transformer models. We show both accuracy and macro f1-score as our performance measures. Since the gender classification is binary and those classes are not extremely imbalanced, these two performance measures show more or less the same, with the f1-score being slightly lower. However, since the sentiment classification is both multiclass and more class-imbalanced, it is useful to have the macro f1-score to show how well the models predict across all classes, not only the largest ones. Since the positive class is the largest, a ternary sentiment classifier could achieve 53% accuracy on the development set simply by predicting the majority class, *positive*, for all samples. That would only give a macro f1-score of 33%, though. Similarly, for rating classification, one could achieve 43% accuracy on the development set by predicting rating 5 for all samples, but the macro f1-score would then be only 17%. Touileb et al. (2021) also used f1-score in their results of sentiment and gender classification on $\text{NoReC}_{\text{gender}}$which eases the comparison between our work and theirs.

4.3.1 Baseline

Table 4.1 shows the performance of the baseline models for classification of author and critic gender. The difference between models here is very small, less than 1.5 pp for any task and split. On the development set, the SVC performs best on author gender classification and the SVR performs best for critic gender classification. It is interesting to note that on the test set, Ridge regression performs best fora all both tasks and performance measures. The most striking feature to notice from Table 4.1, however, is that the performance on critic gender classification increases by around 8 pp from the development set to the test set. The cause of this is not clear, but it could be because the distribution of critic gender for the development set is different from the distribution in the training set, and that the distribution in the test set might be more similar to the training set distribution. In the training set 60% of the reviews are written by male critics, whereas 48% are written by male critics in the development set.

In a similar manner Table 4.2 shows the performance for sentiment classification across the three tasks listed above in Section 4.3. We can see that for ternary sentiment classification, the classifier SVC does better than the regression models by a fair margin. For classification of the rating on the ordinal 1–6 scale, the regression models are better, especially for macro

		dev			test		
gender	metric	SVC	SVR	Ridge	SVC	SVR	Ridge
author	accuracy	**90.6**	90.1	90.1	88.8	89.8	**90.0**
	f1-score	**90.3**	89.7	89.7	88.4	89.4	**89.6**
critic	accuracy	75.7	**75.9**	75.4	83.3	83.0	**83.5**
	f1-score	75.6	**75.9**	75.4	82.8	82.4	**82.8**

Table 4.1: Performance on gender classification for the three used BoW models, support vector classifier, support vector regressor and Ridge regressor, with the best scores for each split and task highlighted.

f1-score where Ridge regression outperforms the SVC by 11.7 pp on the development set and by 8.2 pp on the test set. Ridge regression also has best results for converted sentiment, i.e. the rating predictions transformed back to ternary sentiment predictions, but still not quite as good as the original results of the SVC for the ternary sentiment. It seems that if one wants to classify ternary sentiment, it is best to use a classifier, but for six ordinal classes it is better to use regression.

		dev			test		
sentiment	metric	SVC	SVR	Ridge	SVC	SVR	Ridge
ternary	accuracy	**66,7**	60.5	61.2	**68.2**	63.8	64.1
	f1-score	**61.2**	49.7	53.0	**60.7**	52.7	54.4
rating	accuracy	51.5	**54.7**	54.2	55.3	57.3	**59.0**
	f1-score	26.5	30.2	**37.2**	34.6	27.2	**42.8**
converted	accuracy	60.5	64.3	**65.3**	64.6	66.7	**67.2**
	f1-score	42.9	54.4	**60.5**	47.4	54.8	**59.2**

Table 4.2: Performance for the three used BoW models, support vector classifier, support vector regressor and Ridge regression for sentiment classification, with the best scores for each split and task highlighted.

4.3.2 NorBERT2 and XLM-Roberta

When it comes to the transformer models, the results shown here will be for NorBERT2 and XLM-Roberta. Using these two models can give indications of the power of a large cross-language model compared to a model trained from scratch on Norwegian text, and NorBERT2 is the model we use for further experimentation and interpretation.

Gender classification

Table 4.3 shows that NorBERT2 does better than XLM-Roberta on the development set for both gender classification tasks, but XLM-Roberta does

better on the test set. Like the BoW models, there is a large increase in performance for critic gender classification between the development and test set, especially for XLM-Roberta, with performance increasing by more than 11 pp. It is also interesting to see that while NorBERT2 does better than the BoW models for critic gender classification on the development set and XLM-Roberta does better on the test set, all the BoW models achieve higher performance than NorBERT2 for critic gender classification on the test set. This can be useful, since the BoW models are orders of magnitude smaller than the transformer models and consequently require orders of magnitude less time to train.

		NorBERT2		XLM-R	
gender	metric	dev	test	dev	test
author	accuracy	**97.6**	95.6	96.6	**96.6**
	f1-score	**97.5**	95.5	96.5	**96.5**
critic	accuracy	**79.3**	82.3	73.3	**84.7**
	f1-score	**79.2**	82.2	73.1	**84.2**

Table 4.3: Performance for gender classification between NorBERT2 and Roberta XLM.

Sentiment classification

For sentiment classification using transformer models, we use the same three tasks as for Table 4.2, i.e. ternary sentiment, rating and converted sentiment, but we do not train regression models on ternary sentiment directly. Table 4.4 shows these results. We can see that XLM-R does best for most of the tasks, except that it has quite a bit lower macro f1-score for the rating classification, where it is around 15 pp behind NorBERT2. And while NorBERT2 regression achieves better result for rating classification on the development set, NorBERT2 classification performs better on the test set. Nevertheless, NorBERT2 regression outperforms NorBERT2 classification on converted sentiment, and its converted sentiment results are also better than the results of the classification model trained directly on ternary sentiment across the splits and performance measures. XLM-R does better than both NorBERT2 models on converted sentiment, but it seems it has exchanged some of that performance for a low macro f1-score on the rating classification.

4.3.3 Comparison to previous work on $\text{NoReC}_{\text{gender}}$

Touileb et al. (2021) also trained transformer models on $\text{NoReC}_{\text{gender}}$, but they used NorBERT, since NorBERT2 was not released yet at the time. Since they used binary sentiment, not ternary sentiment, the sentiment classification will be hard to compare, but the gender classification performance can be compared easily. Their standard NorBERT baseline

		NorBERT2				XLM-R	
		classification		regression		regression	
sentiment	metric	dev	test	dev	test	dev	test
ternary	accuracy	74.0	72.1				
	f1-score	70.1	68.0				
rating	accuracy	62.7	64.1	63.9	64.3	**65.1**	**65.5**
	f1-score	45.3	**49.3**	**57.8**	47.3	41.3	34.5
converted	accuracy	74.7	72.3	74.5	74.5	**77.3**	**76.9**
	f1-score	65.9	66.7	72.0	72.5	**73.8**	**73.7**

Table 4.4: Performance for sentiment classification between NorBERT2 and Roberta XLM, and using either a classification or regression head.

achieved 89.6% f1-score for author gender classification on the development set and 90.1% on the test set. This is close to our results using BoW models, shown in Table 4.1, and our transformer models achieve 5 pp or more higher performance across dev and test.

Touileb et al.'s (2021) results for critic gender classification were 70.4% and 63.8% on the dev and test set, respectively. This is where the differences between the performance of our models and theirs are biggest, especially on the test set, where our Ridge model achieves 82.8% macro f1-score and XLM-Roberta 84.2% macro f1-score.

As mentioned, ternary sentiment scores cannot be compared to binary scores easily, but if one removed the fair class from the picture and computed the macro f1-score of only the negative and positive class, XLM-Roberta would have a macro f1-score of respectively 77.5% and 79.6% for the dev and test set. This is close to, but not quite reaching the results of Touileb et al. (2021) at 82.5% and 80.7% mean f1-score for binary sentiment classification using their NorBERT baseline model.

Given these differences in results it seems likely that NorBERT2 is a more powerful model than its predecessor, NorBERT. In the following section, we describe some experiments which regard BERT's document length limitation and the effect of gender on model predictions.

4.4 Experiments

Since BERT is capped at 512 tokens, some analysis was done to investigate the impact of truncating the input texts to different lengths before passing them to the model. This was done both with NorBERT2 and with simple BoW models in order to compare performance and to see how much the BoW models gain when going from a cap of 512 tokens to using all of them. This section contains the result of these analyses.

This section also includes investigations into two of the key questions of the thesis; what is the impact of gender normalization and what is the

impact of adding gender metadata.

4.4.1 Effect of text truncation

The text lengths for truncation that were chosen were the powers of two from 8 to 512 inclusive for NorBERT2. In addition to the mentioned truncation lengths, the BoW models were also at the end trained using the whole texts, to see if there was a big difference between using only 512 and all the tokens. This is why Figure 4.1 has one more data point for the BoW models than for NorBERT2 at the right. Figure 4.1 shows how much the truncation impacts the performance at different text lengths. We can see that NorBERT2 mostly achieves better accuracy than the BoW SVC, even at low truncation thresholds. This is the case for sentiment analysis and for author gender classification, but for critic gender classification the BoW model is performing on par with the BERT model, except at 128 tokens. For the BoW models, increasing from 256 to 512 tokens gives better performance gains than going from 512 tokens to the whole text, suggesting that the 512 token limit of BERT does not significantly affect its performance. However, the reason for the low increase from 512 tokens to the whole text could also be that just around 20% of the texts are actually longer than 512 words. On the other hand, since NorBERT2 uses a subword tokenizer the number of BERT tokens will be larger or equal to the number of words. Because of this, 45% of the reviews contain 512 or more *BERT tokens*. Furthermore, even though XLM-Roberta is not used for any more experiments in this thesis, it can be of interest to note that since it uses a multilingual subword tokenizer, 60% of the reviews contain 512 or more *XLM-Roberta tokens*.

One can see in Figure 4.1 that for BERT, the critic gender classification performance is actually hurt when going from 128 to 256 tokens. This is actually a significant performance drop of 5%. Interestingly, the critic gender classification also shows a performance dip when going from 32 to 64 tokens both for BERT and BoW, and having just the 128 first tokens is better for BERT's performance than having all the first 512 token, suggesting that the text contains little signal for this specific class. Critic gender is, however, BERT's best performing classification task when having just the first 8 tokens.

Another feature to note from Figure 4.1 is that while author gender classification seems to have decreasing marginal benefit each time the truncation length is doubled, both models' sentiment classification get their largest absolute performance increase of 6% when going from 256 to 512 tokens. This could suggest that the critics mostly use the first and middle part of the text to describe the authors, whereas they place most of the sentiment bearing parts of their reviews at the very beginning or towards the end of the text. We can also see that the NorBERT2's critic gender classification performance increases when going from 256 to 512 tokens. These performance increases towards the end of the texts for sentiment classification and critic gender classification are the reasons why we chose to use the **head+tail** tokenization method for these two tasks, and **head-only** for author gender classification.

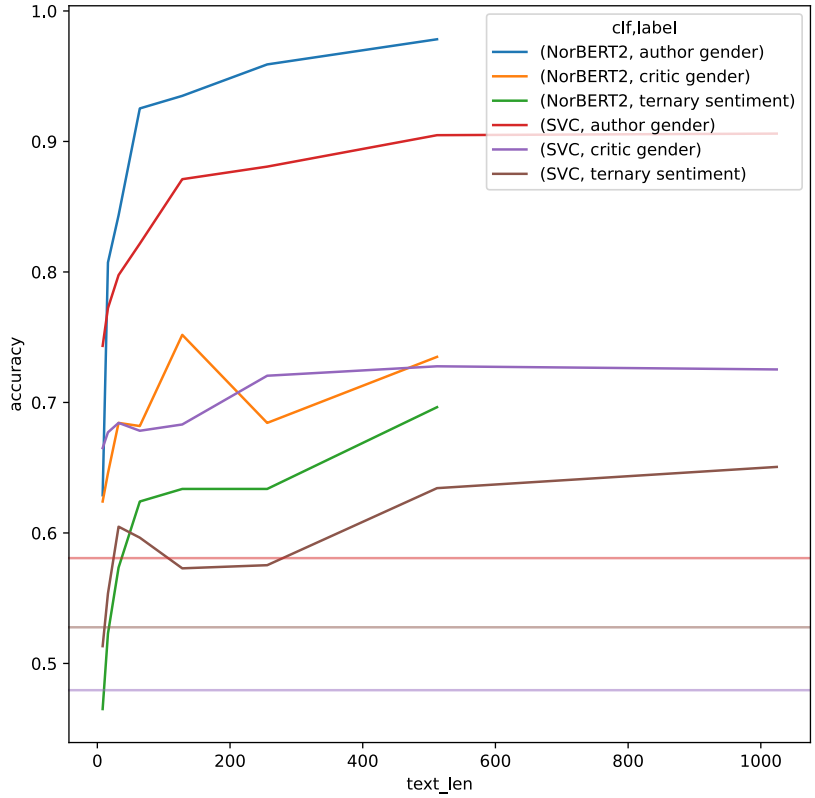

Figure 4.1: Accuracy on the development set for the BoW models and NorBERT2 for different text lengths. The lines that stop at 512 tokens show the NorBERT2 performance, whereas the other lines are for the BoW models. The horizontal lines at the bottom show the majority class baseline for each classification task, using the same color as the BoW line for that task.

The sentiment classification is also the only task that needs more than 8 tokens to pass its majority baseline of 53%, which it does between 16 and 32 tokens. The baseline for critic gender classification accuracy is actually below 50%, which can seem peculiar since it is worse than random chance for a binary task. The reason for this is simply that the majority class for critic gender is different from the training set to the development set.

It is also interesting to see the strange NorBERT2 curve for critic gender classification, stepping up to 75% accuracy on 128 tokens in two large bounds, then falling deep on 256 tokens before ending at 74% at 512 tokens. This performance is worse than the best BoW model for critic gender classification, which is an SVR model achieving 76% accuracy when using the whole input text. It is not clear why simpler BoW models perform on par or even better than NorBERT2 for this task, but it could be that transformers do better when the ratio of signal to noise is higher. Considering that

predicting the gender of a critic just from reviews written by them would be a hard task for humans as well, achieving 75% accuracy is quite impressive. The question, which will be asked again below in Section 4.4.2, is simply why BoW models can perform as well or better than a transformer model for some tasks and inputs.

It is also possible that this task requires separate hyperparameter optimization, which was not done for this truncation impact analysis.

4.4.2 Normalization impact

In order to asses the impact of normalizing away gender information, the models were also tested with input text normalized in different ways. Figure 4.2 shows normalization impact on model performance at 512 tokens. As we have seen before, the ternary sentiment is least affected by the normalization, showing two mostly straight lines with NorBERT2 clearly better than the SVC. For critic gender classification the models perform similarly and not very affected by the normalization, except for BERT when it gets neither names nor pronouns, showing a 4.8 pp decrease in accuracy.

The performance on author gender classification shows both the highest normalization effect and difference between the model types. With no normalization NorBERT2 outperforms the SVC, achieving 98% against 91% accuracy, whereas with full normalization, the SVC gets 0.5 pp higher accuracy than NorBERT2. Another interesting difference is that NorBERT2 performs better with names and no pronouns than the reverse, whereas the SVC is better with pronouns and no names, as can be seen by the divergence between the second and third data point for the upper two lines in blue and orange in Figure 4.2. It also seems like the NorBERT2 model is better than the SVC at using either gendered pronouns or names to replace the loss of the other, while losing both apparently gives the model little to work with, considering NorBERT2's larger relative decrease in performance between half normalization in the middle two columns of Figure 4.2 and full normalization on the right. This could suggest that BERT uses meaningful information better than the BoW models, but when the data contains less signal for the task at hand, the simple models can be sufficient. Since the accuracy is still around 80% for both models, there clearly must still be something to fit the data on. In that manner it is strange that an SVC based on BoW can perform better than NorBERT2. One explanation could be that replacing gendered pronouns with <*PRON*> and person names with <*NAME*> in the input text somehow impedes NorBERT2's ability to use the interaction between the tokens in its contextual embeddings to good effect. For a BoW representation, the interaction between the input tokens is not taken into account, or only to a slight degree between neighbouring words when using n-grams for $n > 1$. Thus having dummy tokens does not affect the impact of other tokens for BoW models as much as it can for transformer models, which uses full self-attention. Still, for ternary sentiment classification, the NorBERT2 performance is barely harmed by the normalization, so it cannot be that the normalization interferes with NorBERT2 in general. We expect gender to carry little signal for sentiment

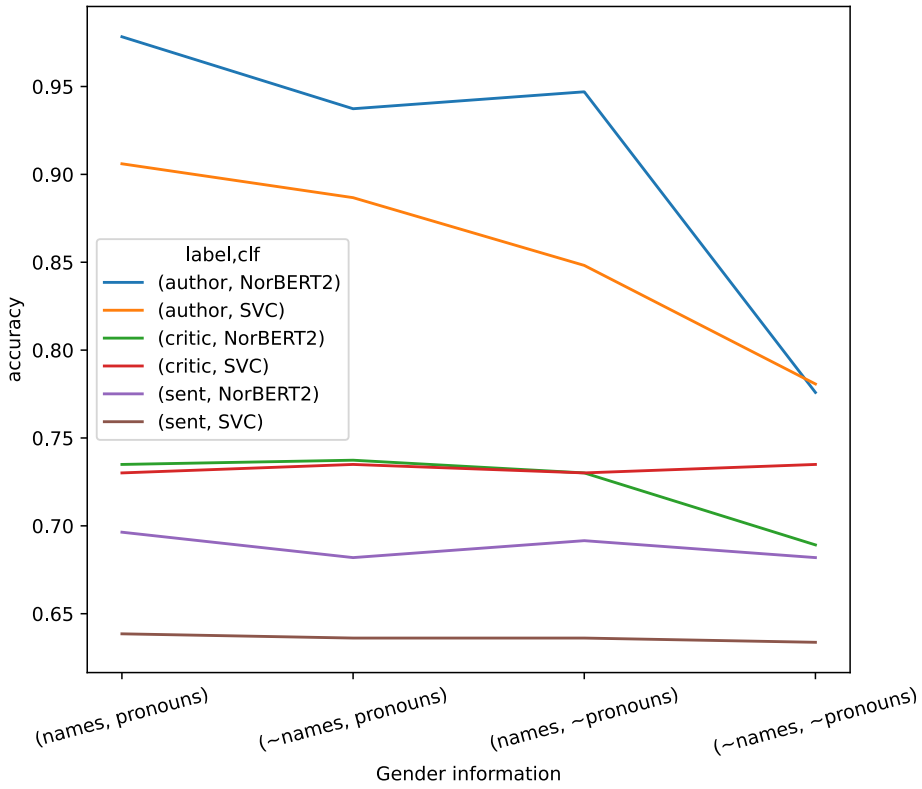

Figure 4.2: Accuracy for the three different tasks when given differently normalized inputs. The x-axis goes from no normalization on the left, to most normalization on the right, where both gendered pronouns and person names are removed.

analysis, and that may be why the sentiment classification performance is hardly impacted by the normalization at all, with NorBERT2 simply being better than the SVC by a fair margin. Normalizing the names seems to have bigger impact on the ternary sentiment accuracy for NorBERT2 than normalizing the pronouns, noting the slightly lower performance in the second and last column of Figure 4.2.

4.4.3 Additional impact of adding gender metadata

Until now we have normalized away the gender information passed to the models to see how much they are affected by gender. Another way of testing how much gender matters to the model predictions is to supply the models with knowledge of gender directly, and see how that affects the performance. Adding gender metadata in this context means to concatenate one-hot representations of author and/or critic gender to the inputs to the classifiers. Using BoW representations of data, we simply concatenate the BoW document vectors and the gender vectors. Using NorBERT2 we first get the pooled output of the hidden layers, and then concatenate the pooled output and the gender vectors before passing it to the output layer.

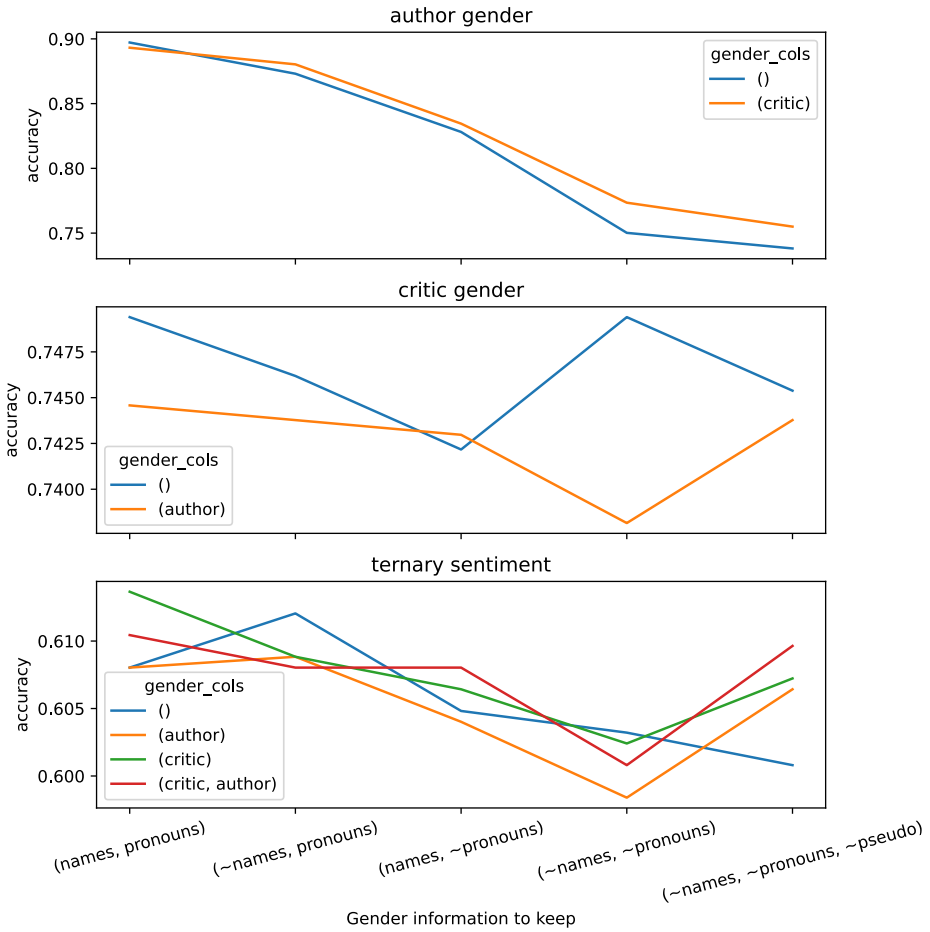

Figure 4.3: Three subplots showing mean accuracy for the three different tasks when given differently normalized inputs and added metadata using BoW SVC, SVR and linear regression models. Each subplot has its own y-axis at different scales. The x-axis is shared between the subplots goes from no normalization on the left to most normalization on the right, where even pseudo-informative features are removed.

Figure 4.3 has three subplots, one for each classification task, and shows mean results across three different linear BoW models; SVC, SVR and linear regression. The five different points on the x-axis represent the degree of gender normalization, from no normalization to the left, to most normalization at the right. For author gender classification, the results are easy to interpret; normalizing gender decreases model performance while increasing the positive effect of adding gender metadata. When removing both names and gendered pronouns, the effect of adding metadata is more than 2 pp, whereas adding no metadata is actually slightly better on the original data.

For critic gender classification and sentiment analysis, the results are less conclusive. Firstly, it should be noted that the scale of the y-axis is a lot smaller for these two subplots at the bottom of Figure 4.3, just a bit

more than 1 pp absolute difference between the lowest and highest value, which was more than 15 pp for author gender classification. Adding author gender metadata seems to harm critic gender classification performance and critic gender classification seems not to have a linear negative relationship between degree of gender normalization and performance. For sentiment analysis, there seems to be a slight negative relationship between accuracy and normalization until one reaches the highest degree of normalization at the right side, where performance increases again. One possibly interesting part to notice is that adding just the critic gender always gives better or equal performance for sentiment analysis than adding just the author gender, although adding both genders is mostly better again.

We also added gender metadata to NorBERT2, the results of which can be seen in Figure 4.4. The model used here for gender classification is a binary classification model. For ternary sentiment, a regression head was trained on top of NorBERT2 using the ratings from 1 to 6. The outputs of that regression model are then thresholded to classify ternary sentiment. The performance of these models is slightly lower than the best performance achieved on each task in Section 4.3. This is because for this experiment the models were trained in total 32 times, so we used a higher effective batch size to expedite the training. We deemed the important factor here to be the relative differences between each normalization scheme and added metadata, not the absolute performance in itself. Since we are going to compare these results with the experiments of Touileb et al. (2021), we also chose to use macro f1-score instead of accuracy as the performance measure in Figure 4.4, as they used f1-score as their measure. For the binary gender classification tasks, the f1-score will not be far away from the accuracy, but the macro f1-score will be lower than the accuracy of ternary sentiment classification due to the label imbalance. Another difference between Figure 4.4 and Figure 4.3 apart from the performance measure is that the latter has one extra column on the x-axis, where pseudo-informative features are removed. These features were only removed for BoW, since we could simply pass them as stop-words to the vectorizer. Removing them could cause trouble for the NorBERT2 models, since the resulting text would not be grammatically correct, and was therefore not done in the NorBERT2 experiment.

In order to add metadata to a classifier based on NorBERT2, we concatenated the pooled output of NorBERT2's hidden layers with one hot representations of the metadata, before passing them to a linear output layer. In other words, the exact same procedure as for the other NorBERT2 experiments in this section, except for the concatenation with the gender vector. This is similar to how Touileb et al. (2021) added the metadata, except that they used a two-layer multilayer perceptron (MLP) instead of just one linear layer after the concatenation.

The results for NorBERT2 in Figure 4.4 are similar to the results for the BoW models in Figure 4.3, but with some differences. For the BoW models, adding metadata improves author gender classification performance when there is gender normalization but not on the original text. The results for NorBERT2 in Figure 4.4 shows that adding metadata *only* improves

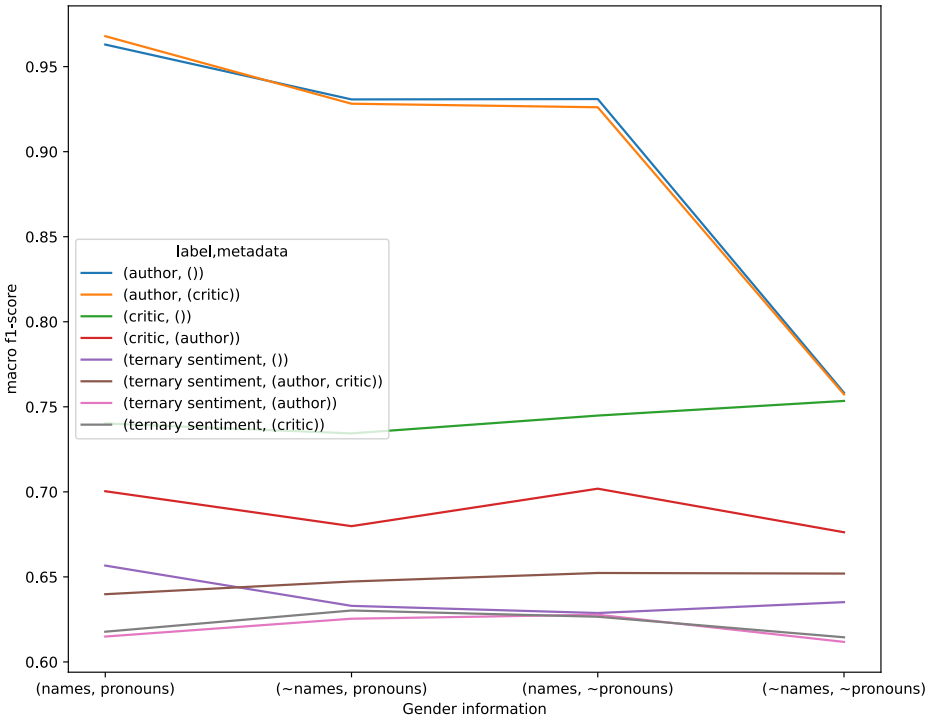

Figure 4.4: Macro average f1-score for the three different tasks when given differently normalized inputs and metadata using NorBERT2. The x-axis goes from no normalization on the left to most normalization on the right, where both person names and gendered pronouns are replaced by dummy tokens.

author gender classification performance on the original text, not when any normalization is done on the input data. Note however that the difference in author gender classification performance between adding critic gender as metadata and not adding metadata is less than 0.5 pp for NorBERT2 in all cases. The critic gender classification performance, indicated by a green and a red line in Figure 4.4, shows by far the highest difference when adding metadata: adding the author gender as metadata decreases accuracy by 3.5 pp on the original text and by more than 7 pp with full normalization. When it comes to sentiment analysis, we could add either gender or both, so more tests were done. Still, Figure 4.4 indicates that not adding any metadata gives best sentiment classification performance on the original text. When gender data is normalized, adding both author and critic gender as metadata gives best performance. The experimental results for metadata impact on ternary sentiment analysis using NorBERT2 are a lot more conclusive than for the BoW models in Figure 4.3, and the only feature shared by the two is that adding critic gender as metadata seems to always be better or as good as adding author gender.

Comparison to earlier work

As mentioned earlier, Touileb et al. (2021) also experimented on adding metadata to the models. The main differences between what they did and what we did here is that we added only gender as metadata, not polarity, and we tested on different normalization schemes, while they only tested on the original text. When we add gender as metadata to predict gender, we only add the other gender, i.e. adding critic gender as metadata when predicting author gender and vice versa. Touileb et al. (2021) instead added the polarity as metadata when predicting gender. There are differences between our results and theirs; most notably their results show increased performance on binary sentiment classification when adding either or both genders as metadata, whereas our results indicate that adding metadata is detrimental to ternary sentiment classification. Of course binary and ternary sentiment are two quite different tasks, and so is using cross-entropy as the loss function compared to mean square error. However, adding metadata for gender classification shows effects in the same direction in both Touileb et al.'s (2021) and our experiments, despite the fact that they added polarity while we added the other gender as metadata. Their experiments shows much higher effect of adding metadata, though, with author gender classification performance increasing by an average of approximately 5 pp between dev and test, and critic gender classification performance dropping by almost 6 pp on average when adding polarity as metadata. In comparison, our experiments showed only an improvement of 0.5 pp for author gender classification performance when adding critic gender as metadata and a decrease of 3.5 pp for critic gender classification performance when adding author gender. The interesting part is that adding metadata increased author gender classification performance while decreasing critic gender classification performance substantially in both cases.

The reason why author gender classification performance increased more for Touileb et al. (2021), could be that the polarity carries more information relevant to the author gender than the critic gender does. Another possible explanation is that the author gender classification model trained by Touileb et al. (2021) had more room for improvement, since its performance went from an average across dev and test of 89.8% to 94.8% when adding polarity, while our model's performance was already 96.3% without adding metadata, increasing to 96.8 when adding critic gender. There are also differences in model implementation: Touileb et al. (2021) write that they use the softmax function as their output layer, suggesting that they use two output labels with cross-entropy loss as their criterion. On the other hand, we classify a single label using the sigmoid function as the output layer and binary cross-entropy loss. As described in Section 4.2, we also try to balance the imbalanced author gender classes by lowering the weight of the male class from 1 to 0.73 in the loss function. No matter the reason for these differences, it is clear that there is more work to be done in order to understand the effects of adding demographic factors to transformer models. Hung et al. (2023) have recently published

work on this topic, and found that while adding demographic factors to a multilingual model is useful, the performance gains are likely due to confounding factors, and not the demographic knowledge itself.

Chapter 5

Interpretability

One of the research questions for this thesis is whether it is possible to use methodology from Explainable Artificial Intelligence to shed more light on what information is used by the models when predicting gender and polarity. In this chapter we explore different methods to investigate this question.

In large, this chapter will be split in two; interpretation of the interpretable BoW models and then the harder task of interpreting the transformer model NorBERT2, which is not interpretable by itself. In this second part, the *Learning Interpretability Tool* (LIT), introduced as the Language Interpretability Tool by Google Research (Tenney et al., 2020), was used to compute the feature attribution scores. LIT is an open-source platform for visualization and understanding of NLP models.

The reason to interpret machine learning models is to be able to tell why they arrived at their predictions. This can help debug and improve the model, build trust in the model, justify model predictions and gain insights, argues Molnar (2022). He includes two definitions of interpretability in his book: *Interpretability is the degree to which a human can understand the cause of a decision*, from Miller (2019), and: *Interpretability is the degree to which a human can consistently predict the model's result*, from Kim et al. (2016).

Notably, both the definitions focus on a human's understanding of the model, and that is the scale on which the interpretability is measured. This also means that there is no objective measure, like multiclass accuracy or mean squared error, to consistently measure the interpretability of a model. The subjectivity of interpretability has also lead to weak theoretical foundations for interpretable machine learning, according to Watson et al. (2021).

In order for humans to interpret a model, one needs a way to get explanations for its predictions, like why the model predicted a positive sentiment for a given input text. An explanation is the answer to a "Why?" question (Miller, 2019), and Pearl and Mackenzie (2018) argue that such questions are actually counterfactual questions in disguise. Since such questions are about causes and their effects, human intuition infers that were it not for the cause, one would not see the effect, otherwise the explanation would not be sufficient. Pearl and Mackenzie (2018) further

argue that this ability is what distinguishes human from animal intelligence and from machine learning. Because of the importance of counterfactuals, a large part of the interpretation of the NorBERT2 models will take place in Section 5.2, dealing with counterfactual analysis. First comes Section 5.1 in which we analyze the most important features for linear models based on a BoW representation of the documents.

Another difference between these two sections, is that interpreting the feature coefficients of linear models is trying to explain the entire model behaviour, at least at a modular level, and not individual predictions. Given an input vector, the coefficients of the model are a complete explanation of its prediction, since the prediction is the dot product of the input and coefficients. However, that is not a useful explanation, since it includes too much information to be interpreted by humans. In a similar vein, we cannot show coefficients for all the features in Section 5.1 below, but only the most salient ones. In the second part of the chapter, Section 5.2, we use methods to explain both global and local behaviour.

5.1 Feature importance for linear models

In this section we will analyze some of the weights for the bag-of-words models, to see what these models use as an indicator for each gender and for sentiment. In his book Interpretable Machine Learning, Molnar (2022, Chapter 4) mentions feature effect plots as a way to analyze linear models, arguing that the weights of the linear regression model can be more meaningfully analyzed when they are multiplied by the actual feature values. Instead of just showing the model coefficients, these effect plots also show how much effect each feature has had across all the predictions. This can be especially helpful for these BoW models, since they use TF-IDF weighting. The feature effect plots for gender classification in this section, Figure 5.1 and Figure 5.2, are plotted from the coefficients of models trained on fully gender normalized data, i.e. with person names and gendered pronouns replaced by dummy tokens and pseudo-informative features for the given class removed. The reason for this is that person names and pronouns are very useful for the model and get weighted highly during training. However, it gives no value to look at the most salient features for a model and see only names, which is why they are normalized away for the interpretability experiments.

29224 gendered pronouns were replaced with <*PRON*>, and 61456 person names were replaced with the placeholder <*name*>. The gendered pronouns replaced were *he, him, his, she, her* ('han', 'ham', 'hans', 'hun', 'henne', 'hennes' in Norwegian). For finding the person names, named entity recognition (NER) was performed with *spacy*[1] (Honnibal, Matthew et al., 2020). There will of course still be words in the text that can reveal gender after normalization, but that is an open list of words that are hard to define in a principled manner, unlike the gendered pronouns and using

[1] spacy version 3.4.2 and the nb_core_news_sm model version 3.4.0

	precision	recall	f1-score
female	0.752	0.592	0.662
male	0.745	0.859	0.798
accuracy			0.747
macro avg	0.748	0.725	0.730
weighted avg	0.748	0.747	0.741

Table 5.1: Classification report for author gender classification using Support Vector Regression

NER. For these analyses 53909 pseudo-informative words were removed for critic gender classification and 54776 for author gender classification.

The next subsections will start with looking at feature effect plots for author and critic gender classification, before doing the same for the sentiment classification.

5.1.1 Author gender

When trying to interpret a model, it is important to know its performance. If the performance of a binary classifier like this is 50% or lower, explaining the predictions wouldn't really make sense, since they would be worse than random chance. That is why this subsection starts with the classification report for author gender classification in Table 5.1. This table shows that the accuracy is 74.7%, lower than the SVC accuracy of 91% when training on the original data, but still better than 50%.

Table 5.1 also shows that the model is skewed towards predicting male gender, with a recall for male author gender of 86% against just 59% recall for female author gender. This discrepancy is likely due to the data imbalance, with 63% of the reviews being about books written by male authors. This also results in a model intercept of 0.57. Since the decision threshold is 0.5, any input text without features highly weighted towards female gender would be predicted as male by default. Weighting the female class higher during the model training increases its recall by a small margin, but not enough to upset the resulting loss in accuracy and macro average f1-score.

When ordered by the maximum effect of each feature, we can see in Figure 5.1 that 'ein' - *a/an/one* and 'en' - *a/an/one* has the highest and third highest effect towards male author genders on the development data. The normalized person name, *name*, has the sixth highest effect, with *name name* a bit further down. Many of the words with high male author effect are quite generic, normal words that occur in most of the texts. Not a single one of them specifically relates to male gender.

For the features that have highest effect towards female author gender, there are a lot more words that are related to the female gender and family, like 'kvinner' - *women*, 'barn' - *children*, 'kvinne' - *woman* and 'moren' - *the mother*. Words like this could also have been removed in the normalization step, but as argued in the start of this section, such gendered words are an open list of words which can not be defined in a principled way. Still, there

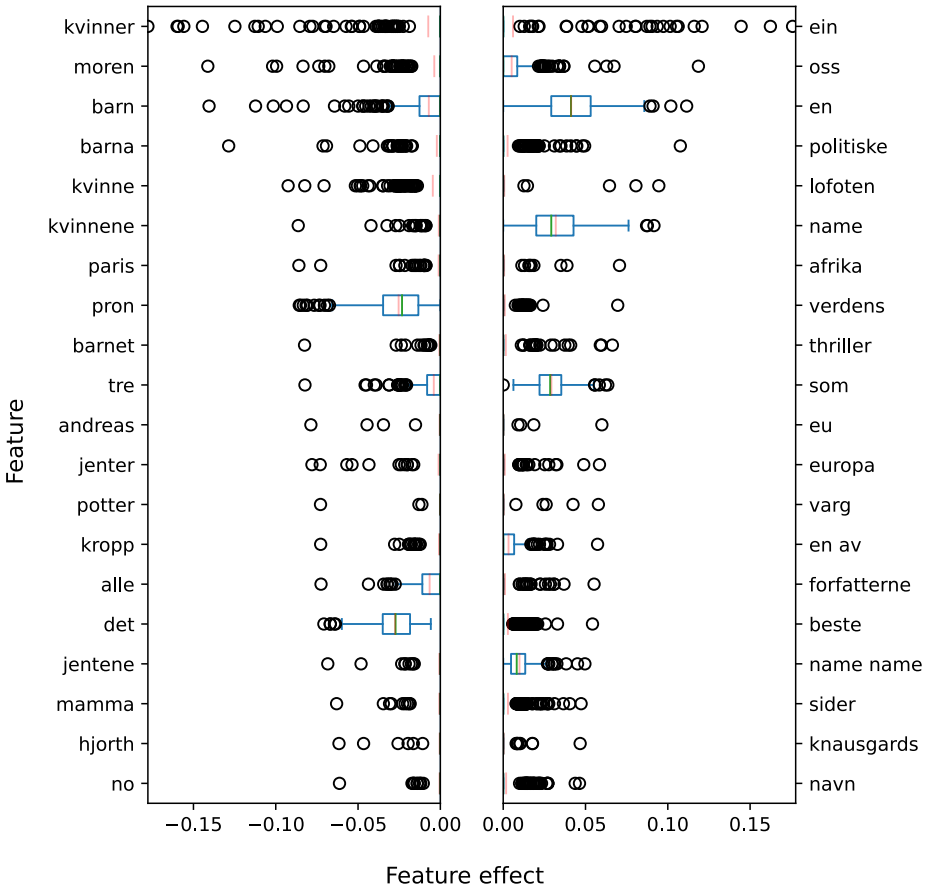

Figure 5.1: A horizontal box plot showing the effects of the 20 most impactful features for author gender classification for each gender, with effect toward female authors on the left side and toward male authors on the right side. The features are sorted by maximum impact across the validation set. The mean impact is marked with a red line for each feature

are quite a few gender-neutral, generic words on the left side as well, like *PRON*, 'det' - *that/it* and 'alle' - *all*. While these generic words do not have the highest maximum effect for a single document, the boxes in Figure 5.1 show that these words have a higher effect on the data set in total. It is also interesting to see *PRON*, which is the normalized substitute of a gendered pronoun, on the left side here, since it replaces both male and female pronouns. Both *name* and *name name* seem to predict male author gender, and *PRON* has a high effect towards female gender, suggesting that male authors may be mentioned more often by name compared to female authors. *PRON* is actually the feature with highest *mean* effect towards female author gender, followed by 'det' - *that/it*, 'er' - *is*, 'boka' - *the book* and 'ikke' - *not*. On the male side there are also some adjectives: 'politiske' - *political*, 'beste' - *best*, and the noun 'forfatterne' - *the authors*. All of this suggests that the male gender is assumed as the default unless anything else is explicitly specified. This 'male default' has also been described in

literature. Perez (2021) writes that men has been seen as the human default for as long as we have data—or rather the absence of data for women. From Aristoteles to Simone de Beauvoir, the woman is defined as 'the other' or as a departure from the default (Perez, 2021, Introduction).

Figure 5.1 is sorted by maximum effect and not mean effect since the maximum effect gives more information about the model's coefficients while still keeping information about each feature's aggregated effect on the whole data set. Sorting by mean would show more words that are used in many of the documents. Those are words like 'en' - *a/an/one*, 'som' - *which* and 'er' - *is*, which come from closed word classes and carry little information. In theory the TF-IDF weighting should make sure that these features, like 'det' - *that/it*, which occur in every document, should not be weighted so highly, but it seems that the smoothing that was used for the TF-IDF vectorization, adding one to every document frequency, might have been too high.

In Figure 5.1 it is simple to spot the features that are used in many documents, since the boxes go from the first to the third quartile, and it is also clear that most of the features present have a median and even 75th percentile of zero. This means that while these features are used in only a small fraction of the reviews, they have a high impact on the classification when they actually are used.

Figure 5.1 also shows that a few names escaped the normalization, giving a high weight to the names 'Andreas', 'Potter' and 'Hjorth'. Exactly why these names were not normalized is not certain.

gender	average word counts								
author	er	en	name	oss	PRON	som	av	alle	det
F	10	7.22	13.64	0.32	7.31	7.79	5.63	0.66	5.70
M	9.20	7.35	14.90	0.39	6.40	7.87	6.03	0.55	5.52

Table 5.2: Average number of times each word has been used per document in the training set, grouped by author gender

As a quick sanity check, Table 5.2 shows how much some of the words high up on the feature effect plot, Figure 5.1, has been used on average when reviewing female or male authors. The word 'er' - *is/are* is not actually in the plot, but has a high mean effect towards female gender. Table 5.2 shows that the words that have an effect toward female author gender are used more in reviews describing women, especially *PRON*, but also 'er' - *is/are* and 'alle' - *all*, whereas the words that have an effect towards male author gender are only used slightly more when male authors are reviewed, like 'en' - *a/an/one*. Even though these are small differences in usage ratio between the genders, they are given a high weight during the training of the models. This shows that the imbalanced nature of the data set, with almost double the amount of reviews of books written by males than books written by females in the training set, means that gender-neutral words that are used as much when describing each gender, will probably be weighted toward

the male gender. Because of this, as well as the mentioned model intercept of 0.57, a gender-neutral text would also probably be assigned male gender. This interpretation is further supported by the model's classification report, shown in Table 5.1, where the recall for female author gender is just below 60% versus a recall of around 86% for male author gender as well as the model's intercept of 0.57.

5.1.2 Critic gender

The classification report for critic gender classification in Table 5.3 shows that the data in this case is a lot more balanced than for the author gender classification and thus the recall discrepancy is a lot smaller here than in Table 5.1. Nevertheless, the model's intercept is 0.56, just 0.01 less than the intercept for the author gender classification model was. From Section 4.3 we know that critic gender classification, i.e. classifying the gender of the writer of the input review, is harder than author gender classification, which in this case is classifying the gender of the person being reviewed. However, after removing features that explicitly reveal gender, critic gender classification actually achieves slightly better performance than author gender classification.

	precision	recall	f1-score
female	0.800	0.704	0.749
male	0.716	0.809	0.759
accuracy			0.754
macro avg	0.758	0.756	0.754
weighted avg	0.760	0.754	0.754

Table 5.3: Classification report for critic gender classification using ordinal support vector regression with a threshold of 0.5 on the development set

Like for the author gender classification, Figure 5.2 shows that most of the features with high impact are common words that occur in a most the reviews, like 'er' - *is/are* and 'det' - *that/it*, which occur in all of them. It also seems that while both *name name* and *name* were used mostly for male authors, here something is different: *name name* as well as *PRON* have high effect for male critics whereas *name* has an impact toward female critics. There are very few explicitly gendered features here, only 'kvinner' - *woman* and 'moren' - *the mother* but these are still only on the left side, i.e. features weighted toward female critic gender.

For the case of critic gender classification, it can be interesting to see the results without normalizing gendered pronouns as well, since these pronouns do not directly give away the gender of the critic as they do for the authors to a much larger extent. Figure 5.3 shows the largest coefficients for each gender when keeping the gendered pronouns unchanged. The figure shows e.g. the probable reason why *PRON* had an effect toward male gender in the Figure 5.2, since 'han' - *he* has the third highest weight for male critics and 'hun' - *she* is not one of the features with the highest weight for

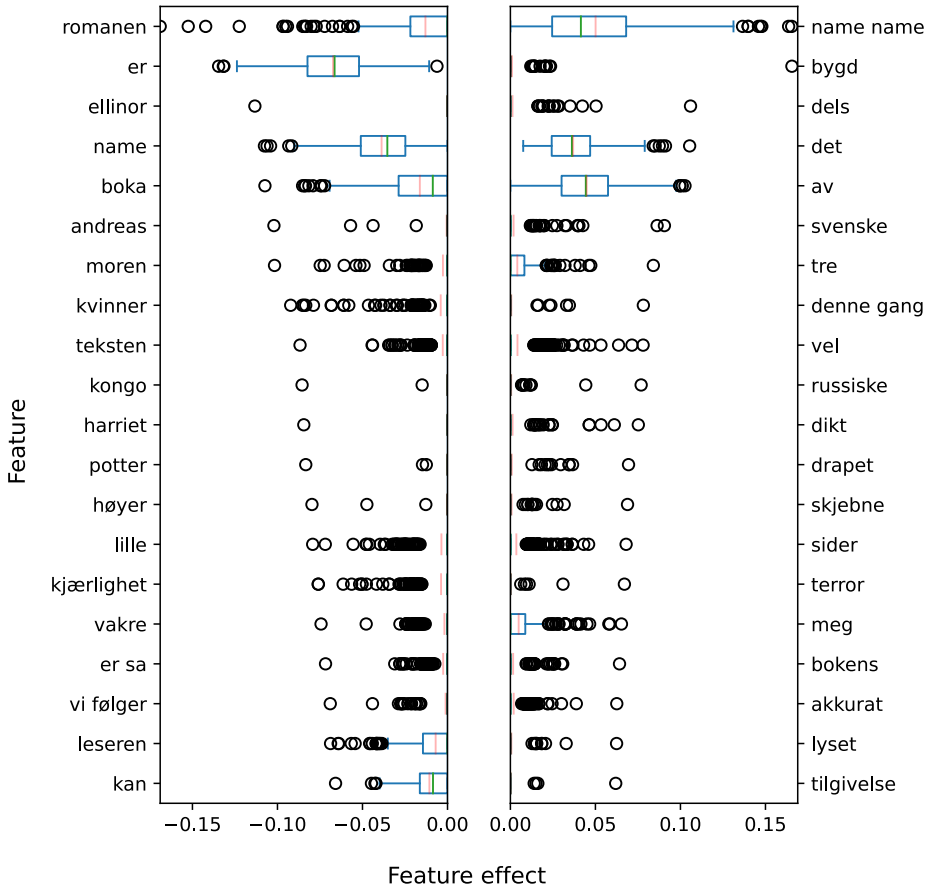

Figure 5.2: A horizontal box plot showing the effects of the 20 most impactful features for critic gender classification for each gender, with effect toward female critics on the left side and toward male critics on the right side. The features are sorted by maximum impact across the validation set. The mean impact is marked with a red line for each feature

female critics. This is probably because male critics mainly review male authors, 'han' - *he* being the word with the highest weight for male authors when it is not normalized away, while female critics review more or less an equal amount of books written by female and male authors.

With regards to the coefficients it is also interesting to see the words that have the highest weights for female critics: 'romanen' - *the novel*, 'boka' - *the book* and 'hovedpersonen' - *the main character* in second place. It is quite striking that the model weights these neutral nouns used to describe books so highly towards the female critic gender. On the male side there are words like that; 'dikt' - *poem*, and 'sider' - *pages/sides*, but 'sider' is mostly used in the text as metadata to tell how many pages the reviewed book consists of. 'Leseren' - *the reader* and 'vi følger' - *we follow* are also both on the left side, while 'meg' - *me* is on the right side. The question here is whether female critics in general use this kind of language more than their male counterparts, or if it is just an artifact of the training data.

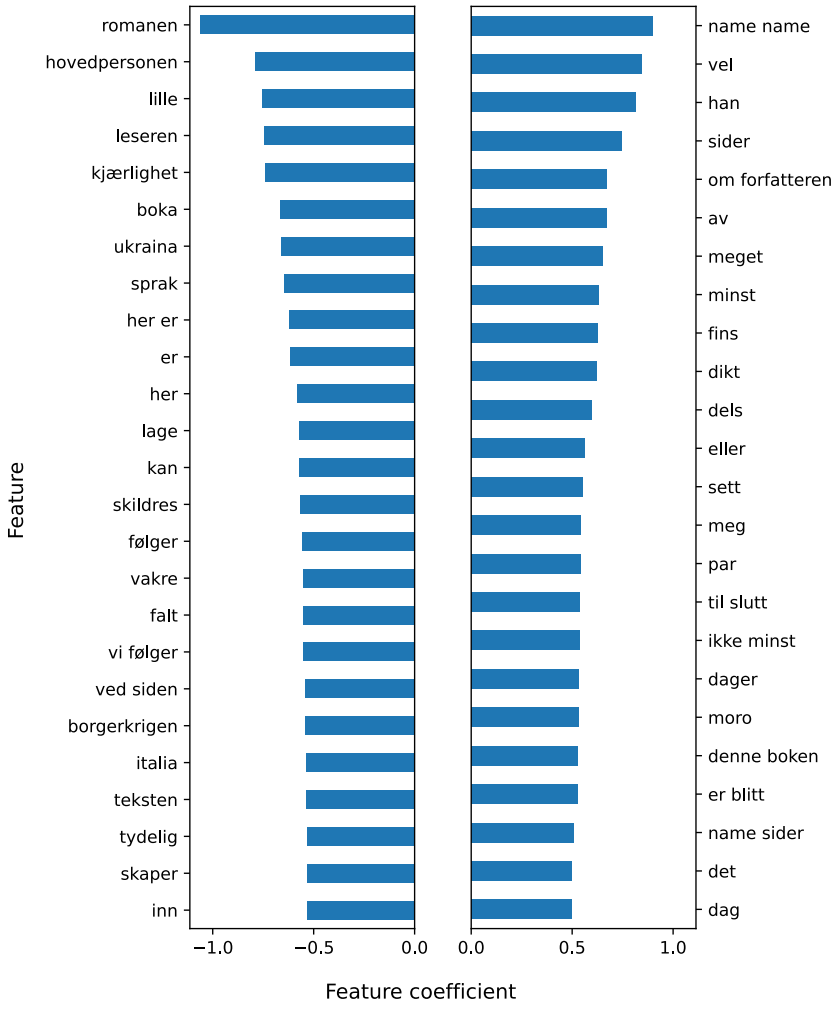

Figure 5.3: A horizontal bar plot plot showing the 25 highest coefficients of critic gender classification for each gender, with effect toward female critics on the left side and toward male critics on the right side.

The common words 'er' - *is/are* and 'her' - *here* are also weighted toward female critics, whereas 'vel' - *well* and 'av' - *of/off* are weighted toward male critics. It is hard to interpret these differences, and they could very well just be spurious correlations in the training data.

The word 'minst' - *least*, most often used in the phrase 'ikke minst' - *not least* (410 of 723 occurrences), has the eighth highest weight for male critic and also high sentiment weight. The same goes for 'meget' - *very (much)*, which is also both an indicator for male critic gender and high sentiment. It is not strange that 'meget' has a high sentiment score, since 'meget' is more often used to modify a positive adjective than a negative one. Coming up with a reason why male critics use this word more than the women would be mostly speculation, but it could be that 'meget', and to a lesser extent its synonym 'svært', belong to a different register than their more general synonym 'veldig', sounding somewhat more old fashioned and pompous

and possibly used more by male critics. According to The Norwegian Academy (n.d.), the use of 'meget' as a degree adverb is 'literary'. The hypothesis of different registers based on critic gender is also somewhat supported by the male critics' use of 'denne boken' - *this book*, shown in Figure 5.3 and 'bokens' - *the book's*, shown in Figure 5.2, with the more formal 'en' suffix as opposed to female critics' use of the more colloquial 'boka' - *the book*, with an 'a' suffix, shown on the left side in both Figure 5.2 and Figure 5.3.

Another difference between the genders is that on the left side of both Figure 5.2 and 5.3 there are expressions like 'leseren' - *the reader* and 'vi følger' - *we follow* whereas on the right side is 'meg' - *me*. This could indicate that the female critics use a more general language describing how *the reader* or an inclusive *we* might experience the book, while male critics might more directly describe how they experienced the book. In all fairness, they are describing the same thing, since it is unlikely that the female critics have done surveys asking several other people about their impression—it is still their own impression of the book they are describing. However, the intent here is not first and foremost to isolate different semantic content, but to identify the different ways and forms by which that content is conveyed.

Table 5.4 shows the sanity check for the critic gender weights. Both 'er' - *is/are* and 'romanen' - *the novel* are used significantly more by female authors. However, it is interesting to see the frequencies of *name* and *name name*. Both *name name* and *name* are used quite a bit more by male critics than female critics. However, female critics use *name* alone more often than male critics do. This explains why *name name* is weighted toward male gender, shown on the top right in both Figure 5.2 and Figure 5.3, whereas *name* is the fourth feature on the left side in 5.2.

critic	er	name/ alone	name name	vel	av	det	meget	romanen
F	10.25	13.82/5.45	3.87	0.12	5.53	5.53	0.03	0.50
M	8.97	14.87/4.61	4.74	0.20	6.12	5.63	0.07	0.28

Table 5.4: Average number of times each word has been used per document in the training set, grouped by critic gender

Table 5.4, along with the Figure 5.2 and Figure 5.3, can give some insight into the difference between the feature weights and their effects. We can e.g. see in the weight plot, Figure 5.3, that 'meget' - *very (much)* has a pretty high weight, but Table 5.4 shows that it is not used often, and neither is it on the effect plot, Figure 5.2. This means that a document containing 'meget' - *very (much)* could push the model towards predicting male critic gender, but assuming that the use of 'meget' - *very (much)* will be somewhat similar in the future, this feature is unlikely to have a large aggregate effect on the model predictions.

The feature effect plots for gender classification Figure 5.1 and Figure 5.2 also show some genre differences between genres. Apparently, male

authors write thrillers, whereas female authors are not associated with a genre in Figure 5.1. Female critics are higly associated with novels, with 'romanen' - *the novel* at the top left of Figure 5.2, while male critics have high weight for poems. We discuss in Section 3.5 whether the reason for the different ratings across genders can be that female and male authors write different genres. However, the feature effect plots do not show enough genres to speculate further on that, especially for author gender classification, since female authors were not associated with a specific genre in Figure 5.1. When it comes to the critics, it could be that novels on average receive lower ratings than poems, but $\text{NoReC}_{\text{gender}}$ does not contain genre information and we can see in Figure 5.1 that 'dikt' - *poem* is not used in many of the reviews.

5.1.3 Sentiment analysis

For sentiment analysis, there will be two classification reports, one for the classification of the rating on the original scale of 1 to 6, shown in Table 5.5, and one for the constructed ternary sentiment classes in Table 5.7

	precision	recall	f1-score
2	1	0.08	0.15
3	0.55	0.34	0.42
4	0.49	0.69	0.57
5	0.60	0.61	0.61
6	0.50	0.12	0.19
accuracy			0.54
macro avg	0.63	0.37	0.39
weighted avg	0.56	0.54	0.52

Table 5.5: Classification report for rating classification using ordinal linear regression, rounding the regression scores to the closest integer to get the predicted class.

The accuracy of 0.54, shown in Table 5.5 indicates that predicting on a scale with six different possibilities is a harder task than binary classification. Table 5.5 only includes five of the six possible ratings since there is no review of rating 1 in the development set and neither were any predictions of rating 1 made. The table shows that the classes on the end of the scale, 2 and 6 have very low recall, compared to the classes in the middle. Rating 4 is regarded as the fair class for ternary sentiment, and this is also somewhat reflected in the model's intercept of 3.91, which by itself would be put in the fair class, like anything else in the interval [3.5, 4.5). As shown in Figure 5.4 the predicted ratings are clumped together in the middle, especially having too many predictions for the fair class, i.e. from 3.5 to 4.5 in the histogram. This class has both highest recall and lowest precision. The model is a bit too conservative and has fewer predictions than the true ratings show for the classes 2 and 6.

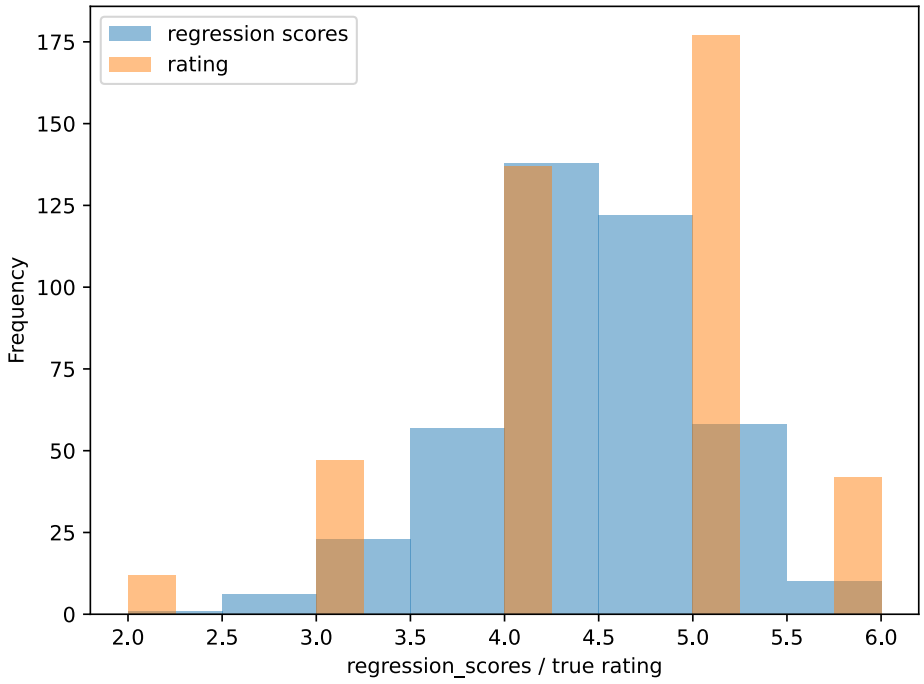

Figure 5.4: Histogram of the predicted regression scores for rating classification overlayed by the true ratings.

This can also be seen in the statistical summary of the predicted regression scores, shown in Table 5.6. The mean of 4.42 is just 0.03 off the true mean of 4.45, but the standard deviation of the true ratings are 0.92, quite far above the standard deviation for the predicted regression scores of 0.59.

count	mean	std	min	25%	50%	75%	max
415	4.42	0.59	2.14	4.07	4.43	4.82	5.80

Table 5.6: Statistical summary of the predicted regression scores for rating classification using Ridge regression.

When one changes the thresholds to get ternary sentiment classification results from the earlier rating predictions, the results look better, as shown in Table 5.7. The score for the fair class obviously stays the same, since that class does not change, but the scores for the positive and negative class increase, resulting in an accuracy of 64%. For a multiclass problem that is good enough to use the weights for interpretation. As the results before thresholding, the fair class has both highest recall and lowest precision.

Figure 5.5 shows the feature effect plot for the rating classification model. The left side, containing features with effects toward negative sentiment is mostly as expected, with words like 'dessverre' - *sadly*, 'lite' - *little* and 'dårlig' - *bad* at the top. The word 'interessant' - *interesting* is also on the left

	precision	recall	f1-score
negative	0.70	0.36	0.47
fair	0.49	0.69	0.57
positive	0.79	0.68	0.73
accuracy			0.64
macro avg	0.66	0.58	0.59
weighted avg	0.68	0.64	0.64

Table 5.7: Classification report for ternary sentiment classification using ordinal Ridge regression, with regression score thresholds of 3.5 and 4.5.

side, which does not intuitively make sense, but that is the only word there that stands out from the rest.

On the right side of Figure 5.5, the issue is the same as in Figure 5.1 and 5.2, where common words with no direct relation to the aspect being classified. In this case 'ein' - *a/an/one*, 'dei' - *they* and 'ein' - *a/an/one*. The first and third word above have the same English translation because they are the same word, but come from each of the two written languages in Norway, *nynorsk* and *bokmål*. The reason the two first words are *nynorsk* is likely because there are just a few critics writing in *nynorsk* and they give higher ratings than the rest of the critics on average.

However, 'en' - *a/an/one* is also there and is the word with the highest mean effect towards positive sentiment. Some of the reason for its high effect might be very simple, the correlation between the rating and the number of times 'en' - *a/an/one* occurs in a text is 0.14, which is just slightly less than the correlation between the rating and the number of words in a text, which is 0.15. It is likely that the word 'en' - *a/an/one* gets some of its high effect simply due to the fact that longer reviews have slightly higher ratings on average. For the sake of completeness, the correlation between the length of a text and the number of times 'en' - *a/an/one* is used is 0.62.

As we know, the TF-IDF vectorizer simply takes the term frequency and multiplies it with the inverse document frequency, meaning that the TF-IDF value for a given word in a document is not relative to the total number of words in the document, but based on the absolute count. Since 'en' - *a/an/one* is such a common word, it would be expected to be used at around the same rate in any document. Thus it is used more times in the longer documents, and since the length of the reviews correlate positively with higher ratings, so does 'en' - *a/an/one*, giving it a high model weight. On the other hand, there are words used almost as much as 'en' - *a/an/one*, like 'det' - *that/it*, which correlates even higher with the length of the document, but has neither a positive correlation with the rating nor a high positive effect on the sentiment. There must also be something else about 'en' - *a/an/one* that gives it the high effect on sentiment, but it is not clear what it is.

Continuing, 'han' - *he* is also placed on the right side of Figure 5.5, just below the middle. It also has the fifth largest mean effect toward positive sentiment. Its coefficient in the model is 0.975, while 'hun' - *she*, which does

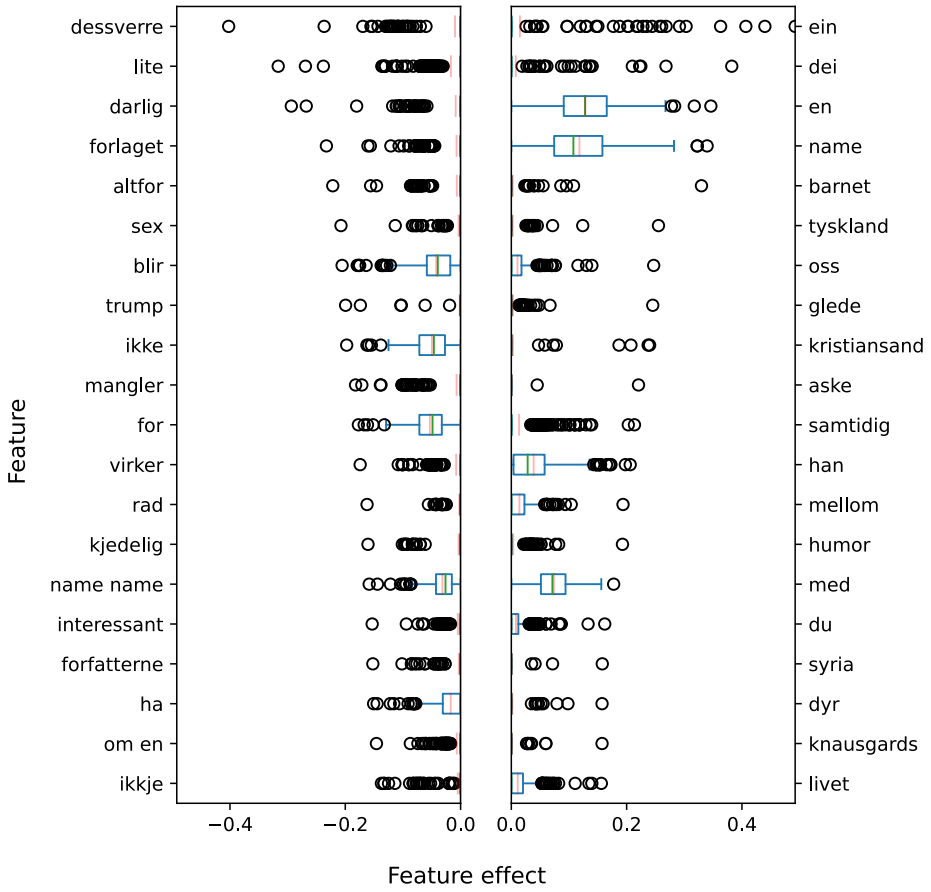

Figure 5.5: A horizontal box plot showing the effects of the 20 most impactful features for sentiment classification for each gender, with effect toward negative sentiment on the left side and toward positive sentiment on the right side. The features are sorted by maximum impact across the validation set. The mean impact is marked with a red line for each feature.

not show up in Figure 5.5, has a coefficient of -0.197. This discrepancy clearly shows the consequences of the fact that male authors are given higher ratings by the critics, which is picked up by the model.

5.2 Counterfactual analysis

As mentioned in the introduction to this chapter 5 and reiterated by several authors (Sundararajan et al., 2017; Pearl & Mackenzie, 2018; Molnar, 2022), one can view any attempt at interpretation as a counterfactual exercise, where humans compare a given cause and effect with the absence of that cause. However, this section is not only implicitly about counterfactual analysis: In Section 5.2.2 we analyze how using counterfactual examples generated by switching the gender of the gendered pronouns impact the model performance. Later, in Section 5.2.5, instances where these

counterfactual examples resulted in another class predictions than the original examples will be analyzed on the token-level.

In Section 4.4.2, similar analyses were made by normalizing the gender information of the input text. However, there are two important differences between what was done there and in this section. Firstly, in Section 4.4.2, the models were re-trained on the normalized text in order to see how much information was left for the models to fit on after normalization. In this section, the models are just trained once on the original text and then tested on the new, normalized text without re-training, in order to see how important the normalized or changed features were to the model predictions. The second difference is that while the gendered pronouns were either kept unchanged or normalized in Section 4.4.2, here they are also switched to the opposite gender. Switching the gender of the pronouns would not make that much sense when re-training the models on the new text, but when using the original model, these gender-switched examples can give insight into how important those pronouns are to the model by seeing how many of the new predictions are different from the original predictions.

5.2.1 Counterfactual generators

Another way to generate new examples with different predictions is to use counterfactual generators. A counterfactual generator creates counterfactual examples for data points by changing them in some way and testing if that changes the prediction of the model (Tenney et al., 2020). HotFlip, proposed by Ebrahimi et al. (2018) is one of the counterfactual generators included in LIT. HotFlip is a white-box adversarial generator that swaps tokens for another based on the gradients of the input embeddings. Originally meant to trick a character-level classifier, it can also be extended to attack word-level classifiers (Ebrahimi et al., 2018), which was done for LIT.

Another counterfactual generator present in LIT is the Ablation Flip, building on ideas from Watson et al. (2021). It generates counterfactuals examples by ablating (removing) one or more tokens and returning the minimal examples that changes the prediction. A generated example is considered minimal if no strict subset of the applied token ablations succeeds in flipping the prediction ('Learning Interpretability Tool (LIT)', 2023).

HotFlip was used more than Ablation Flip, since HotFlip changes the input tokens instead of removing them, thus increasing the chance that the resulting examples are grammatically correct. On the other hand, HotFlip introduces a new factor with the changed tokens, whereas Ablation Flip ensures that any change in the output is just caused by removing some of the input. The following section analyses counterfactual examples generated by switching the gendered pronouns in the input to the other gender

5.2.2 Switching gender

Figure 5.6 shows the effect of switching the gendered pronouns in the input to the model to the opposite gender, without retraining the model on the new input, as was done in Figure 4.2. As in Figures 4.2 and 4.3, the x-axis goes from no gender normalization to maximum gender normalization, but one difference here is that instead of removing pseudo-normal features, the gendered pronouns are switched to the other gender, which to a bigger extent shows the gender's effect on the prediction than only removing the gendered pronouns. Figure 5.6 shows the effect for gender classification, whereas Figure 5.10 shows the effect on the sentiment classification, both using all of the six different ratings as the label and the constructed ternary sentiment classes, with rating 4 as the fair class, rating 5 and 6 as positive and rating 1, 2 and 3 as negative. This division is more or less how these ratings are viewed, with some caveats, as noted earlier, in Section 2.1.2. The BERT model used for this experiment was NorBERT2, upon which a classifier was trained both using multiclass targets and regression targets. For regression targets the class predictions were obtained by rounding the regression scores to the closest integer.

The chosen BoW model for these experiments was Ridge Regression from the scikit-learn library (Pedregosa et al., 2011), which is a regression model whose loss function is the linear least squares function and which uses the l2-norm as regularization. This type of model was chosen for its efficiency and good experimental results on the three classification tasks, especially outperforming multiclass classification models on the rating classification task with all six labels, as we demonstrated in Section 4.3.

Gender classification

For the gender classification tasks one can see two pairs of lines in Figure 5.6; blue and orange for author gender classification and green and red for critic gender classification. The first thing to notice is that critic gender classification is clearly less impacted than author gender classification by the changes in gender information. The green line for NorBERT2 critic gender classification stays more or less straight with some small bumps, whereas the Ridge regression loses up to 10 pp accuracy when both names are normalized and genders switched. For both model types, normalizing names have higher impact than changing the pronouns, 2 pp for NorBERT2 and 8 pp for Ridge. Not surprisingly, the author gender classification exhibits higher impact of changing the gender. Removing the pronouns has some effect, 4 pp down for NorBERT2 and 14 pp down for Ridge, shown between the first and second point on the x-axis in Figure 5.6. The effect is even higher when the names are also removed, where the accuracy for both models drop by 18 pp, shown between the fourth and fifth point on the x-axis. However, the largest impact comes from switching the gender of the pronouns, further dropping the accuracy by approximately 40 pp for Ridge and 24 pp for NorBERT2.

With swapped gender pronouns, the Ridge model's performance is

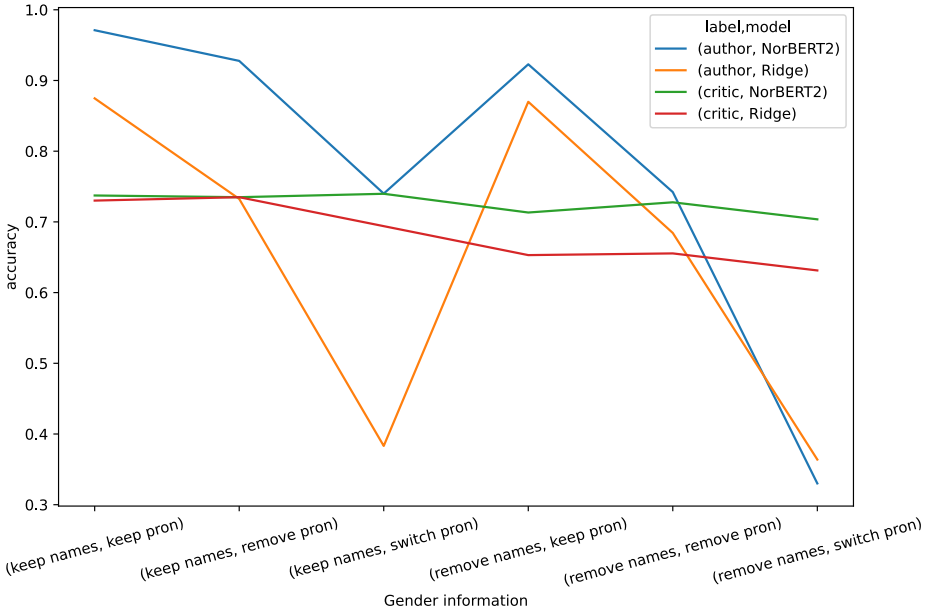

Figure 5.6: Aaccuracy of a Ridge regression BoW model and a NorBERT2 regression model for author and critic gender classification. The models were first trained on the original data, shown on the first tick on the x-axis, and then tested on inputs with varying degrees of gender changes without retraining the model.

below 40%, lower than what a random model would achieve. While NorBERT2 stays above random chance at 74% when names are kept, it drops even further down to 33% when the names are removed. This drop in accuracy is of course expected when one removes or inverts the features that are likely the most important to the model predictions, but the resulting performance is still better than the inverse of the original performance, which means that both of these models must also use some other features than names and pronouns to successfully predict the author gender.

To shed some more light on the accuracy drop, one can look at the Primary Component Analysis (PCA) projections down to two dimensions of the embeddings of the CLS token for the original texts and for the text with switched gender in Figure 5.7 and 5.8. Figure 5.7 shows that the two classes are pushed away from each other—even though the model is not always right, it is certain in its predictions. Furthermore, even at just two dimensions, the PCA explains 86.5% of the total variance, and 84.0% of the total variance is explained by the first component alone. This indicates that the model's representation of author gender to a large degree can be put on a single axis.

Looking at Figure 5.8, one can see that the classes are not pushed apart anymore, here there is a lot more overlap and there is no longer a region in the middle with no predictions. It also looks like the plot flipped along the $y = x$ diagonal, compared to Figure 5.7, since the two classes changed place

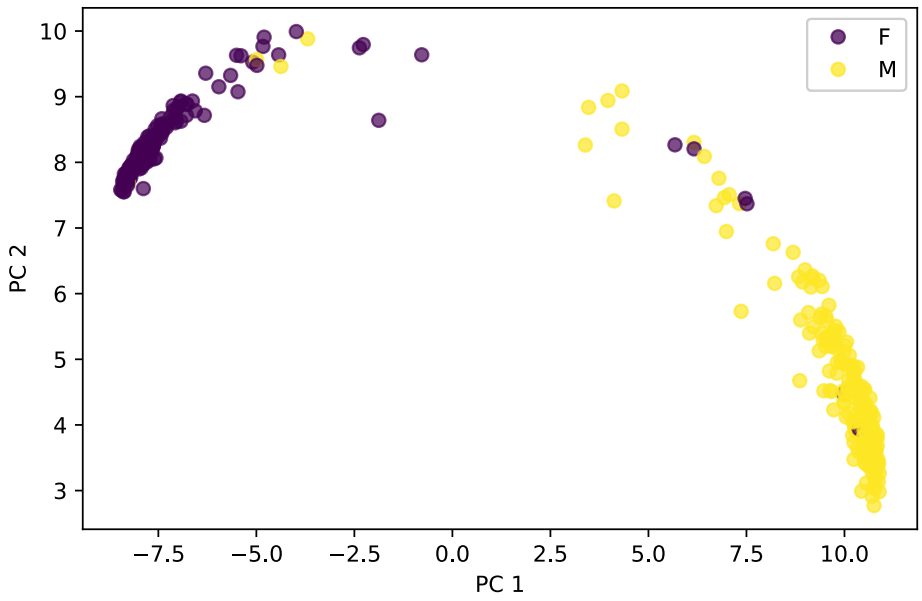

Figure 5.7: PCA of the CLS token embeddings for author gender classification on the original data of the development set, explaining 86.5% of the total variance, with true labels marked by colors.

in the plot, so it could be that what was Primary Component 1 (PC 1) in Figure 5.7 is PC 2 in Figure 5.8 and *vice versa*. Despite the overlap, each class still has the clear majority on one side, which, along with the model accuracy staying above 50%, means that NorBERT2 also uses some information other than the pronouns to predict gender. The PCA in Figure 5.8 still explains as much as 78.0% of the variance, but the second component explains 5.5% of the variance, compared to just 2.5% for the second component of the PCA of Figure 5.7.

Table 5.8 shows a statistical summary of the logits predicted by the model. The first and third quartile show that most of the logits are pushed far away from the decision boundary of 0. The mean of 0.72 and the median of 3.99 reflect that most of the authors reviewed are male. When the gendered pronouns are switched, however, the mean is -0.43 and the median -1.85, which indicates that switching gender also moves the majority of the predictions to the female author gender. Table 5.9 summarizes the same data, but using absolute values of the logits. The minimum of 0.59 for the original data means that all the data points are at least 0.59 away from the decision boundary of 0. This distance from 0 to the predictions likely corresponds to the central gap in the class embeddings seen in Figure 5.7. When the gendered pronouns are switched, the model is not as sure about the predictions, having absolute logits closer to 0 with higher standard deviation, as shown in the second row of Table 5.9.

Lastly, one interesting part to notice in Figure 5.6 is that for the Ridge model, removing names while keeping pronouns reduces the critic

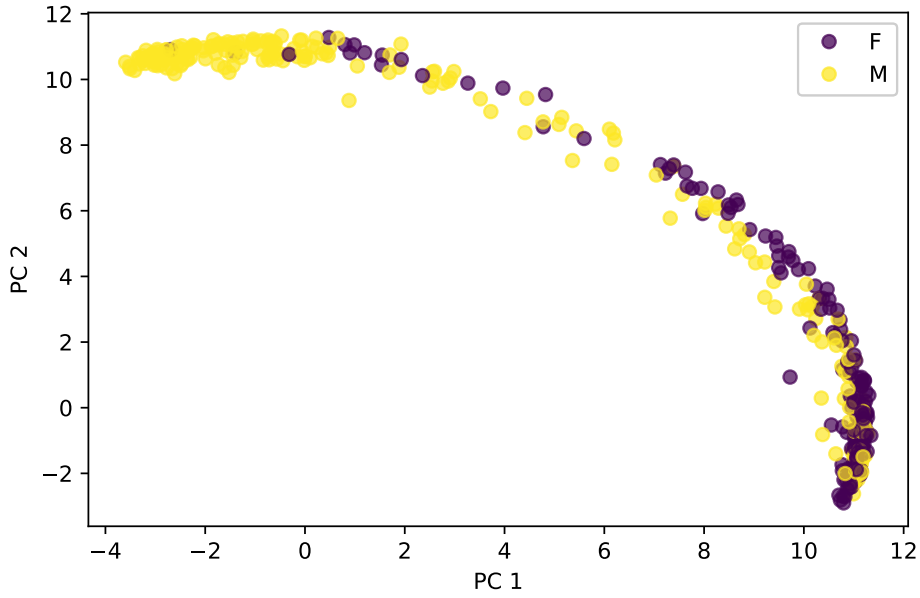

Figure 5.8: PCA of the CLS token embeddings for author gender classification on the gender-switched data of the development set, explaining 78.0% of the total variance, with true labels marked by colors.

normalization	mean	std	min	25%	50%	75%	max
pronouns unchanged	0.72	4.36	-4.77	-4.49	3.99	4.65	4.92
switch gender	-0.43	3.76	-4.71	-4.10	-1.85	4.00	4.87

Table 5.8: Statistical summary of the predicted logits for author gender classification using NorBERT2

gender classification performance by 8 pp but author gender classification performance by less than 1 pp. For NorBERT2 the inverse is the case, with the accuracy dropping by 2 pp for author gender classification and 5 pp for critic gender classification. Since the development set has 415 samples, a change of 1 pp signifies a difference of 4 reviews. It seems strange that there is this difference in the utility of names for the different model types and classification tasks, but it might be due to the way names are tokenized by BERT compared to simple TF-IDF document vectors, as will be further discussed for Figure 5.10, regarding the sentiment classification.

Figure 5.9 shows the two-dimensional PCA projection for the critic gender classification model. Compared to the PCA for author gender classification in Figure 5.7, the data points are much more spread out, meaning that it is harder to place the class embedding for critic gender classification on a single axis. The PCA for critic gender classification explains only 51.5% of the total variance as well, 46.1% by the first component and 5.4% by the second component. This indicates that the class embeddings for critic gender classification are harder to place on one

normalization	mean	std	min	25%	50%	75%	max
pronouns unchanged	4.37	0.65	0.59	4.35	4.58	4.70	4.92
switch gender	3.58	1.22	0.07	3.12	4.07	4.44	4.87

Table 5.9: Statistical summary of the absolute values of the predicted logits for author gender classification using NorBERT2

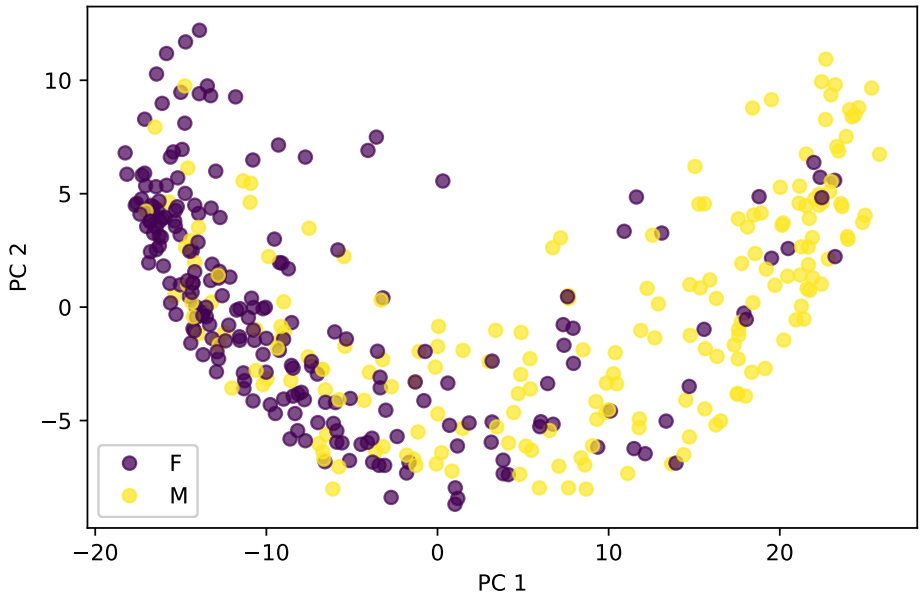

Figure 5.9: Two-dimensional PCA projection of the CLS token embeddings for critic gender classification on the original data of the development set, explaining 51.5% of the total variance, with true labels marked by colors.

axis, which makes sense, since there are no clear-cut features to predict the gender of the critic.

Sentiment classification

Then we continue with the sentiment classification, which in earlier experiments in Section 4.4 have not been much affected by gender normalization. The sentiment plot of Figure 5.10 can be divided into three pairs of lines. In brown and green lines, seen together in the bottom right part of the figure, is the accuracy for ternary sentiment and rating classification for the Ridge regression model. The rating classification accuracy for NorBERT2 regression and multiclass classification are the orange and blue lines in the middle and subplot, and on the top are the ternary sentiment score for those two NorBERT2 models. The accuracy for ternary sentiment will always be as good or better than the accuracy for the rating, since they are computed from the same predictions. The ternary sentiment metrics are computed simply by mapping the predicted and true

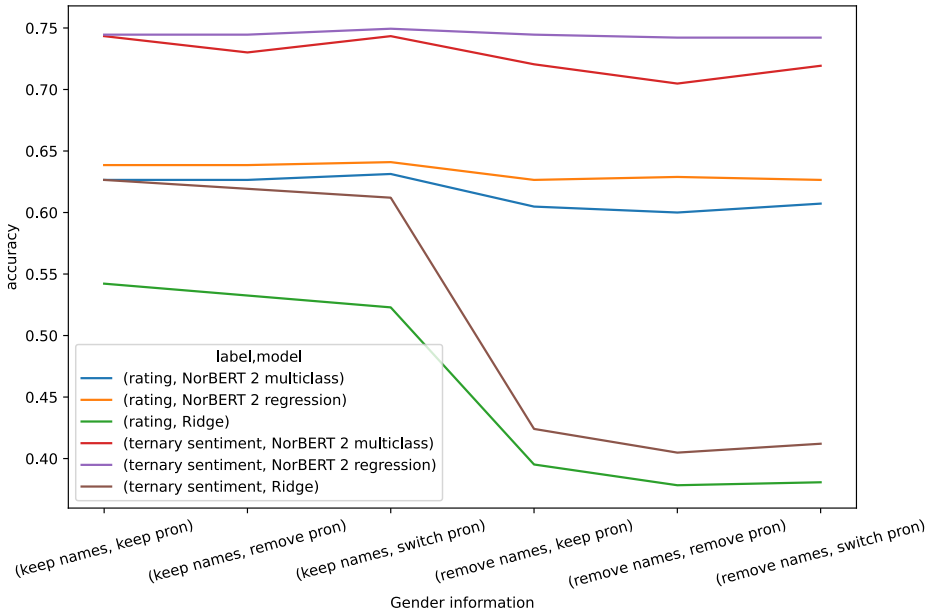

Figure 5.10: Accuracy of a Ridge regression BoW model, a NorBERT2 (ordinal) regression model and a NorBERT2 multiclass model for sentiment classification. The models were first trained on the original text, shown on the first tick on the x-axis, and then tested on inputs with varying degrees of gender changes, without retraining the model.

ratings from the original six classes into three classes as described in the first paragraph of Section 5.2.2.

The first thing to take away from Figure 5.10 is that the NorBERT2 models are mostly unaffected by changes in gender information passed to them as input, whereas the Ridge model loses 15 pp accuracy for rating classification and 20 pp for ternary sentiment classification when names are removed. Comparatively, changing the gendered pronouns has little effect, with less than 2 pp decrease in accuracy for Ridge regression. The NorBERT2 multiclass model also slightly more affected by removing names than changing gender, with a decrease in accuracy of 2.5 pp for removing names against 1.5 pp for removing pronouns. However, switching the gender of the pronouns actually leads to better performance than removing them for the NorBERT2 models.

The reason why the Ridge regression does so much worse when names are removed may be twofold: Firstly, the names of authors are likely a good predictor for the rating, since different critics likely assign, if not equal, then at least similar ratings to the same book, and also to some extent to different books by the same author. This means that some authors should have higher average ratings for their reviewed books than others, and removing the names denies the model information it had been using to better predict the sentiment. However, in that case why are the BoW models impacted so much harder than the NorBERT2 models? This question leads to the second

argument: for BoW models, each name is a separate feature, whereas most of the names are unlikely to be part of the NorBERT2 vocabulary as whole units, but rather that parts of the names are tokenized as subtokens by NorBERT2's subword tokenizer. This means that the embeddings for the names will be shared by whichever subtokens they consist of, which is likely to have a regularizing effect on the BERT models' weights for those names.

Comparing the two NorBERT2 models, the multiclass classifier is slightly more affected by the gender changes than the regression classifier, especially for the ternary sentiment. On the ternary sentiment task, the NorBERT2 regression model performance decreases by less than 0.25 pp when both names are removed and pronouns removed or switched. The multiclass model performance on ternary sentiment, on the other hand, is more impacted by the gender normalization, falling by 4 pp when both names and pronouns are removed. This could indicate that the multiclass model is slightly less robust to changes in the input than the regression model. Another factor to this is that for rating classification, the difference between the two NorBERT2 models fluctuates less, staying between 1 and 3 pp versus the difference between 0 and 4 pp for ternary sentiment. What this could tell us is that when the regression model gets different input and makes mistakes, it is more likely than the multiclass model to change the prediction by just a little and stay within the same ternary sentiment class, since the ordinal regression model can benefit from the ordered nature of the ratings (Gutiérrez et al., 2016).

Figure 5.11 shows the PCA of the class embeddings for the ordinal regression model for sentiment classification. Here we can see that the model manages to put the reviews from the development set on a scale from 2 to 6, though with quite a bit of overlap between the classes. Despite the overlap, Figure 5.11 might give some insight into why using regression for ordinal classification can be helpful, since it shows how the regression model can use the ordered nature of the ratings.

Table 5.10 shows the predicted regression scores of the NorBERT2 regression model for rating classification. Compared to the predicted regression scores for the Ridge model, shown in Table 5.6 in Section 5.1.3, the NorBERT2 model manages to push the ratings further out from the mean, having a standard deviation of 0.83, which gets close to the true standard deviation of 0.92. The minimum and maximum regression scores of 1.67 and 6.08 are outside the range of the true values of the rating, which lie between 2 and 6 inclusive in the development set.

count	mean	std	min	25%	50%	75%	max
415	4.37	0.83	1.67	3.84	4.44	4.98	6.08

Table 5.10: Statistical summary of the predicted regression scores for rating classification using a NorBERT 2 regression model.

After looking at aggregate effects on model performance by counterfactual examples made by switching gender in this section, the following

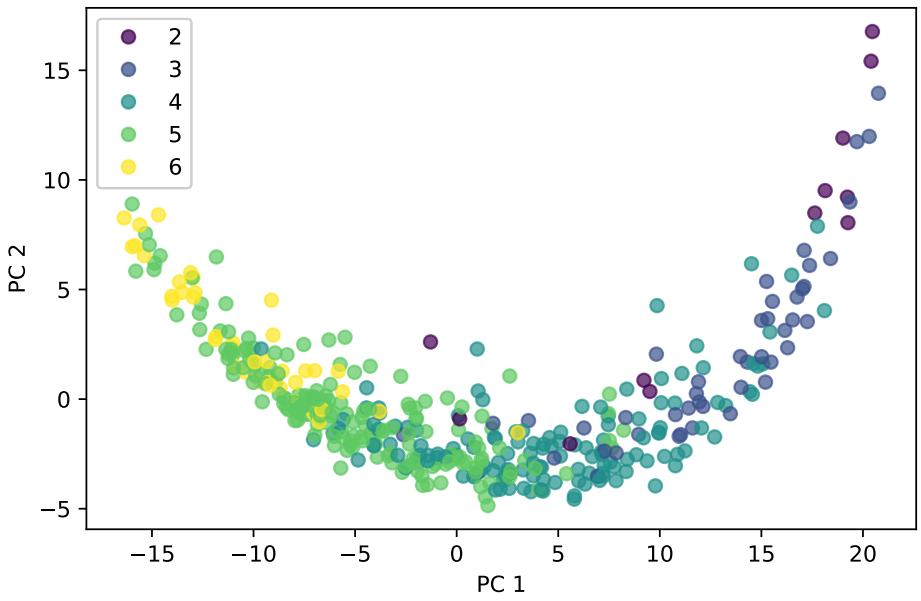

Figure 5.11: Two-dimensional PCA projection of the CLS token embeddings for sentiment using ordinal regression on the original text of the development set, explaining 65.3% of the total variance, with true ratings marked by colors.

section will deal with how the different input features impact the prediction of the model.

5.2.3 Feature Attribution

Feature attribution is to explain individual predictions by attributing each input feature according to how much it changed the prediction (Molnar, 2022). In some articles it is also called input saliency (Bastings & Filippova, 2020) or salience maps (Tenney et al., 2020). LIT contains four different feature attribution methods: gradient norm, Gradient-dot-input, Integrated Gradients and Local interpretable model-agnostic explanations (LIME). Using Molnar's taxonomy (2022, Chapter 9), the three first of these methods are gradient-based, while LIME is perturbation-based. Perturbation-based methods are model-agnostic, and work by changing parts of the input to generate explanations. LIME, proposed in 2016 by Ribeiro et al., uses these variations on the data to train an interpretable model which should be a good approximation to the original model locally, but not necessarily globally. Then, the predictions of the original model can be explained by interpreting the local model.

Gradient-based attribution methods

The gradient-based methods compute the gradient of the prediction with respect to the input features (Molnar, 2022, Chapter 9). This means that

the attributions from these methods will always be relative to the model prediction of a single class, unlike the analysis of BoW model weights in Section 5.1, where the weights are the same regardless of what the model output is for a given input. Gradient-based methods are further divided into gradient-only methods and path-attribution methods. Gradient norm and Gradient-dot-input are gradient only methods. Their interpretation means that if one increases the value of the input feature, the class probability would go up for a positive gradient and down for a negative gradient. For images, whose features are pixels, that would mean increasing or decreasing the color values values of a pixel (Molnar, 2022, Chapter 9). For NLP, it is hard to say what increasing an input feature would mean. However, for a given input token, a positive attribution means that removing that token from the input would likely reduce the model confidence for its current output. Reversely, when the attribution is negative, removing the token would likely increase the model confidence (Tenney et al., 2020).

Integrated Gradients

Integrated Gradients is a path-attribution method, proposed by Sundararajan et al. (2017). Path-attribution methods compare the current input to a baseline input, which can be a black or grey image for object detection and a zero vector embedding for NLP (Sundararajan et al., 2017). The interpretation of path-attribution networks is always with respect to this baseline, using the difference between the classification scores of the actual input and the baseline as the attribution (Molnar, 2022). Sundararajan et al. (2017) make a remark of this baseline as a way for humans to interpret cause and effect by counterfactual intuition: 'When we assign blame to a certain cause we implicitly consider the absence of the cause as a baseline for comparing outcomes'.

Sundararajan et al. (2017) write that a challenge for designing attribution methods is that they are hard to evaluate empirically, since it is hard to distinguish errors stemming from model misbehaviour from misbehaviour of the attribution method. Their Integrated Gradients method was designed using what they call 'an axiomatic approach' in order to compensate for that shortcoming. What they mean by an axiomatic approach is that they first define two axioms, or desirable characteristics an attribution methods should satisfy. Further, they let those axioms guide the design of their method, in such a manner that they can prove mathematically that their method satisfies the axioms they defined. Sundararajan et al. (2017) finally argue that this approach ensures there are no artifacts of their method which affect the attribution. Nevertheless, they acknowledge that their method does not address interaction between the input features or the logic employed by the neural network.

Sundararajan et al. (2017) identified *Sensitivity* and *Implementation Invariance* as the axioms an attribution method should satisfy, and found that most methods did not satisfy them. In order to satisfy *Sensitivity*, for input and baselines that differ in one feature and result in different predictions, that feature must have a non-zero attribution. To satisfy

Implementation Invariance, the attribution for networks that give equal output for all inputs must also be equal. Sundararajan et al. (2017) show that since gradients are invariant to the implementation, *Integrated Gradients* are also implementation invariant.

Alternatives to Integrated Gradients

Despite Integrated Gradient's desirable characteristics, it has both been criticized for not being worth its increased complexity compared to other gradient methods, and it has also been extended to improve its performance. Madsen et al. (2022) argue that attention gives sparser explanations than gradient methods, and are easier to understand. Their results also indicate that using Integrated Gradients, being approximately 50 times more expensive than the *gradient* method, is rarely a worthwhile trade-off.

Other articles again discuss whether attention can really be seen as an explanation of model predictions, arguing that it is unclear toward what attention is used as explanation, and finding input saliency methods more suited (Bastings & Filippova, 2020). Nevertheless, they describe some fundamental limitation of input saliency methods, which is that the flat representations of per-token saliency weights can only be called an explanation in a narrow sense, and it is impossible to fully explain the predictions of a deep non-linear model by only looking at the input tokens. Despite this limitation, Bastings and Filippova (2020) still argue that saliency methods are useful, and they do not find that there are many other possible alternatives at the point of writing. One of the alternatives they mention is counterfactual analysis, and we will explore further down in Section 5.2.5 some of the limitations of the saliency method Integrated Gradients and how counterfactual analysis can help with the interpretation.

Sanyal and Ren (2021) propose an extension to Integrated Gradients, called Discretized Integrated Gradients, due to what they argue is a core limitation of the Integrated Gradients method. Being a path-attribution method, Integrated Gradients uses a straight line interpolation between the zero-vector baseline and the word embedding. However, since the word embedding space is a discrete space, these interpolation points are unlikely to be close to actual word embeddings and may not be representative of the word embedding distribution, putting the faithfulness of the gradients computed from these interpolation points to the question. Discretized Integrated Gradients mitigate this limitation by avoiding the straight line interpolations, instead using a non-linear path with monotonically situated interpolation points between the baseline and input word embedding that are close to real word embeddings (Sanyal & Ren, 2021).

The baseline used by the *Integrated Gradients* implementation of LIT is the zero embedding vector, which is suggested as a good baseline by Sundararajan et al. (2017) in their original paper, even though it does not correspond to a valid input. In the following section we will analyze the feature attributions on the three tasks author gender classification, critic gender classification and sentiment classification, both in the aggregate and on individual data points, i.e. reviews. Since the reviews are long, it would

not make sense to show attributions for all 512 tokens, so some of the most salient sentences have been manually chosen for analysis and discussion. To choose the sentences for analysis, the most interesting reviews were first selected based on counterfactual analysis: did the counterfactual examples lead to a change in model prediction? Secondly the parts of the text with the highest attribution score or containing the counterfactual change are used as examples below, like Example 5.2, Example 5.9 and Example 5.11.

5.2.4 Counterfactual explanations

One of the problems with feature attribution methods, as mentioned by Bastings and Filippova (2020), is that it is a flat representation of attribution scores, trying to give a complete explanation for the predicted class. Not complete in the sense of explaining every factor and interaction that led to the prediction, but in the sense that it gives a score to every input token. However, as Molnar (2022, Chapter 2) mentions, good explanations are contrastive. Humans seldom want to know all the reasons for a predictions, but just why that prediction was made instead of another prediction. Since feature attributions scores always are in relation to a single class, they cannot be contrastive in that sense.

Necessary and sufficient causes

This also leads to the concepts of necessity and sufficiency, which Watson et al. (2021) argue are the building blocks of all successful explanations. In propositional logic, x is a sufficient condition for y iff $x \rightarrow y$, and x is a necessary condition for y iff $y \rightarrow x$ (Watson et al., 2021). The distinction between necessary and sufficient causes has important implications in AI, claims Pearl (1999), where necessary causation is a concept tailored to a specific event under consideration, while sufficient causation is based on the general tendency of certain event types to produce other event types. Pearl (1999) also shows that necessity and sufficiency are independent aspects of causation, both of which should be used when making causal explanations.

In *The Book of Why*, Pearl and Mackenzie (2018, Chapter 1) use a firing squad of two soldiers as an example of a sufficient cause. Either of the two soldier firing at the prisoner is sufficient to cause their death, but neither is necessary, since there are two soldiers. Later, Pearl and Mackenzie (2018, Chapter 8) demonstrate necessary causes by using the example of a fire breaking out after someone struck a match. 'What caused the fire, striking the match or the presence of oxygen in the room?' Both are necessary causes for the fire, since if either of them were not present, the fire would not have happened. Despite that, a human is more likely to explain the fire by the match, since that is the factor that changed in that moment and is not ever-present. This shows how background factors that are normally present in the world, like oxygen, can qualify as explanations if those explanations are based solely on necessary causation.

Feature attribution methods versus counterfactual analysis

The feature attribution methods introduced in Section 5.2.3 give neither sufficient nor necessary causes. A feature can have a very high attribution score, but one still cannot be sure if it is either sufficient or necessary for the prediction. One cannot say, "if feature x is present, the prediction will be y", nor "if feature x is not present, the prediction will not be y", examples of sufficient and necessary causation, respectively, no matter what x's attribution score is. The feature attribution methods are still not useless, though, since they can give insights into the relative importance of different features. Counterfactual examples, on the other hand, can help identify causes. Pearl (2000, Chapter 7) writes "Event $X = x$ may have caused $Y = y$" if:

(i) $X = x$ and $Y = y$ are true; and

(ii) there exists a value u of U such that $X(u) = x$, $Y(u) = y$, and $Y_{x'}(u) \neq y$ for some $x' \neq x$,

where U is a set of background variables. Thus in a counterfactual example, if changing feature X from x to x' causes a change in prediction Y from y to y', everything else being the same, then x may have caused y. One can still only talk in probabilities, and Tian and Pearl (2000) define three probabilities that are relevant: Probability of necessity (PN), probability of sufficiency (PS) and Probability of necessity and sufficiency (PNS). PN stands for the probability that event y would not have occurred in the absence of event x, given that x and y did in fact occur. PS stands for the probability that event y would occur if x would have occurred, given that x and y did not in fact occur. PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y.

The counterfactual examples can also be seen as a way to test the faithfulness of the gradient methods. Knowing that two inputs give different results, the differing input features should have different non-zero attributions for both the class predicted originally and the class predicted for the counterfactual example. This also more or less corresponds to the sensitivity axiom that Sundararajan et al. (2017) define.

5.2.5 Interpreting gradients

Since BERT is a deep neural network with distributed representations, it can be seen as a black box, hiding its reasons for why it outputs what it does, unlike the linear models analysed in Section 5.1. One way to interpret BERT's decision is to use gradients, as described in the previous section. This section covers some analyses made of feature attribution using gradients, first on an aggregate level across the whole development set and then using the manually chosen sentences from the development set.

Gradient Norm

Firstly, Figure 5.12 shows how salient the tokens in each position is across the development data set. This is a way to investigate on aggregate what parts of the texts are most important for the output. We can see that for author gender classification, the salience scales inversely with the token position, meaning that the start is most important for the prediction, gradually decreasing until the end. The critic gender classification salience is also high at the beginning of the texts, dropping quickly to its lowest point at around token 60, before slowly increasing towards the end. At the very end there is a steep increase for critic gender classification salience, likely because some reviewers cite their reviews. The line plot for sentiment classification is similar to the critic gender classification plot, except that the sentiment classification salience increases even slower towards position 500, before a sharp increase at the very end.

Interestingly, but maybe not surprisingly, this corresponds to the discoveries made in Figure 4.1, analysing the impact of truncation of the input text at different truncation lengths (powers of 2 from 8 to 512). Figure 4.1 shows that the author gender classification performance increases sharply to around 0.93 at 64 tokens, before increasing more slowly to 0.97 at 512 tokens. Both critic gender classification and sentiment classification, on the other hand, also gain good parts of their performance with low truncation lengths before stagnating until 256 tokens and finally gaining big performance increases going from truncation length 256 to 512. Figure 4.1 is thus in accordance with the aggregated salience shown in Figure 5.12.

Integrated Gradients

The gradient norm is fast to compute, and suffices for the aggregated analysis above, but since it is a norm, it is always positive and can not include the direction of the gradient, only its size. That is why, for the following analysis on the sentence level, *Integrated Gradients* is the method used. The examples shown in these sections will be tokens from the NorBERT2 tokenizer, colored by their attribution to the class selected for attribution and numbered for easy reference. If tokens start with '##', it means that they are subtokens that are part of the same word as the previous token. The color gradient used for these examples are shown in Figure 5.14. Since there is seldom a good direct word-for-word translation from Norwegian to English, the examples with colored tokens are not translated, but only shown in Norwegian. Still, the tokens whose attributions are discussed will be translated to English.

In order for the attribution scores shown in this section to make any sense, Figure 5.13 shows parts of the distribution of the attribution score for the tokens in the development set. The reason it only shows parts of the distribution is that any bar outside the range shown in the plot is not possible to discern at this scale. Nevertheless, Table 5.11 also shows the minimum and maximum attribution scores of -0.105 and 0.152. What Table 5.11 and Figure 5.13 shows us is that most attribution scores are very close

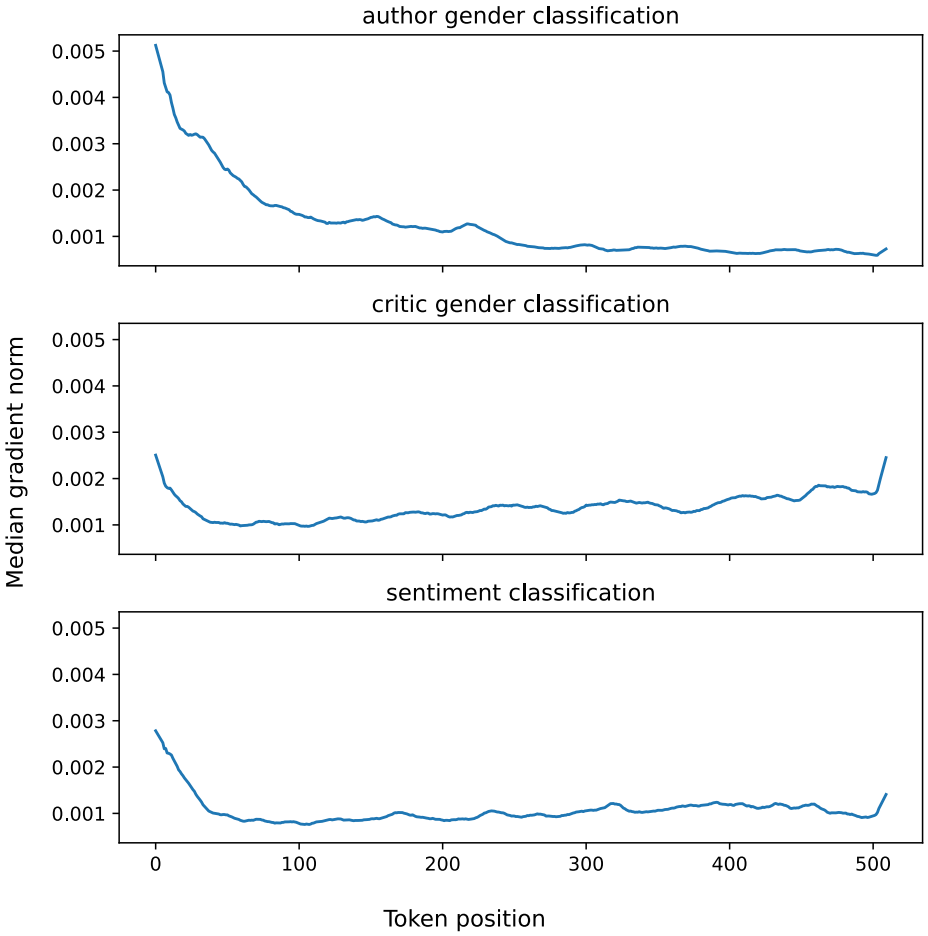

Figure 5.12: This plot shows the median gradient norm across the development data set for each 510 token position (CLS and SEP tokens were excluded) and grouped by the three classification tasks. The subplots share x-axis and the scale of the y-axis is the same for all of them.

to zero, so the few scores that get above an absolute score of just 0.01 can be assumed to have a substantial impact on the prediction. This is in line with human intuition, in that most words in a review are not directly related to the sentiment and especially not to the gender of the critic writing the review nor to the gender of the author being reviewed. Using the definition of statistical outliers as those data points further away from the quartiles than 1.5 times the interquartile range, attribution scores below -3.65e-3 or above 5.32e-3 are outliers. Since the attribution scores are computed with respect to the predicted class, it is natural that the mean of 7.48e-04 is above zero.

Author gender classification

Looking at some generic examples for author gender classification, we have e.g. Example 5.1, which was correctly classified as male author. We can see

mean	std	min	25%	50%	75%	max
7.48e-04	3.32e-03	-1.05e-01	-2.82e-04	9.40e-04	1.96e-03	1.52e-01

Table 5.11: Statistical summary of the 614095 attribution scores for the tokens in the development set and across all three classification tasks

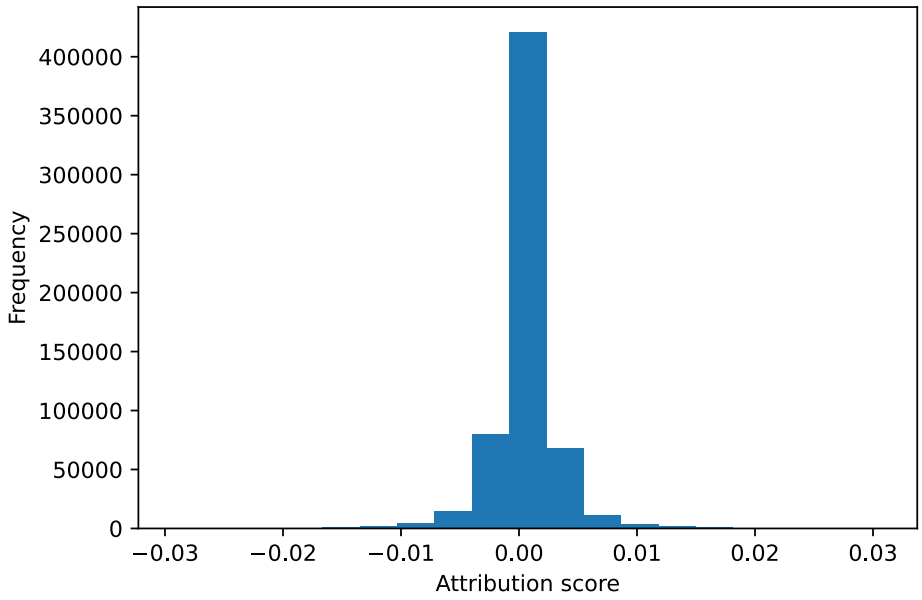

Figure 5.13: Histogram of in total 614095 Integrated Gradient attribution scores for the predicted class across the three classification tasks using the development set.

that 'han' - *he* has a positive attribution towards that prediction.

(5.1) Fu ##mi ##o Sa ##sak ##i har åpenbart rett , men han greier ikke overbevise deg om det .

However, the *Integrated Gradients* method is not always correct, and there is also Example 5.2, where 'han' - *he* has a negative attribution towards the predicted male author gender class. It is, however, clear that 'han' - *he* in this sentence should have a positive attribution for the male author gender prediction, especially since the word that comes after it: 'utgitt' means *published*.

(5.2) Denne gangen har han utgitt en drøy bok ...

Furthermore, the adversarial counterfactual generator HotFlip, which we introduced in Section 5.2.1, found that the way to turn the model prediction for this example from male to female, changing only one token, was to substitute 'han' - *he* for 'hun' - *she*. Even so, after running Integrated Gradients on the substituted text, the attributions for male author gender are still the opposite of what they should be in Example 5.3 below:

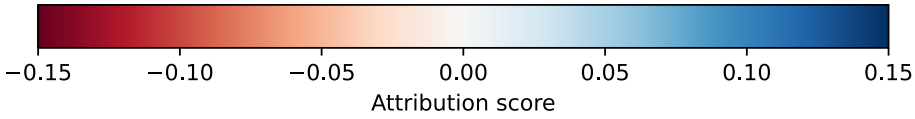

Figure 5.14: The color gradient used to show a token's attribution to the prediction of a given class.

(5.3) Denne gangen har hun utgitt en drøy bok ...

This substitution changed the model's logits from 4.53 to -4.33, which means that the model is quite certain in its predictions in both cases. The gender classification models compute class probabilities with the sigmoid function, so to get the class predictions one can simply take the sign of the logits as the predicted class, negative sign for female gender and positive for male gender.

Looking at the counterfactual examples made by switching the gendered pronouns, there were 110 out of 415 examples where the predicted class was changed. The original predicted class for Example 5.4 was correctly female author gender, but this example, as well as Example 5.5, show attribution scores for the male class for comparison. The original logits was -4.62 and the one for the gender-switched example is 3.73.

(5.4) Her beskrev hun ... 20 år etter hans død

(5.5) Her beskrev han ... 20 år etter hennes død

Like Example 5.2, Example 5.4 shows high attribution scores for male author gender, even though the noun, which in this case actually refers to the author, is 'hun' - *she*. In the counterfactual Example 5.5, where the noun has been changed to 'han' - *he*, the attribution scores are a lot lower, even though the author of this counterfactual example was predicted to be male. In the second part of the examples, we see that 'hans' - *his* in the Example 5.4 has a low score, while its substitution 'hennes' - *her* in Example 5.5 gets the highest score of the counterfactual document. The scores for both of these substitutions show that switching gender does not only change the polarity of the scores, but also their absolute value.

Critic gender classification

It is somewhat harder to interpret the model for critic gender classification since that is a task humans would also have a hard time performing. Given just the text of a review, not knowing which newspaper it comes from nor the name of the critic, a human would be hard-pressed to predict the gender of the critic with any degree of certainty. However, both the BoW models and NorBERT2 achieve around 75% accuracy on this task, which begs the question of what information these models use to perform that well. Given that the interpretability of a machine learning model relates to how well a

human can understand the causes of its prediction or consistently predict the model's result, as defined by Molnar (2022), it will be harder to interpret a model for a task that humans could not do as well themselves. Even so, the following examples might shed some light on the models' predictions.

There were seven reviews among the 415 documents in the development set where switching the gendered pronouns changed the prediction of the critic gender model. In one of them, the logits went from 0.65, predicting the correct male critic gender, to -0.92, predicting female critic gender. This is a large difference, so the change in prediction can not only be attributed to the logits already being close to the decision boundary of 0. Nevertheless, the statistical summary of the changes in attribution score between the two examples, Table 5.12, shows small attribution changes for the female class. Interestingly, even though the gender-switched input changed the prediction to the female class, its mean attribution score to the female class actually decreased compared to the original input. This does not intuitively make sense, but it goes to show that there are factors impacting the prediction of the model that the Integrated Gradient method does not take into account.

mean	std	min	25%	50%	75%	max
-1.7e-5	4.1e-4	-2.4e-3	-1.7e-4	-3.4e-5	1.3e-4	3.2e-3

Table 5.12: Summary of the difference in attribution scores between the gender-switched and original review, i.e. new scores minus original scores

Example 5.6 and Example 5.7, shown below, contain the largest and the fourth largest change in attribution for the female critic class. The largest one is changing 'han' - *he* in Example 5.6 to 'hun' - *she* in Example 5.7,with the attribution score going from -0.0012 to 0.002, just above the third quartile of attribution scores. The fourth largest change is for the token *Arch*, whose attribution goes from -0.0005 to 0.0008, both very small values. It is still notable that this change occurred for a token that was not changed itself, but it could e.g. be due to attention from *Arch* to 'han' - *he* or 'hun' - *she*, if the model has understood that they are coreferences. Still, the absolute values of the attribution scores in these examples are so low tha one can just barely see the differences in color.

(5.6) Arch ##ers liv er mer interessant enn bøkene han skriver .

(5.7) Arch ##ers liv er mer interessant enn bøkene hun skriver .

It is not surprising that the highest change in attribution score occurred for a token which was actually changed. On the other hand it means that the gendered pronouns are also an important feature for the critic gender classification, not only for author gender classification. This is not something that would necessarily be expected, since the gendered pronouns

are used to describe someone other than the critic themselves. The reason behind this is probably the class imbalance, as shown in Section 3.2. While female critics review around the same amount of male and female critics, male critics review almost three times more male authors than female authors. Consequently, features that indicate the gender of the author might also to some degree indicate the gender of the critic, since the model can pick up on the statistical imbalance of the data. This also corresponds to 'han' - *he*'s high coefficient in the BoW critic gender classification model, shown in Figure 5.3.

Sentiment classification

When it comes to generating counterfactual examples for the sentiment classification, HotFlip changes appropriate adjectives to one that is associated with stronger sentiment, likely an adjective that HotFLip has identified as most positive or negative. In this case HotFlip changed 'enkel' - *simple* in Example 5.8 to 'glimrende' - *brilliant* in Example 5.9 to go from fair to positive ternary sentiment, resulting in the following difference in attributions:

(5.8) Rent tekst ##messig er denne boka svært enkel ...

(5.9) Rent tekst ##messig er denne boka svært glimrende ...

In the artificial review containing Example 5.9, 'glimrende' - *brilliant* has the highest attribution score by a factor of two. However, the attribution of 'svært' - *very* changes from a small positive attribution in the original Example 5.8 to a negative attribution of around the same absolute value. Since 'svært' - *very* is an adverb that intensifies the adjective it is used with, one would expect its attribution to be of the same polarity as the adjective. One can only speculate as to why this is not the case here. It could be because the impact of the full self-attention is hard to interpret for such long texts as this, or simply that the Integrated Gradients method is not perfect. For both of the above examples the gradient was computed with respect to the positive class, even though the original example was classified as fair, in order to have better grounds for comparison.

The counterfactual examples generated by switching the gendered pronouns changed the ternary sentiment prediction in just two cases of the 415 reviews in the development set. As shown previously in Figure 5.10, switching gender of the pronouns has a small effect on sentiment classification, slightly improving performance. Figure 5.15 show a histogram of how much the regression score changes when comparing the original input to the gender-switched input. One can see that it is close to a narrow normal distribution, with a slightly longer tail on the left side. Table 5.13 shows that the mean is just below zero, at -0.002, while the median is a straight zero, which means that the new regression scores are slightly lower on average. The largest difference seen is -0.093 which is still less than a tenth of a rating.

count	mean	std	min	25%	50%	75%	max
415	-0.002	0.022	-0.093	-0.013	0.0	0.009	0.068

Table 5.13: Statistical summary of the changes in regression score between original and gender-switched input, i.e. new scores minus old scores.

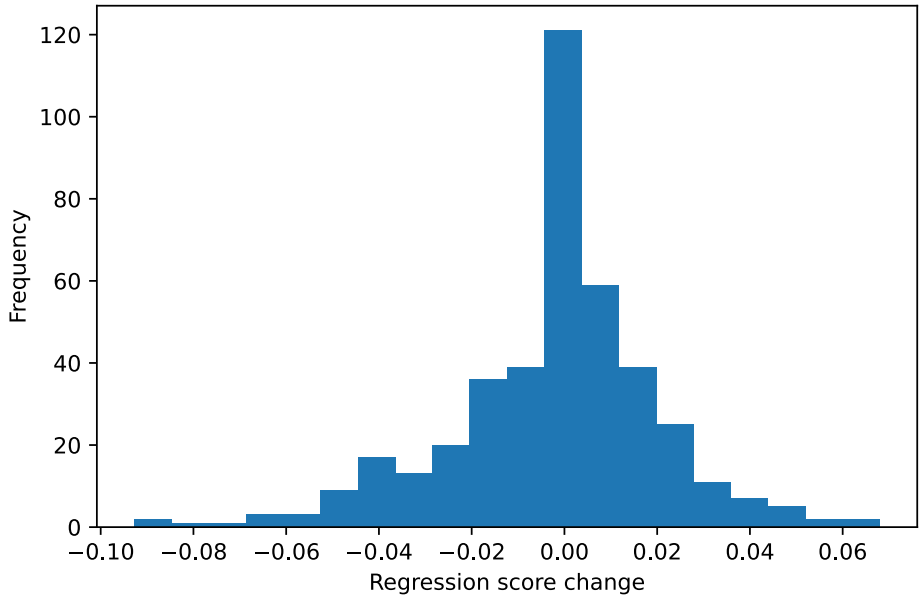

Figure 5.15: Histogram of the difference in regression score between the regression scores when using the gender-switched data as input abd the original regression scores for the ratings, i.e. new scores minus old scores.

In the first of those cases, gendered pronouns are not used a lot, but there is one sentence where using 'hun' - *she* instead of the original 'han' - *he* gives a higher attribution for the negative class. The original and artificial examples are shown below in Example 5.10 and Example 5.11, with the original example first:

(5.10) Nygård ##shaug er ujevn . Han har skrevet ...

(5.11) Nygård ##shaug er ujevn . Hun har skrevet ...

In the above examples, we can see that 'ujevn' - *uneven* has the highest attribution score. Its attribution score of 0.042 for the negative class is actually the highest of that document and in the 99.98th percentile of attribution scores. Figure 5.16 shows the attribution score for the negative class for each token in the original document. The majority of the tokens have an attribution score very close to zero, with just a few tokens standing out, like 'ujevn' - *uneven* with its score of 0.042 around position 280 on the x-axis. This makes the colors of the plot quite bland and hard to distinguish,

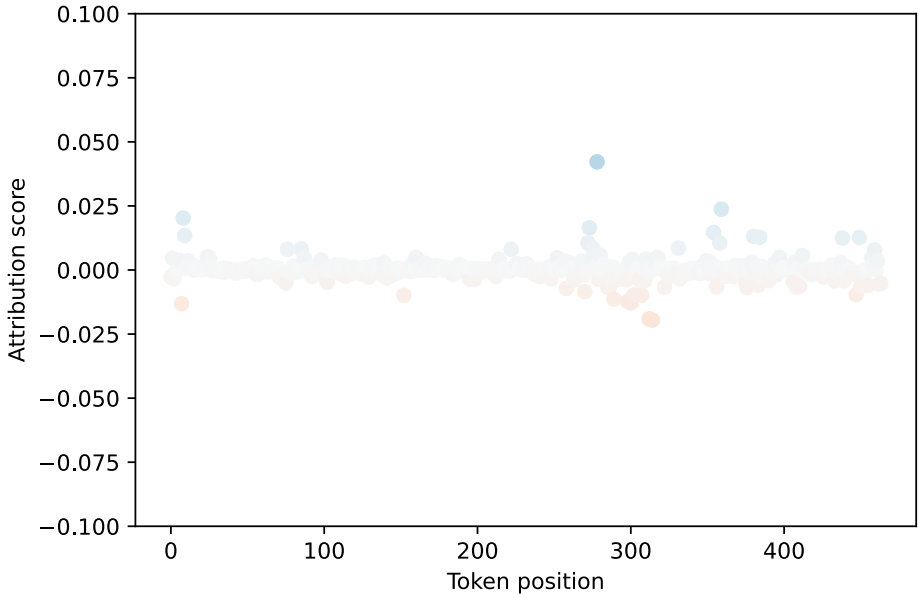

Figure 5.16: Scatter plot of the attribution scores for the negative class.

but that is a deliberate choice in order to use the same color scheme and scale for all the examples.

'Han' - *He* in the Example 5.10 above has a small attribution of 0.0058, while 'hun' - *she* in Example 5.11 has an attribution to the negative class of 0.011, which almost doubles that. This change in attribution score from 0.0058 to 0.011 was the largest change between the original to the gender-switched input text. On the other hand, even though this is the largest change of attribution scores, it still does not have a very high impact on the model prediction. The original regression score was 3.54, on a scale from 1 to 6, and the score with the gender-switched data was 3.49. The new score barely crossed the threshold of 3.5, which means that the real reason for the change in prediction was not an exceedingly high impact of changing the gendered pronouns, but that the regression score already was very close to the decision boundary. Still, a reduction in regression score of 0.05 on the gender-switched input is one of the largest changes, as shown in Figure 5.15 and Table 5.13, four times the first quartile of -0.013. It should also be noted that 3 is the correct rating for this example, corresponding to the negative ternary sentiment class.

The second of the two cases where the prediction was changed is similar to the first one in that the original regression score was 3.56, changing to 3.48 and the correct rating of 3 with the gender-switched input. The attribution score of the gendered pronouns for the negative class also changes similarly to the first example where 'han' - *he* in Example 5.12 is changed to 'hun' - *she* in Example 5.13, both shown below:

(5.12) I etter ##ordet skylder han å gjøre leseren oppmerksom
på at ... Og i note ##apparatet burde han ha opplyst om

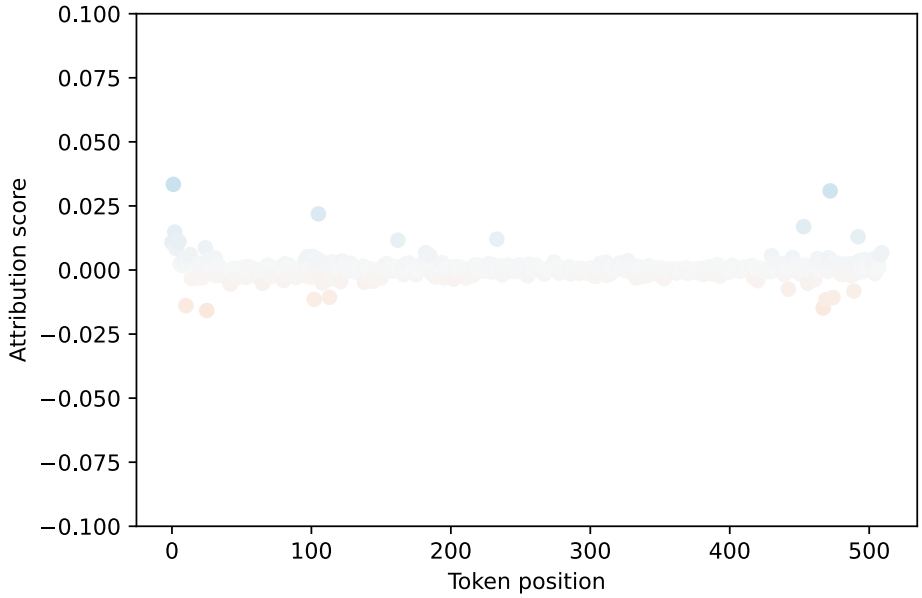

Figure 5.17: Scatter plot of the attribution scores for the negative class for the document containing attribution Example 5.12

hvor han siterer ...

(5.13) I etter ##ordet skylder hun å gjøre leseren oppmerksom på at ... Og i note ##apparatet burde hun ha opplyst om hvor hun siterer ...

While each instance of 'hun' - *she* has around double the attribution score of the replaced 'han' - *he*, except for the last replacement where 'hun' - *she* has six times higher attribution, the absolute difference in attribution is quite small. The attribution score of the pronouns in these examples is also small compared to other words like 'oppmerksom' - *attentive*, with the highest attribution score of the document, and 'burde' - *should*. However, an interesting factor to note here is how the pronouns interact with the other words: In the original text, Example 5.12, the two subsequent tokens 'skylder han' - *he owes*, together has a substantial negative attribution to the negative class, whereas 'skylder hun' - *she owes* in the switched text, Example 5.13, has a small positive attribution to the negative class. So even though the gendered pronouns themselves might not get the highest attribution scores, they affect the attribution of other tokens, likely through the attention mechanism of BERT.

Like Figure 5.16 and earlier in the attribution distribution of Figure 5.13, Figure 5.17 shows that the majority of tokens have an attribution score very close to zero, but a few distinguish themselves from the rest, and these are likely the only ones with substantial impacts on the predictions.

Chapter 6

Conclusion

In this thesis we have investigated the relations between gender and sentiment in Norwegian book reviews. Our data analysis shows that female critics and authors give and receive statistically significantly lower ratings than their male counterparts. This difference in ratings is shown to have effects on the models trained on the data, both the bag-of-word models and the NorBERT2 models. The analysis of the models was done using methods from interpretable machine learning. We have demonstrated that both using interpretable linear models and more advanced methods for interpreting black-box models can give insight into why the models make certain predictions.

6.1 Contributions

The main contribution of this thesis is extensive analysis of the effect of gender on document-level text classification for Norwegian text. Furthermore, our work shows that there are several complementary ways to analyse such gender effects, ranging from simple statistical analysis of the metadata to methods created for interpretation of deep neural networks. Several of these methods can also be used for interpretation of the models and their predictions in general. Our work also demonstrates how one can use interpretation methods for transformer models on Norwegian text.

6.2 Research questions

The research questions we investigated in this thesis were the following:

RQ1 Can a model still predict the gender categories and sentiment if we first normalize gendered words in the texts?

- a How could such gender normalizing pre-processing best be carried out?
- c Does gender normalization affect prediction of author gender?
- c Does gender normalization affect prediction of critic gender?
- d Does gender normalization affect prediction of sentiment?

RQ2 What is the effect of supplying knowledge of the gender during training of the models?

- a What is the effect of supplying knowledge of the author as a variable when attempting to predict the gender of the critic, and vice versa?
- b What is the effect of supplying knowledge of the author and/or critic as a variable when attempting to predict the sentiment?

RQ3 Is it possible to use methodology from interpretable machine learning to shed more light on what information is used by the models when predicting gender and/or sentiment?

- a To what extent does using linear interpretable models satisfy both predictive performance and explainability of the models?
- b To what extent do methods for interpreting deep neural networks give insight into why the models make certain predictions?

Looking back at those research questions, it is definitely clear that the models still can predict the gender of authors when the texts are gender normalized, although with substantially lower accuracy. When the models are re-trained on the fully normalized data, both the BoW SVC and the NorBERT2 model retain around 78% accuracy of the original 91% for the SVC and 98% for NorBERT2. Even when testing the models trained on the original data using the gender normalized data, the BoW model has 69% accuracy, and NorBERT2 has 75% accuracy. For critic gender classification and sentiment classification, gender normalization has just a slight effect, but there is an effect.

The gender normalization used for the experimentation replaces gendered pronouns and person names with dummy tokens. With a different normalization method one could of course expect other results, but this method was chosen due to its simplicity and since it is possible to implement it in a principled manner. One could of course also have removed gendered words like *woman*, *man*, *mother* and *father*, but it would be hard to define in a principled manner which words should be removed and which should not, so further gender normalization was not done.

The replaced names carry more information than just gender, since they refer to people, but without normalizing the names much gender information would be left in the texts. Since we wanted to control for the effect of such information, we chose to normalize the names. As has been argued as well, letting the models use the names of the authors as features is not very interesting for interpretation either, since the models could then just weight the names of authors with high or low ratings highly, instead of using the content of the reviews. We saw that the BoW models likely used the names in such a way for sentiment analysis, since the performance of the Ridge model trained with person names dropped quite a bit when it was given input where the names were normalized. Nonetheless, the gender normalization did not affect critic gender classification or sentiment classification nearly at all, especially when re-training the models on

the normalized data. Only the aforementioned Ridge regression model for sentiment classification, trained on the original data, lost a lot of its performance when given input with normalized names. This is likely not due to the loss of the gender information contained in the names, but the fact that different authors get different average ratings, and the models can no longer use this correlation when the names are normalized.

We also demonstrate in Section 4.4.3 that supplying the models with gender metadata has some effects of performance, but not as much as what Touileb et al. (2021) found and not always in the same direction as their results. Their results indicate that adding either or both genders when predicting sentiment increased the performance, while we found that it decreased the sentiment classification performance. It should be noted that the model architectures were not the same and neither were the classification heads, since they performed binary classification and we performed ternary sentiment classification. With normalized gender, adding metadata to NorBERT2 improved performance of sentiment classification but not for author gender classification or critic gender classification. This was the opposite as the author gender classification results for the BoW models, where adding metadata improved the performance substantially when the gender was normalized.

Chapter 5 demonstrates that there are several methods from interpretable machine learning that can be used to shed light on what information the models use. The methods explored can roughly be divided into three: using interpretable models, gradient-based methods for neural networks, and counterfactual analysis. These methods have different strengths and weaknesses and can complement each other. For example, having access to the gradients of a neural network means having access to more information about the model's prediction. However, as the examples for interpreting author gender in Section 5.2.5 show, the gradient-based methods can at times give strange results, where methods without access to model internals would be more conservative. This means that one should not always take any explanation from an interpretation method at face value, but corroborate the results using other methods as well.

NorBERT2 has almost 10 pp higher performance than the BoW models for author gender classification and ternary sentiment analysis, but for critic gender classification their performance is about equal. This means that there are some tasks where using interpretable, linear BoW models can be beneficial compared to the much larger transformer models. We conclude that linear models satisfy both predictive performance and explainability for the critic gender classification task on the $\text{NoReC}_{\text{gender}}$ data set. However, we cannot conclude that this is the case for predicting the gender of the writer of the text one passes to the model in general.

We discuss the possible gender bias of $\text{NoReC}_{\text{gender}}$ in Section 3.5 and determine that we cannot actually conclude whether the differences between ratings for men and women are caused by gender bias. There could also be other reasons for the differences, like female critics *choosing* to review other books than male critics, or that female authors write books in different genres than male authors. Nevertheless, in one way it does not

matter if these differences stem from bias or not. No matter the reason for the differences, any machine learning model trained on the sentiment data of $\text{NoReC}_{\text{gender}}$ without precaution, would be biased against female authors. To these models, a correlation is a correlation, and they cannot even know what causation is, since they have no notion of counterfactual queries and answers (Pearl & Mackenzie, 2018).

6.3 Limitations

This thesis has only used a very small fraction of the interpretation methods that have been developed to date. Since research in interpretable machine learning has boomed in the latest years, new methods are being developed at a rapid pace (Molnar et al., 2020). The gradient-based methods used in this thesis were originally made for interpretation of image models, before being adapted to work in NLP. It could be that methods developed specifically with NLP in mind would make interpretation easier or more powerful.

In this thesis counterfactual analysis has been used for interpretation of the models. When changing a word in the input text also changes the prediction of the model, it is intuitively apparent that the word is important for the prediction. However, we lack robust theoretical grounds to quantify the counterfactual importance of such a word.

While this thesis may add some insight into the effects of gender in Norwegian book reviews, it does so only for a binary understanding of gender. We recognize that this is not sufficient to reflect all variations of gender, but given the data available and the scope of this thesis, that was the extent of what was possible. The previous work on gender bias that this thesis use as a foundation has also been based on binary gender settings.

We did not do any work on how one could prevent gender bias from having an impact on models used for prediction. Of course, normalizing gender may have an effect, but in theory, removing a sensitive variable like gender may not make the model more fair, since other variables can correlate to the sensitive variable. We could afford to neglect this since none of the models trained during the experimentation for this thesis will be used in the future for predicting anything with an effect on other people. However, what our analysis indicates is that if no measures are taken to prevent or alleviate bias, the bias present in the training data will have impacts on the model predictions. That may be self-evident given that machine learning models use any correlations they find, be they spurious or discriminating, to better fit to the data. This notwithstanding, bias is something machine learning practitioners should keep in mind when they train models to be used for prediction.

6.4 Future Work

As mentioned above, this thesis only touches the surface of the interpretation methods that exist. It would certainly be interesting to see what kind

of results one could get with different and more advanced interpretation methods, which were outside the scope of this thesis.

Another direction to take the research in further work could be to compare the gender bias inherent in different pretrained models and model architectures, for example comparing NorBERT2 to cross-lingual models like XLM-Roberta with respect to gender. Furthermore, it could be valuable to use models that can inherently deal with texts longer than 512 tokens. The approach for document-level classification used in this thesis was to use the first 128 and the last 382 tokens, but this approach clearly loses information that could have value for the models. In a similar manner, it could be beneficial to investigate in what situations interpretable BoW models can perform on par with more expensive and not inherently interpretable deep neural networks, like we saw for critic gender classification in this thesis. Such models also do not have BERT's current limitation of maximum 512 tokens. $\text{NoReC}_{\text{gender}}$ is also a quite small data set in the grand scheme of things, so it would be of interest to investigate whether the results achieved on $\text{NoReC}_{\text{gender}}$ can be reproduced for other data sets, or if there are large differences.

Further research using counterfactual analysis for NLP should find ways to more robustly quantify the changed word's impact, by using e.g. the probability of necessity and probability of sufficiency, in the vein of Watson et al. (2021) and Pearl and Mackenzie's (2018) work on counterfactuals. Molnar et al. (2020) note that causal interpretation of machine learning models is a challenging task, and that making a model work well for causal interpretation can conflict with predictive performance. Keeping this in mind, it would be valuable to learn how much performance one would need to sacrifice in order to make models more explainable, and strike a balance between performance and interpretability.

In order to rectify this thesis' and most previous work's limitation of only using binary gender, a possible research direction could be to include more gender categories. Such an approach should also use self-reported gender instead of having the researchers annotate gender based on names. To some extent, previous work by Lassen et al. (2022) shows that what matters when evaluating the work of a person is actually not their actual gender, but what the audience *perceives* their gender to be. Nevertheless, not letting people define their gender themselves can be seen as an infringement of their rights and should be avoided, if possible.

Bibliography

Barnes, J., Ravishankar, V., Øvrelid, L. and Velldal, E. (2020). *A Systematic Comparison of Architectures for Document-Level Sentiment Classification*. arXiv: 2002.08131.

Barnes, J., Touileb, S., Øvrelid, L. and Velldal, E. (2019). Lexicon Information in Neural Sentiment Analysis: A Multi-Task Learning Approach. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland.

Barry, J. (2017). Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches. *Proceedings of the 25th Irish Conference on Artificial Intelligence and Cognitive Science*. Dublin, Ireland.

Bastings, J. and Filippova, K. (2020). The Elephant in the Interpretability Room: Why Use Attention as Explanation When We Have Saliency Methods? *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online.

Beltagy, I., Peters, M. E. and Cohan, A. (2020). *Longformer: The Long-Document Transformer*. arXiv: 2004.05150.

Bergem, E. A. (2018). *Document-Level Sentiment Analysis for Norwegian*. MA thesis. University of Oslo.

Briscoe, E. and Feldman, J. (2011). Conceptual Complexity and the Bias/Variance Tradeoff. *Cognition* 118.1.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Belanger, D., Colwell, L. and Weller, A. (2020). *Masked Language Modeling for Proteins via Linearly Scalable Long-Context Transformers*. arXiv: 2006.03555.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online.

Conneau, A. and Lample, G. (2019). Cross-Lingual Language Model Pretraining. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y. and Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA.

De Amicis, C., Falconieri, S. and Tastan, M. (2021). Sentiment Analysis and Gender Differences in Earnings Conference Calls. *Journal of Corporate Finance* 71. Elsevier.

Dettmers, T., Lewis, M., Shleifer, S. and Zettlemoyer, L. (2022). 8-Bit Optimizers via Block-wise Quantization. *Proceedings of the 10th International Conference on Learning Representations*. Online.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805.

Ebrahimi, J., Rao, A., Lowd, D. and Dou, D. (2018). HotFlip: White-Box Adversarial Examples for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia.

Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. *Communications of the ACM* 56.4. Association for Computing Machinery.

Friedman, B. and Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems* 14.3.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F. and Ureña-López, L. A. (2021). A Survey on Bias in Deep NLP. *Applied Sciences* 11.7. Multidisciplinary Digital Publishing Institute.

Gutiérrez, P. A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F. and Hervás-Martínez, C. (2016). Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowledge and Data Engineering* 28.1.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem and Adriane Boyd (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. URL: https://doi.org/10.5281/zenodo.1212303.

Hovy, D. and Prabhumoye, S. (2021). Five Sources of Bias in Natural Language Processing. *Language and Linguistics Compass* 15.8. Wiley.

Hung, C.-C., Lauscher, A., Hovy, D., Ponzetto, S. P. and Glavaš, G. (2023). Can Demographic Factors Improve Text Classification? Revisiting Demographic Adaptation in the Age of Transformers. *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia.

Katharopoulos, A., Vyas, A., Pappas, N. and Fleuret, F. (2020). Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. *Proceedings of the 37th International Conference on Machine Learning*. Online.

Kim, B., Khanna, R. and Koyejo, O. O. (2016). Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability. *Advances in Neural Information Processing Systems*. Vol. 29.

Kitaev, N., Kaiser, L. and Levskaya, A. (2020). Reformer: The Efficient Transformer. *Eighth International Conference on Learning Representations*. Online.

Kummervold, P. E., De la Rosa, J., Wetjen, F. and Brygfjeld, S. A. (2021). Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online).

Kutuzov, A., Barnes, J., Velldal, E., Øvrelid, L. and Oepen, S. (2021). Large-Scale Contextualised Language Modelling for Norwegian. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online).

Lapponi, E., Søyland, M. G., Velldal, E. and Oepen, S. (2018). The Talk of Norway: A Richly Annotated Corpus of the Norwegian Parliament, 1998–2016. *Language Resources and Evaluation* 52.3. Springer.

Lassen, I. M. S., Bizzoni, Y., Peura, T., Thomsen, M. R. and Nielbo, K. L. (2022). Reviewer Preferences and Gender Disparities in Aesthetic Judgments. *Proceedings of the Computational Humanities Research Conference 2022*. Antwerp, Belgium.

Learning Interpretability Tool (LIT) (2023). PAIR code. URL: https://github.com/PAIR-code/lit.

Luxburg, U. von and Schölkopf, B. (2011). Statistical Learning Theory: Models, Concepts, and Results. *Handbook of the History of Logic*. Ed. by D. M. Gabbay, S. Hartmann and J. Woods. Vol. 10. Inductive Logic. North-Holland.

Lyu, C., Foster, J. and Graham, Y. (2020). Improving Document-Level Sentiment Analysis with User and Product Context. *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online).

Madsen, A., Meade, N., Adlakha, V. and Reddy, S. (2022). Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates.

McDonald, G. C. (2009). Ridge Regression. *WIREs Computational Statistics* 1.1.

Merriam-Webster (n.d.). *Bias*. *Merriam-Webster.Com Dictionary*. URL: https://www.merriam-webster.com/dictionary/bias (visited on 12/05/2023).

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G. and Wu, H. (2018). Mixed Precision Training. *Proceedings of the 6th International Conference on Learning Representations*. Vancouver, Canada.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2021). Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys* 54.3. Association for Computing Machinery.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. URL: https://christophm.github.io/interpretable-ml-book/.

Molnar, C., Casalicchio, G. and Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *ECML PKDD 2020 Workshops*. Ed. by I. Koprinska et al. Communications in Computer and Information Science. Cham.

Mukherjee, A., Mukhopadhyay, S., Panigrahi, P. K. and Goswami, S. (2019). Utilization of Oversampling for Multiclass Sentiment Analysis on Amazon Review Dataset. *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*. Morioka, Japan.

Øvrelid, L., Mæhlum, P., Barnes, J. and Velldal, E. (2020). A Fine-grained Sentiment Dataset for Norwegian. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France.

Pang, B. and Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, Michigan.

Papazian, E. (2012). Norge - riket uten rikstalemål? *Norsk Lingvistisk Tidsskrift* 30.1.

Park, H., Vyas, Y. and Shah, K. (2022). Efficient Classification of Long Documents Using Transformers. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland.

Paszke, A. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*. Vol. 32.

Pearl, J. (1999). Probabilities Of Causation: Three Counterfactual Interpretations And Their Identification. *Synthese* 121.1.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Causality: Models, Reasoning, and Inference. New York, NY, US: Cambridge University Press.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. 1st edition. New York: Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12.85.

Perez, C. C. (2021). *Invisible Women: Data Bias in a World Designed for Men*. New York: Harry N. Abrams. 432 pp.

Pires, T., Schlinger, E. and Garrette, D. (2019). How Multilingual Is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA.

Rosenthal, S., Farra, N. and Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada.

Sanyal, S. and Ren, X. (2021). Discretized Integrated Gradients for Explaining Language Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic.

Shah, D. S., Schwartz, H. A. and Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online.

Sohoni, N. S., Aberger, C. R., Leszczynski, M., Zhang, J. and Ré, C. (2022). *Low-Memory Neural Network Training: A Technical Report*. Version 2. arXiv: 1904.10631. preprint.

Stanczak, K. and Augenstein, I. (2021). *A Survey on Gender Bias in Natural Language Processing*. arXiv: 2112.14168. preprint.

Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Proceedings of the 18th China National Conference on Chinese Computational Linguistics*. Kunming, China.

Sundararajan, M., Taly, A. and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Sydney, Australia.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics* 2.1. Annual Reviews.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S. and Metzler, D. (2021). Long Range Arena : A Benchmark for Efficient Transformers.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E. and Yuan, A. (2020). The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online.

The Norwegian Academy for Language and Literature (n.d.). *Megen*. *Norwegian Academy Dictionary*. URL: https://naob.no/ordbok/megen (visited on 28/01/2023).

Thelwall, M. (2018). Gender Bias in Sentiment Analysis. *Online Information Review* 42.1. Emerald.

Tian, J. and Pearl, J. (2000). Probabilities of Causation: Bounds and Identification. *Annals of Mathematics and Artificial Intelligence* 28.1.

Touileb, S., Øvrelid, L. and Velldal, E. (2020). Gender and Sentiment, Critics and Authors: A Dataset of Norwegian Book Reviews. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Barcelona, Spain (Online).

Touileb, S., Øvrelid, L. and Velldal, E. (2021). Using Gender- and Polarity-Informed Models to Investigate Bias. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Online.

University Centre for Information Technology (n.d.). *Machine Learning Infrastructure (ML Nodes)*. Norway: University of Oslo.

University of Oslo (2020). *AI hub-node project*. URL: https://www.uio.no/tjenester/it/forskning/kompetansehuber/uio-ai-hub-node-project/index.html (visited on 12/05/2023).

University of Oslo (2023). *ML nodes*. URL: https://www.uio.no/tjenester/it/forskning/kompetansehuber/uio-ai-hub-node-project/it-resources/ml-nodes/index.html (visited on 12/05/2023).

Velldal, E., Øvrelid, L., Bergem, E. A., Stadsnes, C., Touileb, S. and Jørgensen, F. (2018). NoReC: The Norwegian Review Corpus. *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*. Miyazaki, Japan.

Watson, D. S., Gultchin, L., Taly, A. and Floridi, L. (2021). Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice. *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online.

Yadav, A. and Vishwakarma, D. K. (2020). Sentiment Analysis Using Deep Learning Architectures: A Review. *Artificial Intelligence Review* 53.6. Springer.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L. and Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. *Proceedings of the 33rd Conference on Advances in Neural Information Processing Systems*. Vol. 33. Online.