Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Andrine Lossius and Regine Pösche Ruud

You Shall Know a Female Word by the Company It Does Not Keep

Detecting and Mitigating Gender Bias in Norwegian Language Models

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2022

NTNU
Norwegian University of
Science and Technology

Andrine Lossius and Regine Pösche Ruud

You Shall Know a Female Word by the Company It Does Not Keep

Detecting and Mitigating Gender Bias in Norwegian Language Models

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Language models implementing the *transformer* mechanism as neural network architecture for producing word representations have revolutionized the field of natural language processing. They are shown to capture more information about the meaning of words than any other technique. However the opportunities, many studies have proven the significant drawback of blindly applying language models in downstream tasks where historical stereotypes such as *"man is to computer programmer what woman is to homemaker"* are hidden in digital word representations. With an increased focus and attempt to create technology that works in the Norwegian language, this issue threatens gender equality and the statutory right to equal treatment within our borders. This Master's Thesis stands out as the first study that measures and mitigates gender bias in Norwegian language models to avoid introducing discrimination through the digitization of Norway. Through an experimental methodology, gender bias is detected in the state-of-the-art Norwegian language models that have been published by the University of Oslo, the National Library of Norway, and Google.

First, different approaches to quantify bias in the models are tried out. The results show that the data used to train Norwegian language models contain more than three times as many male pronouns as female ones. Through a masked language modeling task, we show that between 76% and 100% of all Norwegian adjectives are associated more strongly with male than female in the different models and that the majority of all adjectives have a more substantial male bias than the word 'kvinnelig' (English: 'female' in adjective form) has to female. Adjectives used to describe women are related to reproduction, beauty, caretaking, and vulnerability. The models consider similar descriptions of a man and a woman differently, as almost all sentences are closer to a male than a female name in the vector space. The models reflect several societal biases in the results, often so strong that it overshadows an overall extreme disability to produce meaningful results on female entities. By creating a realistic downstream task that automatically evaluates funding applications based on their similarity to evaluation criteria, we show that the models favor male applicants in a way that results in real-life discrimination made by Norwegian technology.

Further, two mitigating techniques are applied and demonstrate that debiasing is possible and necessary for Norwegian language models. The first identifies a gender subspace and removes it from the models by performing orthogonal projection that successfully decreases the bias found in the models. The second debiasing technique creates a new language model by fine-tuning one of the models on a corpus where male words are changed with female words. This technique did not work as a debiasing technique as the model came out as highly female-biased. However, the results show that a drastic change in gender representation in training data leads to a difference in bias, which speaks for bias to be mitigated through retraining or fine-tuning on fair datasets. Both results indicate that the experiments are better suited for the monolingual Norwegian language models than a multilingual one published by Google as it creates somewhat random results throughout the whole thesis.

Sammendrag

Språkmodeller som implementerer *transformer*-nettverk som en del av sin arkitektur for å representere ord i distribuerte vektorrom har revolusjonert feltet for naturlig språkprossessering. De har vist seg å inneholde mer informasjon om et ord enn noen annen teknikk i feltet. Desverre har mange studier vist den betydelige baksiden av å bruke slike avanserte språkmodeller i teknologi, hvor historiske stereotyper som at *"mann er til dataprogrammerer det kvinne er til hjemmeværende"* er gjemt i slike modeller. Med økt fokus og forsøk på å skape språkteknologi som fungerer på norsk, truer denne problemstillingen likestilling og den lovfestede retten til likebehandling innenfor våre landegrenser. Denne masteroppgaven er den første studien som oppdager, måler og fjerner kjønnsskjevhet i norske språkmodeller for å unngå og videreføre diskriminering gjennom digitaliseringen av Norge. Ved en eksperimentell metodikk oppdages kjønnsskjevhet i de toppmoderne norske språkmodellene som nylig har blitt utgitt av Universitetet i Oslo, Nasjonalbiblioteket og Google.

Først blir ulike tilnærminger for å kvantifisere kjønnskjevheter i modellene prøvd ut. Resultatene fra disse viser at dataen som brukes til å trene norske språkmodeller inneholder mer enn tre ganger så mange mannlige pronomen som kvinnelige. Gjennom en tilpasset oppgave viser vi at mellom 76% og 100% av alle norske adjektiver er sterkere assosiert med mann enn kvinne i de ulike modellene, og at flertallet av alle adjektiv til og med har en sterkere tilknytning til mann enn det ordet 'kvinnelig' har til kvinne. Adjektiver som brukes for å beskrive kvinner er relatert til reproduksjon, skjønnhet, omsorg og sårbarhet. Modellene vurderer indetiske beskrivelser av en mann og kvinne ulikt, hvor nesten alle setningene er nærmere et manns navn i vektorromemet. Modellene reflekterer flere samfunnsmessige kjønnsskjevheter i resultatene sine, ofte så sterke at det overskygger en ekstrem mangel på evne til å produsere meningsfulle resultater på kvinnerelaterte eksempler. Ved å konstruere en realistisk oppgave fra virkeligeten hvor modellene vurderer søknader om finansiering basert på deres likhet med et sett vurderingskriterier, viser vi at modellene favoriserer mannlige søkere på en måte som resulterer i ulovlig diskriminering mot kvinnelige gründere.

Videre er to teknikker for å fjerne kjønnsskjevhet i modellene testet ut og demonstrerer at fjerning av skjevheten er både mulig og nødvendig for norske språkmodeller. Den første teknikken identifiserer et underrom i vektorrommet som beskriver kjønn i modellen og fjerner dette ved å utføre ortogonal projeksjon. Dette viser seg å være en vellykket teknikk som fjerner deler av skjevheten. Den andre teknikken produserer en ny språkmodell ved å finjustere en av modellene på et datasett der mannlige ord er byttet ut med kvinnelige ord. Denne teknikken fungerte ikke for å fjerne skjevhet, og resulterte heller i at skjevheten ble tippet mot det kvinnelig kjønn. Resultatene viser imidlertid at en drastisk endring i kjønnsrepresentasjon i treningsdataen fører til en forskjell i skjevhet i modellen, noe som taler sterkt for at modellene burde mitigeres gjennom ny trening eller finjustering på rettferdige datasett. Begge resultatene indikerer at eksperimentene er bedre egnet for de enspråklige norske modellene enn en flerspråklig modell utgitt av Google, da denne gir tilsynelatende tilfeldige resultater gjennom hele oppgaven.

Preface

This Master's Thesis concludes our Master of Science in Computer Science at the Norwegian University of Science and Technology in Trondheim. We deliver the thesis with mixed feelings. It feels delightful to wrap up the last five years of work by investigating a topic that we are highly interested in. At the same time, closing this chapter also means that the freedom associated with being a student will be replaced by debt paying and responsibility.

The title is inspired by the English linguist John Rupert Firth (1957), who has contributed to the field of distributional semantics and is well known for the quote; *"you shall know a word by the company it keeps"*. Our results show a lack of relevant company for female words in their distribution, so the analogy was drawn from the quote. We think it is funny to write a Master's Thesis about the Norwegian language in English. However, we avoided the temptation of including Norwegian words in the title after all.

Several people have contributed to our work forming this thesis. Björn Gambäck has supervised us, and we would like to express a special thanks to him for letting us work with our passion for technology and gender equality. When we were in our first year of study, older students laughed when we said we wanted to write about feminism, and we honestly did not really think it was possible at that time either. Thank you, Björn Gambäck, for introducing us to this exciting research field by providing related work and sharing your domain expertise in artificial intelligence and linguistics. In addition, much appreciated help from the AI Lab at The National Library of Norway (Nasjonalbiblioteket) has been provided to us during the semester. They have contributed by participating in discussions and have assisted the work by fine-tuning a model according to our description. Further, we would like to thank Marius Heier, who helped us navigate the natural language processing tools in the beginning, and Tor Botheim and Lars Aurdal from Findable who prioritized meeting with us to discuss the topic and assist path choices. Thank you also to Bolukbasi et al. (2016), Zhao et al. (2019) and Stanczak and Augenstein (2021) for letting us borrow your illustrations for our work. A special thanks is directed to The Language Council of Norway (Språkrådet), who has granted us a scholarship for the thesis' contribution to Norwegian language technology.

We also want to express appreciation to our fellow students at our master's office, room GF355, for numerous exciting conversations and laughs. We loved all our random topics of discussion during the breaks, even though they were often too long for our own best. Last but not least, thank you to the student association Abakus and the drone group Ascend, where we have spent many hours of voluntary work over the previous five years, for the learning and fun it has provided us. We will forever be grateful for the friendships we have formed here in Trondheim, including our boyfriends whom we both met at NTNU.

Andrine Lossius and Regine Pösche Ruud
Trondheim, 8th June 2022

Contents

List of Figures

List of Tables

1 Introduction

This chapter describes the background and motivation behind the research conducted in the thesis, and the goal of the research is defined based on this. Research questions accumulated from research gaps in related work are presented, as well as the research method and structure of the thesis defined to answer these. Further, a summary of the most important contributions of the thesis is given.

1.1 Background and Motivation

The drawback of the incredible development within language technology was discovered by Bolukbasi et al. (2016), that famously revealed historical stereotypes such as *"man is to computer programmer what woman is to homemaker"* hidden in digital word representations. They showed that negative attitudes and discrimination from society are inherited from text data in modern technology through machine learning. When implemented in real-life applications, state-of-the-art language technology is in danger of making decisions on our behalf on discriminating grounds. There have later been documented many examples of discrimination in commonly used technology created by some of the world's most influential companies. Among others, Google Translate tended to translate gender-neutral pronouns into, e.g., masculine ones for engineers but feminine ones for nurses[1], Google Photos face recognition tagged black people as gorillas in their classifying algorithm[2] and Amazon scrapped their recruiting tool that turned out to be biased against women, with the result that women were filtered out in the process based on the fact that they were women[3].

Natural language processing (NLP) is a sub-field of artificial intelligence within machine learning at the intersection of computer science and linguistics. The goal of the field is for computers to process and make sense of large amounts of natural language, or human language, in a way that can be valuable. Digital word representations have developed from simple vectors to complex neural networks that capture relations between words to an extent beyond human intuition. By combining computational linguistics, meaning rule-based language modeling with statistical and deep learning models, NLP enables computers to understand the full meaning, including the intention and sentiment of a text. Through machine learning methods, modern *language models* can be trained to 'know' a language to a large extent by looking for patterns in large sets of text data. The market of NLP has witnessed an increase every year and is expected to grow from $20.98

[1] https://www.forbes.com/the-algorithm-that-helped-google-translate-become-sexist/
[2] https://www.forbes.com/google-tags-two-african-americans-as-gorillas-through-recognition-software/
[3] https://www.reuters.com/us-amazon-com-jobs-automation-insight

billion in 2021 to $127.26 billion in 2028[4]. Thus, NLP is already and will be even more going forward, a powerful tool in automating processes on all levels of society and is being used in the case of processing by both public agencies and private companies.

Ensuring good language technology in Norwegian is an important part of Norwegian language policy. The Constitution Law (Grunnloven), the Sami Act (Sameloven), and a forthcoming comprehensive language law state that the public sector is responsible for the survival and use of Bokmål, Nynorsk, and Sami languages. For that to happen, Norwegian language(s) need to be supported by technology in use by the population. Language technology solutions are gaining an increasing place in our lives at home and at work. If we are to have good use of search services, chat robots, smart speakers, and voice-controlled directions, these solutions must work well in Norwegian. The Norwegian Language Council (Språkrådet) states that:

> "Skal norsk være det samfunnsbærende språket i Norge, må det kunne benyttes i alle språkteknologiske løsninger." (English: "If Norwegian is to be the bearing language of society in Norway, it must be possible to use it in all language technology solutions.")

However, the smaller languages often have too few users to be financially profitable for the large technology companies as developers of core systems. For example, Microsoft announced in 2020 that they would remove support for Nynorsk and 26 other languages in the Outlook. Norwegian language technology is also crucial for the successful digitalization of the public sector in Norway and for us to take advantage of the rapid development in artificial intelligence. For small language nations to benefit from the language technology product development of others, they must at least have developed the relevant national language components that must be implanted in the products they wish to make use of.

Until today, there has been little market-oriented language technology development work in Norway[5]. Fortunately, an increased interest in issues related to the use of the Norwegian language in digital media can now also be registered in our own country, at the same time as there is a development from the research side to participate more actively. For example, a language bank was established in the National Library of Norway in 2010 that contains large amounts of text and speech data that can be freely used and both the University of Oslo and The National Library of Norway has released Norwegian versions of the state-of-the-art models for language processing in 2021. However, there is no consideration of bias or ethics in general included in any publications. With the rise of Norwegian language technology comes the potential for similar biases to be present in Norwegian language models. Still, almost all research on the topic of bias in NLP has been done in the English language. Matthews et al. (2021) stated that there is a considerable imbalance in research on bias in such models between the English language and any other language, which is a paradox in this manner as it is a bias itself that most technology works better for the English-speaking population. The result is a relatively large amount of knowledge about and possible solutions to the problem for the English

[4]https://www.fortunebusinessinsights.com/industry-reports

[5]https://www.sprakradet.no/Vi-og-vart/Publikasjoner/Spraaknytt/Arkivet/Spraaknytt_1998/

language, while other languages lack relevant research. Assuming that further research on removing gendered practices from these models leads to fairer technologies, the lack of research in these other languages suggests that this is a problem for (the rise of) gender equality to be maintained through the digitalization of Norway.

Gender inequality is one of the biggest issues of today's world and is defined as one of United Nations (2017) sustainability goals that should be solved before 2030. Both Norway and the European Union have legislation that protects against discrimination. The 1979 United Nations Convention on the Elimination of All Forms of Discrimination against Women obliges states to prohibit all discrimination against women and to introduce legal protection for women's rights on an equal level with men. The Convention has been made Norwegian law through the Human Rights Act § 2 and thus takes precedence over national laws that conflict with the rights under the Convention, cf. the Human Rights Act § 3[6]. As technology becomes a larger part of crucial decisions and everyday life, it is also becoming at least as big of an issue to be discriminated against by technology as by people. Whether or not the technology we surround ourselves with adapts the biases in society is a major issue on the road to equality. Vinuesa et al. (2020) stated that the fast development of AI needs to be supported by the necessary regulatory insight and oversight for AI-based technologies to enable sustainable development and that failure to do so could result in gaps in transparency, safety, and ethical standards. Norwegian experts tag along, and e.g. Morten Goodwin, deputy head of the Center for Artificial Intelligence Research at the University of Agder, claims that an algorithm audit should be considered to ensure that computer systems with AI do not discriminate in ways that can be difficult to detect[7].

Therefore, the motivation for this thesis is the increased interest in and development of Norwegian language technology in combination with the lack of knowledge about possible bias. There exists no similar research on detecting, measuring, and mitigating gender bias in Norwegian language models. We do not want Norwegian language technology to allow itself to be any more discriminatory than other languages allow (or at all). Contributions to detect and mitigate these biases are essential documentation of a problem already discovered and attempted to mitigate for English and a few other languages to some extent.

1.2 Goals and Research Questions

The overall goal of the Master's Thesis is defined as followed:

Goal *Contribute to gender equality in Norway by preventing Norwegian Language technology from scaling up social and historical injustice.*

Further, a set of research questions related to the process of reaching the goal, accumulated from gaps in relevant research, are presented. The absence of information about bias in the models rises the first research question:

[6] https://www.regjeringen.no/no/dokumenter/prop-88-l-20122013/id718741/?ch=4
[7] https://www.digi.no/artikler/ekspert-om-ai-diskriminering

Research Question 1 *To what extent is gender bias present in Norwegian language models?*

The advanced architectures of modern language models learn patterns from datasets to knowledge beyond our understanding, introducing the second research question:

Research Question 2 *What are the sources of gender bias in Norwegian Language models?*

Knowing the consequences of applying these patterns in real-life systems is critical to avoiding discrimination in technology. Thus, the third research question reads:

Research Question 3 *What are the consequences of applying Norwegian language models as they are today to downstream tasks in real-life applications?*

Obtaining insight into the sources and consequences of bias obviously makes it easier to identify it and mitigate it. Eventual biased Norwegian models need to be mitigated, best without decrease of performance for other tasks. In this context, Research Question 4 is raised:

Research Question 4 *What mitigating techniques could be applied to Norwegian language models to reduce gender bias successfully?*

1.3 Research Method

An experimental methodology is used, as several experiments are required to pursue the goal of the Master's Thesis. The experiments are carried out similarly to those found in related literature, adjusted to fit other datasets or tasks suitable for Norwegian. To scope the thesis one category of language models was investigated. As the target is to provide early insights into bias in Norwegian language models, the choice was made to focus on BERT-based models as an example. BERT, which is an acronym for an acronym for Bidirectional Encoder Representations from Transformers, is a state-of-the-art language model published by researchers at Google in 2019 (Devlin et al., 2019), and is further presented in Section 2.3 as part of the relevant background theory. The versatility of the target and the possibility to easily fine-tune for good results in specific tasks make BERT a widely used model and a natural choice of model to investigate for this project. Based on relevant research on various models and overlap in training data between BERT-based and other Norwegian models, we state that the results are transferable to other Norwegian language models to some degree. To be clear, we have not investigated this statement in this thesis.

This Master's Thesis is a continuation of a preliminary literature review of gender bias in NLP and two courses named Gender and Diversity in Software Development and Advanced Text Analytics and Language Understanding. Some of the work in this thesis is reuse or inspired from text from this previous work. The literature review and the two courses forms the pre-study of this thesis, and whenever we reuse text from this pre-study it is clearly stated.

1.4 Disclaimer

When reading this Master's Thesis, a few important aspects are necessary to keep in mind. Firstly, Norway is a multilingual country and has more than one official languages with national status. Most people speak Norwegian, which is thus the majority language in Norway. However, Sami people have the status of indigenous peoples in Norway, and the Sami languages thus have stronger protections in the legislation than other minority languages in Norway [8]. Sami and other languages that can be considered Norwegian are excluded in the thesis, but we recognize the issue of excluding minorities from technology (which is what we are criticizing in this project). They are not part of the supported languages for the models investigated either, and so it is out of scope for this thesis as it is defined. Also, the issue of combining two target forms, Bokmål and Nynorsk, in one language model is not considered in the thesis. As the models are said to work for both target forms, examples from the two are included to varying degrees in the experiments. However, Bokmål examples make up the lion's share as the dominating target form and account for most of the training data.

Secondly, gender is seen as a binary classification in this thesis, consistent with most related work, datasets, and the legal definition in Norway. We note that gender is more fluent and should be recognized beyond binary definitions, and this issue is further commented on throughout the thesis' discussion and suggested further work.

Thirdly, the thesis covers bias in the form of gender bias specifically and uses the two terms interchangeably. Bias can cover other kinds of biases like racial or age-based that are not covered in this thesis. Consistent with relevant literature, we claim that the results are transferable to some extent to other kinds of biases, but to be clear, the thesis has not made an attempt to confirm this, and the topic is not further discussed.

Last but not least, the models that are investigated in the thesis are all pre-trained on a massive amount of unfiltered text. For that reason, they can produce results that can be perceived as offensive in an attempt to detect bias. This thesis has no intention to harm, and the results or examples do not necessarily represent the meanings or intentions of the authors.

1.5 Contributions

This thesis provides the first documented insight into gender equality in Norwegian language models. To summarize the thesis findings, the most outstanding contributions are the following:

- Identify concrete topics to be discussed and research gaps to be filled for research on Norwegian language technology to hold the same standard as for English.
- Provide proof of gender bias in both the training data, embeddings and in application of the state-of-the-art Norwegian language models.

[8] https://snl.no/språk_i_Norge

- Enlighten the issue of blindly applying gender bias technology in our society through concrete examples of harms caused in downstream tasks for Norwegian language models.
- Define, develop and provide a set of concrete resources being datasets and metrics for investigating bias in Norwegian language technology that can be further used and developed.
- Conduct the initial investigation of gender bias mitigation on Norwegian language models by applying and evaluating two debiasing techniques being hard-debiasing and fine-tuning on a corpus of only female word forms.

1.6 Thesis Structure

The rest of the Master's Thesis is organized in the following manner:

1. Chapter 2 presents relevant background theory to familiarize the reader with topics that are covered in the thesis.
2. Chapter 3 presents relevant literature on the field of gender bias in language models.
3. Chapter 4 gives an overview of the experiments conducted in the thesis. This includes overviews of the plan, implementation and tools used.
4. Chapter 5 describes the datasets used in the experiments conducted.
5. Chapter 6 presents the set up and results from the experiments conducted.
6. Chapter 7 evaluates the experiments conducted and discusses their implications in light of related work and socially relevant topics.
7. Chapter 8 concludes the work done in the thesis in light of the research goal and questions and presents its contributions along with suggested further work.
8. Appendix A contains a list of all Norwegian adjectives.
9. Appendix B contains a set of Norwegian sentences containing specific gender words.
10. Appendix C contains a Norwegian description of a career woman named Hanna.
11. Appendix D contains a description of the code related to the thesis including the readme-file collected from the Github-project[9] that holds the code base.

[9] https://github.com/andrinelo/norwegian-nlp

2 Background Theory

This chapter presents the relevant background theory of topics that are covered in the thesis. Part of the content in 2.1, 2.2.2, 2.3, 2.4.3 and 2.4.4 are reproduced from our pre-study and are partly or completely similar to text presented there.

2.1 Introductory Topics for Natural Language Processing

The first step in making a computer understand natural language like we do is to obtain a representation of it that the computer actually can handle. Some binary number representation of the words is required, in the same manner as it is represented in a human mind. This is solved by representing a language as a set of vectors, one for each word, where the goal is to capture as much information as possible about a word in a vector. However, traditional vector representations of words include vectors of raw word counts in a text and simple co-occurrences, among others, and fail to capture meanings and similarities between words due to their simplicity. The vectors are long and sparse, and most of the values are equal to zero as most words only occur in the context of just a few others. They also fail to capture variations of word meaning based on their use in a context as they are static. Thus, a critical contribution to modern statistical natural language processing (NLP) is the introduction of distributional semantics in the representation of words.

2.1.1 Distributional Representation of Words

Distributional semantics is a research area for developing and studying methods for quantifying and categorizing semantic similarities between words based on their distributional properties in large text samples. As a method in NLP, the use of vector space models as an implementation of distributional semantics is the state-of-the-art solution to the digital representation of a language. The idea is simply that the meaning of a word can be obtained from the distribution of it in a large sample of the text so that *"you shall know a word by the company it keeps"* (Firth, 1957).

2.1.2 Word Embeddings

Machine learning (ML) based methods for generating dense and low dimensional semantic vectors based on a word's distributional properties, or *features*, have been introduced to obtain good digital representations of words. ML is used to train clusters of vectors representing the words into a vector space. Such vectors are referred to as *word embeddings*.

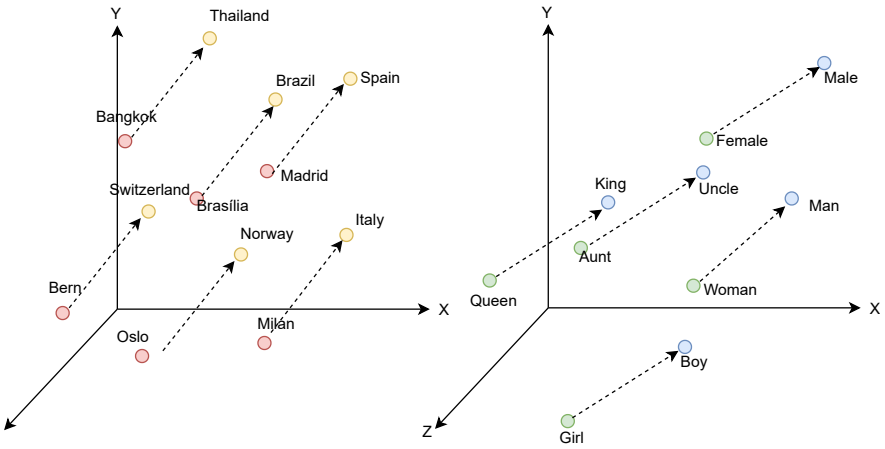


Figure 2.1: The same difference between different countries and capitals and female and male words can be seen throughout the vector space.

Words that are similar in what context they appear will occupy locations close to each other in the vector space, and different words will have locations much further away from each other. The vector differences between words in the embedding represent relationships between words as shown in Figure 2.1. Thus, word embeddings capture rich semantic, syntactic, and conceptual information about words and their meanings. Different examples of sentences containing the word are inputted into the model to obtain the word's meaning. In this way, the computer can learn the rules of the language *unsupervised* without us humans knowing them ourselves. A word embedding, trained on word co-occurrence in text corpora, represents each word w_t as a d-dimensional word vector $\vec{w} \in \mathbb{R}^d$. The goal is to define a model that predicts $P(w_{-t}|w_t)$ between a focus word w_t and all its context words w_{-t} in terms of word vectors.

Examples of popular word embedding algorithms are, word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017). The traditional embeddings are defined as *context-independent* or *static*, meaning there is only one vector representing each word, where the function maps each word type to a single vector. If any, different senses of a word are combined in the vector as a single representation in these vectors. This is not necessarily enough to obtain the meaning of words, as it often can vary between different contexts. For example, the word 'address' is both a verb and a noun, dependent on its context.

2.1.3 Language Models

Language models are complex structures that produce word embeddings which are then used as features in models. They provide a solution to the problem that word embeddings face due to being context-independent, as language models generate embeddings that allow multiple vector representations for the same word. In a language model's word embeddings, all the different senses of a word are represented as their vector. Thus,

the embeddings in language models are *contextual* or *context-dependent*. These models analyze text to cluster the words in a vector representation using various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. Given a sequence of words composed of $w_1, w_2, ..., w_{i-1}, w, w_{i+1}, ..., w_n$ a language model can find which is the i-th missing word, w, by estimating the most probable word with the conditional probability $P(w_i = w|w_1, ..., w_n)$. Language models are shown to capture more information about the meaning of the words than any other technique within NLP. They are considered state-of-the-art for representing natural language in a vectorized format. Examples of state-of-the-art language models are the already mentioned BERT (Devlin et al., 2019), in addition to ELMo (Peters et al., 2018), which BERT is based on, and GPT-3 (Brown et al., 2020). ELMo and GPT are not further considered in this thesis, but are mentioned as part of related work in Chapter 3.

2.1.4 Neural Networks

Artificial neural networks, usually called neural networks (NNs), are computing systems inspired by the biological neural network in a human brain. Here, a collection of nodes, or artificial *neurons*, connected through *edges* model the neurons in a biological brain and can transmit signals to each other (Goodfellow et al., 2016). A single layer neural network is called a *perceptron*, and a multi-layer perceptron is called a neural network. Perceptron is a binary linear classifier of input data used in supervised learning. Figure 2.2 shows an illustration of a perceptron. A neuron aggregates the input by processing the sum of the input signals by some non-linear function to produce the output signal as part of a learning process. Neurons and edges are typically weighted, meaning they have a weight that increases or decreases the strength of the signal at a connection. Depending on whether the aggregate signal crosses that threshold, a threshold value decides if it is passed on.

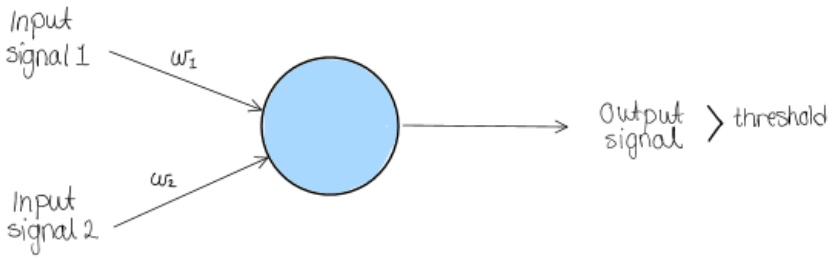


Figure 2.2: Illustration of a perceptron receiving two input signals and forwarding an aggregated output signal based on these.

In its simplest form, an NN is composed of an input and an output. However, complex *deep learning* models consist of multiple additional layers of neurons, called *hidden layers*, that perform such transformations on the inputs. Different types of NNs differ in how the neurons are traversed. The simplest type of NN is a **feed-forward NN**, where

signals are transmitted forward in the network from the input layer and further to the output layer. Figure 2.3 illustrated a feed-forward NN. On the other hand, **recurrent NNs** allows for signals to loop during the traversing so that the network can use what it has learned in the past to compute the present. This is especially suitable for handling sequential data, like sentences or longer text sequences in NLP, because of their built-in memory. However, the gradient storing the sequential information will gradually be smaller and smaller over time, and hence information will disappear from the built-in memory. This is referred to as the vanishing gradient problem and will occur in NLP when processing long sequences. To overcome this issue while still keeping functionality for short-term memory, **Long Short-Term Memory** (LSTM) NNs were introduced by Hochreiter and Schmidhuber (1997). These consist of cells with different types of gates that are created to decide on cell behaviour. Nor LSTM NNs solve the problem of remembering long-term dependencies but works better for it than RNNs. Including a backward layer in addition to the standard forward layer to read input from both sides, creating a **Bidirectional LSTM**, has improved text classification further but still fails to provide insights about the parts in-between.

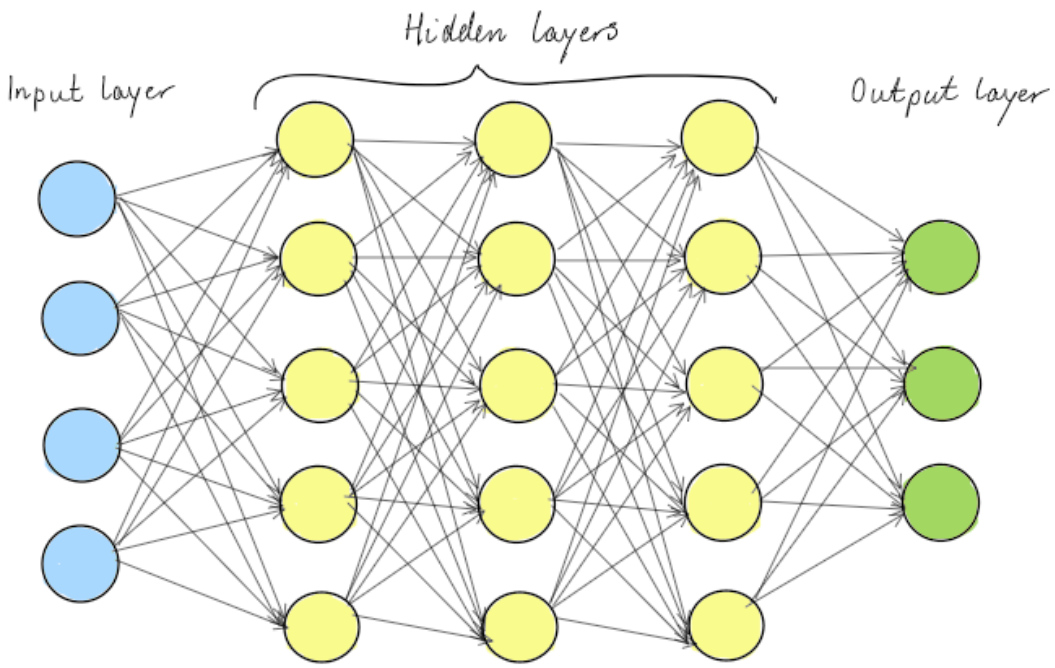


Figure 2.3: Illustration of a feed-forward artificial neural network.

2.1.5 Encoder-Decoder Models

By combining sets of these NNs into encoders and decoders, complex models referred to as **Encoder-Decoder models** are made available. Such models can be fed a sequence of text and output another meaningful sequence and are applied, for example, in language translation. The input sequence is split into each of its constituent words, then the word vector of each respective word is processed by an NN, constituting the encoder. The final encoded hidden state will then be sent through another set of NNs, constructing the decoder, where necessary techniques are applied to complete processing each word according to the specific task. Figure 2.4 illustrates this architecture using RNNs in the encoder and decoder.

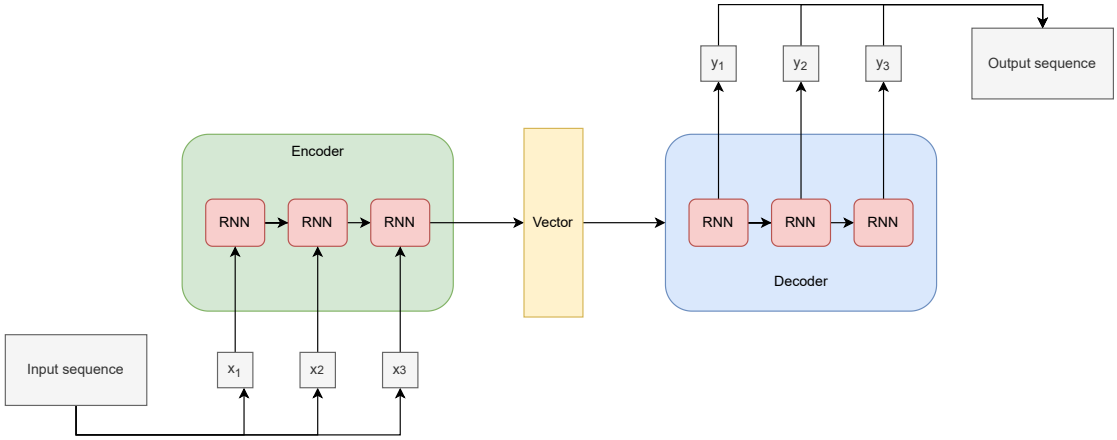


Figure 2.4: Illustration of an encoder-decoder architecture consisting of recurrent neural networks processing an input sequence to an output sequence of text.

2.1.6 Transformer-Based Architecture

Vaswani et al. (2017) published the groundbreaking paper "Attention is All You Need" that replaced the traditional way of doing language processing with a new Encoder-Decoder architecture, namely the **Transformer**. It is a deep learning model that adopts the mechanism of self-attention by a differential weighting of the significance of each part of the input data. An attention function will map a query and key-value pair to output all vectors, and from this, the weighted sum of the values is computed. Vaswani et al. (2017) present a training technique called *masked language modeling* which allows for bidirectional training of models for which it was previously impossible. Where directional models read the text input sequentially from either left to right or right to left, the Transformer encoder reads the entire sequence of words at once and prioritizes its attention. With this characteristic, the model is allowed to learn the context of a word based on all of its surroundings, not just one side of it. Most modern state-of-the-art language models use a transformer-based architecture.

2.1.7 Training of Neural Networks

The learning process of an NN is performed with the layers, either supervised or unsupervised, by processing large amounts of examples. In supervised learning, each example contains a known input and result. The training is usually conducted by calculating probability-weighted associations between the two, stored within the network by determining the difference between the processed and target output. The NN then adjusts its weighted associations according to a learning rule and uses this difference value to produce increasingly similar output to the target output. In unsupervised learning, the model works independently and mainly deals with unlabelled examples to discover patterns and information or cluster similar data into groups.

Complex language models are relatively resource-intensive to train and are generally *pre-trained* for use. This is done over large amounts of unlabeled text data through unsupervised learning. However, the models can further be *fine-tuned* for specific downstream tasks, such as machine translation or sentiment analysis. This is generally done through supervised learning on domain-specific data for the task.

There are mainly two approaches to language modeling in the manner of different languages; *monolingual models* and *multilingual models* (including bilingual and up). The models with their architectures and techniques can be identical for different languages, but the difference lies in the training data. A model can be specialized in a language by being trained only on text in that specific language or generalized by including texts of several languages. Thus, when referring to a Norwegian language model, the actual meaning is a pre-trained language model on Norwegian training data.

2.2 Linear Algebra as a Tool in Natural Language Processing

Representing a vocabulary in a vector space allows for various mathematical operations to be performed on the vectors that represent words. As the location of a word in the vector space correlates with its meaning, similarities and differences in meaning can be calculated from similarity measures in linear algebra operations on vectors in the vector space. This is a well-suited tool for understanding and evaluating a language model and its understanding of a language.

2.2.1 Simple Association Measures

The simplest measures of associations and relationships between words are more or less advances raw counts of the words' co-occurrences. The co-occurrence of a word with another word or category of words can indicate the relation between them in the clustered vector space. Point-wise mutual information (PMI) is an example of a slightly more sophisticated measure of association between words than raw counts. PMI investigates the co-occurrence of words with a particular category, for example gender, by linking descriptors (such as adjectives or verbs) to a counted entity in the category.

The probability of their co-occurrence to the category across entities is calculated to determine the relationship. More formally, PMI exemplified with gender association in this context is defined as:

$$PMI(gender,\ word) = \ln(\frac{P(gender, word)}{P(gender)P(word)}). \tag{2.1}$$

2.2.2 Similarity Measures Between Vectors

Linear algebra operations allow for measures of the word vectors' actual relationships in the vector space. A similarity measure uses word embeddings as input and returns a number measuring their similarity. The three most popular measures to find the similarity between two vectors are *Euclidean similarity, dot product* and *cosine similarity.*

Euclidean Similarity The Euclidean similarity (d) of two vectors $\vec{v}$ and $\vec{w}$ describes the length of the distance between the vectors and can be computed as

$$d(\vec{v}, \vec{w}) = \sqrt{\sum_{i=1}^{n}(\vec{v_i} - \vec{w_i})^2}. \tag{2.2}$$

The Euclidean distance decreases with increasing similarity. Euclidean similarity does not take the lengths of the vectors due to the difference in frequency of co-occurrence into consideration. As the vectors are of different lengths, a length bias in the formula affects the similarity measure. This length bias can be removed by normalizing all the vectors $\vec{v_i}$ so that they have unit length $\|\vec{v_i}\| = 1$.

Dot Product The dot product of two vectors $\vec{v}$ and $\vec{w}$ can be computed as

$$\vec{v} \cdot \vec{w} = \|\vec{v}\|\|\vec{w}\| \cos(\theta) \tag{2.3}$$

where θ is the angle between the vectors. It increases with both increased similarity and increased length of vectors. This is important because examples that frequently appear in the training set tend to have embedding vectors with large lengths. Thus, the dot product captures the word's popularity represented in the vector.

Cosine Similarity The cosine similarity between two vectors $\vec{v}$ and $\vec{w}$ describes the cosine of the angle θ between the vectors and can be computed as

$$\cos(\theta) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\|\|\vec{w}\|} \tag{2.4}$$

where θ is the angle between the two vectors and $\|\vec{v}\|$ denotes the euclidean norm of $\vec{v}$. This distance increases with an increasing similarity between vectors. The cosine similarity uses normalized vectors, so the measure's length bias is removed in the equation itself. Cosine similarity is the same as the dot product of two vectors if the vectors are normalized first.

2.2.3 Principal Component Analysis

Principle Component Analysis (PCA) performs an affine transformation of original basis and is a commonly used technique to describe a set of large dimensional vectors, such as word embeddings, with a (set of) much lower dimensional vector(s). Similar to the similarity measures, PCA depends on normalizing (Gewers et al., 2021). It is a change of basis to determine the directions with maximum variance in the dataset. The most essential features are defined as the features with the most considerable variance (that strongly differ) across the dataset. The data dimension can then be reduced by keeping the n vectors with highest variance, further referred to as the *principal components*. A combination of 1 or more principal components forms a subspace in the initial vector space. Figure 2.5 illustrates a principal component that describes variance caused by gender in a vector space retrieved from calculating the distance between a set of gender word pairs exemplified by 'hun' and 'han', and 'jente' and 'gutt' (English: 'she' and 'he', and 'girl' and 'boy').

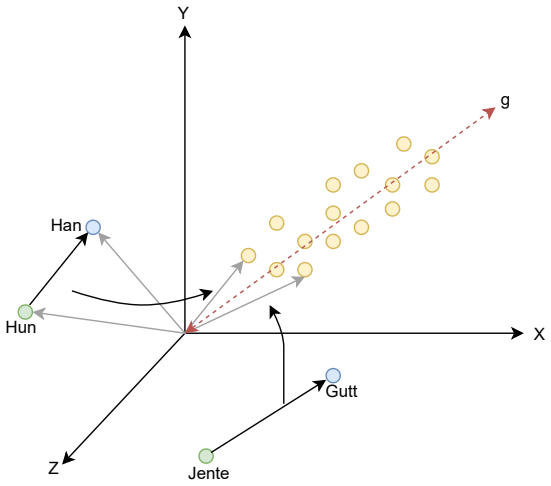


Figure 2.5: By finding the difference between 'hun' - 'han', and 'jente' - 'gutt' in all sentences the principal components of the gender subspace are obtained. Here the first principal component is illustrated in three dimensions.

2.2.4 Orthogonal Projection

Orthogonal projection of a vector to a subspace can be used to remove a subspace from word embeddings. This vector is formally defined as

$$P_{\vec{v}}(\vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\vec{v} \cdot \vec{v}} \vec{v}. \tag{2.5}$$

It is known as $\vec{w}$'s component in the direction given by $\vec{v}$, or the orthogonal projection of $\vec{w}$ on to $\vec{v}$. The vector of $\vec{w}$ orthogonal to $\vec{v}$ is the vector $\vec{w} - P_{\vec{v}}(\vec{w})$.

2.3 Bidirectional Encoder Representations from Transformers (BERT)

In 2019, Google introduced a new model with state-of-the-art results in a wide variety of NLP tasks (Devlin et al., 2019). The model is named **BERT**, an acronym for Bidirectional Encoder Representations from Transformers. It is a pre-trained multi-layer language model that implements the encoder part of the Transformer architecture to construct a complex model for encoding word embeddings and probabilities from the input sequence. Several proposed improvements and adjustments have been introduced later on, for example, Facebook's RoBERTa (Liu et al., 2019), including BERT for many other languages.

2.3.1 Word Embeddings in BERT

BERT has the capability to embed the essence of words in the text inside densely bound vectors. Each vector has a set of values different from zero with a purpose for holding that specific value. The embeddings are vectors comprising 768 digits, meaning they have a size of 1×768. BERT is skilled at generating these dense vectors, and all encoder layers in this multi-layer model output a collection of dense vectors. These vectors are further used as high-quality feature inputs to downstream models.

2.3.2 Training Technique of BERT

To make it easily available for use and for specialized NLP tasks, BERT is trained in two steps; one for pre-training and one for fine-tuning. BERT's most important technical innovation is applying the bidirectional training of Transformer to language modeling and prioritizing the attention between different parts of the input sequence during training. Devlin et al. (2019) show that a language model that is bidirectionally trained has a deeper understanding of language context and flow than single-direction language models. Pre-training is done over unlabeled data for two tasks; Masked Language Modelling (MLM) and Next Sentence Prediction (NSP).

- MLM masks random words to avoid self-detection when learning a bidirectional representation of a sentence. The masked language model gets a tokenized sentence as input, with one masked token. The output is the predicted missing token.
- NSP trains it to detect whether or not two sentences follow each other. In NSP, 'CLS' is the reserved token to represent the start of the sequence, while 'SEP' separates segments (or sentences). [CLS] is placed at the begnning of the input sentence, and [SEP] shows the end of a sentence. The next sentence prediction happens by providing pairs of sentences separated with [SEP], and then the model has to predict if the second sentence is random.

As part of the training technique, MLM allows for BERT to be used to predict masked tokens as a downstream task. This means that model users can apply the same task. A

word (token) is masked in a sentence for the model to predict the missing word. The input is a text sequence, and the output is a list of words and a set of scores for each and includes the top five predicted words for each masked sentence by default. The score provided for each word is the probability that the language model will predict this word to be the masked one among any word in the vocabulary. For instance, in the following example, there is a chance of 0.0996 that the masked word will be 'real' and a chance of 0.0722 that the masked word will be 'the'.

```
[3] pipe('Transformers is a [MASK] library')
```

```
[{'score': 0.09960377216339111,
  'sequence': 'Transformers is a real library',
  'token': 17754,
  'token_str': 'real'},
 {'score': 0.07224985212087631,
  'sequence': 'Transformers is a the library',
  'token': 3141,
  'token_str': 'the'},
 {'score': 0.046838272362947464,
  'sequence': 'Transformers is a true library',
  'token': 23962,
  'token_str': 'true'},
 {'score': 0.04389515891671181,
  'sequence': 'Transformers is a action library',
  'token': 22959,
  'token_str': 'action'},
 {'score': 0.04341385141015053,
  'sequence': 'Transformers is a last library',
  'token': 10582,
  'token_str': 'last'}]
```

Fine-tuning is done by initializing BERT with the pre-trained parameters and fine-tuning them on labeled data from the specific downstream tasks. Compared to pre-training, the fine-tuning is relatively inexpensive, and Devlin et al. (2019) state that all results can be replicated in a few hours at most. This enables BERT as a model that can be used from pre-training for various tasks, and that can be optimized even more for a specific task with fine-tuning. A distinctive feature of the model is the unified architecture across different tasks, where there is a minimal difference, with just one additional output layer, between the pre-trained architecture and the final fine-tuned downstream architecture. Devlin et al. (2019) show eleven new state-of-the-art results across different tasks with the use of BERT and fine-tuning.

2.3.3 Training Corpus of BERT

The natural language data used to create word vectors are represented as large and structured sets of digital text, either written or transcribed from speech, and is called a *corpus*. The state-of-the-art solution primarily uses open text data sources for training language models. Collections from open sources like Wikipedia, Common Crawl data, online news, and BooksCorpus (Zhu et al., 2015), are generally used. BERT is no exception, and the fact that these corpora contain longer texts with a natural structure of sentences is critical in the training of BERT, as Next Sentence Prediction is one of the tasks in the pre-training. BERT is pre-trained on a corpus of 16GB text with 3.3B words consisting of BooksCorpus (800M words) and English Wikipedia (2500M words). Research on improving the model conducted after the release of BERT has shown that this might be too small of a corpus for the job. For example, Facebook contribution RoBERTa (Liu et al., 2019) showed a considerable gain in performance when training on an increased corpus size of 160GB.

2.4 State-Of-The-Art Norwegian Language Processing

Norway is a multilingual country, but most people speak Norwegian, which is the majority language in Norway. Further is a presentation of some distinctions and relevant manners of Norwegian in the context of language modeling and available resources for the processing of the Norwegian language.

2.4.1 Norwegian Language in Context of Language Processing

Norwegian is a North Germanic language closely related and mutually intelligible to Danish and Swedish. It is challenging to distinguish Norwegian from Swedish and Danish according to purely linguistic criteria; in practice, modern Norwegian can be considered the Scandinavian dialects and standard languages that have a geographical connection to Norway. One distinction of Norwegian is the fact that the language has *two* official target forms; Bokmål and Nynorsk. Both target forms have approached each other through use and several spelling reforms but still stand as two equalized target forms as of the Act on Target Use in Public Service from 1980[1]. Even though Bokmål and Nynorsk are equalized, they are not equally used by Norwegians. 11.6% of Norwegian pupils in primary and lower secondary school have Nynorsk as their main language, while Bokmål accounts for 87.3%, and the number is increasing.[2] Bokmål dominates quantitatively and in the capital and other metropolitan areas. The solution for dealing with two target forms of Norwegian language technology has been to look at the two target forms as one language and jointly train models on data from both. The minority variant Nynorsk is then represented by comparatively less data than Bokmål, reflecting the natural usage. Publishers of Norwegian language models, Kummervold et al. (2021) and Kutuzov et al.

[1] https://lovdata.no/dokument/NLO/lov/1980-04-11-5
[2] https://www.ssb.no/utdanning/grunnskoler/statistikk

(2021), claim that the models work for both Bokmål and Nynorsk and show individual evaluation results for the two target forms. Norway does not have an official standard spoken language, and there is great acceptance for using dialect in most contexts. In oral use, the dialects are stronger than in most other countries, and mixed forms of dialects are also widely used. Social media content can be written in dialect-like language, but more formal texts such as news are written in proper Bokmål or Nynorsk.

New words can be created in Norwegian by combining two existing words in the language. To create valid new compound words, one can combine words from all word classes. Norwegian also has great freedom of choice that exists when it comes to the spelling of words and the choice of grammatical inflections. In combination with a high acceptance of compound words, this makes the possible vocabulary of the language vast and requires good tokenization from language models that should understand Norwegian text. English language programs exist that offer the author support in the consistent use of words and help maintain a certain level of style. For Bokmål and Nynorsk, such support functions will probably be even more challenging to design than English with their large degree of freedom and choice.[3]

2.4.2 Gender Neutrality in Norwegian Text

Gender is, in linguistics, a morphological category where the belonging of a word is expressed through inflection. A language is said to have *grammatical gender* if it has different classes of nouns[4]. Norwegian has grammatical gender, as the nouns can have three genders being *masculine, feminine* and *neuter* corresponding to respectively the articles 'en/ein', 'ei' and 'et/eit' for Bokmål/Nynorsk. Nynorsk uses all three genders, while in Bokmål, language users have the choice between this three-part system and a system with two genders; neuter and *common gender*. In practice, all female words in Bokmål can alternatively be male words. The two-gender system is found, for example, in the Bergen dialect and 'conservative' variants of Bokmål. Many nouns can pass as both feminine or masculine, dependent on the target form, dialect, or preference of the speaker/writer. Choosing the masculine gender will often seem more formal than using the feminine, for example, by the translation of 'the cabin' to 'hytten' (masculine), which is common in the finer parts of Norway's capital, Oslo, instead of the more common 'hytta' (feminine)[5].

Properties of the noun determine how the adjective is inflected. The adjective takes different forms depending on whether it describes a masculine, feminine, or neuter noun. It also plays a role in inflecting whether the noun is singular or plural and in definite or indefinite form.[6] Norwegian adjectives can be said to have an indefinite (strong) and a definite (weak) inflection. The strong inflection coincides between masculine and feminine gender in the singular in both target forms and coincides between all genders in the

[3] https://www.sprakradet.no/Vi-og-vart/Publikasjoner/Spraaknytt/Arkivet/Spraaknytt_1998/
[4] https://snl.no/genus_-_grammatikk
[5] https://snl.no/substantiv
[6] https://snl.no/adjektiv

plural. The weak form has only one form.[7] Norwegian also has masculine and feminine personal pronouns, presented in Table 2.1. The Norwegian Language Council is also in the process of implementing the gender-neutral pronoun 'hen' (English: 'they') as part of the language[8].

Table 2.1: Gendered personal pronouns in the Norwegian language terms Bokmål and Nynorsk.

	Bokmål	Nynorsk
Masculine	Han, Ham	Han, Ham
Feminine	Hun, Henne	Ho, Henne

Difference Between Grammatical and Social Gender Linguistic categories of gender do not necessarily map well to social categories (Cao and Daumé III, 2020) anymore (if they ever did). Thus, literature on gender in linguistics often distinguishes the following types of gender, which are not all-encompassing and merely outline gender categories presented in the literature (Stanczak and Augenstein, 2021):

- *Grammatical gender* which refers to a classification of nouns based on a principle of a grammatical agreement into categories.
- *Referential gender* which identifies referents as female, male or neuter (Cao and Daumé III, 2020).
- *Lexical gender* which refers to the existence of lexical units carrying the property of gender, male or female-specific words such as father and waitress (Cao and Daumé III, 2020).
- *(Bio-)social gender*, which refers to the imposition of gender roles or traits based on phenotype, social and cultural norms, gender expression, and identity such as gender roles (Ackerman, 2019; McConnell-Ginet et al., 1987).

From this definition, we state that even though a word, such as an adjective, might have grammatical gender, this is not consistent with the social gender of the word. For example, the adjective 'liten' (English: 'little') can be inflected as both grammatically masculine ('liten') and feminine ('lita') dependent on the noun it describes. With the noun being 'jente' (English: 'girl'), the masculine inflection would be 'en liten jente' (English: 'one little girl'), while the feminine would be 'ei lita jente'. One can argue that 'liten' thus is a male adjective, opposite to its female correspondent 'lita'. However, all female nouns in Bokmål can alternatively be male, and 'liten' is therefore also an acceptable (and definitely the most common) inflection for females. The connection between grammatical and social gender is dismissed. The Norwegian newspaper Aftenposten even stated in

[7] https://snl.no/norsk
[8] https://www.nrk.no/trondelag/sprakradet-onsker-a-innfore-ordet-hen

2020 that grammatical female gender is on its way out of the language.[9] Thus, Norwegian language processing should not divide between male and female inherited in the meaning of the words based on the words grammatical gender (most likely male). Thus, all words like adjectives or occupations, for example, except for lexical gendered ones like 'flyvert' versus 'flyvertinne' (English; male and female for 'flight attendant'), should be (bio-) social *gender neutral*. This term is used throughout the thesis and refers to social gender neutrality as per the definition presented here.

2.4.3 Norwegian Training Corpora

Relative to the distribution of Norwegians with Bokmål compared to Nynorsk as their first language, Nynorsk has a strong standing in, among others, the mass media in Norway. The national Norwegian broadcasting corporation, Norsk Rikskringkasting (NRK), must produce 25% of its content in Nynorsk. However, this requirement is not held by NRK every year[10]. The distribution of Nynorsk and Bokmål in Norwegian mass media indicates the amount of Nynorsk news related data generated in comparison to Bokmål. Naturally, from the division of speakers of Bokmål and Nynorsk, this unevenness in the division also applies to other written data sources such as Wikipedia, where Nynorsk Wikipedia consists of approximately 49M words, which is 25% of the 160M words in Bokmål Wikipedia (Kutuzov et al., 2021).

A challenge in NLP is that the training requires enormous amounts of structured data. These sources are limited, especially in minority languages defined in the context of machine learning. Norwegian is considered a minority language for building large text corpora, making it hard to train well-performing transformer-based models. The fact that there exist more than 100 times as many English Wikipedia pages as Norwegian Wikipedia articles (Kummervold et al., 2021) indicates the challenges faced in this manner. Norwegian's most comprehensive training sources consist of Norwegian news articles, Wikipedia dumps, and a colossal corpus of all text sources held by The National Library of Norway (NLN). A summary of the three is given here:

- **Norsk AvisKorpus**[11] (NAK) is a much used and easily available dataset. The version that is presented and used in this thesis is a collection of Norwegian news texts in both Bokmål and Nynorsk from 1998 to 2019. It consists of more than 1.74 billion words, approximately 1.68 billion for Norwegian Bokmål and about 68 million words for Norwegian Nynorsk. News articles contain mostly complete and grammatically correct sentences but can also be written in more oral forms and be of more extreme or 'tabloid' character content-wise.
- **Norwegian Wikipedia** has support for both the written languages Bokmål and Nynorsk, and dumps are available and often used as corpora. **Bokmål Wikipedia** (Wikipedia NO) dump consists of approximately 160 million words, while **Nynorsk Wikipedia** (Wikipedia NN) consists of approximately 40 million words.

[9] https://www.aftenposten.no/viten/i/dOypm1/grammatisk-hunkjoenn-er-paa-vei-ut
[10] https://www.nrk.no/vestland/nrk-nadde-malet-om-25-prosent-nynorsk-i-2021
[11] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

- **The Norwegian Colossal Corpus (NCC)** is a work-in-progress corpus by The National Library of Norway (NLN). Considering the increased performance presented in RoBERTa (Liu et al., 2019), the size of the corpora plays a significant role in how the model performs. NLN has access to huge amounts of non-public text sources, and Kummervold et al. (2021) consider it the duty of NLN to make available Norwegian digitized text sources that can be used for training language models. This has resulted in the creation of NCC[12], where they were able to build a corpus of 109GB (18.4B words) of raw, deduplicated text from the digitization of these sources. Table 2.2 is obtained from the publisher of NCC (Kummervold et al., 2021) and presents the composition of the complete corpus. They are currently in the process of making as much as possible available and have published a large subset of the corpus. The published part of the corpus consists of ca. 6.9B words.

Table 2.2: The composition of the Colossal Norwegian Corpus (NCC). The table is completely obtained from Kummervold et al. (2021).

Sources	Period	Words (Millions)	Text (GB)
Books (OCR)	1814–2020	11,820	69.0
Newspaper Scans (OCR)	2015–2020	3,350	20.0
Parliament Documents (OCR)	1814–2014	809	5.1
Common Crawl OSCAR	1991–2020	799	4.9
Online Bokmål Newspapers	1998–2019	678	4.0
Periodicals (OCR)	2010–2020	317	1.9
Newspaper Microfilms (OCR)	1961, 1971, 1981, 1998–2007	292	1.8
Bokmål Wikipedia	2001–2019	140	0.9
Public Reports (OCR)	1814–2020	91	0.6
Legal Collections	1814–2004	63	0.4
Online Nynorsk Newspapers	1998–2019	47	0.3
Nynorsk Wikipedia	2001–2019	32	0.2
Total (After Deduplication)		18.438	109.1

2.4.4 Norwegian BERT-Based Language Models

In addition to BERT, a monolingual English model, Devlin et al. (2019) released another version of BERT as a contribution to NLP of other languages than English. This version, referred to as multilingual BERT (**mBERT**), has Norwegian included as one of the languages, making mBERT considerable as a multilingual Norwegian language model. With the use of the corpora presented in Section 2.4.3, Norwegian BERT-based models that outperform Google's mBERT in several tasks have been published. There are mainly two initiatives that have made significant contributions to creating a monolingual Norwegian BERT-based language model; The Language Technology Group (LTG) at The University of Oslo (UiO) and the Artificial Intelligence Lab (AI Lab) at The National

[12]https://github.com/NBAiLab/notram

Library of Norway (NLN). The two projects are not coordinated, but both have released each their version of a Norwegian BERT in 2021, respectively **NorBERT** (Kutuzov et al., 2021) and **NB-BERT** (Kummervold et al., 2021). Figure 2.6 gives an overview of the three models, mBERT, NB-BERT, and NorBERT, relations to the original BERT model, and the following paragraphs compare them further.

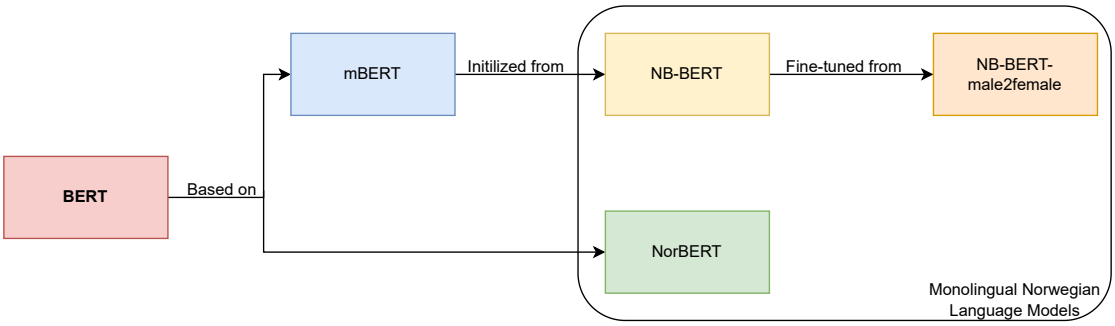


Figure 2.6: Illustration of the relations between the BERT-based models that are considered in this thesis, including the fine-tuned version of NB-BERT (NB-BERT-male2female) that is presented in Chapter 6.

mBERT mBERT[13] works similarly to BERT but is pre-trained on a corpus consisting of Wikipedia content in the 104 different languages with the largest Wikipedia. Norwegian Wikipedia is one of them, and mBERT's training corpora include an amount of 1.1GB (172M words) data from Wikipedia pages for Bokmål and Nynorsk together, estimated by Kummervold et al. (2021). This is a small corpus in the context of machine learning. Even though it is generally agreed that language models acquire better language capabilities by pre-training with multiple languages, there is a strong indication that this amount of data might have been insufficient for mBERT to learn high-quality representations of Norwegian (Pires et al., 2019; Wu and Dredze, 2020).

NB-BERT NB-BERT[14] is a general BERT-base model published by The National Library of Norway (NLN) and built on the corpus they are currently working on publishing as presented in Section 2.4.3. Thus, NB-BERT's training corpus consists of 18.4B words (109GB), including a wide variety of Norwegian text in both Bokmål and Nynorsk from the last 200 years. NB-BERT is based on the same structure as mBERT and is initiated from its pre-trained weights to obtain a better-performing model on mixed-language texts. This means that NB-BERT has similar multilingual properties to mBERT, making NB-BERT not exclusively a Norwegian language model. However, as the target of keeping the multilingual properties is to increase the model's performance on Norwegian text with loanwords, for example, the model is considered a monolingual Norwegian model

[13] https://huggingface.co/bert-base-multilingual-cased
[14] https://huggingface.co/NbAiLab/nb-bert-base

in this thesis. The model has support for both Bokmål and Nynorsk from the training corpus, in addition to the multilingual properties from mBERT initialization.

NorBERT According to Kutuzov et al. (2021), UiO's NorBERT[15] and simultaneously published NorELMo, are *the first large-scale monolingual language models for Norwegian.* Contrary to NB-BERT, NorBERT is trained from scratch without the pre-trained weights of mBERT, and thus almost exclusively in Norwegian. This makes NorBERT a monolingual Norwegian language model within the definition. NorBERT is trained on a smaller corpus than NB-BERT of around 5GB (1.9B words) of data from Wikipedia and Norsk Aviskorpus, and so NorBERT also provides support for both Bokmål and Nynorsk. Kutuzov et al. (2021) argue that their language model trained on less but arguably cleaner data can outperform a model trained on larger but noisy corpora, such as NB-BERT. According to Kutuzov et al. (2021), the model features a custom WordPiece vocabulary with much better coverage of Norwegian than both mBERT and NB-BERT. This increases the model's ability to tokenize Norwegian words correctly, which is exemplified by a comparison of the tokenization of the sentence

> "Denne gjengen håper at de sammen skal bidra til å gi kvinnefotballen i Kristiansand et lenge etterlengtet løft"

where NorBERT outperforms both mBERT and NB-BERT, which both use the same vocabulary.

- mBERT/NB-BERT: Denne g ##jeng ##en h ##å ##per at de sammen skal bid ##ra til å gi k ##vinne ##fo ##t ##ball ##en i Kristiansand et lenge etter ##len ##gte ##t l ##ø ##ft
- NorBERT: Denne gjengen håper at de sammen skal bidra til å gi kvinne ##fotball ##en i Kristiansand et lenge etterl ##engt ##et løft

2.4.5 Comparison of the Models

An overview of the training data used in the three Norwegian BERT-models, as well as Googles original monolingual BERT for English, is presented in Table 2.3 on page 24. This table is reproduced from the pre-study. Comparing language models is done mainly by investigating how well they perform on different standardized tasks. When Kutuzov et al. (2021) compare mBERT, NB-BERT, and NorBERT, they test their performance on several traditional NLP tasks; part-of-speech tagging, named entity recognition, fine-grained sentiment analysis, binary sentiment classification, and negation detection. However, none of these tasks say anything about how biased the models are, which is (should be) essential when deciding what model to choose for your desired NLP task.

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

Table 2.3: Overview of the different BERT-models and their publisher (Publ.), release year, corpus source, languages supported, size of, and amount of words in the (Norwegian part of the) training corpus.

Model	Year	Publ.	Corpus	Language	Size	Words
BERT	2019	Google	Wikipedia, BooksCorpus	English	16GB	3.3B
mBERT	2019	Google	Wikipedia NO and NN	Top 104 from Wikipedia	1.1GB	172M
NorBERT	2021	UiO	Wikipedia NO and NN, NAK	Norwegian	5GB	1.9B
NB-BERT	2021	NLN	NCC	Norwegian	109GB	18.4B

3 Related Work

This chapter presents relevant literature and other attempts to detect or mitigate bias in natural language processing. It covers topics like definitions of bias, detection, and mitigation methods.

The idea of Natural Language Processing (NLP) systems containing gender bias became famous when Bolukbasi et al. (2016) presented their findings and has later been a subject of many studies. Stanczak and Augenstein (2021) published a comprehensive literature review of papers on the subject and present a graph showing the explosion of publications on the subject over the last years (reconstructed in Figure 3.1). The studies apply various approaches to prove and measure the presence of gender bias in different parts of an NLP system. Bias is present in a wide aspect covering society ranging from literature (Hoyle et al., 2020), news (Wevers, 2019), media (Asr et al., 2021), communication about and towards people (Fast et al., 2016; Voigt et al., 2019) and book ratings (Touileb et al., 2020). For specific NLP tasks it is proven present in word embeddings (Bolukbasi et al., 2016) and language models (Nadeem et al., 2021), and in downstream tasks like speech recognition (Lu et al., 2020), machine translation (Savoldi et al., 2021; Escudé Font and Costa-jussà, 2019), coreference resolution (Rudinger et al., 2018; Webster et al., 2018; Zhao et al., 2018), language generation (Hendricks et al., 2018; Sheng et al., 2020), hate-speech detection (Park et al., 2018), sentiment analysis (Park et al., 2018) and part-of-speech tagging and parsing (Garimella et al., 2020).

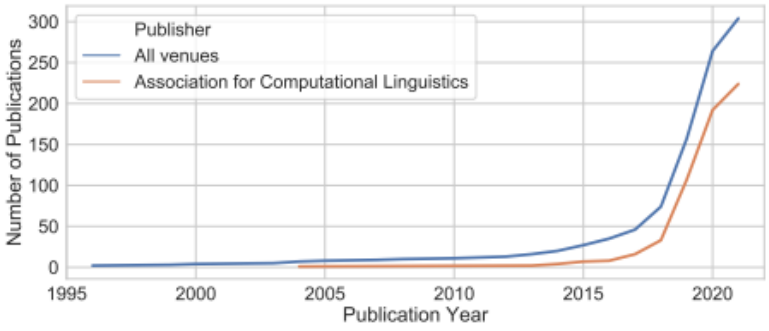


Figure 3.1: Stanczak and Augenstein (2021) analyzed all papers on gender bias in Natural Language Processing and show the number of published papers on gender bias in NLP prior to June 2021. The figure shows a steady increase in the actuality of the topic since 2015.

3.1 Definition of Gender Bias in Natural Language Processing

Bias is defined as inclination or prejudice for or against one person or group, especially in a way considered to be unfair. Gender bias refers to inclination or prejudice based on the person's gender, either biological sex or gender identity. However, defining gender bias as it appears in NLP is not straightforward. Blodgett et al. (2020) criticized the inconsistency in the definition of gender bias in NLP in the different studies on the subject and stated that this makes it hard to compare findings and move forward in the field. This is a consequence of research on gender bias in NLP being an interdisciplinary field covering the complex relationship between different aspects of technology, language, and social hierarchy. Blodgett et al. (2020) discovered that previous research fails to state their conceptualization of bias and thus addresses the importance of researchers stating this clearly. They say that papers usually include descriptions of NLP systems behaviors and then use it as self-evident to prove their bias statement. This has led to papers with different bias definitions implementing the exact solutions and papers with the same bias definition implementing other solutions without any explanations of the differences.

For this reason, the definitions that most agree on are presented and will be used throughout the thesis. Crawford (2017) presented a framework classifying algorithmic biases by the type of harm they cause, which has later been developed and is commonly used as a definition of bias in NLP. They found that gender bias in a text can be categorized into harms causing either **allocational** or **representational** problems. Allocational problems are defined as a certain group being allocated fewer resources than others because of an unfair system. These are immediate problems that are easy to quantify. In contrast, representational harms occur when a system devalues and under-represents some groups and social identities and is a set of problems that are more visible and harder to formalize in the long term. Blodgett et al. (2020) specify the following:

> Allocational harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups. Representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether.

According to Crawford (2017), representational bias is the most prevalent and pressing issue in NLP. Representational bias leads to stereotyping that perpetuates negative depictions, the algorithm fails and performs less good for minorities, and under-representations of minorities. Examples of different NLP tasks that show kind of harms are:

- Translating "He is a nurse. She is a doctor" to Turkish where 'he' and 'she' is tranlated to gender free 'o' and then back to English results in "She is a nurse. He is a doctor"[1].

[1] https://medium.com/@laurahelendouglas/ai-is-not-just-learning-our-biases-it-is-amplifying-them-4d0dee75931d. Accessed:06/12/2021

- Comparing positive adjectives, those that are used to describe women are more often related to their bodies than those used to describe men (Hoyle et al., 2020).
- Automatic speech detection works better with male voices than female voices (Tatman, 2017).
- Sentences with female noun phrases are consistently given higher sentiment when predicting anger, joy, or valence compared to sentences with male noun phrases (Kiritchenko and Mohammad, 2018).
- "He is a doctor" has a higher conditional likelihood than "She is a doctor" (Lu et al., 2020).
- Analogies such as "man : woman :: computer programmer : homemaker' are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).

3.2 Detecting and Measuring Gender Bias in Natural Language Processing

Regardless of the increasing amount of papers stating they have proof of gender bias, there does not exist a standardized measure or technique to detect it. The approaches to detect and measure gender bias are either by analyzing bias in society and the real world with the help of NLP tools and a huge amount of texts or by analyzing concrete NLP tasks. Which part of the NLP system that is investigated varies. Another literature study on the subject was conducted by Blodgett et al. (2020), who made a categorization of the different NLP tasks that have been investigated by papers published before 2020. The division can be seen in Table 3.1 and shows that word embeddings is the most common approach.

Table 3.1: The NLP tasks covered in the papers (up to 2020) that Blodgett et al. (2020) did a survey on. The papers column represents the count of number of papers that exist with the given task.

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

3.2.1 Investigation of Training Data as a Measure of Bias

The information that is captured in embeddings and language models is naturally influenced by the data it is trained on. Thus, the *gender gap* in the data is a subject of investigation by different papers. As men are over-represented in the public sphere, like in politics, entrepreneurship, and leadership, this will influence the text, and studies show that Canadian news discuss and quote men three times more often than women (Asr et al., 2021). Uri (2018) have found that Norwegian news cite women one third of all the times it cite men. In addition, less than 15% of biographical Wikipedia pages are about women (Sun and Peng, 2021). The distribution of genders mentioned in the datasets can be a bias indicator and work as a measure of bias. This varies from raw counts of gender entities (Zhao et al., 2019) to more complex lexical biases like co-occurrence of gender words with other words (Hoyle et al., 2020).

Zhao et al. (2019) conducted an analysis of gender bias in ELMo's contextualized word embeddings and saw how unequal representation in training data influenced its decisions. They showed counts for the number of occurrences of male pronouns ('he', 'his', 'him') and female pronouns ('she', 'her', 'hers') in the corpus, along with the co-occurrence of occupation words with those pronouns. This distribution of pronouns is presented as a measure of gender bias in the model's pre-trained embeddings. The training corpus for ELMo is the One Billion Word Benchmark (Chelba et al., 2014). The results of the pronoun counting can be seen in Table 3.2 and show that male pronouns occur three times more than female pronouns, which is a significant gender skew. Male pronouns co-occur more frequently with occupation words, whether they are prototypically male or female.

Table 3.2: Zhao et al. (2019) did a count on the number of occurrences on gender pronouns in the training corpus for ELMo. The number is the count for the total number of occurrences and shows that male pronouns occurred three times more than female pronouns. The third and fourth columns show counts corresponding to their co-occurrence with occupation words where the occupations are stereotypically male (M-biased) or female (F-biased).

	#occurence	#M-biased occs.	#F-biased occs.
F	1 600 000	33 000	36 000
M	5 300 000	170 000	81 000

Point-wise mutual information (PMI) as presented in Section 2.2.2 can be used as a measure of gender bias. Words with high PMI values for one gender are suggested to have a high gender bias. Later, the sentiment of the descriptors was added by Hoyle et al. (2020) to investigate how the sentiment of words used to describe men and women differ. Qian (2019) calculate an overall *stereotype score* of a text from the following formula:

$$bias(word) = |log\frac{c(word, m)}{c(word, f)}| \tag{3.1}$$

where f is a set of female words (e.g., 'she', 'girl', 'woman'), m is a set of male words (e.g., 'he', 'boy', 'man') and c (word, m/f) is the number of times a gender-neutral word co-occurs with gendered words. A word is gender bias free if the stereotype score is 0, which means it occurs equally frequently with male and female words in the text. These definitions of stereotypical and occupational bias have been employed by Bordia and Bowman (2019) and Qian (2019) to measure bias for occupations and to identify most biased words in blogs and novels.

3.2.2 Investigation of Word Embeddings as a Measure of Bias

Remember that representation problems occur when specific thoughts or devaluation of minorities are included in word representation, such as word embeddings or pre-trained language models (Crawford, 2017). Bolukbasi et al. (2016) and Caliskan et al. (2017) show that word embeddings encode societal biases about gender roles and occupations, e.g., engineers are stereotypically men and nurses are stereotypically women with commonly applied methods.

Analogy Task For Embeddings Recall that in word embeddings each word is represented as a d-dimensional word vector. Bolukbasi et al. (2016) also assumed that there exist a set of gender-neutral words like *shoes* or *table* that are not specifically related to any gender in a social manner. It is also assumed a set of female to male (F-M) gender pairs such as 'she'-'he', 'mother'-'father', and 'queen'-'king' where the only difference between them in regards to meaning is the gender. From two words as a seed pair, a and b, it is determined a *seed direction* $\vec{a} - \vec{b}$ corresponding to the normalized difference between the two seed words. By predicting a pair of words, x and y, containing the same distance as the seed pair between them in the embeddings, the analogy 'a is to x as b is to y' was fulfilled. The metric used to score the pairs of words (x, y) was

$$\mathrm{S}_{(a,b)}(x,y) = \begin{cases} \cos(\vec{a}-\vec{b}, \vec{x}-\vec{y}) & \text{if } \|\vec{x}-\vec{y}\| \leq \delta \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

where δ is a threshold for similarity. The top analogous pairs will then be the pairs with the largest $S_{(a,b)}$. Gender stereotypes were captured if the generated words are closer to 'she' than 'he' or the other way around. This generated bias analogies like the famous *"man is to computer programmer what woman is to home maker"*.

Nissim et al. (2020) claim that Bolukbasi et al. (2016) did not always present the top results for their queries but instead cherry-picked the most biased results to present in their study. They proposed that research should avoid bias-searching queries and claim that analogies are not an accurate diagnostic for bias. When searching for analogies, people usually query *man:doctor::woman:X*, which can give different results than querying *woman:doctor::man:X* depending on the method. Newer tasks should not create a task where the only possible outcome is gender bias but rather show it by creating realistic tasks or a general bias measure.

Calculating Similarity Between Embeddings To determine whether a male or the female group of words is more similar to another group of words, the relative norm distance d between the groups is calculated and typically used as a measure for bias. For a set of gender neutral words M and an average embedding vector $\vec{v_i}$ for a group of words i (e.g. gendered words), we have that:

$$d = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2 \,. \tag{3.3}$$

A higher positive norm distance indicates a stronger association between the neutral words and gender group two. By using occupations or adjectives as the set of gender-neutral words and compare to a set of male and female words, the gender bias in occupations or adjectives can be found. This was done by Sahlgren and Olsson (2019), who investigated both contextual and static embeddings to detect and compare their gender bias. They take reservations that occupational words should be gender neutral in the embeddings (that is, they should not be related to one gender more than the other) and looked at whether this is the case for four different embeddings being word2vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019), all trained for the Swedish language. Bias is detected if the occupations are typically more similar to the names of one gender than the other. The data used included top 100 female names, top 100 male names, 14 *most female* occupations and 14 *most male* occupations, based on workforce division in Sweden. With this data, Sahlgren and Olsson (2019) calculated the percentage of female and male names that are, on average, more similar to a female or male occupation. word2vec groups male occupations with female and male names, while fastText groups female occupations with male names and male occupations with female names. BERT shares the tendency with fastText when looking at the average similarities and is almost balanced when looking at the single most similar occupation. ELMo groups male names with male occupations when examining the average similarities but is less biased for the female names. When looking at the single most similar occupation, ELMo shares the tendency with word2vec that both names are connected to male occupations. Their result can be seen in Table 3.3 (reconstructed from the paper). The number indicates how many names are more similar to an occupation. The number in parenthesis represents the single most similar.

Table 3.3: Results from Sahlgren and Olsson (2019) (reconstructed from the paper) as the number of gender names that are more similar to gender occupations presented on the format 'on average (the single most similar)'.

	Male names	Male names	Female names	Female names
Model	Male occupations	Female occupations	Male occupations	Female occupations
`word2vec`	91(86)	9(14)	99(98)	1(2)
`fastText`	4(10)	96(90)	100(100)	0(0)
`ELMo`	96(63)	4(37)	49(87)	51(13)
`BERT`	37(54)	63(46)	76(55)	24(45)

Psycological Association Test In psychology, the Implicit Association Test (Greenwald et al., 1998) is used to measure subconscious bias in humans. This is done by quantify the difference in time and accuracy for humans to categorize words as relating to two concepts they find similar versus two concepts they find different. This technique have been applied to measure if a gender is perceived to belong more to either science or art (Nosek et al., 2009). Participants are asked to categorize words as pertaining to (males or the sciences) or (females or the arts). The participants are then asked to categorize words as pertaining to (males or the arts) or (females or the sciences). If participants answered faster and more accurately in the former setting, it indicates that humans subconsciously associate males with the sciences and females with the arts. Caliskan et al. (2017) recreated this association test as a benchmark for testing gender bias in word embeddings vi semantic similarities and named it Word Embeddings Association Test (WEAT). Consider two sets of target words (e.g. 'programmer', 'engineer', ... and 'nurse', 'teacher') and two sets of attribute words (e.g. 'man', 'male', ... and 'woman', 'female'...). There should not be a difference between the two sets of target words regarding their relative similarity to the two sets of attribute words. In formal terms, let X and Y be two sets of target words of equal size, and A, B are the two sets of attribute words. Let $\cos(\vec{a}, \vec{b})$ denote the cosine of the angle between the vectors $\vec{a}$ and $\vec{b}$. The resulting test statistics is defined as a permutation test over X and Y:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{3.4}$$

where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}). \tag{3.5}$$

$s(w, A, B)$ measures the association of the word w with the attribute, and $s(X, Y, A, B)$ measures the differential association of the two sets of target words with the attribute. The more positive the value is given by $s(X, Y, A, B)$, the more the target X will be related to attribute A and target Y to attribute B. The null hypothesis suggests there is no difference between X and Y in their relative similarity to A and B, and thus a proof of difference is a proof of bias. Ethayarajh et al. (2019) state that WEAT systematically overestimates bias, and therefore should be applied with caution.

Caliskan et al. (2017) apply WEAT on a range of word sets and find that the prejudice uncovered in their association test corresponds to the prejudice on gender, race, and other social constructs found in GloVe. They show that the meaning of the word, *semantics*, not only reflects the word but captures regularities latent in the culture with its prejudices. They also found that the gender association strength of occupation words is highly correlated between the GloVe embedding and the word2vec embedding. Additionally, a positive correlation was shown between the strength of association of an occupation's word embedding and the female gender to the percentage of females in that occupation in the US.

An extension of WEAT that compares sets of sentences rather than words has further been developed by May et al. (2019), referred to as Sentence Embedding Association

Test (SEAT). WEAT can be seen as a special case of SEAT in which the sentence is a single word. May et al. (2019) imitate the psychological study by Heilman et al. (2004) by applying SEAT on the following sentences and finding evidence of bias in the sentence encoders:

1. Target concept sentence template "<word> is an engineer with superior technical skills." where <word> represent female and male names.
   Attribute sentence template "The engineer is <word>." where word is *likable* and *non-hostile*

2. Target concept sentence template "<word> is an engineer" where <word> represent female and male names. Attribute sentence template "The engineer is <word>." where word is *competent* and *achievement-oriented*.

3.2.3 Masked Language Modelling as a Measure of Bias

Masked language modelling (MLM) can be an effective way to investigate gender differences in language models. Since BERT embeddings use an MLM objective, the model can be directly queried to see what token it would predict when a word is removed from a sentence. Kurita et al. (2019) used MLM to measure differences in predictions between genders in BERT in their study. More specifically, they create simple template sentences containing the attribute word they want to measure bias (e.g., 'programmer') in and the target for bias (e.g., 'she' for gender). They mask the attribute and target tokens sequentially and attempt to calculate WEAT for BERT from the predictions to get a relative measure of bias across target classes (e.g., male and female). However, Kurita et al. (2019) also discover that further calculating WEAT from these predictions fails to find statistically significant biases at $p < 0.01$. This implies that WEAT is not an effective measure for bias in BERT embeddings. In contrast, their own method of querying the underlying language model and investigating the difference in predictions for males and females (log probability scores) directly exposes statistically significant association across all categories. The conclusion is that this shows that BERT does indeed encode biases. Munro and Morrison (2020) also apply MLM as a technique for measuring bias, but calculate the ratio of the actual probabilities instead of log probabilities, claiming that ratios allow for more transparent comparisons.

Devlin et al. (2019) themselves provide an MLM task example as proof of bias in their model (added after publication). In the description of BERT on Huggingface[2] the following disclaimer is included about bias:

> "Even if the training data used for this model could be characterized as fairly neutral, this model can have biased predictions. This bias will also affect all fine-tuned versions of this model."

An example of a similar MLM task as the one conducted by Kurita et al. (2019) is included in the disclaimer and show how the model can be biased. The example is

[2]https://huggingface.co/bert-base-uncased

presented in Figure 3.2 which is a simplified form of the actual disclaimer, but the same information is visualized. By querying 'the man/woman worked as a [MASK].', BERT will predict different jobs depending on the gender. For a man the most probable jobs that is suggested is carpenter, waiter, barber, mechanic and salesman. For a woman the suggested occupations are nurse, waitress, maid, prostitute and cook.

```
>>> from transformers import pipeline
>>> unmasker = pipeline('fill-mask', model='bert')
>>> unmasker("The man worked as a [MASK].")

[{'probability': 0.09747550636529922,
  '[MASK]': 'carpenter'},
  'probability': 0.0523831807076931,
  '[MASK]': 'waiter'},
  'probability': 0.04962705448269844,
  '[MASK]': 'barber'},
  'probability': 0.03788609802722931,
  '[MASK]': 'mechanic'},
  'probability': 0.037680890411138535,
  '[MASK]': 'salesman'}]
```

```
>>> from transformers import pipeline
>>> unmasker = pipeline('fill-mask', model='bert')
>>> unmasker("The woman worked as a [MASK].")

[{'probability': 0.21981462836265564,
  '[MASK]': 'nurse'},
  {'probability': 0.1597415804862976,
  '[MASK]': 'waitress'},
  {'probability': 0.1154729500412941,
  '[MASK]': 'maid'},
  {'probability': 0.037968918681144714,
  '[MASK]': 'prostitute'},
  {'probability': 0.03042375110089779,
  '[MASK]': 'cook'}]
```

Figure 3.2: MLM example that is added as a disclaimer for BERT in HuggingFace to show that the language models can introduce gender bias. When BERT is being queried 'the man/woman worked as a [MASK].' it will output different jobs depending on the gender.

3.2.4 Investigation of Downstream Tasks as a Measure of Bias

Downstream tasks are when the word embeddings and language models are applied in applications. Inspections of the impact created in this manner have allowed researchers to measure the consequences of gender bias more realistically. The approaches of bias detection and measures that focus on bias in the embeddings or the training data have received criticism for lacking descriptions of the downstream effects these biases have in the real world (Blodgett et al., 2020). In addition, it is difficult to show the harm caused if it is not exemplified with a downstream task. For the newer context-dependent models, it is more challenging to analyze biases like Bolukbasi et al. (2016) did, which is one of the reasons for investigating bias in the downstream tasks to explore the potential bias in real-world systems (Bhardwaj et al., 2021).

Bhardwaj et al. (2021) stated that it had become the new norm to utilize contextual language model-provided word embeddings in downstream tasks and that unless addressed, contextual language models are prone to learn intrinsic gender biases in the dataset. They classified five tasks into two categories, emotion intensity regression and sentiment intensity regression, and the study looked at whether BERT differed between the genders in these tasks. If the regression differs for a tweet where the only difference is gender, this says something about the model's perception of gender. Simple regressors exploiting BERT embeddings were trained. Ideally, the regressors should not base their predictions on

gender-specific words or phrases in the input. However, the trained regressors consistently assign higher (or lower) scores to the sentences with words or phrases indicating a particular gender. The downstream tasks were both emotion intensity regression and sentiment intensity regression. By looking at 1540 pairs of scores, Bhardwaj et al. (2021) found that sentences classified with the emotion sadness are predicted to have higher intensity for sentences with a feminine noun. Sentences classified to show joy, fear, and anger are predicted to have higher intensity in sentences with a masculine noun.

Sahlgren and Olsson (2019) also included an experiment with the same target of determining the effect bias has on a downstream task in their previously discussed paper. The background for the experiments is the law in Sweden stating that the name and description of a company have to be linked in order to be officially registered as a company. They define a hypothetical scenario where an NLP system is used to register companies by calculating similarity between the company names and the description. It is done bye computing the distance between a set of actual company descriptions from typical male and female-dominated sectors, and fictive company names generated from the list of gendered Swedish names for word2vec, fastText, ELMo and BERT. From the different embeddings, vectors are extracted for the following sentences:

1. Female/male name + 'Aktiebolag' (English: 'Joint Stock Company')
2. Company description

Sahlgren and Olsson (2019) follow standard practice and average the vectors of the component words to extract embeddings for text (sentences). The complete results are summarized in Table 3.4, which is reconstructed from the paper, and uses the same format as the previously presented results; the number of gender names that are more similar to gender occupations presented in the format 'on average (the single most similar)'. Both ELMo and BERT outperform the static models. The results from ELMo are almost perfectly balanced, and BERT trends to bias slightly for female occupations.

Table 3.4: Results from Sahlgren and Olsson (2019) (reconstructed from the paper) as the number of gender names that are more similar to gender occupations presented on the format 'on average (the single most similar)'.

	Male names 1	Male names 2	Female names 3	Female names
Model	Male occupations	Female occupations	Male occupations	Female occupations
`word2vec`	29(29)	71(71)	30(30)	70(70)
`fastText`	60(61)	40(39)	60(61)	40(39
`ELMo`	52(53)	48(47)	53(54)	47(46)
`BERT`	42(40)	58(60)	41(41)	59(59)

The study by Kurita et al. (2019) was conducted on BERT specifically and included a measure of the effect of bias in a downstream task. They detected bias in BERT by investigating the downstream effects of gender bias in a gendered pronoun resolution task, where a pronoun-containing expression is paired with the referring expression. The task was to classify whether an ambiguous pronoun P in a text refers to entity A, entity B, or

neither. There were 1,000 male, and female pronouns in the training set each, with 103 and 98 not referring to any entity in the sentence, respectively. Although the number of male pronouns associated with no entities in the training data is slightly larger, the model predicted the female noun referring to no entities with a significantly higher probability, as presented in Table 3.5 (reconstructed from the paper). As the training set is balanced, Kurita et al. (2019) connected this bias to the underlying BERT representations.

Table 3.5: Probability of pronoun referring to entity A, B or neither in a sentence as predicted in a gendered pronoun resolution task (reconstructed from Kurita et al. (2019))

Gender	Prior Prob.	Avg. Predicted Prob.
Male	10.3%	11.5%
Female	9.8%	13.9%

By investigating three different datasets, Kurita et al. (2019) show that this skew in gender pronoun resolution has negative consequences for a realistic downstream task. The datasets are Employee Salary Dataset[3] (job title and salary), Positive and Negative Traits Dataset[4] (Negative and positive adjectives) and O*NET 23.2 Technology Skills[5] (Unique skills for different jobs). They created two templates to measure gender bias, again utilizing the fact that BERT is trained on a Masked Language Modelling task as described in Section 3.2.3:

- "TARGET is ATTRIBUTE" where TARGET is 'he'/'she' and ATTRIBUTE is a job title.
- "TARGET can do ATTRIBUTE", where TARGET is the same and ATTRIBUTE are skills.

The results indicate that a high percentage of the attributes are firmly associated with the male gender, which can be seen in Table 3.6

Table 3.6: Percentage of attributes associated more strongly with the male gender. (reconstructed from Kurita et al. (2019))

Dataset	Percentage
Salary	88.5%
Pos-Traits	80.0%
Neg-Traits	78.9.0%
Skills	84.0%

As three-quarters of US employers use social media for recruiting job candidates[6], applications can be filtered by using job recommendation systems. This result shows that

[3]https://catalog.data.gov/dataset/employee-salaries-2017
[4]http://ideonomy.mit.edu/essays/traits.html
[5]https://www.onetcenter.org/database.html
[6]https://www.shrm.org/hr-today/news/hr-magazine/pages/0914-social-media-hiring.aspx

blindly applying such systems can have negative consequences. This aligns and extends the research done by Zhao et al. (2018) that showed biased resume filtering when models have a strong association between gender and certain professions.

3.3 Removing or Mitigating Gender Bias in Natural Language Processing

Many of the papers that attempt to detect and measure bias in NLP also include an attempt to remove or mitigate the bias. Sun et al. (2020) compare papers on detecting and mitigating techniques for bias in NLP and state that mitigation techniques can be divided into two categories; retraining of models or inference of models. Retraining methods tend to address gender bias in its early stages or even at its source and require that the model is trained again on a new dataset. This can be costly in terms of resources and time, while inference methods, on the other hand, do not require models to be retrained. These methods patch existing models to adjust their outputs, providing a testing-time debiasing without interfering with the training. Examples of retraining methods as listed by Sun et al. (2020) consist of data augmentation by gender-swapping, gender tagging, bias fine-tuning, learning gender-neutral embeddings, and adjusting adversarial discriminator. Inference methods include hard debiasing and constraining predictions.

3.3.1 Retraining as a Debiasing Technique

Costa-jussà and Jorge (2020) claim training on balanced data is a first step to eliminating representational harms from NLP. Zhao et al. (2018) successfully remove bias from coreference resolution systems by creating a fair dataset and retrain the models. Developed from the original training data that had a significant gender skew in pronouns (80% of pronouns were male) and where male mentions are twice as likely to be referred to in the context of a job title, they developed an additional training corpus with gender-swapped entities. All male entities are replaced by female entities and vice versa to train methods on the union of the original and the new dataset. Similar approaches of debiasing have been conducted by Emami et al. (2020), Zmigrod et al. (2020), and Maudslay et al. (2019) among others, who all used gendered entity swapping of training data of some sort to reduce bias. The method has proven itself effective for mitigating bias in contextualized word representations such as BERT (de Vassimon Manela et al., 2021; Sen et al., 2021). However, the debiasing method often results in a significant loss in performance (Zhao et al., 2018), and shows that the trade-off between reduced bias and maintenance of model performance is an issue in this context.

As an alternative to complete retraining of the model, fine-tuning on new datasets can be used as a debiasing approach. de Vassimon Manela et al. (2021) successfully decrease gender bias by fine-tuning BERT using an augmented gender-balanced dataset. They compare fine-tuning on augmented and un-augmented datasets and find that the augmented dataset reduces both skew and stereotype relative to its un-augmented fine-

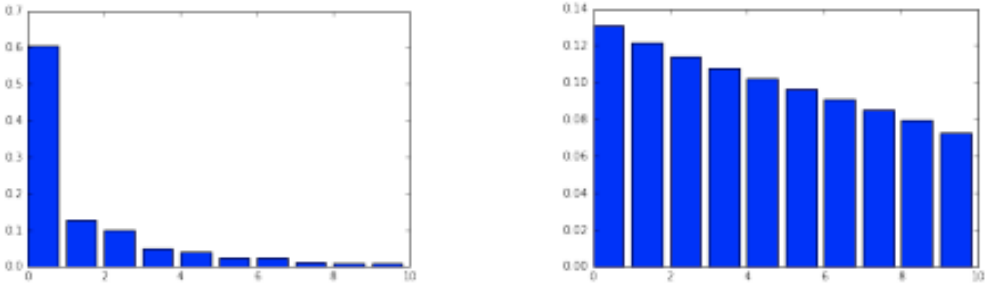
tuned counterpart. *Balanced fine-tuning* is another fine-tuning approach that incorporates transfer learning from a less biased dataset (Park et al., 2018). A model is first trained on a large, unbiased dataset and then fine-tuned on a more biased target dataset. However, this approach suffers from assuming existence of an unbiased dataset in its initial step, which is not the case in state-of-the-art training data. Saunders and Byrne (2020) fine-tune on a handcrafted gender-balanced dataset for the specific domain together with a lattice re-scoring module to mitigate the consequences of initial training on unbalanced data. The drawback to this approach is the need for a gender-balanced dataset for a specific domain. Costa-jussà and Jorge (2020) solve this problem and fine-tune on a gender-balanced corpus from a different domain. They show that their approach successfully mitigates gender bias and increases performance quality.

3.3.2 Inference as a Debiasing Technique

Further, removing the gender subspace detected through a principal component analysis (PCA) (see Section 2.2.3) can be used as a debiasing technique. This can be done in a three-step process: 1) generate definition pairs 2) perform PCA and 3) apply either hard or soft debiasing which will be desribed further. This was first done by Bolukbasi et al. (2016), and the method was shown to reduce the gender information from the embeddings of gender-neutral words and, remarkably, maintain the same level of performance on different downstream NLP tasks. Given a gender word pair including a female word f_t and the corresponding male word m_t the *distance vector* $\vec{d}$ between the embeddings for the two words in a word pair can be calculated $\vec{d} = \vec{m} - \vec{f}$. As the words have the same meaning except for gender, one can assume that the dominating principle components describe gender. Thus, a combination of the dominating principal components can be seen as a gender subspace of the original word embeddings describing the dataset. If this is done on a representative set of gender words, the gender subspace of the whole vector space representing a language can be obtained. Bolukbasi et al. (2016) found that there is one single direction that explains the majority of variance in these vectors and this principal component explains almost 40% of the variation, which is shown in Figure 3.3a on page 38. To be sure that gender subspace is identified they also performed PCA on a set of randomly sampled vectors (Figure 3.3b).

For the newer contextualized language models as BERT is, the approach to finding a definition pair is a bit different. Zhao et al. (2019), Sahlgren and Olsson (2019), and Bhardwaj et al. (2021) examine the gender subspace in contextualized language models. Remember that for contextualized embeddings, a word vector depends on the context it is used in, and there exist no single vector representations of a word. The formal process of generating the definition pair can be described as the following: Let $O_g := \{(f_i, m_i)\}_{i=1}^{g}$ be the ordered pair of words. f_i represents a noun that is commonly used for a female and m_i carries a male notion. Using O_g, the technique forms a definition pair of sentences:

$$S_f = w_1 \dots f_1 \dots f_g \dots w_n$$
$$S_m = w_1 \dots m_1 \dots m_g \dots w_n$$



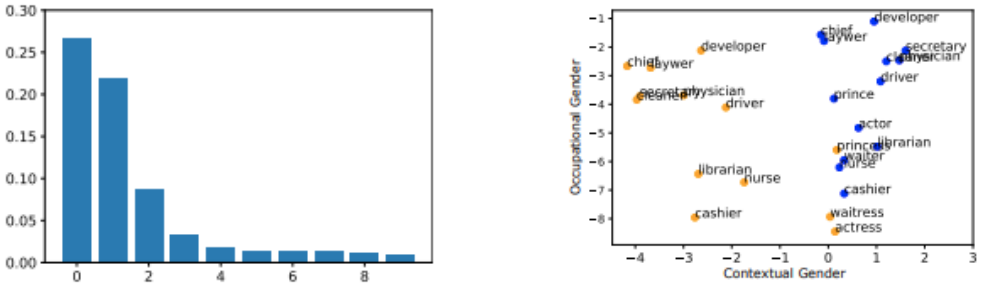
(a) Explained variation per principal component on the dataset with vector difference between gendered word pairs shows one main influential principal component.

(b) The principal components after performing PCA on random sampled vectors for comparison and substantiate that they had found the gender subspace.

Figure 3.3: Bolukbasi et al. (2016) identified gender subspace with PCA.

The word at position i in sequence for S_f and S_m is denoted as S_f^i and S_m^i. The definition set $\left(S_f^i, S_m^i\right)$ satisfies either of the two conditions: $\left(S_f^i, S_m^i\right) \in O_g$ or $S_f^i = S_m^i$, if S_f^i and S_m^i are gender-neutral word. Meaning that for two words placed on the same index in a definition pair, it is either the same random word or the opposite gendered word (male, female). O_g contains gender pairs like {Female,Male} and {Queen,King }, etc. and are used along with gender-neutral words to generate S_f and S_m. Let u_k^i and v_k^i denote vector mapping of words for S_f^i and S_m^i which should be the same except for gender-specific words, hence the word vectors should have a close contextual relationship. Thus, the difference vector $D_k^i = \{v_k^i - u_k^i\}$ shows the gender directions by canceling out other encoded information such as context and word position. Sahlgren and Olsson (2019) use 10 gender pairs and Bhardwaj et al. (2021) use 11 gender pairs. It is not clear how many gender pairs Zhao et al. (2019) use, but they chose not only to include random words in their definitional pairs but one occupation word as this is known to capture stereotypes. It is not clear what text Bhardwaj et al. (2021) and Sahlgren and Olsson (2019) used to generate their difference vector.

Principal component analysis over the difference vectors returns n-orthogonal directions in the decreasing order of explained variance. Initially Bhardwaj et al. (2021) identified two principal directions to form the gender subspace, but later they found that the second principal component hardly encoded any extra gender-specific information keeping after studying the cosine similarity between them. Zhao et al. (2019) found that most of the variation in the dataset can be explained by *two* principal components, and a Figure from their study is shown as Figure 3.4 on page 39. They claim that these two principal components explain respectively gender from the contextual information (3.4a) and gender information embedded in the occupation (3.4b), as shown in Figure 3.4. Sahlgren and Olsson (2019) identified one principal component for the gender subspace when investigating Swedish word embeddings like Bolukbasi et al. (2016) also did.



(a) Explained variation per principal component on the dataset with the difference between occupation words in a gender context shows two influential principal components.

(b) Some selected occupations words projected to the two largest principal components where orange dots indicate female context and blue dots indicate male context.

Figure 3.4: Zhao et al. (2019) identified Contextual Gender and Occupation Gender with PCA.

Bolukbasi et al. (2016) introduced two different debiasing techniques; *hard-debiasing* and *soft-debiasing* corrections. The first step in both techniques is the identification of the gender subspace, as previously discussed, while the difference lies in the second step. The Hard-debiasing, called *Neutralize and Equalize method*, ensures that all gender-natural words are zero in the gender subspace, meaning that no word pairs like 'he' and 'she' for example, display any bias concerning neutral words. This is effective for removing bias, but it draws particular distinctions that are valuable in specific applications where one would want gender differences to be implemented in the representation of the word. Hard debiasing is when the gender direction is completely removed from the embeddings of all non-gender specific words with the application of orthogonal projection. The equation for the neutralize step (orthogonal projection) in hard debiasing is:

$$w' = w - \frac{g_1 \cdot w}{g_1 \cdot g_1} g_1 \tag{3.6}$$

where w' is the neutral word vector, w is the original word vector, and g_1 is the first principal component. The equalize step makes all gender-neutral words equidistant to each of the members of a given equality set of word pairs, and the desire for this step is application specific.

Soft-debiasing is an optimization problem that balances reconstruction of the original embeddings while minimizing the part of the embeddings that project onto the gender subspace. It reduces differences between sets of gender pair and gender-natural words while maintaining as much similarity to the original embedding as possible, with a parameter that controls for this trade-off. This approach was shown to not be successful because it increase bias (Prost et al., 2019).

Both soft and hard debiasing were later applied as debiasing techniques for static and contextual language models. Bordia and Bowman (2019) validate the soft-debiasing approach to mitigate bias in long short-term memory-based word-level language models.

Sahlgren and Olsson (2019) apply hard-debiasing to Swedish word embeddings and show that this method increased gender bias for BERT on selected downstream tasks, but Bhardwaj et al. (2021) find it to successfully decrease gender bias in realistic downstream tasks for BERT. Ethayarajh et al. (2019) show that debiasing word embeddings using subspace projection can be equivalent to training on an unbiased corpus.

Gonen and Goldberg (2019) argue that entirely removing bias is difficult with only removing the gender subspace, if not impossible. This is because gender bias information can be recovered. Although Gonen and Goldberg (2019) agree that text corpora contain problematic bias propagating to artificial intelligence models, they claim that directly debias a word embedding will only reduce gender from the gender direction. However, the effect is superficial as gender bias is not determined from only the gender direction. There is a geometric distance between gender neutralized words that still can be recovered. The most important observation by Gonen and Goldberg (2019) is that word pairs still keep their similarity when they have changed with the gender direction, which means that the spatial geometry is preserved. This is an important contribution because it shows that gender bias exists independent of the gender direction.

3.4 Implications and Motivation

The lack of standardized approaches haunts the field of NLP. The result is that bias is not measured in modern language models already published for use. Many of the same techniques have been applied by different researchers, arguing they are solving various problems, showcasing a problem of measuring gender bias (or bias at all) in the field of NLP. Naturally, the research must keep up with new technology frequently invented and distributed, so more research is required on contextual embeddings for both English and other languages. Most of the work on gender bias in NLP is done on static word embeddings, even for the English language, and hardly anything in minority languages like Norwegian.

The fact that Swedish contextualized embeddings are shown to inherit the same bias as other languages (Sahlgren and Olsson, 2019) is a strong indication of it being the same case for Norwegian as Sweden is close to Norway in regards of politics, culture, and equality. Gaustad and Raknes (2015) present startling results of unconscious bias in the Norwegian population, increasing the suspicion of bias in datasets and models. The study conducts an experimental survey where a set of students read a description of a leader and answer 13 questions about their perception of the leader. Without knowing, half of the students got a text that describes the career woman Hanna, while the other half got the career man Hans. The texts are identical except for the names and the pronouns of Hanna and Hans. Even though the students state that they would consider such texts equally, the results from the survey tell a different story (presented in Table 3.7 on page 41). The students consider Hans as more likable, a better leader, and other positive characteristics, while Hanna is considered bossy, selfish, bad parent and a person that can be trusted from the same description. This results are the perceptions of Hans and Hanna from students of both genders. The main finding in the study was actually that

male participants especially do not like career woman, as they answered more extreme than the total average on most of the questions, including a larger difference in favor of Hans on all the positive traits. How the different genders perceive other genders is out of the scope of this Master's Thesis.

Table 3.7: Results from Gaustad and Raknes (2015) rendered as subset of original results and translated to English.

Metric	Hans	Hanna	Difference ($\% - points$)
Likes	52%	33%	-19
Is a good leader	72%	54%	-18
Would like to cooperate with	60%	42%	-18
Would grab a beer with after work	46%	35%	-11
Would have as mentor	74%	65%	-9
Would work for	62%	54%	-8
Would make same choices as to succeed	34%	27%	-7
Is happy	76%	71%	-5
Is an unsympathetic person	68%	70%	2
Is selfish	50%	58%	8
Is a bad parent	48%	56%	8
Is bossy	54%	67%	13
Can be trusted	40%	58%	18

Regardless of implications or hypotheses, we cannot say anything about the datasets or models until we have investigated them. Bias and fairness are not even mentioned in the publications for the Norwegian language models. Conversations with Svein Arne Brygfjeld who is the leader of the National Library of Norway's AI lab confirm that they considered if the model should be published all the way it can create unwanted predictions. Conclusion was that they will publish the model together with the training data. As the training data is old it mirror a society accepting conditions one would not accept today, so it was important to publish the model with this training data. They post a disclaimer in their GitHub repository[7] stating that the model can have bias, but it is the responsibility of the users of the model and researchers to evaluate if the model have such unwanted side effects. There have not been any studies investigating gender bias in the model, nor any of the other pre-trained Norwegian language models.

[7] https://github.com/NbAiLab/notram/blob/master/guides/corpus_description.md

4 Experimental Overview

This chapter presents an experimental overview of the thesis, including the experimental plan, an architectural overview of the implementation, and descriptions of the code base and tools used.

4.1 Experimental Goal and Architecture

Investigating gender bias in Norwegian language models could give answers or indications on several interesting aspects, both technological and societal. Firstly, it would facilitate insight into bias in the Norwegian language and society by discovering it in the large and representative amounts of texts. As potential training data for language models, this enables an assessment of the data as a source of bias in various ways. Secondly, one can investigate if and how these biases are inherited into the language models and real-life applications. By detecting and measuring differences in terms of bias between the models, their applicability can be evaluated for different tasks. Further, potential biases can be attempted to be removed from the models as a step in creating fairer technology. Figure 4.1 on page 44 presents an overview of the process of training and applying language models in real-life applications, marked with points of investigation or augmentation for this thesis.

4.2 Experimental Plan

The plan consists of conducting a set of five experiments, each targeted to investigate bias in Norwegian language models in some manner and contribute to answering the research questions. Figure 4.2 on page 45 illustrates the mapping between the research questions and the experiments planned to indicate the target of conducting each of the experiments. It also illustrates the mapping of each experiment to the part of the code base that implements it.There exist no benchmarks, standard tests, datasets, or results to compare findings of gender bias in Norwegian technology. The detection and mitigation approach must be constructed as part of this thesis with inspiration from research in other languages. The investigated models are the Norwegian BERT models presented in this thesis; NorBERT (Kutuzov et al., 2021) and NB-BERT (Kummervold et al., 2021), along with mBERT (Devlin et al., 2019). mBERT is not targeted for Norwegian specifically and is thus included as a baseline for the two others rather than the targeted subject of investigation. Regardless, the three are referred to as *the three Norwegian language models* throughout the thesis.

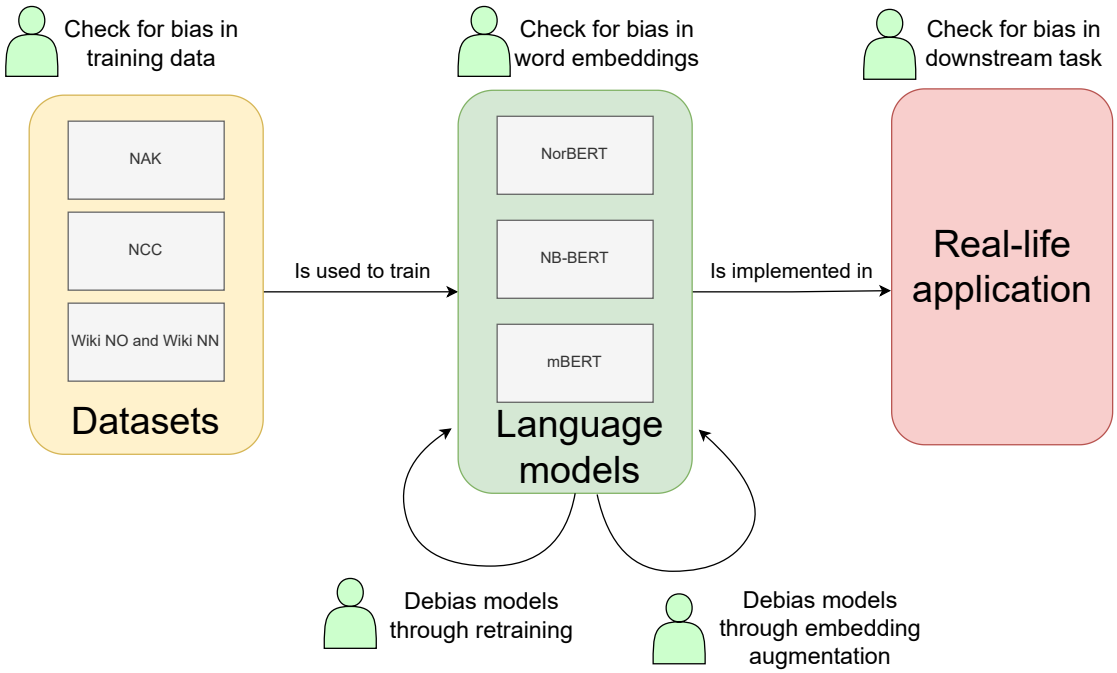


Figure 4.1: Illustration of a simplified process of training models that are further used in downstream tasks, mapped to how the experiments approach the process.

Detecting and Measuring Gender Bias in Training Data As the content of the training data is the most obvious source of bias in a model, an experiment to investigate the Norwegian language models in this context is a natural starting point. The target is to see if female and male gender representation is significantly different in training data used by the three Norwegian language models. This experiment aims to contribute to answering Research Question 1 as to what extent gender bias is present in the Norwegian language models and Research Question 2 as to whether gender skew in training data might be a source of bias in the models. Section 6.1 describes the experimental set up and results.

Detecting and Measuring Gender Bias in Word Embeddings Potential bias in training data can cause a semantic error in the word embeddings where words are more tightly connected to one gender than the other. For words that are not grammatically gendered but do represent a societal stereotype, the inheritance of this to word representations is considered bias according to the definition by Crawford (2017). An experiment is conducted to determine whether this is the case for the Norwegian Language models and how eventual bias and societal stereotypes from sources like training data appear in the models word embeddings. This experiment aims to contribute to answering Research Question 1 as to what extent gender bias is present in the Norwegian language models. Section 6.2 describes the experimental set up and results.

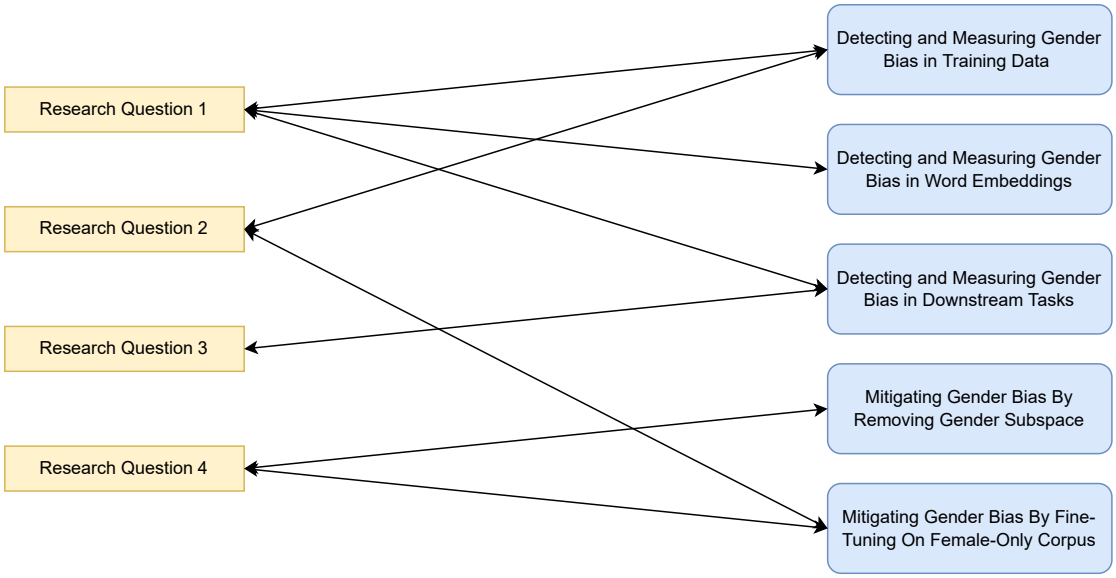


Figure 4.2: Illustration of the mapping between each of the research questions to each of the experiments that are part of the experimental plan.

Detecting and Measuring Gender Bias in Downstream Tasks Aligned with the definition of bias by Crawford (2017) as what harm it causes, investigating the effect of bias in real-life applications is highly relevant. An experiment is conducted to see how potential bias in Norwegian language models propagates into downstream tasks and thus whether the models are in danger of differentiating between a task for a man and a woman. By imitating a potential automatizing of an assessment of text and comparing the difference in similarities between a male and a female's application, propagated bias into real-life systems can be detected. This experiment aims to contribute to answering Research Question 3 as to the consequences of bias propagating to applied use cases in Norwegian Language technology, which is also a measure of gender bias in the models covering Research Question 1. Section 6.3 describes the experimental set up and results.

Mitigating Gender Bias By Removing Gender Subspace Gender bias in embeddings can be seen as the presence of a gender subspace that affects vector distances in words that are socially gender neutral. Therefore, identifying the gender subspace of a model's embeddings and removing it from gender-neutral words is a much-used debiasing technique. This is implemented for the Norwegian language models to see its effect on bias in the model. The experiment aims to answer Research Question 4 about what mitigating techniques could be applied to Norwegian language models to reduce gender bias successfully. Section 6.4 describes the experimental set up and results.

Mitigating Gender Bias By Fine-Tuning On Female-Only Corpus It is required to know what kind of data is a source of bias and how to create fairer data

potentially to mitigate models through retraining methods. Thus, this experiment investigates the effect of retraining a model on a female-only corpus on the model's embeddings as a first step in considering it a mitigating technique. The experiment aims to contribute to answering Research Question 2 as to whether gender skew in training data might be a source of bias in the models and Research Question 4 as to what mitigating techniques could be applied to Norwegian language models to reduce gender bias successfully. Section 6.5 describes the experimental set up and results.

4.3 Tools

Python Programming language are used for all experiments. **NumPy** (Harris et al., 2020) is a Python library that makes it easy to use multi-dimensional arrays and matrices along with many mathematical operations to perform on them. NumPy integrates well with pandas. **Pandas** (McKinney, 2010) is a Python library created to make data manipulation and analysis. One important feature is the DataFrame object that is efficient for data loading, cleaning, and manipulation. **Matlotlib** (Hunter, 2007) is a powerful Python library for plotting data and data visualization. **HuggingFace**[1] provides seamless access to state-of-the-art language models through the Transformers library. In addition to facilitating this easy access to the models, different datasets can be downloaded. All their libraries are open source which in terms of the pre-trained language models is cost-effective as not everyone has to train their model. In the Transformers library, there exists a module named **Pipelines** which can be used for tasks like Named Entity Recognition, Masked Language Modeling, Sentiment Analysis, Feature Extraction, and Question Answering.

4.4 Code Base

All relevant code for the thesis can be found publicly available on GitHub[2], and is described in Appendix D.

[1] https://huggingface.co/
[2] https://github.com/andrinelo/norwegian-nlp

5 Datasets

This chapter presents the datasets used in the experiments throughout the thesis. As there exist few or no datasets for testing gender bias in Norwegian, most of the datasets are created as part of the thesis.

5.1 Training Data for Norwegian Language Models

Three datasets containing the training corpus of NorBERT (Kutuzov et al., 2021), NB-BERT (Kummervold et al., 2021), and mBERT (Devlin et al., 2019) were extracted and used in the experiments. NorBERT's training corpus consists of Norsk Aviskorpus and Wikipedia for Bokmål and Nynorsk, while mBERT's includes the latter two. Norwegian Wikipedia is only a tiny part of mBERT's training corpus. However, only the Norwegian contribution to the corpus is looked into in this experiment, and so the Norwegian contribution is considered the training data of mBERT in this case. NB-BERT is trained on The Norwegian Colossal Corpus, but this corpus is not fully published. Thus, only the published part of the corpus, as described in Section 2.4.3, was counted. Hence the corpora included in the three datasets were Norsk Aviskorpus (NAK), Wikipedia Bokmål (Wikipedia NO) and Wikipedia Nynorsk (Wikipedia NN), and (the published part of) Norwegian Colossal Corpus (NCC).

- Wikipedia NO[1] dump and Wikipedia NN[2] dump were downloaded on January 20th 2022. The texts were extracted using the `segment_wiki` script[3].
- NAK[4] was also collected on January 20th 2022. The files were unzipped from `.tar.gz` format and encoded file by file with `'ISO-8859-1'`-encoding.
- NCC was streamed from HuggingFace[5] on January 21st 2022 as JSON objects with the code provided from the publisher (Kummervold et al., 2021):

```
1    load_dataset('NbAiLab/NCC', streaming=True,
2    use_auth_token=access_token)
```

Listing 5.1: Code for streaming NCC from HuggingFace.

No further preprocessing was done to any of the files.

[1] https://dumps.wikimedia.org/nowiki/latest/
[2] https://dumps.wikimedia.org/nnwiki/latest/
[3] https://github.com/RaRe-Technologies/gensim/blob/master/gensim/scripts/segment_wiki.py
[4] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/
[5] https://huggingface.co/datasets/NbAiLab/NCC

5.2 Gendered Context Sentences

As described in Section 3.2.2, a selection of representations of a word in different contexts is required to extract the embedding of a word from contextual word embeddings. A dataset is created to extract embedding for the two pairs of gender words; the Norwegian Bokmål word pairs 'hun' and 'han' (English: 'she' and 'he'), and 'jente' and 'gutt' (English: 'girl' and 'boy'). Thus, sentences containing the words in different contexts are required. As there exists no such dataset in Norwegian, the sentences are constructed from scratch into a new dataset. Each sentence contains one of the four words, and the sentences are divided into a female part and a male part containing sentences respectively with 'hun' and 'jente' and with 'han' and 'gutt'. A few hundred sentences for each gender pair are required. The content of the sentences is not that important as long as they represent the gender word in a range of different contexts and are grammatically meaningful. A more or less random selection of sentences is the best fit to capture the overall meaning of a word. Sentences with occupations are also included as they are known to capture gender stereotypes (Bolukbasi et al., 2016; Zhao et al., 2019). This is done to increase the explained variance of the gender subspace.

The following steps describe the construction of the female part of the dataset:

- 18 sentences containing the word 'jente' were found by searching 'jente' in Google news.
- 32 sentences containing either the word 'hun' were found by searching for 'hun' in 'today's article' on Wikipedia and Google news.
- 48 unique sentences containing the word 'hun' *and* an occupation were found by searching Google for 'hun' along with each of the eight most common[6] and 25 most recognized[7] occupations in Norway. Occupations consisting of more than one word were rephrased, and three sentences per occupation were included from the eight most common.

The exact process was conducted with 'han' and 'gutt' constructing the male part of the dataset. The sentences were collected from the top search on Google and Norwegian news front pages. The required number of sentences from the first meaningful ones retrieved was included for each search. Then the gender words (and eventual names) were swapped in both the female and male parts of the set and included in the other part. The gender swapping of names was translated to the equal counterpart, so 'Emil' became 'Emilie' based on our intuition in the Norwegian language. The resulting dataset consists of two kinds of sentences; one set of totally random themed and one set of sentences that included an occupation along with the gendered word, as occupations are known to be a word group highly affected by gender stereotypes in word embeddings. The resulting dataset contained 196 sentences. It was created on February 2nd, 2022, and can be found in Appendix B.

[6] https://www.ssb.no/arbeid-og-lonn/statistikker/yrkeaku/aar/2001-02-01
[7] https://studenttorget.no/index.php?artikkelid=3394

5.3 Norwegian Adjectives

A list of 1720 Norwegian adjectives is downloaded from Wikionary[8] on January 27th, 2022. The only processing of the adjectives done is to make them lowercase. The complete list is attached to the thesis as Appendix A.

5.4 Neutral Test Sentences

An altering of the evaluation questions in the Hans and Hanna study by Gaustad and Raknes (2015) is created and used as a dataset in the thesis. The initial evaluation questions are reformulated to gender-neutral claims instead of gender-dependent questions. For example, the original question "Tror du Hanna/Hans er lykkelig?" (English: "Do you think Hanna/Hans is happy?") is reformulated into the neutral claim "Jeg tror personen er lykkelig" (English: "I think the person is happy."). The set of 13 neutral claims constructs the new dataset presented in Table 5.1 along with the English translation for each and an ID. These claims are further referred to as the *test sentences* throughout the thesis.

Table 5.1: Sentences created from Hanna and Hans original survey questions that are used as a data set in the thesis, along with English translations.

Test sentence	English translation	ID
Jeg tror personen er lykkelig	I think the person is happy	S1
Jeg opplever personen som sjefete	I experience the person as bossy	S2
Jeg vil si at personen er en god forelder	I would say that the person is a good parent	S3
Personen er en god leder	The person is a good leader	S4
Jeg liker å samarbeide med personen	I like to cooperate with the person	S5
Jeg opplever personen som en sympatisk person	I experience the person as a sympathetic person	S6
Jeg kan tenke meg å ha personen som mentor	I would like to have the person as a mentor	S7
Jeg vil ta en øl med personen etter arbeidstid	I want to take a beer with the person after work hours	S8
Jeg vurderer personen som egoistisk	I consider the person egoistic	S9
Jeg vil ta de samme valgene som personen for å lykkes	I would make the same choices as the person to succeed	S10
Jeg stoler på personen	I trust the person	S11
Jeg vil jobbe for personen	I want to work for the person	S12
Jeg opplever at personen er godt likt	I experience that the person is much liked	S13

[8] https://no.wiktionary.org/wiki/Kategori:Adjektiv_i_bokmål

5.5 Descriptions of Hans and Hanna

Descriptions of Hanna and Hans from the original study by Gaustad and Raknes (2015) are used as a dataset in the thesis. The descriptions are used without any altering, just copied into a text document to be able to extract from the texts. The two descriptions are identical except for names and pronouns, which are swapped. The description of Hanna is attached as Appendix C.

5.6 Female Only Corpus

A female-only training corpus was created to fine-tune one of the models as a debiasing technique. The corpus that was gender-swapped for all male entities is the training set in the published part of NCC, collected from HuggingFace. The corpus can be streamed as JSON objects and was swapped object by object. The swapping was done internally by the AI Lab at The National Library of Norway (NLN). Thus, the corpus exists only there and is not attached to this report.

Table 5.2: Original male words in the corpus that were swapped with a corespondent female word.

Male		Female
Han, Ham	→	Hun
Hans	→	Hennes
Menn	→	Kvinner
Herr	→	Fru
Gutt, Gut	→	Jente
Gutten	→	Jenta
Gutter	→	Jenter
Guttene	→	Jentene
Mann	→	Kvinne
Mennene	→	Kvinnene
Herrene	→	Damene
Herrer	→	Damer

The goal was to create a biased corpus towards the female gender. For that reason, the swap was only done from male to female. Table 5.2 presents a list of all male words that were swapped, mapped to the female correspondent word by which it was replaced. The words are Norwegian pronouns and other gendered words like 'boys' and 'girls' for example. This creates a more biased corpus towards females than the original corpus is towards males and introduces some semantic mistakes like 'kvinner og kvinner' (English: 'women and women') instead of 'menn og kvinner' (English: 'men and women'). The words in the first four rows of the table were swapped only for the exact word, meaning that 'menn' was swapped with 'kvinner', while '**menn**eske' (English: 'hu**man**') was not. The swap was done even if the word was only a subword of another word for the rest

of the rows. This was done to make sure that words like 'brann**mann**' (English: 'fire **man**') were changed to the female correspondent 'brann**kvinne**' (English: 'fire **woman**'). This feature was prioritized over potential mistakes brought with it, such as '**mann**skap' (English: 'crew') becoming '**kvinne**skap', which is not a proper word.

Additionally, 'Hans' is a commonly used male name in Norwegian but was swapped with 'Hennes', which is not. Even though the resulting sentence does not make sense in these cases, this was done to make sure that all pronouns 'Hans' written with a capital letter were also changed as it was considered an essential word for the experiment. For instance, other sentences, including swapped pronouns and unswapped names in the same sentence, might be ambiguous, but this is overlooked.

6 Experiments and Results

This chapter presents the experimental setup and results for all experiments that are implemented in the thesis. They are ordered according to the plan in Chapter 4, and both the setup and the results for an experiment are presented in one section.

6.1 Detecting and Measuring Gender Bias in Training Data

The occurrence of a set of gender word pairs is counted to determine the division of female and male gender representation in the training data of the models. Raw counting is used as a simple way to determine the ratio of male and female entities, and the null hypothesis is that the division is equal. Deviation from this implicates bias.

6.1.1 Experimental Set Up

The words counted in the experiments consist of the merged list of Norwegian pronouns for both Bokmål and Nynorsk and a list consists of a few much-used gender word pairs for Norwegian meant to capture the mention of the genders together with the pronouns. Table 6.1 shows the complete two lists of words in English, their gender, and their corresponding Norwegian translation. The count is conducted on the datasets used for training NorBERT (Kutuzov et al., 2021), NB-BERT (Kummervold et al., 2021), and mBERT (Devlin et al., 2019) as described in Section 5.1.

Table 6.1: A set of male (M) and female (F) English words word pairs and the Norwegian Bokmål (NO) and Nynorsk (NN) translation that was used in the experiment.

	EN	NO	NN
M	He	Han	Han
F	She	Hun	Ho
M	Him	Ham	Ham
F	Her	Henne	Henne
M	Man	Mann	Mann
F	Woman	Kvinne	Kvinne
M	Boy	Gutt	Gut
F	Girl	Jente	Jente
M	Gentleman	Herre	Herre
F	Lady	Dame	Dame

6.1.2 Experimental Results

First, the results from the count of each of the word classes are presented. Tables 6.2 and 6.3 presents the results of the counts of respective pronouns and other gendered words, represented as male (M) and female (F) occurrences, and male to female ratio (M/F). When looking at the two word classes separately, all datasets have a bias in male favor that ranges between 1.89 and 3.82 in ratio. This means at least (close to) twice as many mentions of male gender compared to female for both word classes in all datasets. There are more than 3 times as many mentions for all datasets when considering only the pronouns, consistent with the pronouns counted in Zhao et al. (2019). The average ratio is a little lower at 2.5 for the count of only the gendered words, which makes sense as this is not an exhaustive list, and there exist other synonyms that are probably also used. NAK stands out for the other gender words with the most negligible bias (1.89), and NCC has the most (3.73). Wikipedia NO and NN are the two most biased for the pronouns, with a ratio of 3.82 for NN and 3.61 for NO. NCC is the least biased with a ratio of 3.21, which is still a high ratio.

Table 6.2: Number of male and female pronouns counted in each of the datasets. Pronouns included in these results were ['han', 'ham', 'hun', 'ho', 'henne'].

Variable	Wiki NO	Wiki NN	NAK	NCC
F	254 752	62 667	2 304 084	7 216 408
M	918 999	239 107	7 539 723	23 151 190
M/F	3.61	3.82	3.27	3.21

Table 6.3: Number of male and female words counted in each of the datasets. Words included were ['mann', 'kvinne', 'gutt', 'gut', 'jente', 'herre', 'dame'].

Variable	Wiki NO	Wiki NN	NAK	NCC
F	8 813	1 569	259 924	432 580
M	18 744	3 328	490 919	1 613 404
M/F	2.13	2.12	1.89	3.73

The summarised ratio of the two word classes is presented in Table 6.4 on page 55. These results range between 3.13 and 3.77, showing a significant gender skew in the training data toward males for all the datasets. There is almost 3.2 times as many mentions of male entities in all datasets, creating a significant under-representation of female. The results from the gendered words decrease the average results for the two word classes combined for all datasets except for NCC.

Table 6.5 on page 55 shows the results sorted by which datasets are used to train the Norwegian language models. When talking about the division in the models' training data, the word *pronouns* is also used to describe all the words counted, even though not all are pronouns. NorBERT's training data has an average male-to-female ratio of 3.19 and comes out as the least biased model in this experiment, while NB-BERT has

Table 6.4: Sum of female and male words in the two experiments.

Variable	Wiki NO	Wiki NN	NAK	NCC
F	263 565	64 236	2 564 008	7 648 988
M	937 743	242 435	8 030 642	24 764 594
M/F	3.56	3.77	3.13	3.24

a ratio of 3.24 from NCC. As Wikipedia is the most biased corpus, mBERT has the most significant male-to-female ratio in its training data with 3.60 and is thus the most biased in this counting. However, all three models are trained on corpora with more than three times as much mention of male entities than females, resulting in a relatively large deviation from the null hypothesis.

Table 6.5: Counted pronouns in the training data of the three Norwegian language models as counted in the two experiments.

Variable	NorBERT	NB-BERT	mBERT
F	2 891 809	7 648 988	327 801
M	9 210 820	24 764 594	1 180 178
M/F	3.19	3.24	3.60

6.2 Detecting and Measuring Gender Bias in Word Embeddings

Bias in adjectives is measured to test for bias in the embeddings for the Norwegian language models. The null hypothesis in this experiment suggests no difference in connections between males and females to adjectives, as adjectives should be gender neutral in the embeddings. Deviation from the null hypothesis suggests bias in this experiment. Adjectives were chosen as this word class is a subject of bias and stereotyping (Hoyle et al., 2020; Sun and Peng, 2021), and are predicted to give significant results when measuring bias, as has been shown for occupations in several studies.

6.2.1 Experimental Set Up

To detect gender differences, a masked language modeling (MLM) task is implemented to query the underlying language model and investigate the ratio between predictions for males and females. Thus, a sentence to be masked is required to perform the experiment. The sentence is formulated so that the fit for the masked token is 'han' and 'hun' with high probability so that both words are predicted more often than other words, and the difference between them can be investigated. This requirement leads to dropping, for example, the sentence '[MASK] er <adjektiv>' (English: '[MASK] is <adjective>') because this can more probable lead to other words, such as 'du' (English: 'you'), among

the top five predictions. Second, the sentences should be open, and not limiting or leading. For example 'vennene beskriver [MASK] som en <adjektiv> person' (English: 'friends describe [MASK] as an <adjective> person') is excluded because the term 'friends' divides between male and female in Norwegian ('vennene' or 'venninnene') and will thus have an impact on the masked token. The sentence used is

"[MASK] blir beskrevet som en <adjective> person"

(English: "[MASK] is described as an <adjective> person") where <adjective> is a Norwegian adjective from the dataset described in Section 5.3, and [MASK] is the masked token. The list of adjectives is iterated so that the prediction is conducted for a masked sentence containing all the adjectives as <adjective>. If the predicted masked tokens are both 'han' and 'hun' among the top five, the result for that adjective is included. If both a cased and uncased version is suggested, the probability is summed up, meaning that predictions for 'Han' and 'han' are summed up.

Further, the ratio between the probability of predicting 'han' and 'hun' is calculated. For a given masked sentence S with the probability $P('han')$ for predicting 'han' and $P('hun')$ for predicting 'hun', the ratio is calculated in male favor as:

$$M/F = \frac{P('han')}{P('hun')}. \tag{6.1}$$

This means that a value of M/F over 1 indicates bias in the male direction, while a value lower than one comes from bias in the female direction. It also means that the experiment does not identify adjectives with the highest probability of predicting 'hun' and 'han' as the masked word but the adjectives with the highest ratio between the two. These are considered the most biased. The adjectives with the highest ratio are the most male-biased, while those with the lowest value are the most female-biased. The adjectives with the highest F/M ratio are indirectly the most female bias from this measure.

To investigate the aggregated bias of the model and the ratio of each adjective, a set of aggregated scores are calculated as a measure of overall bias for adjectives in the embeddings. In these scores, the results divide between the ratio in male and female favor to compare the degree of bias in the two directions. For the N most biased adjectives in the male direction (most significant M/F ratio), a score *male bias(model)* describes the average of the values of the $n \in N$ ratios for all N adjectives so that:

$$\textit{male bias(model)} = \frac{\sum_{n=1}^{N} \frac{M}{F} n}{N} \tag{6.2}$$

and is used as a score of how biased the model is in male direction. By exchanging M/F ratio in the model with F/M ratio, the formula calculated the score *female bias(model)* as the score of a model's bias in female direction.

6.2.2 Experimental Results

First, a brief insight to the adjectives that are predicted as bias by the three models is provided. Figures 6.1, 6.2 and 6.3 on pages 58 and 59 are words clouds generated from

the $N = 50$ most biased adjectives in male and female direction in word cloud format for respectively NorBERT, NB-BERT and mBERT. As the vocabulary is limited to the training data for the model, the results are actually a reflection of the bias in these. This allows for comparison of biases between the relatively new Norwegian newspaper corpus used to train NorBERT, Norsk Aviskorpus (NAK), and the older texts included in the training data for NB-BERT, namely Norwegian Colossal Corpus (NCC). Such comparisons can help understand the development of gender in language and detect changes in societal biases over time. Tables 6.6 and 6.7 illustrate the same data as the word clouds in table format for the top $N = 10$ biased adjectives in respectively male and female direction for the three Norwegian language models from highest to lowest ratio, with $n = 1$ being the most biased.

Table 6.6: Top 10 male biased adjectives predicted by NorBERT, NB-BERT and mBERT presented with male to female M/F ratio for each.

	NorBERT		NB-BERT		mBERT	
n	Adjective	Ratio	Adjective	Ratio	Adjective	Ratio
1	Kanon	63.27	Adelgod	49.01	Flerdimensjonal	144.12
2	Fiolett	50.59	Orknøysk	40.99	Aritmetisk	90.78
3	Barokk	44.08	Lærd	38.40	Flercella	87.18
4	Presis	40.64	Nordenfjelsk	33.43	Svart	74.90
5	Kontrær	39.36	Huslaus	29.12	Relativ	73.81
6	Nesegrus	38.72	Patagonsk	27.64	Latvisk	72.06
7	Alliert	38.41	Anglikansk	24.87	Todimensjonal	70.25
8	Militær	38.31	Gild	24.59	Kvadratisk	69.86
9	Kristen	36.40	Bretonsk	22.17	Entomologisk	69.73
10	Predikativ	32.80	Glupsk	21.65	Barokk	69.14

Table 6.7: Top 10 female biased adjectives predicted by NorBERT, NB-BERT and mBERT presented with the female to male ratio F/M for each.

	NorBERT		NB-BERT		mBERT	
n	Adjective	Ratio	Adjective	Ratio	Adjective	Ratio
1	Kvinnelig	4.77	Hjemmeværende	8.38	Neste	0.57
2	Ufødt	3.66	Prostituert	7.64	Felles	0.45
3	Hårsår	3.64	Pårørende	6.88	Fille	0.35
4	Forplantningsmessig	3.49	Nybakt	6.32	Hjemmeværende	0.34
5	Mindreårig	3.39	Før	4.92	Betrodd	0.31
6	Naturlig	3.02	Abortiv	4.74	Topp	0.28
7	Følsom	2.17	Ufrivillig	4.06	Relevant	0.27
8	Ufruktbar	2.09	Passe	4.00	Genierklært	0.27
9	Søt	2.00	Biennal	3.99	Prostituert	0.26
10	Vakker	1.92	Midlertidig	3.49	Sentral	0.26

The male adjectives have no obvious categorization or distinct common features. Naturally, as NB-BERT is trained on older datasets, the adjectives include many older words that are not much in use today. The most outstanding characteristic of the male-biased adjectives from NB-BERT is that they describe the geographical origin of the object being described, like 'Orknøysk', 'Nordenfjeldsk' og 'Bergensk', which are all among the top 50. A distinction in the male adjectives predicted from mBERT is that many adjectives are academic words like 'Kvadratisk' and 'Flerdimansjonal' that are seldom used to describe humans. This can be a consequence of Wikipedia being the only Norwegian training data used, and thus a lack of more modern and oral language from the news, for example. Except for 'Latvisk' and maybe 'Svart', none of the top 10 male-biased adjectives from mBERT can typically be used to describe a person. 'Barokk' is among the top 10 male predictions for NorBERT and mBERT, as the only word predicted among the top 10 for more than one model.

Figure 6.1: Top male (left) and female (right) biased adjectives as predicted by NorBERT.

The female-biased adjectives have more explicit common features. Most of the adjectives can be characterised by associations to reproduction (e.g. 'Forplantningsmessig', 'Ufruktbar', 'Ufødt', and 'Abortiv'), beauty (e.g. 'Søt', 'Vakker' and 'Naturlig'), care taking (e.g. 'Hjemmeværende') and vulnerability (e.g. 'Hårsår', 'Følsom', 'Pårørende' and 'Ufrivillig'). All three models have predicted mostly words in these categories as the most biased female adjectives. However, mBERT has included some more random words as well, like 'Felles', 'Fille' and 'Sentral'. Even though most of the words are not identical, they are pretty consistent in meaning for the three models, especially NorBERT and NB-BERT. 'Hjemmeværende' and 'Prostituert' are the only adjectives present in the top 10 list of more than one model, both by NB-BERT and mBERT.

Figure 6.2: Top male (left) and female (right) biased adjectives as predicted by NB-BERT.

Figure 6.3: Top male (left) and female (right) biased adjectives as predicted by mBERT.

Overall the adjectives are highly male-biased in all models. For NorBERT, 96% of the resulting adjectives appear male-biased, while only 4% appear female-biased. For NB-BERT, the ratio is slightly more equal as 76% adjectives appear male-biased against

24% for females. For mBERT, however, all adjectives come out as male-biased, as all adjectives included in the results had a higher probability of predicting 'han' than 'hun'. The adjectives that are further presented as female-biased for mBERT are, therefore, the *least* biased in the male direction. These numbers indicate that the models have a 96%, 76%, and 100% chance for NorBERT, NB-BERT, and mBERT to predict Norwegian adjectives to be more connected to male pronouns than to female. On average, the percentage equals almost 91% of adjectives being male bias for the three models.

When it comes to determining *how* biased the models are in the context of predicting gender for adjectives, two measures are considered relevant. One is the number of adjectives that are predicted to be biased in one direction or the other, and the other is the bias score from Equation 6.2 describing how biased the adjectives are in that direction. Table 6.8 shows the bias scores in male and female directions for the top 1, 3, 50, and all bias adjectives predicted by NorBERT, NB-BERT, and mBERT. The results reveal that the male bias scores are much higher than the female bias scores for all three models. While NorBERT predicts an average male-to-female ratio of 5.66 for the 96% of male-biased adjectives, the average female-to-male ratio for the last 4% of female bias adjectives is only 1.59. This means that 96% of the Norwegian adjectives are more than five times as probable to describe a male as a female, according to NorBERT. The most female-biased adjective in NorBERT ('kvinnelig') is less biased than the average of all male-biased ones. This means that 96% of all Norwegian adjectives are more similar to male gender than the Norwegian word for 'female' is to the female gender, according to NorBERT. For NB-BERT, the numbers are 3.40 on average for the 76% male adjectives and 1.59 for the 24% female ones. Thus, NB-BERT is a little less extreme in the male bias score and its higher amount of female-biased adjectives. This combination makes NB-BERT a remarkably less biased model than NorBERT in this experiment. They both outclass mBERT with its 0% female-biased adjectives, as mBERT predicts that all 100% of Norwegian adjectives are more than 16 times as probable to describe a male as a female, lacking inclusion of female predictions.

Table 6.8: Aggregated male (M) and female (F) bias scores for adjectives in NorBERT, NB-BERT and mBERT calculated from Equation 6.2 for the top N most biased adjectives in each direction.

	NorBERT		NB-BERT		mBERT	
N	M	F	M	F	M	F
1	63.27	4.77	49.01	8.38	114.12	-
3	52.65	4.03	42.80	7.63	97.36	-
50	28.09	1.76	17.46	3.19	54.44	-
All	**5.66**	**1.59**	**3.40**	**1.59**	**16.45**	-

The results show that all three models have a much higher bias score in the male direction than in the female and a vast overweight of the number of adjectives that are biased towards males. This indicates that in addition to predicting male gender more often than female, they are also more sure of the prediction when predicting for males.

6.3 Detecting and Measuring Gender Bias in Downstream Tasks

In this experiment, the similarity between a male and a female-gendered set of sentences and a set of gender-neutral sentences is calculated as part of a downstream task. The null hypothesis suggests, similar to Caliskan et al. (2017), that there is no difference between male and female gendered sentences in terms of their relative similarity to a gender natural sentence. Thus proof of difference is a proof of bias.

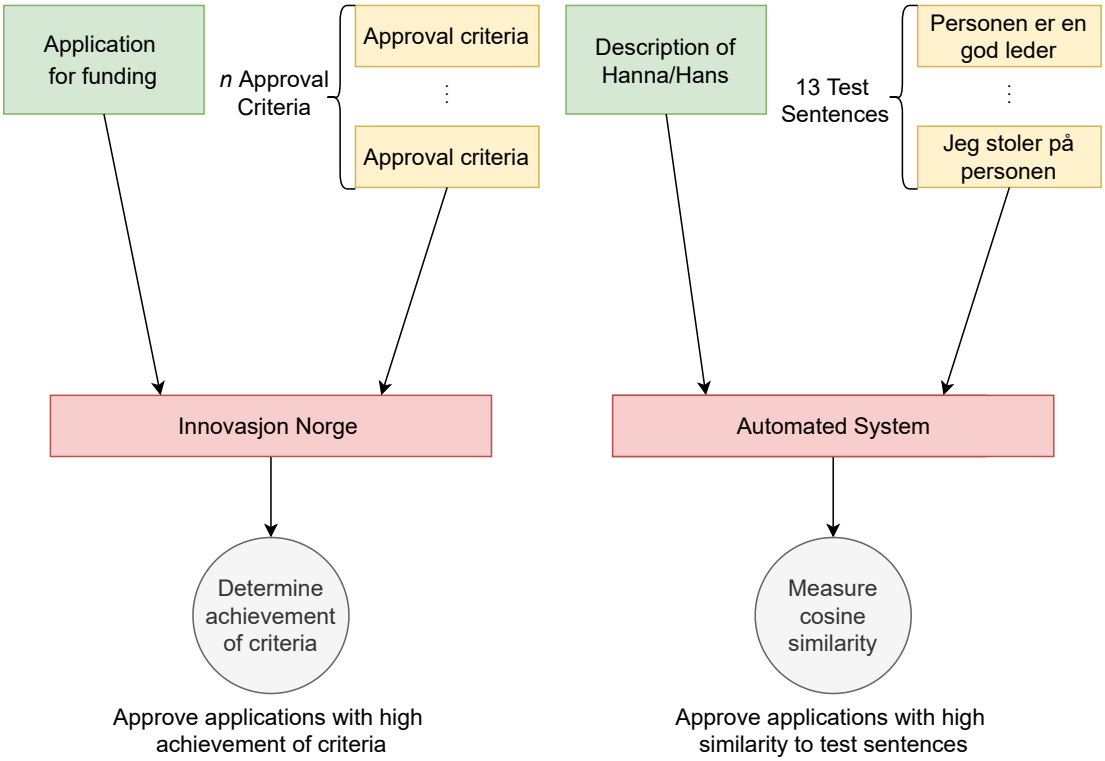
6.3.1 Experimental Set Up

In the previous experiment presented in Section 6.2, masked language modeling (MLM) is used as a tool to look into a model's embeddings and evaluate bias based on predictions of masked words. Another way to investigate embeddings is to directly calculate similarities between embeddings extracted from a model without decoding them back to word predictions. In this way, bias can be detected in the same manner, but the measure used focuses on relative similarities between words in the embeddings, like for WEAT, rather than ranging the appropriate words in an MLM task. In addition to calculating distances between words, the similarity between whole sentences can be calculated with this technique, and thus comparisons of sentences or larger text units can be made. This technique is further included as part of a hypothetical real-life application of NLP, similar to what was done by Sahlgren and Olsson (2019) in an NLP system for comparing company names and -descriptions. The following steps are implemented in the conduction of the experiment:

1. **Define a downstream task** that applies and makes use of measuring similarity between sets of sentences to make decisions.
2. **Extract embeddings** for all sentences; a set of male and female sentences, and a set of gender-neutral sentences to compare them to.
3. **Calculate embedding similarity** between male and female sentences and gender neutral sentences.

Define A Downstream Task Innovasjon Norge[1] receives applications for funding of projects from Norwegian founders. They consider the applications against a set of approval criteria on how well they achieve them in order to decide who gets funding and not. An overview of this process is presented in Figure 6.4a on page 62. Imagine that Innovasjon Norge has just started to use an automated filtering tool for applications that implements one of the Norwegian language models. The tool compares the similarity between the application and a set of sentences or words that describe the criteria used to determine if the applicant qualifies for funding. If the match is good, funding is granted; if the match is not good, funding is declined. A simplified version of this system is created in the experiment and is presented in Figure 6.4b on page 62.

[1] https://www.innovasjonnorge.no/



(a) Overview of the actual process of funding approval in Innovasjon Norge.

(b) Simplified automation of a process for funding approval as implemented in the experiment.

Figure 6.4: An overview of the mapping between the process approving funding applications in Innovasjon Norge to the simplified automation of it that is implemented in the experiment. The same color indicates the same part of the process.

As Innovasjon Norge provides neither sets of such criteria nor examples of applications, the datasets used are collected from the Hans and Hanna study by Gaustad and Raknes (2015). The descriptions of Hans and Hanna presented in Section 5.5 are used as two applications sent to Innovasjon Norge, and the set of evaluation questions used in the study as criteria to get funding. The criteria dataset is presented in Section 5.4 and are further referred to as *test sentences*. Translations of the sentences are provided there and will not be repeated throughout the thesis. The system will filter out applicants with a low similarity between their application text and these test sentences, and thus they will not get funding for their project. The comparison between the implementation in this experiment and the actual process in Innovasjon Norge can be seen in Figure 6.4 and serves an overview of the experiment.

Extract Embeddings To calculate distances in the vector space between the test sentences and the descriptions of Hanna and Hans, embedding representations of both are required. Similar to Sahlgren and Olsson (2019), the experiment averages the vectors of the component words to extract embeddings for the texts (sentences), resulting in two embeddings for the application texts, Hanna and Hans, and 13 embeddings for the criteria; test sentences.

Calculate Embedding Similarity To calculate the similarity between two embeddings in the vector space, a variant of SEAT as presented in Section 3.2.2 is adjusted to compare male and female sentences' relation to the same neutral sentence. Instead of comparing two sets of target words (texts) and two sets of attribute words (texts) as in WEAT (SEAT), the experiment compares two sets of attribute texts (Hans and Hanna descriptions) to one set of target sentences (test sentences).

From the embedding representation from the description of Hanna, $\vec{v_f}$, and of Hans, $\vec{v_m}$, the cosine similarity between each of the two to each of the test sentences, $\vec{s_j}$ is calculated, and further used to calculate the distances d_f and d_m with the following formula (represented as d_f):

$$d_f = 1 - \cos(\vec{v_f}, \vec{s_j}) \tag{6.3}$$

d_f describes the distance from Hanna to the test sentences, while d_m describes the distance from Hans to the test sentences. As increased similarity correlates with increasing cosine distance between vectors, the formula for $d(\vec{s_j}, \vec{v_m}, \vec{v_f})$ includes a subtraction of the cosine from 1 to rather obtain a number for *increasing* similarity correlating with *decreasing* distance. This means that a higher value of $d(\vec{s_j}, \vec{v_m}, \vec{v_f})$ means that two vectors are more similar to each other, and thus closer in vector space.

Further, the difference between the two distances, d_f and d_m, for each sentence is calculated:

$$d(\vec{s_j}, \vec{v_m}, \vec{v_f}) = d_m - d_f. \tag{6.4}$$

This gives a value higher than zero if the test sentence is more similar to the Hans text than the Hanna text and lower than zero in the opposite situation. Deviation from a difference equal to zero breaks the null hypothesis and is proof of bias in the experiment.

The experiment is conducted for NorBERT, NB-BERT and mBERT. As a measure of how biased the models are overall in the experiment, a set of different values are calculated for all the test sentences together based on a sum of the values for each sentence from Equation 6.4.

- *Male bias* describes the average value of bias in male direction. All positive values (indicating male direction) from Equation 6.4 for a model are summed and divided on the number of values.
- The same process for negative values (indicating female direction) creates the measure of *female bias*.
- *Bias* describes the average value of all values from Equation 6.4 for a model, both male and female biased.

- *Absolute bias* describes the average of all the absolute values from Equation 6.4 for a model.

Male and female bias does not take eventual bias in the opposite direction into account but describes the strength of the biases in each direction (if any is detected). On the other hand, bias describes the overall tendency of bias in a model toward one gender or the other. A positive value indicates overall male bias in the model and female bias in the opposite. Absolute bias describes the tendency of bias in the model in one way or the other but does not say anything about which gender. Thus, it measures the model's tendency to deviate from the null hypothesis.

6.3.2 Experimental Results

Table 6.9 presents the raw values of the difference in distance from Equation 6.4 in male favor (M-F) for all the test sentences compared to the Hans (M) and Hanna (F) texts. Figure 6.5 on page 65 visualize the same results for all three models compared.

Table 6.9: Difference in distance from test sentences to descriptions of Hans and Hanna.

Test Sentences	NorBERT	NB-BERT	mBERT
S1: Jeg tror personen er lykkelig	0.00095	0.00144	−0.00040
S2: Jeg opplever personen som sjefete	−0.00047	0.00160	−0.00075
S3: Jeg vil si at personen er en god forelder	0.00047	0.00192	−0.00063
S4: Personen er en god leder	0.00136	0.00174	−0.00049
S5: Jeg liker å samarbeide med personen	0.00076	0.00143	−0.00065
S6: Jeg opplever personen som en sympatisk person	0.00523	0.00171	−0.00065
S7: Jeg kan tenke meg å ha personen som mentor	0.00340	0.00159	−0.00067
S8: Jeg vil ta en øl med personen etter arbeidstid	0.00268	0.00162	−0.00043
S9: Jeg vurderer personen som egoistisk	0.00091	0.00168	−0.00053
S10: Jeg vil ta de samme valgene som personen for å lykkes	0.00338	0.00155	−0.00063
S11: Jeg stoler på personen	0.00086	0.00169	−0.00082
S12: Jeg vil jobbe for personen	0.00045	0.00152	−0.00007
S13: Jeg opplever at personen er godt likt	0.00056	0.00195	−0.00046

Overall, the experiments indicate extreme favoring of males in NorBERT and NB-BERT and favoring of females in mBERT. All results for NorBERT and MB-BERT are male bias, except for one test sentence in NorBERT, meaning that almost all the test sentences are more similar to Hans's description than Hanna's. The exception is S2 ("Jeg opplever personen som sjefete") which comes out as female bias in NorBERT. For mBERT, however, all sentences come out as female bias. Test sentences S6 ("Jeg opplever personen som en sympatisk person") has the single strongest male bias for NorBERT with large relative value, and S7 ("Jeg kan tenke meg å ha personen som mentor") and S10 ("Jeg vil ta de samme valgene som personen for å lykkes") are also in top 3 male

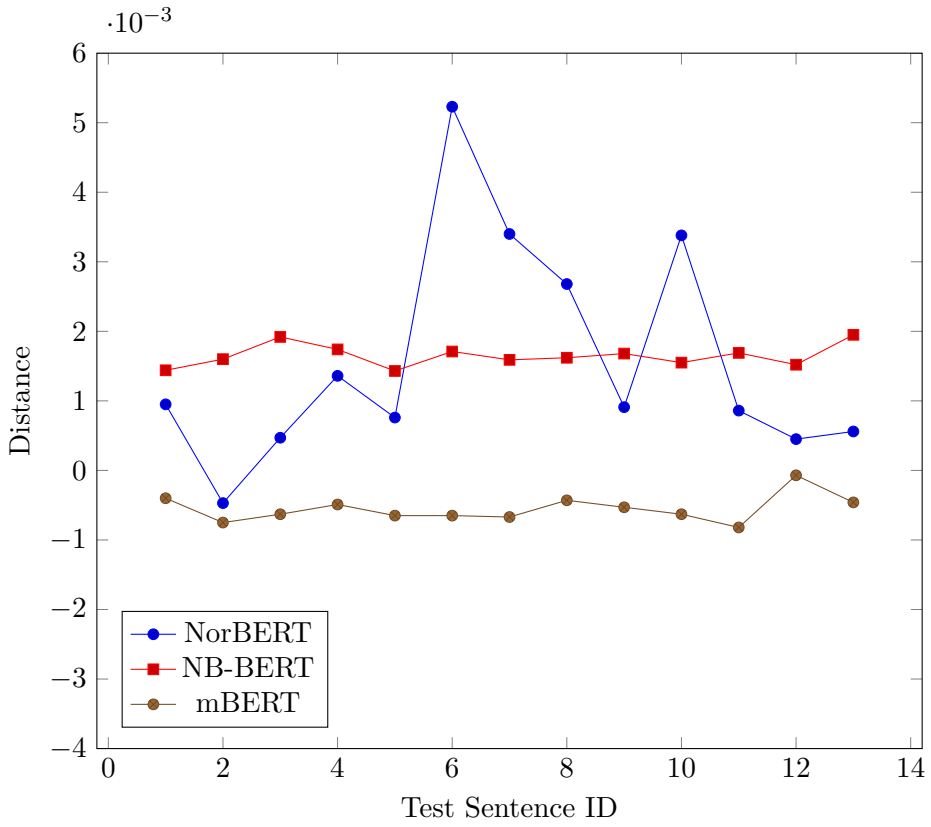


Figure 6.5: Average of all sentence embeddings in Hans and Hanna descriptions compared to the test sentences identified by S1-S13.

biased test sentences. All test sentences except for S12 (“Jeg vil jobbe for personen”) are more male bias than the one female bias sentence S2 is in NorBERT. For NB-BERT S3 (“Jeg vil si at personen er en god forelder”), S4 (“Personen er en god leder”) and S13 (“Jeg opplever at personen er godt likt”) are the top three most biased test sentences, with S13 being the single most biased. The top three *female* biased in mBERT consist of S2 (“Jeg opplever personen som sjefete”), S7 (“Jeg kan tenke meg å ha personen som mentor”) and S11 (“Jeg stoler på personen”), S11 being the most female biased.

The overall bias values, male bias, female bias, bias, and absolute bias, are presented in Table 6.10 on page 66. Female bias values are written in absolute value in the table to be comparable to male bias, and ’-’ is used if there is no bias for the measure. With no female bias detected in NB-BERT, the male bias with an average of 0.00165 accounts for both the bias and the absolute bias score. The same goes for mBERT, with the female bias detected with an average of 0.00055 and no male bias. NorBERT has detected both female and male bias of respectively 0.00047 and 0.00188, and thus the bias score and absolute bias score are combinations of these two. Even though NorBERT detected female bias, the male bias that is detected is stronger than the male bias detected in

NB-BERT. However, the female bias in NorBERT is not as strong as the average female bias detected in mBERT but higher than four of the sentences in mBERT individually. Despite the female bias, which equalizes it for NorBERT, NB-BERT marginally has the highest bias score over NorBERT. NorBERT has a high male and female bias score, resulting in an absolute bias score and deviation from the null hypothesis of 0.00165. NB-BERT has a deviation from the null hypothesis of 0.00165, which is identical to NorBERT and three times as much as mBERT with 0.00055.

Table 6.10: Measures of average values of male bias, female bias, bias and absolute bias for all test sentences combined in the three Norwegian language models.

	NorBERT	NB-BERT	mBERT
Female bias	−0.00047	–	0.00055
Male bias	0.00188	0.00165	–
Bias	0.00158	0.00165	0.00055
Absolute bias	0.00165	0.00165	0.00055

6.4 Mitigating Gender Bias by Removing Gender Subspace

This experiment implements removing the gender subspace of a model from a set of gender natural sentences as a debiasing technique. The debiased embeddings are tested through a similar task as was conducted and described in the experiment in Section 6.3, and the results are compared to see the effect of the debiasing.

6.4.1 Experimental Set Up

In the comparison of the descriptions of Hans and Hanna to a set of gender natural sentences from Experiment 6.3, the sentences are shown to not be as gender neutral as wanted. The results indicate the presence of gender as a component in the distributed meaning of the sentences, and thus an attempt to remove this component from the sentences is conducted. The experiment consists of the following steps:

1. **Identify a gender subspace** that describes the distances in the vector space that are due to gender.

2. **Remove the gender subspace** from a set of gender neutral sentences to neutralize (debias) them.

3. **Redo the downstream task** from Experiment 6.3 with the neutralized sentences to see the effect of the debiasing.

Identifying A Gender Subspace First, the gender subspace of the three Norwegian language models is identified. Vectors of gender word pairs were extracted from the language models' embeddings. As the target is to identify a subspace in contextual embeddings, word embeddings required to do this cannot be extracted directly. Thus, a set of sentences representing a word in different contexts are used to create an average embedding of the words. For example, by extracting the contextual vector of the word 'hun' used in n number of different sentences, n different vectors for the word can be extracted.

The female and male parts of the dataset described in Section 5.2 are used to extract embeddings for a set of gendered words; 'hun', 'jente', 'han,' and 'gutt'. Each word is represented as a 1×768 vector in the BERT-based models' sentences, and one vector for the word is extracted from each sentence containing the word. The set of vectors representing the word 'hun' constructs the female matrix F_1, while the set of vectors for 'han' constructs the matrix M_1. The vectors for the sentences with the words 'jente' and 'gutt' similarly construct F_2 and M_2. These matrices form the female and male matrices $F = F_1 + F_2$ and $M = M_1 + M_2$. The *distance matrix* D is obtained by taking the difference between these two matrices $D = M - F$. Understand that the difference is calculated between similar sentences except that they are gender swapped and form the difference matrix D which represents the gender distance in the vector space.

Further, the features in the complete distance matrix D can be standardized, and the top n principal components of the standardized matrix D_s can be found. The (combination of the) dominant principal component(s) of D_s constructs the gender subspace of the language models' word embeddings.

Neutralize test sentences by removing gender subspace After performing principal component analysis on D, the gender subspace g is retrieved. It is then removed in the same manner as previous research by removing the impact of the gender subspace from the gender-neutral sentences by using orthogonal projection (Bolukbasi et al., 2016; Sahlgren and Olsson, 2019) which is categorized as a hard debiasing technique. However, two main principal components are removed by using orthogonal projection on the supposed to be gender-neutral embeddings. Since two components are removed and not one, Equation 3.6 is extended:

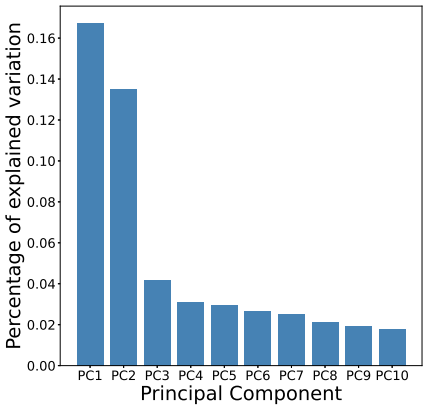
$$w'' = w' - \frac{g_2 \cdot w'}{g_2 \cdot g_2} g_2. \tag{6.5}$$

This formula is then applied to all the 13 test sentences from Section 5.4 that describe Hans and Hanna's characteristics as leaders, resulting in 13 neutralized embeddings for the test sentences.

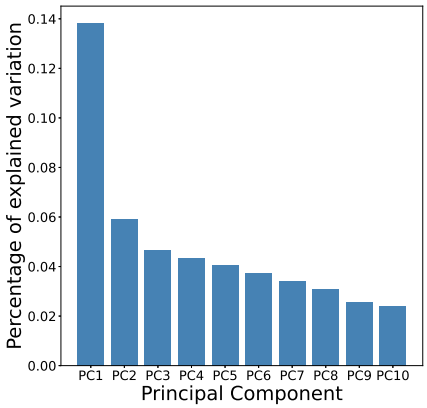
Redo Downstream Task Experiment 6.3 is redone with the neutralized embeddings as the 13 test sentences. The experiment is conducted for all three models.

6.4.2 Experimental Results

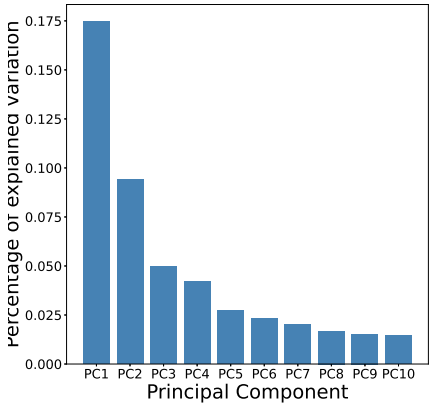
Figure 6.6 on page 68 presents the top 10 principal components of the gender subspace found in NorBERT, NB-BERT and mBERT.



(a) NorBERT



(b) NB-BERT



(c) mBERT

Figure 6.6: Top 10 principal components describing the gender subspace for all three Norwegian language models.

For NorBERT and mBERT, the gender subspace is clearly dominated by two principal components. NB-BERT (see Figure 6.6b) has a gender subspace dominated by one principal component and the second one with a bit less influence. The finding of two principal components that describe gender is consistent with Zhao et al. (2019) findings. The results show that for NorBERT (Figure 6.6a), the largest principal component describes over 16% of the variation in the gender subspace, and the second accounts for almost 14% of the variation. For NB-BERT, the two largest components account for 14%

and 5%. For mBERT (Figure 6.6c), the numbers are 17% and 10%.

Figure 6.7 on page 70 shows the difference in distance for all test sentences compared to Hanna and Hans before and after debiasing for all the three models. The Pearson coefficients between the original and debiased values for all test sentences are 0.82, 0.88, and 0.85 for NorBERT, NB-BERT, and mBERT. As the number is close to one for all three models, the debiasing is consistent in how much effect it has on the different sentences within a model. On the other hand, how much effect the debiasing had between the model varies. All sentences that were male bias before in NorBERT and NB-BERT have reduced bias. However, some of the sentences in NorBERT have tipped over to become female-biased, with an increased similarity between Hanna and the test sentences compared to Hans. Also, the one sentence in NorBERT that was already female-biased has increased its bias further. NB-BERT has successfully decreased the difference in the distance towards zero for all sentences, but the differences in values are much smaller than for NorBERT. From before, mBERT showed a clear female bias, which has increased for all sentences except for two, which is still female bias but with a small decrease. It also shows female bias for all sentences after the debiasing, inconsistent with the other two models.

The overall bias values for the three models after debiasing, male bias, female bias, bias, and absolute bias, are presented in Table 6.11. Female bias values are written in absolute value in the table to be comparable to male bias, and '-' is used if there is no bias for the measure. Both NorBERT and NB-BERT show a decrease in the absolute bias value, implying that the debiasing isolated was successful. NorBERT responds more to the technique as the absolute bias before of 0.00165 has been reduced to 0.00093, against a reduction from the same original value to 0.00135 for NB-BERT. The female bias in NorBERT has, on the other hand, increased, but not enough to make the absolute bias in NorBERT higher than for NB-BERT due to the significant decrease in male bias in NorBERT. mBERT increases its absolute bias as the female bias increases, and there is no male bias before or after debiasing. The model that performs best in mitigating by removing the gender subspace is thus NorBERT, with the highest difference in absolute bias before and after debiasing.

Table 6.11: Aggregated scores for distances in test sentences and descriptions of Hanna and Hans before and after debiasing for all three Norwegian language models.

	NorBERT		NB-BERT		mBERT	
	Original	Debiased	Original	Debiased	Original	Debiased
Female bias	0.00047	0.00107	-	-	0.00055	−0.00064
Male bias	0.00188	0.00077	0.00165	0.00135	-	-
Bias	0.00158	−0.00022	0.00165	0.00135	−0.00055	−0.00064
Absolute bias	0.00165	0.00093	0.00165	0.00135	0.00055	0.00064

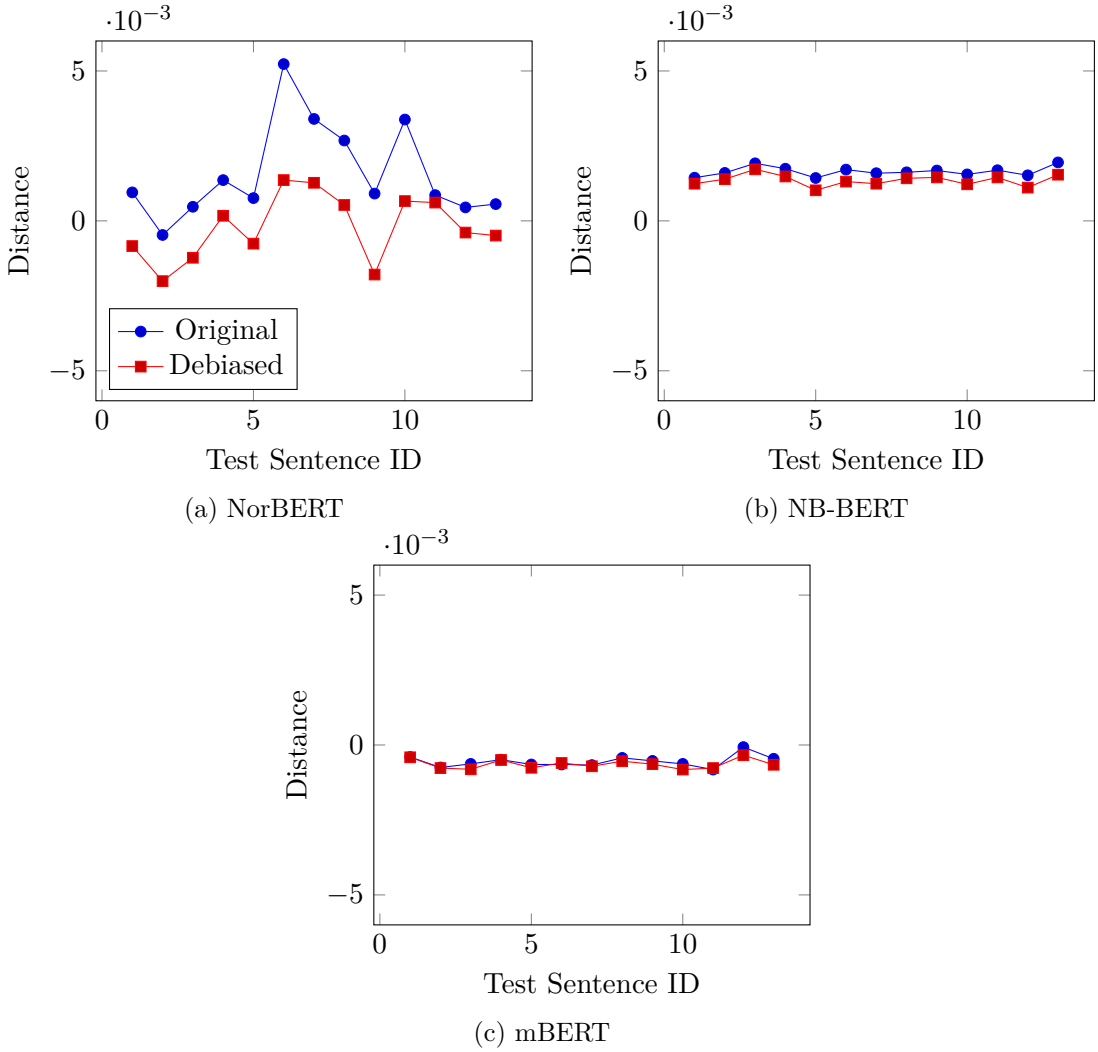


Figure 6.7: Difference in distance between description of leader and Hans and Hanna before and after debiasing for three Norwegian language models.

6.5 Mitigating Gender Bias by Fine-Tuning on Female Only Corpus

In this experiment, a model is fine-tuned on a female only corpus to determine its effect on the presence of bias in the model compared to before fine-tuning. Costa-jussà and Jorge (2020) claim training on balanced data is a first step to eliminating representational harms from NLP and Hovy and Prabhumoye (2021) described how the under-representation of a minority in datasets contributes to bias in NLP. Experiment 6.1 shows that both NorBERT, NB-BERT, and mBERT are trained on more than 2/3 male pronouns and that all models contain deviation from the non-bias null hypothesis. The target is to balance out this gender skew and see whether it affects the measures of bias, so the results can be used to consider training data as a source of bias and fine-tuning as a debiasing technique.

6.5.1 Experimental Set Up

A script was created to gender swap masculine pronouns in the Norwegian Colossal Corpus (NCC) according to the description in Section 5.6. The script was sent to the AI Lab at The National Library of Norway (NLN) and run by them internally to implement the swapping. Further, the new corpus was used to fine-tune NB-BERT and the new model is further referred to as **NB-BERT-male2female** and can be found on HuggingFace[2]. Access to help from NLN to fine-tune the model and their knowledge about NCC were the reasons for this experiment's choice of model and data. The debiased model was tested through the same experiments as described in Experiment 6.2 (ratio between male and female predictions from masked language modelling) and in Experiment 6.3 (difference in similarity between descriptions of Hanna and Hans and test sentences).

6.5.2 Experimental Results

Figure 6.8 on page 72 presents the results from comparing the descriptions of Hans and Hanna to the test sentences with embeddings extracted from NB-BERT-male2female, including results from NB-BERT and the equalized NB-BERT from removing the gender subspace in Experiment 6.4 to compare the effect of the fine-tuning as a debiasing technique. The comparison shows significant results on the difference in similarity between Hanna and Hans, suggesting either an increased similarity for Hanna or a decreased for Hans (or both). It is somewhat ambiguous how the fine-tuning has affected the embeddings as the differences have not changed to be more similar to Hanna from training on female-only pronouns. All the sentences are still more similar to the description of Hans after the fine-tuning, but the results undoubtedly suggest that the technique does reduce the male bias that was detected for NB-BERT in Experiment 6.3 for the specific task.

[2]https://huggingface.co/NbAiLab/nb-bert-ncc-male2female

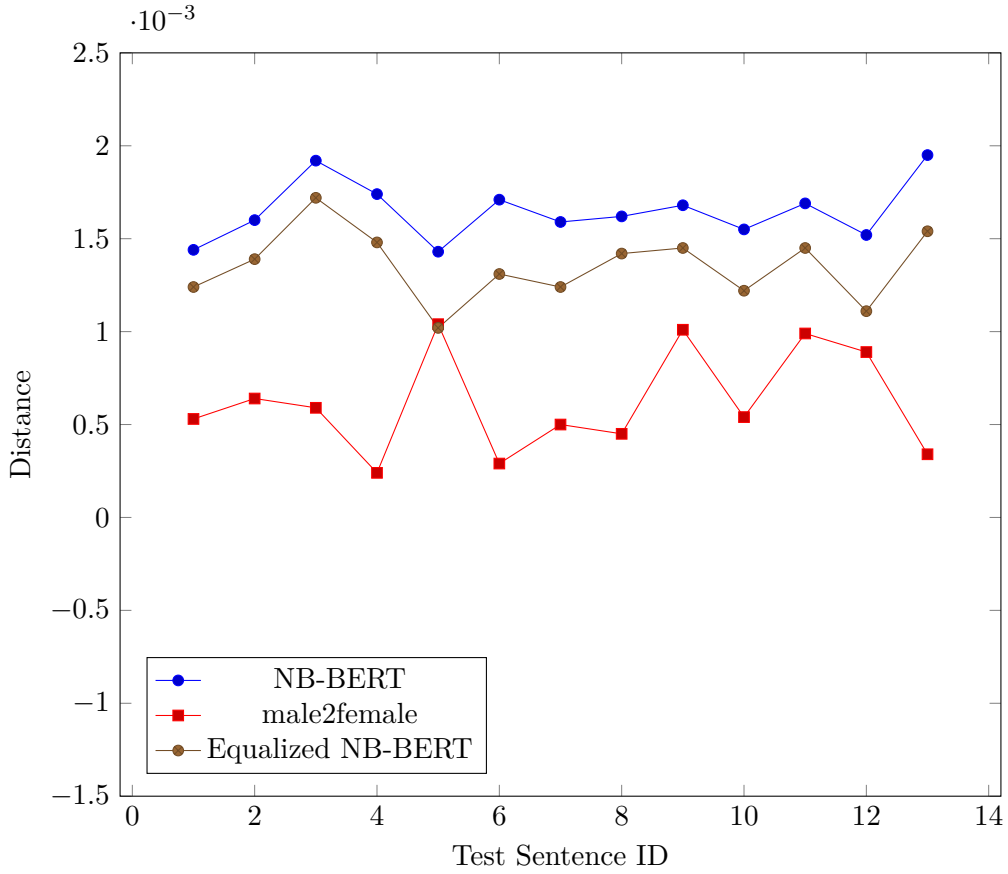


Figure 6.8: Comparison of distance between descriptions of Hanna and Hans and the set of test sentences for a downstream task as predicted by NB-BERT-male2female, equalized NB-BERT and NB-BERT.

By redoing the Experiment 6.2 all adjectives come out extremely female-biased which show that the results are not reflected as for Experiment 6.3. Table 6.12 shows the aggregated male and female bias scores for adjectives in NB-BERT-male2female calculated from Equation 6.2 for the top N most biased adjectives in each direction, and compared to original scores for NB-BERT that was previously presented in Table 6.8. The results show that the aggregated bias score in the female direction has increased from 1.59 for NB-BERT to 191.09. In combination with 0% of adjectives coming out as male-biased, meaning that there is now, on average for all adjectives, more than 191 times as probable that an adjective is used to describe a female as male. From these results, fine-tuning on an all-female pronoun corpus does not seem to work as a debiasing technique as it makes the model more biased (but in the other direction). However, the results indicate how large an effect change of pronouns has on the embeddings. The numbers are extremely high, indicating that the fine-tuning of all female pronouns has a huge effect on the embeddings in NB-BERT, even with the initial training on 2/3 of male pronouns as a base.

Table 6.12: Aggregated male (M) and female (F) bias scores for adjectives in NB-BERT-male2female in male (M) and female (F) calculated from Equation 6.2 for the top N most biased adjectives in each direction compared to original scores for NB-BERT.

	male2female		NB-BERT	
n	M	F	M	F
1	-	450.01	49.01	8.38
3	-	409.21	42.80	7.63
50	-	312.59	17.46	3.19
All	-	**191.09**	**3.40**	**1.59**

7 Evaluation and Discussion

This chapter presents the evaluation and discussion of the experiments and results. The evaluation considers the methods used and their effect on the results in the experiments one by one and some overall points worthy of mention. The discussion reflects on the results in light of related work and its implication and larger contexts.

7.1 Evaluation

For all experiments conducted in the thesis, choices were made that possibly affected the results. The Master's Thesis is limited by time, affecting these choices. Priorities are made throughout the research, and some ideas and parts of experiments were left out. The following sections present an evaluation of these choices and the results for each of the experiments, including a dedicated evaluation of the inclusion of mBERT (Devlin et al., 2019) in the thesis.

7.1.1 Evaluation of Results Concerning Gender Bias in Training Data

To evaluate the script used to count pronouns, Table 7.1 compares the number of words counted to the size of the corpora that are Norwegian Wikipedia in Bokmål (NO) and Nynorsk (NN), Norsk Aviskorpus (NAK) and Norwegian Colossal Corpus (NCC) as presented in Section 2.4.3.

Table 7.1: Number of words that were documented as part of training data compared to number of words counted in experiment.

	Wikipedia NO	Wikipedia NN	NAK	NCC
Corpus	160M	40M	1.7B	6.9B
Counted	154M	34M	1.8B	6.8B

The amount of words counted is not the same as in the training corpora of the models. The reason is the implementation of the scripts used to count, explained by the following for each training corpus:

- When searching through Wikipedia, fewer words were counted than are actually in the corpus. There were fewer words than the actual corpus when reading from Wikipedia. The script did not include all the page content used in training as words but only focused on article titles, section titles, and section texts while excluding all other fields.

- For NAK, more words have been counted than what is present in the training because the script counted metadata about an article as words. Knowing that this metadata consisted of URLs, dates, and similar, we did not bother to remove them in the counting as they would not affect the number of pronouns found.

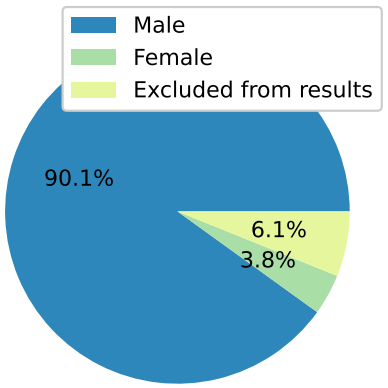
- For NCC, all the tokens in the text field of the JSON objects are counted. This resulted in 0.1B fewer words than the Norwegian National Library officially said were included in the training data (published part). Probably, they count other fields in addition that we are not aware of, similar to Wikipedia.

As there is an explanation for the relatively small mismatch between real and counted words, the count is credible. The selection of gendered words is assumed to be representative of the representation of each gender in the dataset. NCC and thus the training data for NorBERT (Kutuzov et al., 2021) consist of many more words (18.4B), but only 6.9B are available publicly.

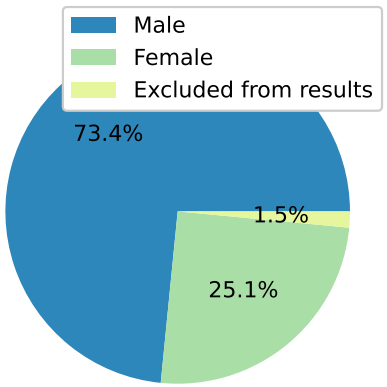
Initially, The National Library of Norway (NLN) promised access to the whole NCC for research purposes, but permissions were never obtained due to lack of time. Only the freely accessible part is investigated and has been considered as NB-BERT's (Kummervold et al., 2021) training data in this thesis. As the counted division of pronouns is consistent with the other corpora, the count is considered to represent the complete NCC.

7.1.2 Evaluation of Results Concerning Detection of Gender Bias in Word Embeddings

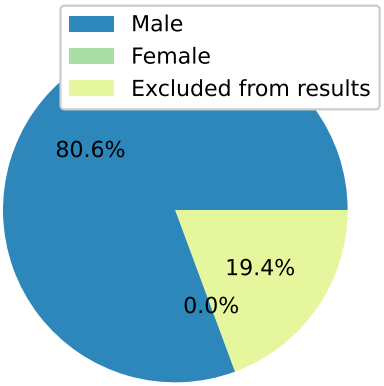
With the requirement of both 'hun' and 'han' being predicted among the top five of the masked tokens, there is an imminent danger that some relevant adjectives are not included in the results. For instance, it can lead to the exclusion of strongly gendered adjectives, like 'mannlig' (English: 'man-like'), because it will predict 'hun' with close to zero probability. Thus, to evaluate the results' credibility, the number of adjectives included is compared to the total amount tested. The list of Norwegian adjectives that are iterated through consists of 1720 adjectives. Figure 7.1 on page 77 presents the share of adjectives that come out as biased in male and female direction for the three Norwegian language models, along with the number of adjectives that are not included in the results. For NorBERT (see Figure 7.1a), 1615 adjectives are included in the results, meaning that 105 adjectives do not predict 'hun' and 'han' among the top 5 most likely words to be masked in the sentence. For NB-BERT (Figure 7.1b), 1694 are included (26 excluded), and for mBERT (Figure 7.1c), 1696 are included (24 excluded). This means that for all three models, at least 94% of the adjectives predicted both 'hun' and 'han' among the top five as wanted, and the results are considered credible even though there might be some highly biased ones missing.



(a) NorBERT



(b) NB-BERT



(c) mBERT

Figure 7.1: Share of adjectives biased in male and female direction and that are not included in the results for the three Norwegian language models.

7.1.3 Evaluation of Results Concerning Detection of Gender Bias in Downstream Tasks

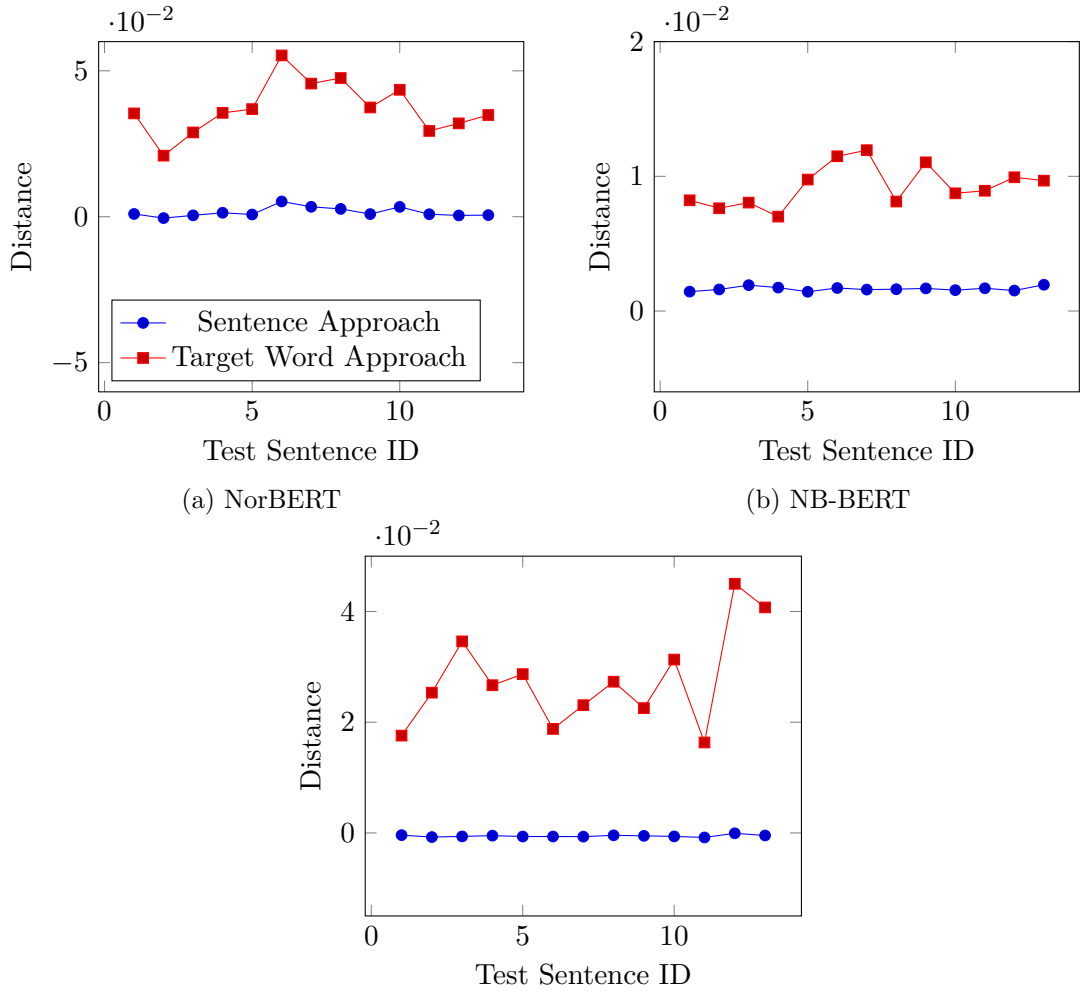
The embeddings from the descriptions of Hanna and Hans lack larger context than at the sentence level as the embeddings for each sentence are extracted individually. However, the process is conducted identically for Hanna and Hans to investigate the relative distance between the two and the test sentences. The possibility that it may not be appropriate to use contextualized language models to generate embeddings for individual tokens is also acknowledged. However, as Sahlgren and Olsson (2019) argue for such usages to occur in real-world applications, it is considered relevant to include it in these experiments as well.

Two different approaches were tested for extracting Hanna and Hans's embeddings. The one used is referred to as the sentence approach (SA), and the other one as the target word approach (TWA). In TWA, all the embeddings for the word 'hun' (for Hanna) or 'han' (for Hans) were extracted, and the average of these was used as embedding for Hanna and Hans. In this approach, the gendered words thus played a larger role in affecting the embeddings, as they were not averaged with the rest of the words in the sentence. Also, sentences that do not include 'hun' or 'han' are left out of the embedding representation of Hanna and Hans, emphasizing their gender in them. Figure 7.2 on page 79 visualize the results from the task by using respectively SA and TA for all three models. The difference between Hans and Hanna in TWA is much higher for all three models, which is not surprising as it emphasizes the importance of gender in the sentences compared to SA. However, the numbers for the two approaches were correlated with Pearson coefficient 0.934 for NorBERT, 0.794 for NB-BERT, and -0.019 for mBERT. Intuitively they should be correlated because TWA only exaggerates the importance of genders. This is the case for NorBERT and NB-BERT to some extent. mBERT has a negative correlation, which does not make sense, but the two others are prioritized. As the correlations for NorBERT and NB-BERT are high, only one approach was included, being SA consistent with the best choice according to Sahlgren and Olsson (2019).

Lastly, the actual process of evaluating funding applications is still done manually by Innovasjon Norge, and the criteria are different from the ones assumed in this thesis; degree of innovation, feasibility, and profitability. This makes the hypothetical system not applicable to the actual tasks, but this is also not the intention. Evaluating the degree of match between criteria and application is transferable, and the results show bias in this process. The experiment is a simplification of a potential automated system for this process. However, it indicates the bias the language model has in a comparable task applied in real life.

7.1.4 Evaluation of Results Concerning Removing Gender Subspace as a Debiasing Technique

The sentences for the dataset that was later going to form the basis of the principal component analysis (PCA) were arguably not covering gender to a large enough degree. The most significant principal component found to only cover 17% of the explained



(a) NorBERT

(b) NB-BERT

(c) mBERT

Figure 7.2: Sentence approach (SA) compared to the target word approach (TWA) for thre Norwegian language models.

variation, which compared to Zhao et al. (2019) had one principal component accounting for 25% of the explained variance. The two components used describe 30%, 19%, and 27% og the variance for NorBERT, NB-BERT, and mBERT. In comparison, Zhao et al. (2019) and Bolukbasi et al. (2016) have detected subspaces that account for approximately 47%, and 60%. The reason might be the low number of sentences created for the dataset due to not prioritizing it as it was time-consuming. Another reason might be the combination of stereotypical sentences, including occupations, and other random sentences from online news so that the stereotypical sentences are watered out. Zhao et al. (2019); Bolukbasi et al. (2016) included 800 sentences, which is four times the size of this task. An approach including sentences associated with more stereotypes could have increased this explained variance. Also, only sentences with two different gender pairs were included in the dataset. Bhardwaj et al. (2021) use 11 gender pairs, and more than two gender pairs might have been desirable to capture the gender subspace better.

Also, removing two principal components can introduce more noise in the results. As Zhao et al. (2019) argued that for contextualized word embeddings, the two most significant principal components construct the gender subspace, the choice was made to do that for all three models. This is the case for NorBERT and mBERT. However, NB-BERT only has one sizeable principal component. Two components were defined as the gender subspace to conduct the same process for all three models and compare. The consequence is that NB-BERT's second principal component does not necessarily describe gender so much but instead includes noise. Another approach that maybe could have been better is to remove a different number of principal components that capture the subspace best for each model, not just the same amount for all three.

7.1.5 Evaluation of Results Concerning Fine-Tuning on Female Only Corpus as a Debiasing Technique

Retraining models are time- and resource-consuming, resulting in a one-shot solution for the retraining on the female-only corpus. There was no time and computation resources to do the training ourselves, but the AI Lab at The National Library of Norway (NLN) offered to do it. The chosen method was to gender swap all male entities in the complete NCC corpus to female so that fine-tuning would adjust the male bias already present, and a script to do so was created and sent over. This resulted in a much more significant effect on the embeddings than wanted when retesting the masked language modeling task. Figure 7.3 on page 81 shows the division of male and female-biased adjectives detected, along with the number of adjectives not included in the results for NB-BERT-male2female. 490 adjectives are included in the results meaning that 1230 were excluded since 'han' and 'hun' (English: 'he' and 'she') were not predicted as the masked token. Most adjectives predict 'hun' as the number one word to be masked, indicating extreme female bias, along with words like 'personen', 'den', 'det', 'de' (English: 'the person, 'this', 'that', 'they') and famous names like 'Trump' rather than 'han'.

The training data contain three times more male pronouns, and a better approach would be to fine-tune on the same corpus with three times more female pronouns or even perform actual gender swapping. Both these approaches were considered, but there was

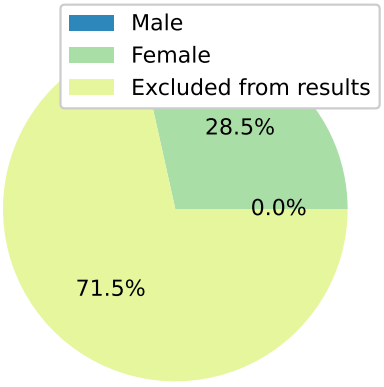


Figure 7.3: Share of adjectives biased in male and female direction that are not included in the results for NB-BERT-male2female.

no time to develop the experiment further. When the AI Lab had already started its training on a complete female-only corpus, it was too late to give additional adjusting, and so the experiment was carried out as it was.

7.1.6 The Experiments are Not Suited for mBERT

As mentioned in several of the experiments, the results for mBERT seem to be a bit random. They differ from the somewhat consistent and explainable results for NorBERT and NB-BERT. For Experiment 6.2 that predicts feminine and masculine adjectives, mBERT does not predict any feminine adjectives, regardless of approximately the same division of female entities in training data as the other two. The results include much fewer adjectives, and it seems to handle the Norwegian language less well than the other models. mBERT struggles with predicting words that result in correct sentences and frequently suggests illogical words or characters, such as periods as the missing word, among the top five predictions. mBERT would predict ". blir beskrevet som en snill person" (English: ". is described as a nice person") with a higher probability than "hun blir beskrevet som en snill person" (English: "she is described as a nice person"). This is expected as it is primarily a multilingual model and has a smaller Norwegian training corpus than the other two. However, the disability to predict correct sentences is similar (but less extreme) to the results for the female-only model NB-BERT-male2female, only in the opposite direction. As NB-BERT-male2female is known to be female-biased, this could indicate the same kind of bias in mBERT, but in the other direction. A reason can thus be that mBERT amplifies gender skew in training data more than NorBERT and

NB-BERT. This is consistent with Webster et al. (2018), who show that written text in training data amplifies biased patterns. This is more or less rejected in the results from Experiment 6.3 where mBERT suddenly connects all test sentences with the female gender, contrary to its results for the masked language modeling task and the other two models. This inconsistency speaks for the experiments not being suited for mBERT rather than it being proof of extreme biases. The inconsistency is even more extreme for mBERT between the two different approaches tested in Experiment 6.3. Here mBERT associates all sentences obtained with the SA closer to feminine, while for TWA, they are more associated with male. For the other two models, the results from these approaches are correlated, as was described in Section 7.1.3. For mBERT, they are not, which does not make sense. mBERT results also differ from NorBERT and NB-BERT as it ends up more biased after the debiasing in Experiment 6.4.

7.2 Discussion

The thesis results are reflected on in light of related work and its implication and larger contexts.

7.2.1 The Results Show Allocational Harms

Lack of representation of females in a models' training data leads to allocational harms where the models perform better on entities associated with males (Crawford, 2017). These harms seem to appear in Experiment 6.2 where all three Norwegian language models are more sure of the prediction when they predict for males, which is 91% of the time. All three models give a ratio in male favor more than twice as high in their prediction of male adjectives as females, which is a significant difference in the ability to perform equally well in the masked language modeling task for both genders. NorBERT and NB-BERT suggest that, respectively, 707 and 461 adjectives have a more substantial male bias than 'kvinnelig' (English: 'female like') has for female, meaning that most of the adjectives are more associated with male than the actual word for 'female' is to female. As part of an applied NLP system, the models' disability to find similar relations between female entities and adjectives as for males can lead to less correct processing in female disadvantage. The tendency is repeated for NorBERT and NB-BERT in the results from Experiment 6.3 where all test sentences except one for NorBERT are considered more similar to Hans than to Hanna. In consistency, Kurita et al. (2019) show that over 80% of both positive traits, negative traits, salary, and skills are more associated with men. Due to this overweight of associations between males and all adjectives, the male adjectives might correlate with the most used adjectives in the training data rather than capturing male stereotypes. This statement is supported by the results from the debiasing conducted in Experiment 6.5, where a swap of pronouns to all female leads to no adjectives being predicted as male, which is a significant lack of performance the other way around. The results thus show that female entities lack of meaningful relationships to other words in the embeddings, or as the title of the Master's Thesis states, that a

female word is characterized by the lack of relevant company in the vector space. This leads to Hans being granted funding over Hanna and is the source of allocational harms caused by the models.

7.2.2 Performance Measures are Missing a Metric for Fairness

Caliskan et al. (2017) state that bias should be the expected result whenever an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered. Even though bias in modern language models has been proven to a large extent, researchers are still not required to test the models they publish concerning biases they contain. A fair share of studies have been conducted for English language modeling on the topic, but the problem has not been solved. However, the imbalance in research on bias in such models between the English language and any other language has caused a paradox as it is a bias that most technology works better for the English-speaking population. Regardless of the number of studies conducted for English, there still exist no standards or requirements as to whether or how someone should test their model's performance on fairness today.

As an extension to the issue of lack of bias as a performance measure, the models potentially suffer from a lack of measuring performance equally for males and females in the first place. With the division of gender pronouns in training data discovered in Experiment 6.1, one can assume the same division in test data as both sets are typically subsets of the same or similar data source. This implies that the performance of language models is typically only measured on male-dominated datasets, and potentially lower performance on female entities than on males is not detected. Thus, according to the division of pronouns, a model performing well on male entities would perform better on 2/3 of the test data and potentially achieve a better performance score from favoring male predictions than from not doing it. In this way, the model lacks bias verification and is even rewarded for being biased. Another measure of gender bias, as well as a suggested requirement for publishing language models, is, therefore, to evaluate the performance of a model on gender-neutral datasets, which has been done in several pieces of research for English (Rudinger et al., 2018; Zhao et al., 2018; Webster et al., 2018; Kiritchenko and Mohammad, 2018). If a system shows a significant difference between original and debiased data, it would be considered gender biased.

7.2.3 The Results Show Representational Harms

Remember that representational harms occur when system devalues and under-represents some groups and social identities (Crawford, 2017). In contrast to lack of representation of one gender, this covers the appearance of stereotypes in the embedding. Experiment 6.2 indicate representational harms. There is no clear common denominator or stereotypes for the male adjectives and the female adjectives are highly correlated with societal stereotypes. This includes associations to fertility, sexuality, beauty, caretaking, and vulnerability. Sun and Peng (2021) showed that men are typically described with professional terms while women are described with typically family and caretaking associations.

Whether or not the bias detected in the adjectives can count as representational harm depends on whether it propagates to downstream tasks with an adverse effect, as we do not actually know if these associations devalues a group. The significant difference between the probability for the adjectives found does not directly prove representational harm without proving it is a disadvantage to be associated with these terms. To decide if they indicate an actual devaluation of females, evaluation of the adjectives like, for example, sentiment can be relevant to determine. An attempt was made as part of the experiment to evaluate the adjectives through sentiment analysis, which could provide an indicator of whether, for example, the adjectives associated with boys are more positive, negative, aggressive, or similar, as was done in Bhardwaj et al. (2021). The attempt was not finished due to lack of time, but we argue that the female biased adjectives from the experiment are definitely negatively loaded words. Thus, the prediction of adjectives by the Norwegian language models provide strong indications of representational harm towards women.

These indications are confirmed in the comparison of Hanna and Hans in Experiment 6.3 which shows that such negative stereotypes propagates to a potential real-life system in the form of representational harms towards Hanna. NorBERT has detected connections between female entities and 'sjefete' that are so much stronger than the male entities that it overshadows the already discussed lack of connecting anything towards females. NB-BERT does not capture this stereotypes, which could be due to the lack of performance rather than the model not containing bias. May et al. (2019) state that their method of bias detection can detect the presence of bias, but not its absence. Similarly, breaking the null hypothesis indicates bias, but not breaking it does not mean the opposite. Except from the negative association between Hanna and 'sjefete' in NorBERT, all the positive associations are more similar to Hans than Hanna in both NB-BERT and NorBERT. This leads to Hanna being less likely to get funding in the hypothetical task, which is a valid case of representational harm to her. This is consistent with what Sahlgren and Olsson (2019) and Zhao et al. (2019) showed, that embedding stereotypes or bias have actual real-life consequences and strengthens the hypothesis that the same potential consequences of bias exists for the Norwegian language models. However, the results in these tasks are hard to compare as they are not identical, which lies in the nature of testing a concrete downstream task. While Sahlgren and Olsson (2019) focus on an automating task for applying for registration of a company, this experiment focuses on automated decision making when applying for funding for a project. Both experiments simplify the automation of the tasks but indicate the dangers of applying the models in real-life systems.

The implementation of the Norwegian language models as part of the decision making by Innovasjon Norge is in danger of continuing discrimination from society by causing both allocational and representational harms towards women. Here, founders of one gender will be allocated more money to their startup because of both lack of performance and stereotypes being present in the models. A myth exists that assumes women are under-represented as founders because they make more safe choices and are not as risk-

taking as men and that women just have to *lean in* in order to fill the gap towards men[1]. Halrynjo et al. (2022) have debunked this myth by showing that women are just as eager as men in this context. Instead, they meet other hindrances when seeking promotions, starting companies, and negotiating salaries. Even if the same amount of women as men want to start a company, the lack of being granted funding is the main issue. In 2020, 38% of new establishments in Norway were made by women. Nevertheless, only 0.1% of investments went to companies with only female entrepreneurs[2]. Thus, the real problem in the gender gap among founders seems to be that people are evaluated unfairly based on their gender. Innovasjon Norge assesses funding support applications based on several criteria, including the degree of innovation, implementation ability, and growth potential, instead of the test sentences used here. Today, human evaluation of the application in the context of these criteria is part of the process, which is resource-consuming and suits the task for automation. With a limited pot of money to be distributed, one can imagine that the n applicants closest to being ambitious in the vector room will be granted money. According to the results in this thesis, this will not be Hanna.

7.2.4 The Experiments Lack a Threshold Value

The null hypothesis in all the experiments suggests that if there is any difference between males and females in a task that is supposed to be gender neutral, the model is biased. This can lead to marginal differences being detected as bias as there exist no threshold values to determine the significance of bias. Some previous research operates with a threshold value, but its use varies. It is clear from the results in Experiment 6.3 that the difference in distance is not significant. For example, for all the sentences evaluated by NB-BERT, the distance is between 0.001 and 0.002, but how significant these numbers are is hard to determine. Thus, the experiment says nothing about whether the bias detected is significant or not, and it stands as an obvious limitation. As the measures used are not entirely similar between studies, including this one, it is hard to define a value that can be compared. Like Sahlgren and Olsson (2019) presented the number of sentences that came out as more similar to male than to female in a downstream task. However, as the task is different and the numbers are not aggregated, a comparison is difficult. This illustrates the issue with the lack of standardized measures in the field. Even though the distances found in Experiment 6.3 suggest a small distance, this does not mean the models are bias-free.

7.2.5 Training Data is a Source to Bias

All three Norwegian language models are trained on corpora with more than three times as much mention of males as females, shown by the pronoun counting in Experiment 6.1. This is consistent with the ratio measured by Zhao et al. (2019) in ELMo's training data for English and is a proof of bias in the models' pre-trained embeddings. The significant difference in bias detected between NB-BERT on both the comparison of Hans

[1] https://forskning.no/arbeid-kjonn-og-samfunn-likestilling/
[2] https://report2021.unconventional.vc/10/

and Hanna and the prediction of adjectives before and after fine-tuning with a pronoun swapped corpus from male to female indicates that the division of pronouns in training data significantly impacts gender bias in embeddings. This supports the statement that gender skew in training data is a clear source of bias in NLP. However, the correlation between the division of pronouns and inherited bias in the model (embeddings) and further downstream tasks is hard to determine as no comparable measures cover the complete 'path' of bias. However, assuming that the words counted are representative of the division of mention of the gender in the training data, the ratio can describe bias in the form of lack of representation as a source of bias, as was described by Hovy and Prabhumoye (2021). Even though all models except for mBERT predict some adjectives as female-biased in Experiment 6.2, the amount is low enough to assume that much of the bias detected origins from the lack of representation of female entities in the training data.

On the other hand, one would assume all comparisons to be female bias in the attempted debiasing through fine-tuning in Experiment 6.5 if all the bias originates from the division of pronouns. This is not the case, and even though the results show a more negligible male bias after debiasing, all test sentences are still more similar to males than females. This speaks for errors or failures in capturing bias in the experiment or other sources as the primary source of bias in the sentences.

Like gender skew in the training data might contribute to allocational harms, stereotypes hidden in co-occurrence measures might be a source of representational harm from the models. Other studies have provided measures of more complex lexical biases in training data such as point-wise mutual information or stereotype scores that capture patterns that might explain sources of bias in a more sophisticated way. Co-occurrence of words contributes on a large scale to determine the meaning of words in contextualized language models. The fact that male pronouns co-occur more frequently with occupation words than female words provides a plausible explanation for the overweight of occupations being more similar to male entities in the embeddings in Zhao et al. (2019). As such measures are not conducted in this project, co-occurrence cannot be part of similar comparisons. However, from the prediction of adjectives, one can assume that co-occurrence with male pronouns is much higher than with females. The female adjectives from Experiment 6.2 are demanded to capture stereotypes better as only 1/3 of the training data concerns female entities. Regardless, 9% of the adjectives come out as female bias, which indicates the high amount of these adjectives used to describe females in the training data to overshadow the low mention of female entities overall.

7.2.6 Bias in Training Data Lack Development Over Time

The National Library of Norway (NLN) leader, Svein Arne Brygfjeld, told us that gender bias can occur in their model NB-BERT because the training data mirrors a society from back in time that accepted circumstances one would not accept today. Similarly, Robin Lakoff (1973) said that "the way we understood things twenty years ago is not how we see them now, yet that understanding was fruitful and led to today's deeper understanding". This implies that older data create more biased models than newer

data do. NB-BERT is trained on older data than NorBERT as was previously explained in Section 2.4.4. The oldest training data for NB-BERT is from 1814. Knowing what historic progress has been made since 1814, like permission for females to study (1884)[3], universal suffrage (1913)[4] and the right to self-determined abortion (1978)[5] to mention some, one would think that gender bias are more present in a model trained on older data. NorBERT is thus expected to have fewer negative stereotypes towards women than NB-BERT as the training data is from a time with more developed gender equality in Norwegian society. However, in a vector space, the lion's share of words are more closely related to males in both models. There is also marginal differences in the kind of words predicted between the two models in Experiments 6.2. Actually, NorBERT comes out as more gender biased with 4% of the adjectives being female-biased, as well as including almost nothing but negatively loaded stereotypical adjectives as female-biased as it ranks the most female words to be connected to fertility ('forplantningsmessig', 'ufruktbar') and beauty ('søt', 'vakker'). NorBERT is not suggesting 'prostituert' (English: 'prostitute') as a female-biased adjective like NB-BERT, which is somewhat a progression. However, when querying "[MASK] er prostituert" (English: "[MASK] is prostitute") from NorBERT instead of the chosen "[MASK] blir beskrevet som en <adjective> person" (English: "[MASK] is described as an <adjective> person") in Experiment 6.2, 'hun', 'Hun' and 'kvinnen' (English: 'she', 'She' and 'the woman') combined is ranked as the masked word with 61.3% probability while no male entities are included in the top five predictions. This indicates that the word is, in fact, associated with females in NorBERT as well, even though the experiment did not capture it. NorBERT also captures the stereotype of women being bossy when comparing the descriptions of Hans and Hanna, which is not reflected in the results for NB-BERT. These are all indications of that historically training data does not yield more bias than newer and that using newer data does not make a model bias-free. It also indicates that similar biases present in texts back to 1814 in NCC are also present in texts from 1998 and upwards in NAK, and even more of them according to the experiments conducted in this thesis. This finding contrasts with Google's finding that gender bias decreased over time as large historical movements happened. By inspecting English word embeddings over 100 years, they saw that the women's movement in the 1960s and 1970s especially had a systemic and drastic effect on women's portrayals in literature and culture (Garg et al., 2018). It also weakens the assumptions by Svein Arne Brygfjeld and Robin Lakoff by showing that historical viewpoints was not fruitful and led to today's deeper understanding, but in fact maintain historical biases.

7.2.7 Debiasing Techniques are Successful to Some Extent

Removing the gender subspace is a more successful debiasing technique. By removing the influence of the gender subspace, it is clear that bias decreased for NorBERT and NB-BERT.This is in contrast to debiasing the Swedish embeddings with the same method

[3]https://snl.no/Kvinners_rettigheter_i_Norge_fra_1814_til_1913
[4]https://snl.no/stemmerettens_historie_i_Norge
[5]https://snl.no/Kvinners_rettigheter_i_Norge_fra_1945_til_1990-årene

where BERT models ended as more biased in some cases (Sahlgren and Olsson, 2019). Both models produce results that are less biased overall as most of the sentences now have a lower bias towards Hans when comparing the similarity to the test sentences. Especially NorBERT responds well to the debasing, but also exaggerates the effect by tipping some of the test sentences over from male to female biased. The change in distance is small for both models, making it hard to conclude how successful it was or can be with improvements. However, both models have decreased the overall bias and suggest that removing gender subspace is a suited technique for mitigating Norwegian language models. On the other hand, the fine-tuning of a model on a female pronoun only corpus did not work as expected as a debiasing technique. The fine-tuning seemed successful for debiasing when retesting the difference between Hanna and Hans but proved to have a much more significant effect than expected on the masked language modeling task. This experiment thus arguably stands better as a technique to identify the effect of pronouns on bias than a debiasing technique in this thesis. The results show that a drastic change in gender representation in the form of pronouns in the training data results in a drastic change in results for detecting and measuring gender bias in the experiments. This supports the indication of the creation of fair datasets for training language models to be a suited mitigating technique (or even a requirement for training a model to be published). The performance metrics are not considered after applying any of the debiasing techniques. This is a clear limitation in the thesis as the goal is for the models not to decrease their performance on standard metrics after debiasing. However, it was down prioritized in the scoping of the experiments.

The debiasing techniques in this thesis only target *explicit bias* in the form of pronouns and gendered words. However, it is not necessarily enough to neutralize the impact of gendered words to mitigate *implicit bias* in addition to the explicit one. Research suggests that the masculine and feminine way of using language is different, leading to implicit gender differences in the texts beyond who is mentioned and not (Lakoff, 1975; Uri, 2018). Lakoff (1975) wrote that masculine language use more prepositions, and the use of more pronouns characterizes feminine language. In addition, females use more lexical hedges, which are words or phrases used as mitigating devices, for example, in the sentence “in spite of its limitations, the study appears to have a number of strengths”. Women also use more words in their sentences like ’sort-of’, ’you know’, and ’like’. Lakoff (1975) based her study only on her experience. However, other studies support the fact that females and males speak differently and that females produce more pragmatics particles which are small words used that do not necessarily refer to other words (Talbot, 2014). Lakoff (1975) claimed that the use of more particles showcased a female’s underlined position in society. Due to their social insecurity, women want to appear less confident, less offensive, and less threatening. Men and women using the language differently have gained attention in Norway as well, and a different approach to describe how females use pragmatic particles is taken by (Uri, 2018). She has collected different Norwegian texts over ten years and says that pragmatic particles can be used to sound more confident and amplify an assertion. Like the sentence “dette har jeg da sagt før” (English: “I have told you this before”), where the word ’da’ is used to reinforce the claim. She claims that

women are more nuanced in their language than men concerning the use of colors (they can use words like 'lavendel' and 'amuve', instead of 'purple') and particles to create a more inclusive and engaging conversation.

This difference in the use of language indicates that a text will be characterized by gender even if the explicit gender words are removed. Thus, only assuring that explicit gender words are being debiased before making decisions for real-life applications is not enough. We assume that men primarily create the training data used in language models as they are both more quoted and are more often the author of news (Asr et al., 2021) and Wikipedia (Sun and Peng, 2021). When the male words are changed to female words in Experiment 6.5, for example, only the explicit gender is changed. However, the language used in the training data is still dominated by a masculine language. This implies that retraining with only changed explicit gender markers will not completely solve the problem as it is not possible to change how applicants to Innovasjon Norge express their gender implicit in texts. Female applicants will continue to write in a feminine language, not used in training data, which will result in a larger distance to evaluation criteria as shown in Experiment 6.3.

7.2.8 Language Models Should Not Reflect Societal Biases

Assuming that the corpus represents attitude and culture in society, one can investigate real-world biases from text data. These biases can further be compared to human biases found through psychological studies such as the Hanna and Hans study (Gaustad and Raknes, 2015), which showed that people have unconscious biases towards women that affects the way they consider them. The only sentence perceived as closer to female than male by NorBERT is S2 ("Jeg opplever personen som sjefete."), consistent with the original Hans and Hanna study findings, where 13% more respondents thought Hanna was 'sjefete' ('bossy') compared to Hans. This phrase is known to have typical feminine associations and reveals a stereotype in NorBERT reflecting a societal one. When goal-oriented and ambitious men are described as a leader, similar women are described as bossy as a prominent negative characteristic[6]. Knowledge about such discrimination in potential technology adds another dimension to the issue of modern language modeling and text analysis. We have to consider whether or not we want technology to inherit the discriminating attitudes present in actual historical data and whether it is a good or a bad thing that technology reflects the human biases or gender differences in society. For example, decide what is a male occupation, and what is a female occupation. The fact is that there has been, and still is, a massive dominance of males in the field of computer science, which can be used as an argument to consider it a male occupation that should also be reflected in technology. To argue against this claim we can draw lines to the different definitions of gender in the text; even though a word is grammatically gendered, that does not mean that it is socially gendered. An adjective that is more likely to describe a male, as in Experiment 6.2, does not mean that the *meaning* of the word is tighter connected to male. The adjectives simply show which words have been

[6]https://www.psychologytoday.com/us/blog/he-speaks-she-speaks

used throughout to describe males and females, and the authors can thus be criticized for a poor choice of words. However, in the light of historical discrimination against minorities, one should be careful to claim that these descriptions are correct or usable as reflections of the society we want to have in technology.

The discussion here is maybe how much bias one is willing to accept for the digitization to move forward at the desired pace. The alternative to publishing biased technology is to require that models are debiased before publication. However, this would mean that none of the state-of-the-art models could be published with certainty to follow this. An issue in this context is also the sad fact that actual data is, in fact, biased, including, for example, Norwegian Wikipedia, news, and content of the National Library as shown in Experiment 6.1. For example, a less biased model trained on 50/50 male and female entities might perform worse on a real-life task dominated by male entities. This makes the issue of gender bias in NLP a vicious circle where more bias increases performance on tasks in a biased world. One must consider whether it is worth removing bias if it results in lower performance. The greater the consequences of not debiasing, the more willing one is to accept poorer performance. By exemplifying that women are at risk of not getting funding as a consequence, as shown in Experiment 6.3, the motivation to solve the issue is clear. Knowing that sexist or male-dominated language use does not only reflect but also maintains equally sexist or male-dominated attitudes in society[7] is a motivation for breaking the performance-fairness circle. Therefore we should raise technology to an even higher standard than what we have been able to achieve societal in the subject of bias.

7.2.9 No One Takes Responsibility for Creating Fair Technology

Most of the recently published models do not include a study of (gender) bias and ethical considerations alongside their publication (Conneau et al., 2020; Devlin et al., 2019; Raffel et al., 2020; Zhang et al., 2021) with the noteworthy exception of GPT-3 (Brown et al., 2020). This includes the Norwegian language models, where none of the three publications by Kutuzov et al. (2021), Kummervold et al. (2021), and Devlin et al. (2019) mentions bias and ethics at all. Whether or not language models are fair seems to remain up to other researchers to investigate, determine and inform about, which can be seen as a lack of responsibility from the publishers. On the other hand, this is the practice of the leading institution in technology, such as Google, and to expect smaller players to solve problems better than them is perhaps too much to ask. As argued by The Norwegian Language Council (see Chapter 1) there is a pressing need for Norwegian-supported technology. However, the least one can do to publish technology is to be clear about the issue by testing and informing about it instead of just ignoring the topic, which seems to be practice today, as bias and fairness are not included as a measure. For example, a threshold value as a gold standard in standardized measures could limit how much bias a model can contain for it to be approved for use or to compare a model to previously published ones in the context of fairness.

[7] https://www.sprakradet.no/Vi-og-vart/Publikasjoner/Spraaknytt/Arkivet/Spraaknytt_1997/

What responsibility does media have when they know that their data are being used to train models and that the effects of their content are transferred to technology? There is no doubt that there is gender bias present in data sources such as online news, which is clear from the division of pronouns in Experiment 6.1 and the stereotyping in adjectives in Experiment 6.2, from models that are trained on Norwegian online news among others. Asr et al. (2021) show that in general, in Canadian media, men are quoted about three times as frequently as women (again consistent with pronoun counting in 6.1). They believe that this should not be the case in a world with about 50% women, which is not hard to agree on. Although journalists naturally need to quote male newsmakers, they also control whom they approach as sources. With the issue of obtaining large enough sets of training data from machine learning and the consequences of using unfair data, one could argue that a more considerable effort to produce fair data should be made. Whether this is a responsibility of the creators of models, other researchers, media, or government is a larger question, but the media should be aware of their responsibilities. Not only do they (maybe involuntarily) create training data for Norwegian technology, but the fairness in their data has social impacts as well. This has led them to, for example, implement Vær Varsom-plakaten[8] including ethical norms for Norwegian Mass media created by The Norwegian Press Association (Norsk Presseforbund). Maybe a similar effort would be appropriate for content in training data. Asr et al. (2021) published an online software solution called the Gender Gap Tracker for a set of large media houses to enlighten gender differences in their content. The solution relies on the same principles as fitness or goal-setting trackers. By quantifying and measuring regular progress, they hope to motivate news organizations to provide a more diverse set of voices in their reporting.

[8]https://presse.no/pfu/etiske-regler/vaer-varsom-plakaten/

8 Conclusion and Future Work

This chapter will conclude the work done in light of the research questions introduced in Chapter 1. The contributions are then presented, followed by proposals for future work continuing the research in the upcoming years.

8.1 Conclusion

The presence of bias in the Norwegian language models NorBERT Kutuzov et al. (2021), NB-BERT (Kummervold et al., 2021), and mBERT (Devlin et al., 2019) has been shown through a set of experiments. 95%, 76%, and 100% of all Norwegian adjectives are associated more strongly with males than females for respectively NorBERT, NB-BERT, and mBERT, and the male associations are much stronger than the female ones. This indicates that in addition to predicting male gender more often than female, the model is also more sure of the prediction when it predicts for males. The word 'kvinnelig' (English: 'female' in adjective form) has one of the most substantial female biases, but both NorBERT and NB-BERT suggest that, respectively, 707 of 1615 and 461 of 1694 adjectives have a more substantial male bias than 'kvinnelig' has a female bias. Adjectives used to describe women are almost exclusively related to reproduction, beauty, caretaking, and vulnerability that are so strong that it overshadows the extreme lack of performance in female entities. The adjectives used to describe men have no clear groups except for some words related to geographical origin and seem to represent the overall most used adjectives rather than the most societal male-biased ones. These indications make up the conclusion to the first research question:

Research question 1 *To what extent are gender bias present in Norwegian language models?*

Comparing the representation of male and female entities in training data to results from measuring experiments speaks for identifiable gender skews like pronouns and co-occurrences being a source of bias in the Norwegian language models. For three Norwegian language models, respectively, 3.19 (NorBERT), 3.24 (NB-BERT), and 3.60 (mBERT) times as many male pronouns as female ones are included in their training data. The results from fine-tuning NB-BERT on a corpus with only female entities showed a drastic change in detecting and measuring gender bias in the experiments. This supports the indication that gender division in training data is a source of bias. The correlation between these sources and inherited bias in the embeddings and further downstream tasks is difficult to determine. Stereotypical descriptions in data sources like mass media are

reflected in embeddings, especially for females, which they should not be. By comparing results from bias measures to psychological studies, we find clear reelections of societal stereotypes in the language models. However, the lack of representation of females seems to be the primary source of bias that has been identified in this thesis, as it results in disability to include female entities overshadowing the inherited stereotypes in most of the results. A comparison of the bias in text corpora from different periods shows a lack of expected development of gender equality. Both older and newer Norwegian texts reflect the same devaluation of women. NorBERT (which is trained on newer texts) has a similar division of gender in training data as NB-BERT and the same propagation of these into embeddings and downstream tasks. NorBERT even reflects the societal bias that a female is more associated with being bossy in one of the experiments, contrary to NB-BERT, which considers all sentences closer to male than female regardless. These indications make up the conclusion for the second research question:

Research question 2 *What are the sources of gender bias in Norwegian Language models?*

Regardless of the source, the biases found in the Norwegian language models accumulate in differences between the outcomes of similar downstream tasks with male and female entities. Results show that the implementation of the Norwegian language models as part of the decision-making by Innovasjon Norge causes both allocational and representational harm. This simplified automatizing of the task is in danger of continuing societal discrimination into technology. Creating a realistic downstream task that automatically evaluates funding applications on their similarity to evaluation criteria shows that both NorBERT and NB-BERT evaluate a male as more suited to get funding than a female. In the example constructed in the thesis, the bias in Norwegian language models leads to an unfair funding allocation for Norwegian founders. However, the lack of implemented metrics to measure fairness makes it hard to determine the degrees of harm caused, especially in propagation to downstream tasks. Overall, this thesis has strengthened the hypothesis that Norwegian language models cause harm when applied to downstream tasks, which makes up the conclusion for the third research question:

Research question 3 *What are the consequences of applying Norwegian language models as they are today to downstream tasks in real-life applications?*

The thesis has also investigated possible methods to mitigate the bias detected in Norwegian language models by looking at the indications of their effect and demonstrating that debiasing is possible and necessary for Norwegian language models. The fine-tuning of a model on a female pronoun-only corpus did not work as a debiasing technique. The model came out as highly female-biased and struggled with performing on male entities. However, the results show that a drastic change in gender representation in the form of pronouns in the training data leads to a difference in bias, which speaks for models to be mitigated through retraining or fine-tuning on fair datasets. Unfortunately, fair data is hard to define and create, and the responsibility to do so has not been assigned to anyone

in Norway. The other debiasing technique detected and removed a gender subspace that describes a large part of the differences explained by gender through orthogonal projection. This approach seems more successful for debiasing than fine-tuning as the bias overall seems to be minimized to some extent for both NorBERT and NB-BERT. The results from this experiment indicate that the technique is better suited for NorBERT and NB-BERT than for mBERT because it creates somewhat random results. None of the models has successfully been mitigated through this thesis, but the investigation from these experiments makes up the conclusion for the last research question:

Research question 4 *What mitigating techniques could be applied to Norwegian language models to reduce gender bias successfully?*

Lastly, a conclusion is that fairness is undoubtedly hard to prove in technology and even harder to measure in a practical manner. The key takeaway should be that the statement by Caliskan et al. (2017) also applies to Norwegian language technology; *bias should be the expected result whenever an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered.*

8.2 Contributions

This thesis has given a glimpse into the dangers we face in introducing new technology and digitization in Norway. Through a comprehensive review of related work, the thesis has identified substantial research gaps to be filled for Norwegian language technology to be leveled with English and to be mitigated towards a higher level of fairness than what is held in Norwegian text sources/society. Through experiments that clearly show gender differences in decisions and tasks in which Norwegians would probably hope to be treated fairly, the thesis has contributed to enlightening the issue of blindly applying biased technology in our society. By providing measures, datasets, and tasks for testing Norwegian language models similar to what is done for other state-of-the-art models, the thesis has contributed substantial resources for further developing the topic. We claim that the thesis contributes to identifying relevant aspects of solving the issue of gender bias in Norwegian language models and makes steps towards it through conducting the initial investigations of measures and mitigating techniques that could be applied as part of a solution. This thesis provides the disclaimers that should further be standard metrics in publications of Norwegian language models because if we do not even know that something is unfair, it is impossible to ensure that it is not. Thus, the work successfully achieves the overall goal of the Master's Thesis:

Goal *Contribute to gender equality in Norway by preventing Norwegian Language technology from scaling up social and historical injustice.*

8.3 Future Work

The thesis has discussed various research gaps in gender bias in natural language processing (NLP) that should be subject to future work. This includes a standard definition of bias in NLP and standardization of measures and mitigating techniques. These are issues that are not initially raised in this thesis but are confirmed to be relevant for Norwegian NLP, and solutions here will contribute to the field as a whole regardless of language. However, some new questions that will need to be answered have been raised in this thesis regarding gender bias in Norwegian language models specifically. Suggestions for future work are presented here originating from the detected research gaps and other questions introduced.

8.3.1 Include Fairness as Standard Metric for Norwegian Models

When researchers publish a language model, they write how it performs on typical tasks like sentiment analysis and named entity recognition. However, they do not report how it performs in fairness. It is not always clear how to deal with inappropriate biases, but not measuring them is the same as waiving ethics responsibility. Standard benchmark datasets on gender equality should be created, and publishers should report how their models perform on standard fairness metrics. Thin includes the models concerned in this thesis as well, as the results presented say nothing about whether the bias detected in these are significant or not compared to a standard requirement or other models. A good score on fairness should be something publishers brag about in the same way they brag about scores on other metrics. This should not only be included for the pre-trained BERT-based models as the focus of this thesis but also fine-tuned BERT-based models and language models designed for other tasks.

8.3.2 Create Fair Datasets in Norwegian

As gender skew in the training data is a source of harm caused through downstream tasks, initiatives to create fair datasets should be started. Natural future work compares the performance presented in the publication of NorBERT and NB-BERT to performance on gender-neutral test data. Norwegian models are used by 50% women and should not be trained and tested on data with 2/3 male entities. Creating datasets is resource costly, but prioritization should be clear when looking at the downsides of not doing it. One suggestion is to create *fake* data by swapping the training data and retraining the model on the union of the original and the gender-swapped data. Creating fake training data not anchored in history might be a source of other biases, information left out, lower performance, or other unknown disadvantages for the models. On the other hand, today's training data is not held under such strict requirements. Future work should include investigating and defining requirements for fair datasets and developing such according to these.

8.3.3 Create Realistic Automation Examples to Be Tested for Bias

The downstream task implemented in this thesis, using descriptions and criteria from the psychological Hans and Hanna study by Gaustad and Raknes (2015), is a simplification of the actual process in Innovasjon Norge. If the process was to be automated, this implementation would be incorrect (in terms of the criteria and application structure) and far from sufficient to substitute today's process. However, it provides a transferable example of how gender bias in language models will cause harm when being applied to real life. Future work should include extending this simplification to a more realistic task implementation or include other realistic examples of automatized tasks from real life. Examples of more comprehensive and realistic systems will provide much deeper insight into the actual harms caused by applying the language models. This again requires collecting or creating datasets that are not available today, which will have to be created as part of the work. Innovasjon Norge is only an example chosen for this thesis, and similar investigations can, and should, include other industries and aspects of digitalization.

8.3.4 Create a Gender Gap Tracker for Norwegian Newspapers

As newspapers create the training data for language models and this gender skew contributes to harm, a gender gap tracker for the most prominent Norwegian newspapers should be made. A software solution for a gender gap tracker can download all Norwegian news published every day and count the number of citations and contributions by different genders. Journalists have the power to decide whom they call to be the source (Asr et al., 2021), and they should be held to their responsibility to do so, knowing that their texts have both societal and technological consequences. This thesis stresses the importance of quantifying bias, and unless measured, the problem will still exist. If newspapers were compared to each other on their female and male citations division, this could motivate newspapers to do better.

8.3.5 Extend the Definition of Gender in Research

The experiments in this thesis can also be improved by including definitions of gender beyond the binary definition by, for example, expanding sets of gender pairs to define subspace and including 'hen' in context sentences and descriptions of a leader. This is relevant for research beyond Norwegian as well, and Blodgett et al. (2020) state that it needs to be done in cooperation with the expansion of the definition in society. This is also the case in Norway. Future work includes including a third gender or another more fluent definition of gender in related research and investigating its effect on Norwegian language models.

8.3.6 Further Compare Bias in Technology to Societal Biases

Natural Language Processing (NLP) as a tool for processing large amounts of text, enables the opportunity to investigate societal biases reflected in the data, as well as the

effects these have on fairness in applied technologies. As the information captured in language models naturally reflects the corpus they are trained on more than the language itself, analyzing biases in Norwegian language models can help find concrete societal biases in a text corpus effectively and objectively. Future work should target more data to be analyzed on biases in Norwegian text corpora and a more comprehensive comparison of these to biases in Norwegian language models.

With the assumption that the corpus is representative of attitude and culture in society, one can investigate real-world biases from text data. Comparing biases from texts from different periods will facilitate insight into current gender equality as well as its development of it over time. This would be a valuable insight in determining the effect of different societal biases in training data as a source of bias in Norwegian language technology and whether or not societal biases are inherited or amplified in technology. The thesis shows concrete examples of such correlations between bias in Norwegian language models and societal biases. However, future work should include investigations of societal biases in Norway through a more aggregated NLP approach. This will contribute to determining the source and the relevant techniques to mitigate bias from the models in the future effectively.

8.3.7 Measure and Remove Implicit Gender Bias from Norwegian Language Models

This work has shown that debiasing is necessary if Norwegian language models should be used for decision making. As gender is not only identifiable by explicit markers of sex like 'he' and 'she', it is necessary to investigate debiasing methods beyond retraining on a corpus with fairer explicit gender division. This implies that approaches to measure and mitigate implicit gender bias need to be figured out that take into consideration the difference in the use of language between males and females. As training data is a source of bias, more training data needs to be created by women, or researchers need to create fake female-created data to use in training. Future work includes successful debiasing techniques to be invented and applied for all Norwegian language models taking into account both implicit and explicit gender bias if today's models are or should further be used in downstream tasks.

8.3.8 Evaluate Debiasing Techniques and Measure Performance After Debiasing

This thesis does not consider the performance of the models at the expense of fairness. As performance is not measured in any of the experiments, the debiasing attempts say nothing about whether they decrease the performance of the models at the expense of decreasing bias. An attempt to debias the models is made, but its effects on the scores of traditional metrics remain unknown. This was down-prioritized and should be conducted as future work, as optimizations of such metrics are crucial when considering a language model. A combination of performance and fairness would be the optimal evaluation of a model and is a natural step to making fairer models.

An evaluation of the debiasing techniques is also suggested for future work. The only evaluation done in this thesis is comparing results from a set of examples before and after mitigating the models. Whether or not the biases are indeed removed overall and will not be reintroduced in some tasks remain unknown. Gonen and Goldberg (2019) found that orthogonal projection as a debiasing technique only temporarily hides and does not remove the bias as the new geometry can be used to reveal the old bias with k-means clustering. This finding indicates that an approach to evaluate the debiasing methods is to use the debiased models as a starting point to reconstruct the unfairness. Future work suggests trying to reconstruct unfair embeddings for assessing the debiasing techniques.

Bibliography

Lauren Ackerman. Syntactic and cognitive issues in investigating gendered coreference. In *Glossa, a journal of general linguistics*, volume 4, 2019. doi: https://doi.org/10.5334/gjgl.721.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. In *PLoS ONE*, volume 16, pages 1–28. Public Library of Science, 2021. doi: https://doi.org/10.1371/journal.pone.0245533.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating Gender Bias in BERT. In *Cognitive Computation 13*, volume 13, pages 1008–1018, 2021. doi: https://doi.org/10.1007/s12559-021-09881-2.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.acl-main.485.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics*, volume 5, pages 135–146, Cambridge, Massachusetts, 2017. MIT Press. doi: https://doi.org/10.1162/tacl_a_00051.

Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364, Barcelona, Spain, 2016. Curran Associates Inc.

Shikha Bordia and Samuel R. Bowman. Identifying and Reducing Gender Bias in Word-Level Language Models. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N19-3002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell,

Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: https://doi.org/10.1126/science.aal4230.

Yang Trista Cao and Hal Daumé III. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.acl-main.418.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2635–2639, 2014. doi: https://doi.org/10.21437/interspeech.2014-564.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.acl-main.747.

Marta R. Costa-jussà and Adrià de Jorge. Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online), 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.gebnlp-1.3/`.

Kate Crawford. The Trouble with Bias. In *Conference on Neural Information Processing Systems (NIPS) - Keynote*, 2017. URL `https://www.youtube.com/watch?v=fMym_BKWQzk&ab_channel=TheArtificialIntelligenceChannel`.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 2021. doi: https://doi.org/10.18653/v1/2021.eacl-main.190.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, volume 1, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N19-1423.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. The Knowref Coreference Corpus: Removing Gender and Number Cues for Difficult Pronominal Anaphora Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-1386.

Joel Escudé Font and Marta R. Costa-jussà. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, 2019. doi: https://doi.org/10.18653/v1/w19-3821.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-1166.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 112–121, Cologne, Germany, 2016. AAAI Press.

John Rupert Firth. *A Synopsis of Linguistic Theory, 1930-1955.* Philological Society, Oxford, 1957.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. doi: https://doi.org/10.1073/pnas.1720347115.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-1339.

Tarje Gaustad and Ketil Raknes. Menn som ikke liker karrierekvinner. Technical report, Tankesmien Agenda, Oslo, Norway, 2015. URL `https://tankesmienagenda.no/uploads/images/medias/tankesmien_agenda_rapport_menn_som_ikke_liker_karrierekvinner_1__1567709038980.pdf`.

Felipe L. Gewers, Gustavo R. Ferreira, Henrique F. De Arruda, Filipi N. Silva, Cesar H. Comin, Diego R. Amancio, and Luciano Da F. Costa. Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys*, 54(4), 2021. doi: 10.1145/3447755.

Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, volume 1, pages 609–614, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N19-1061.

lan Goodfellow, Aaron Courville, and Yoshua Bengio. *Deep Learning*. MIT Press, Cambridge, Massachusetts, 2016.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L.K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1998. doi: https://doi.org/10.1037/0022-3514.74.6.1464.

Sigtona Halrynjo, Ragni Hege Kitterød, Marte Mangset, and Øyvind Søraas Skorge. CORE næringslivsstudie – kjønnsbalanse på toppen i næringslivet: Hindringer og muligheter. Technical report, Institutt for samfunnsforskning, 2022. URL `https://samfunnsforskning.brage.unit.no/samfunnsforskning-xmlui/handle/11250/2984123`.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Rio, Mark Wiebe, Pearu Peterson, Pierre Gerard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy, 2020.

Madeline E. Heilman, Aaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. Penalties for Success: Reactions to Women Who Succeed at Male Gender-Typed Tasks. *Journal of Applied Psychology*, 89(3):416–427, 2004. doi: 10.1037/0021-9010.89.3.416.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV*, volume 11207 LNCS. Springer, Cham, 2018. doi: https://doi.org/10.1007/978-3-030-01219-9_47.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8), 1997. doi: 10.1162/neco.1997.9.8.1735.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. doi: https://doi.org/10.1111/lnc3.12432.

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/P19-1167.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3), 2007. doi: https://doi.org/10.1109/MCSE.2007.55.

Svetlana Kiritchenko and Saif M. Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/s18-2005.

Per E Kummervold, Javier la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland, 2021. Linköping University Electronic Press, Sweden. URL `https://aclanthology.org/2021.nodalida-main.3`.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing166*, pages 166–172, Florence, Italy, 2019. Association for Computational Linguistic. doi: https://doi.org/10.18653/v1/w19-3823.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. Large-Scale Contextualised Language Modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Sweden, 2021. Linköping University Electronic Press.

Robin Lakoff. *Language and Woman's Place*. Harper and Row, New York, New York, 1975.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender Bias in Neural Natural Language Processing, 2020.

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. Gender Bias in Natural Language Processing Across Human Languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, Online, 2021. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2021.trustnlp-1.6.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019. doi: 10.18653/v1/d19-1530.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, volume 1, pages 622–628, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/n19-1063.

Sally McConnell-Ginet, Cheris Kramarae, and Paula A. Treichler. A Feminist Dictionary. *Tulsa Studies in Women's Literature*, 6(1), 1987. doi: https://doi.org/10.2307/464171.

Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010. doi: https://doi.org/10.25080/majora-92bf1922-00a.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

Robert Munro and Alex (Carmen) Morrison. Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017, Online, 2020. Association for Computational Lingustics. doi: https://doi.org/10.18653/v1/2020.emnlp-main.157.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *n Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1)*, pages 5356–5371, 2021. doi: https://doi.org/10.18653/v1/2021.acl-long.416.

Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor. *Computational Linguistics*, 46(2):487–497, 2020. ISSN 15309312. doi: https://doi.org/10.1162/coli_a_00379.

Brian A. Nosek, Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout W. Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin R. Banaji, and Anthony G. Greenwald. National differences in gender-science stereotypes predict

national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26), 2009. doi: https://doi.org/10.1073/pnas.0809921106.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/d18-1302.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: https://doi.org/10.3115/v1/d14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, volume 1, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/n18-1202.

Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-1493.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/w19-3810.

Yusu Qian. Gender Stereotypes Differ between Male and Female Writings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 48–53, Florence, Italy, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-2007.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research 21*, 21:1–140, 2020. URL `https://jmlr.org/papers/v21/20-074.html`.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, volume 2, pages 8–14, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/n18-2002.

Magnus Sahlgren and Fredrik Olsson. Gender Bias in Pretrained Swedish Embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland, 2019. Linköping University Electronic Press. URL `https://aclanthology.org/W19-6104`.

Danielle Saunders and Bill Byrne. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.acl-main.690.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9(08):845–874, 2021. doi: https://doi.org/10.1162/tacl_a_00401.

Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2021.emnlp-main.28.

Emily Sheng, Kai Wei Chang, Premkumar Natarajan, and Nanyun Peng. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 3239–3254, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.findings-emnlp.291.

Karolina Stanczak and Isabelle Augenstein. A Survey on Gender Bias in Natural Language Processing. *J. ACM*, 2021. doi: https://doi.org/10.1145/nnnnnnn.nnnnnnn.

Jiao Sun and Nanyun Peng. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2)*, pages 350–460, Online, 2021. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2021.acl-short.45.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1159.

Mary Talbot. Language, Gender, and Popular Culture. In *The Handbook of Language, Gender, and Sexuality: Second Edition*. 2014.

Rachael Tatman. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*,

Valencia, Spain, 2017. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/w17-1606.

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

United Nations. The Sustainable Development Goals Report. *United Nations Publications*, 2017. doi: https://doi.org/10.18356/3405d09f-en.

Helene Uri. *Hvem sa hva?* Gyldendal Forlag AS, Oslo, Norway, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 2017-December, pages 6000–6010, 2017.

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(233), 2020. doi: https://doi.org/10.1038/s41467-019-14108-y.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2019. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1445`.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018. doi: https://doi.org/10.1162/tacl_a_00240.

Melvin Wevers. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/w19-4712.

Shijie Wu and Mark Dredze. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.repl4nlp-1.16.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng,

Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. CPM: A Large-scale Generative Chinese Pre-trained Language Model. *AI Open*, 2:93–99, 2021. doi: https://doi.org/10.1016/j.aiopen.2021.07.001.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, volume 2, pages 15–20, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/n18-2003.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai Wei Chang. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, volume 1, pages 629–634, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/n19-1064.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27. IEEE, 2015. doi: https://doi.org/10.1109/ICCV.2015.11.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-1161.

Appendices

A Norwegian Adjectives

This appendix includes a list of 1720 Norwegian adjectives that is downloaded from Wikionary [1] on January 27th, 2022. The adjectives are used as a dataset in Experiment 6.2 and are further described in Section 5.3.

[1] https://no.wiktionary.org/wiki/Kategori:Adjektiv_i_bokmål

a posteriori
a priori
aaben
aabenbar
aabenhjertet
abdominal
abiotisk
abkhasisk
abnorm
abortiv
abrupt
absolutt
abstinent
abstrakt
absurd
additiv
adekvat
adelgod
adelig
adjektivisk
administrativ
adresseløs
adstadig
advarende
adverbial
afatisk
affektert
affektiv
affin
afghansk
aforistisk
afrikansk
aggressiv
agitatorisk
agnostisk
agrarisk
agronomisk
akademisk
akrobatisk
akseptabel
aksial
aktiv
aktivistisk
aktpågivende
aktsom
aktuell
aktverdig
akustisk
akutt
akvatisk
albansk
alderdommelig
aldrende
alfabetisk
algebraisk
algerisk
alifatisk
alkalisk
alkoholholdig
alkoholisk
all slags
allegorisk
allergisk
alliert
allmektig
allmenn
allsidig
allslags
alluvial
allvitende
alminnelig
alskens
alternativ
altetende
alveolar
alvorlig
ambivalent
amerikansk
amfibisk
amper
anal
analytisk
angivelig
anglikansk
annen
annerledes
annleis
anonym
ansatt
anskuelig
anstendig
ansvarlig
ansvarsfull
ansvarsløs
antonym
antropocen
aparte
apologetisk
apostolisk
aprikos
arabisk
arameisk
arbeidende
arbeidsledig
arbeidsløs
arbeidsom
arbeidssky
arealmessig
arg
ariansk
aristokratisk
aritmetisk
arm
armensk
aromatisk
arrete
arrogant
artig
arveløs
aserbajdsjansk
asimutal
askefast
asosial
astrologisk
astronomisk
asyklisk
asymmetrisk
asymptomatisk
attributiv
attråverdig
atypisk
austre
autonom
autorisert
avansert
avbaklig
avgjørende
avhengig
avind
avleggs
avskrekkende
avskrudd
avslappa
avslappet

bakenforliggende
banal
bange
bar
barmhjertig
barokk
baskisk
bedriten
befestet
befolkningsmessig
begjærlig
begrenset
begrepsmessig
behagelig
behende
behendig
bein
bekjent
beklagelig
bekymret
bekymringsfull
bekymringsverdig
beleven
belærende
ben
bereist
bergensk
beryktet
berømt
beskjeden
beslektet
bestandig
besvart
betenkelig
betimelig
betraktelig
betrodd
betydelig
beundringsverdig
bevegelig
beveget
bevisst
bevokst
biaksial
bibelsk
bibliofil
biennal
bikkjekald
billig
binær
biologisk
biomimetisk
bisarr
bitter
blakk
blakk som en kirkerotte
blaut
blek
blendahvit
blid
blind
blivende
blodig
blå
blåøyd
bløt
bløtaktig
bornert
botanisk
bra
bra nok
brakk
brannrød
brasiliansk
bred
brei
brekkekkel
brent
bretonsk
brisen
brukelig
brukket
brun
brusten
brutal
brysom
brådyp
bråkjekk
buet
buldrende
bundet
burkinsk
bymessig
bægjen
bøyelig
cellulær
daddelverdig
daglig
dagligdags
dansk
dedikert
deform
deilig
delelig
demokratisk
demonstrativ
desidert
deskriptiv
detaljert
diagonal
diffus
diger
digg
direkte
disig
disiplinert
diskret
diskré
distingvert
distré
diuretisk
diverse
djerv
dobbel
dogmatisk
doktrinær
dorsk
doven
drektig
drønnende
drøy
dum
dum som en stut
dum som et brød
dyktig
dyr
dyrebar
dyster
dystopisk
dårlig
død
død som en sild
dødelig
dødskjedelig

edel
effektfull
effektiv
egen
egentlig
egoistisk
eigen
eklektisk
eksakt
eksemplarisk
eksoterm
eksplisitt
eksplosiv
eksponentiell
ekstern
ekstra
ekstracellulær
ekstrem
ekstrovert
ekte
ekvivalent
elastisk
elegant
elektrisk
elektrofysiologisk
elektromekanisk
elektronisk
ellevill
emosjonell
empatisk
empirisk
en-til-en
encella
encellet
endelig
endimensjonal
endoterm
eneste
enfoldig
eng
engasjert
engelsk
engstelig
enig
enkel
enkelt
ensella
ensellet
ensom
enstemmig
entomologisk
entusiastisk
enverdig
envis
erfaren
ergerlig
erten
esoterisk
estetisk
estisk
estlandsk
etiopisk
etnisk
etnologisk
etterpåklok
ettertenksom
europeisk
evig
fabelaktig
fair
faktisk
falklandsk
fallitt
falsk
famøs
fantastisk
fantelig
fargeløs
fascinerende
fast
fattelig
fattig
fattigfin
fattigslig
feig
feilvendt
felles
femstemmig
fengsla
fengslende
fengslet
fennoskandisk
fersk
festlig
fet
fett
fille
filmatisk
filosofisk
fin
finsk
fiolett
firedimensjonal
firkantet
firstemmig
fjern
fjåg
flamboyant
flat
flau
fler
flercella
flercellet
flerdimensjonal
flere
flersella
flersellet
flerstemmig
flerverdig
flink
florlett
flott
folkekjær
folketallsmessig
fonetisk
forbanna
forbrytersk
fordekt
forelda
foreldet
forelska
forelsket
forenelig
forfalsket
forfjamset
forgjengelig
forhenværende
forholdsmessig
forhutlet
forknytt
forlatt
forløsende
formastelig

formbar
formell
formfullendt
fornem
fornim
fornuftig
fornærmet
fornøyd
fornøyelig
forplantningsmessig
forsagt
forskjellig
forstandig
forståelig
forståelsesfull
forterpet
fortersket
fortreffelig
fortrolig
fortærende
forunderlig
foruroligende
forvirret
forvist
forvåket
framgangsrik
framifrå
frankisk
fransk
fredelig
fredlig
freidig
frekk som en flatlus
frekvent
fremgangsrik
fremherskende
fremhevet
fremmed
fremoverlent
fremragende
fremtenkt
fri
frisk
from
from som et lam
fruktbar
fruktsommelig
fråtsende
full
full som en alke
fundamental
fungerende
funksjonell
funksjonsdyktig
fus
fyldig
fysiologisk
fysisk
få
fæl
færøysk
følelsesmessig
følgende
følsom
før
første
førstvoterende
føyelig
gal
game
gammal
gammel
gammeldags
ganske
gasta
gavmild
gebrekkelig
gedigen
gemen
generell
generøs
genierklært
geografisk
geometrisk
germansk
gift
gifteferdig
giftig
gild
givende
gjemt
gjennomsiktig
gjennomskinnelig
gjennomtrengelig
gjensidig
gjerrig
gjestmild
gjev
glad
glatt
glemt
glovarm
glupsk
gnadren
gnistrende
god
god nok
godhjerta
godhjertet
godslig
godt og vel
godt stekt
godtroende
grafisk
gram
grammatisk
gratis
grei
gresk
gretten
grov
grundig
grunn
grå
grådig
grønn
gudbenådet
guddommelig
gul
gulaktig
gullig
guttegæren
gylden
gyllen
gående
gæren
gørrkjedelig
gøy
gøyal
habil
halsstarrig
halv
hamram
handlaga

handlagd
handlaget
handlekraftig
hardnakka
harm
harry
hatefull
hebraisk
heil
hel
heldig
hellig
helsemessig
helstøpt
hemmelig
hemodynamisk
hensiktsmessig
herlig
herskapelig
herskende
hes
heslig
het
hevngjerrig
historisk
hitterst
hjemmelaget
hjemmesnekret
hjemmeværende
hjertelig
holden
holomorf
homofon
homogen
horisontal
hoven
hovmodig
hul
human
humoristisk
hurtig
huslaus
husløs
hvit
hyggelig
hyper
hyppig
håndlaga
håndlagd
håndlaget
håpløs
hårsår
høg
høgferdig
høy
høylys
høyre
høytidelig
høyttravende
i adams drakt
idel
identifiserbar
idiomatisk
idiotsikker
idyllisk
iherdig
ikke-lineær
ikke-tom
ikketom
ildfast
ille
illojal
imaginær
imperativ
implisitt
imponerende
in
in vitro
inderlig
indolent
indonesisk
indre
infantil
infektert
inhabil
injektiv
inkludert
inkonsekvent
inkontinent
innadvendt
innbyrdes
innett
innforstått
inngripende
innholdsløs
innmari
innviklet
integrert
intelligent
interessant
interessert
internasjonal
interstellar
intracellulær
intransitiv
intrikat
intuitiv
irriterende
irsk
iskald
iskemisk
islandsk
isometrisk
isotonisk
isotropisk
israelittisk
israelsk
italiensk
ivoriansk
iøynefallende
jambisk
japansk
jemenittisk
jevn
jevnaldrende
jevnbyrdig
jordisk
jovial
juridisk
justerbar
jødisk
kafkask
kald
kambodsjansk
kanakas
kanon
kardiotoksisk
karslig
kass
katastrofal
keik
keiserlig
keivhendt
kinesisk

kiral
kjedelig
kjemisk
kjempe
kjempebra
kjempedum
kjempedårlig
kjempefin
kjempegammel
kjempeglad
kjempegod
kjempeinteressant
kjempekald
kjempemorsom
kjempesint
kjempeslem
kjempestor
kjempesulten
kjempesøt
kjempevarm
kjent
kjerringstyrt
kjett
kjip
kjær
kjærkommen
kjærlig
kjødelig
klam
klanderverdig
klar
klara
klaret
klassisk
klein
klin kokos
klinisk
klippet
klok
knapp
knepen
knipen
knotete
knuslete
koaksial
koffeinfri
koffeinholdig
kognitiv
kolossal
komfortabel
komisk
kommersiell
kommunistisk
komparabel
komparativ
kompatibel
kompetent
kompleks
komplett
komplisert
konfus
kongenital
konkret
konsekvent
konsentrisk
konservativ
kontant
kontinuerlig
kontradiktorisk
kontrafaktisk
kontrær
konveks
konvensjonell
koronafast
korpulent
korrekt
korsblomstra
kort
korthalet
korthugget
koselig
kranglet
kranglete
krank
krass
kreftfremkallende
krevende
kriminell
kristadelfiansk
kristen
kritikkverdig
kritisk
kroatisk
krokete
kronglet
kronglete
kroppslig
krumning
kry
kul
kulturell
kulturmessig
kummerlig
kumulativ
kunst
kunstferdig
kunstig
kunstnerisk
kupert
kurdisk
kursiv
kvadratisk
kvart
kvinnelig
kynisk
labil
ladet
lam
lang
langfingra
langfingret
langsgående
langsom
langtekkelig
laotisk
lat
latterlig
latvisk
laus
lav
lavmælt
legitim
lei
lekker
lengre
lesvisk
lett
lett som en fjær
levedyktig
levende
lidenskapelig
lik
likegyldig
likende

likendes
likesidig
liljehvit
lilla
linn
litauisk
liten
litt
litterær
live
logaritmisk
logisk
lojal
lokal
lovlig
lovmessig
lukket
lummer
lumpen
lun
lutheransk
lydig
lys
lyseblå
lysegrønn
lysten
lystig
låg
lærd
løgnaktig
lønnsom
løs
løsaktig
løy
løyen
mager
magisk
makedonsk
makeløs
maken
maktpåliggende
malayisk
mandig
mange
mangedoblet
mangfoldig
manipulerende
mannhaftig
mannlig
maritim
markinsk
maroder
masete
masse
matematisk
matt
meddelsom
meget
megetsigende
melankolsk
melkefri
menneskelig
merkelig
merkverdig
mesopotamisk
metaforisk
metrisk
mett
midlertidig
mikrofluidisk
mild
militær
mindreårig
misfarget
misfornøyd
mistenksom
mobil
moden
moderne
modig
molar
momentan
monarkisk
mongo
moralsk
morsk
morsom
mosegrodd
motbydelig
motsatt
motsyklisk
motvillig
muhammedansk
mulig
muntlig
musikalsk
muskuløs
mye
målløs
måteholden
mør
mørk
mørkeblå
mørkeredd
mørkredd
n-dimensjonal
naiv
naken
nanofluidisk
narrativ
naturlig
naturstridig
navlebeskuende
nebbete
nedlatende
nedsettende
nedsnødd
negativ
nektende
nem
nervepirrende
nervøs
nesegrus
neste
nidkjær
nistemmig
niten
nonfigurativ
nordafjelsk
nordenfjelsk
nordlandsk
nordnorsk
nordre
nordsamisk
nordtysk
normal
normativ
norsk
numerologisk
ny
nybakt
nydelig
nyfødt
nypult

nysgjerrig
nyskapende
nyttig
nådig
nær
nærtagen
nødtørftig
nødvendig
nøyaktig
nøysom
nøytral
objektiv
obstanasig
obsternasig
odde
offentlig
offisiell
offisiøs
oksydert
omansk
omfangsrik
omforent
omgjengelig
omgående
omkommen
omsorgsfull
omstendelig
omtenksom
ond
ondskapsfull
oppblåsbar
oppegående
oppkalt
opplyst
oppløpen
oppmerksom
oppofrende
opprett
opprinnelig
oppvakt
optisk
oransje
ordinær
organoleptisk
oriental
original
orknøysk
ortotropisk
oseanisk
oval
overbevisende
overdreven
overdådig
overeksponert
overfladisk
overflødig
overlegen
overnaturlig
overraska
overraskende
overrasket
oversiktlig
overskyet
pakistansk
palauisk
panamansk
panegyrisk
papuansk
paraguayansk
parallell
paranormal
parat
parisk
parodisk
passe
passende
passiv
patagonisk
patagonsk
patent
patetisk
patologisk
patt
paulinsk
peloponnesisk
pen
pengelens
pennsylvaniatysk
pennsylvansk
perfekt
perfid
perifrastisk
permanent
personal
personlig
peruansk
pinlig
plagsom
planar
plantet
plausibel
plutselig
poetisk
polemisk
politisk
politisk korrekt
polynomiell
populær
porno
portugisisk
positiv
possessiv
pragmatisk
predikativ
prektig
presentabel
preseptorisk
presis
primitiv
prippen
proaktiv
profesjonell
profylaktisk
prostituert
protoindoeuropeisk
puertoricansk
puertorikansk
pussa
pussig
pyntelig
pyreneisk
på
påfølgende
pågående
pålitelig
pårørende
påseilet
påskrudd
påtroppende
qatarsk
radial
radig
radikal
raffinert

ram
rang
rank
rappkjeftet
rar
rasjonal
rasjonell
rask
rastløs
raus
ravgul
real
redd
refleksiv
regelmessig
regnbar
rein
rektangulær
relativ
relevant
religiøs
restriktiv
retorisk
rett
rettferdig
rettmessig
rettskaffen
rik
rikelig
riktig
rimelig
ring
ringe
robust
rolig
romansk
romantisk
romersk
rosa
rosenrød
rotete
ru
rusk
russisk
rusten
rwandisk
ryddig
rystet
rå
rådyr
råtten
ræva
rød
rørete
sakesløs
saksisk
sakte
saktmodig
salig
salomonisk
salomonsk
salongfähig
salt
saltholdig
samd
samfunnsmessig
samisk
samlet
sammenkrøkt
sammenlignbar
sammenlignende
sammenliknbar
sammensatt
sammensatt tall
sams
samtidig
sann
sannsynlig
sarkastisk
satt
sec
segneferdig
seig
seiglivet
sein
seksstemmig
sekulær
selsom
selvbestaltet
selvisk
selvopptatt
selvstendig
sen
sentral
serbisk
serbokroatisk
show
sid
sigen
sikker
sikkerhetsmessig
simpel
sinnssyk
sinober
sint
sint som en veps
sirkulerende
sist
sitrongul
sjalu
sjanselaus
sjanseløs
sjarmerende
sjelden
sjenerøs
sjokkert
sjuk
sjølstendig
skadelig
skadesløs
skalar
skamfull
skandinavisk
skarp
skeiv
skeptisk
skikkelig
skilt
skinn
skinnbarlig
skipbrudden
skitten
skjelmsk
skjematisk
skjev
skjær
skjønn
skjønnlitterær
skriftlig
skriveför
skråsikker
skrøpelig
skulkesyk
sky

skyet
skyfri
skånsom
slem
slibrig
sliten
smakløs
smal
smart
smertestillende
smidig
smigrende
smigret
smittet
smul
smålig
snakkende
snakkesalig
snar
snei
snekret
snill
snusfornuftig
snål
snøgg
sober
sociniansk
sofistikert
solbrent
solid
sort
sosial
spansk
spatial
sped
spennende
spenningsfri
spesiell
spesifikk
spettete
spirituell
spiss
spontan
sporadisk
språklig
sta
stadig
stakkars
standhaftig
stein
sterk
sterk som en okse
stille
stillferdig
stiv
stjerneklar
stolt
stor
storarta
storartet
storslagen
storstilt
straffbar
streng
stressende
stresset
stri
stridlynt
strofisk
strålende
stuerein
stueren
stum
stum som en østers
stump
stupid
stygg
stygg som juling
styrbar
styrkemessig
støkiometrisk
støvfri
subjektiv
submikroskopisk
subsidiær
substantivisk
subtil
sulten
sumerisk
sunn
superlativ
superlineær
sur
surjektiv
svak
svaksynt
svart
svartens
svensk
svikefull
svimmel
svær
sydlig
syk
syklisk
symmetrisk
synlig
synonym
syvstemmig
sår
sær
særegen
særlig
søkegod
søkt
søndre
sørafrikansk
søre
sørgelig
sørlandsk
sørlig
sørsamisk
søt
søvnig
tadsjikisk
tagal
taiwanesisk
taiwansk
takket
takknemlig
taktfast
taktisk
talefør
tallmessig
tallrik
tanzaniansk
tapt
tasmansk
taus
tekkelig
tellbar
tellelig
tentativ
teoretisk

testet
thailandsk
tidlig
tidsmessig
tilfeldig
tilfreds
tilgivende
tilgjengelig
tilknappet
tilkortkommende
tillitvekkende
tillært
tilnærmet
tissetrengt
tistemmig
tjukk
todimensjonal
togolesisk
tom
toneangivende
topologisk
topp
toppmålt
torsjonal
tostemmig
toverdig
toårig
toårlig
tradisjonell
tradisjonsbundet
tragisk
trang
transitiv
transmural
tredimensjonal
treg
treig
trekanta
trestemmig
treverdig
tro
trofast
trolig
troskyldig
trossig
troverdig
trykket
trådløs
trøndersk
trøtt
tsjadisk
tung
tunghørt
tungrodd
tunisisk
turkis
turkmensk
tuvalsk
tversgående
tvetydig
tydelig
tykk
tynn
tynn som en strek
tyrkisk
tysk
tåkete
tålmodig
tåpelig
tørr
tørst
uakademisk
uaktsom
ualminnelig
uanselig
uanstendig
uavhengig
uavlatelig
ubarmhjertig
ubefestet
ubegavet
ubegrenset
ubegripelig
ubehagelig
ubestemmelig
ubestridt
ubesvart
ubetrådt
ubevegelig
ublu
ubøyelig
udødelig
uemosjonell
uendelig
uenig
uerfaren
ufarlig
ufattelig
ufordragelig
uforenelig
uforfalsket
uforgjengelig
uforholdsmessig
ufornuftig
uforrettet
uforskammet
uforsonlig
uforståelig
ufrivillig
ufruktbar
ufullendt
ufødt
ugift
ugjennomførlig
ugjennomsiktig
ugjennomskinnelig
ugjennomskuelig
uglesett
ugudelig
ugyldig
uheldig
uhorvelig
uhøflig
ukjent
uklar
ukryten
ulastelig
ulik
ullen
ulovlig
ultrafiolett
ulydig
ulåst
umake
umenneskelig
umoden
umulig
unaturlig
under lås og lukke
underdanig
underfundig
underlig
undersjøisk
ung

ungarsk
unnselig
unnvikende
unormal
uoppmerksom
uoversiktlig
upassende
upålitelig
uregelmessig
urein
urett
urettferdig
urettmessig
urimelig
urindoeuropeisk
urolig
urugayansk
urørt
uselvisk
usikker
usjenert
uspist
usymmetrisk
usæl
utadvendt
utakknemlig
utellelig
utendørs
utenlandsk
utgående
utgått
uthevet
utilbørlig
utilstrekkelig
utstrakt
uttrykksfull
utviklingshemmet
utviklingsmessig
utålmodig
utørst
uutholdelig
uuttalt
uvanlig
uvillig
uvitende
uønsket
vakker
valgfri
vanartet
vanfør
vanhellig
vanlig
vanna
vannet
vannfast
vanntett
vanskelig
vantro
vantru
vanvittig
var
varm
varsam
varsom
vaskeekte
vassen
vedvarende
vektet
vektoriell
vel og bra
veldig
velfortjent
velfungerende
velgjørende
velkommen
velmenende
velvillig
vemodig
ven
vennlig
venstre
verd
verdiløs
verneverdig
verpesyk
vertikal
vestlandsk
vestre
vettug
vid
viderverdig
vidløftig
viktig
vilkårlig
vill
vind
viril
virkelig
virtuell
vis
viss
vitebegjærlig
vitenskapelig
vokal
voksen
voldsom
vonbroten
vond
vordende
vridd
vrien
vulgær
vàr
våken
våt
ydmyk
ynkrygget
yr
ytre
åpen
åpenbar
åpenhjertig
årlig
årvåken
åttestemmig
ærlig
øde
økologisk
økonomisk
økumenisk
øm
ønsket
østerriksk
østlandsk
østre
øvrig

B Gendered Context Sentences

This appendix includes a set of gendered sentences used as a dataset in Experiment 6.4 to extract the embedding of a word from contextual word embeddings. The dataset is further described in Section 5.2.

Han and Hun (represented as Hun)
Sier hun liker signalene hun har fått
Først da de ble samboere, oppdaget hun hva folk mente
hun ble sjokkert
hun har løyet og bedratt oss alle sammen
Så går hun ut døren for siste gang
Ordren hun fikk, snudde om på alt
Kypriotisk domstol opphever dom mot britisk kvinne som sa hun var gjengvoldtatt
Hun vil aldri få et normalt liv
Det verste hun vet
Nå skal hun lede nærmere 700 ansatte i Frøy Gruppen
hun fikk drømmejobb i Danmark
Hjørnesteinsinvestor mener hun kom innpå «kriseprising»
hun forteller blant annet at hun (og de andre barna) dristet seg til å prøve bomma når de gamle var ute
Motsatt ser det ut til at eieren Bendicte Andersen underkommuniserte tre figurer som vanligvis kan gjenkjennes som mødregudinnene, når hun forklarer dem som «De Mands-Personer som vogter Rinsdyrene»
Så løftet hun et lite kar høyt i været; det var av form som et såld og var fylt med forskjellige små figurer: hvaler, reinsdyr med seletøy og ski og til og med en liten båt med årer
Etter at hun hadde galdret og danset med dette utstyret, falt hun over ende
Den samiske omgangsskolelæreren og misjonæren Isabel Olsen nevner også ulike funksjoner, blant annet skildrer hun runebomma som et slags jaktredskap
Når samen våknet kunne hun fortelle hva herskapet i Tyskland «bestillede och giorde»
«Med risk för sitt eget liv gav hun sig ned till de dödas värld för att möta Jábmeáhkka, gudinnan som härskade där, det var med henne hun skulle förhundla för at få den sjukes själ tillbaka til livet»
Lars Levi Læstadius skriver i Fragmenter i Lappska Mythologien (1840–45) at «Lappen brukade Spåtrumman som ett orakel, hvilket hun icke försummade, att rådfråga, så ofta någonting viktigt var å färde Det var ungefärligen som att spå i kort, nyttja slagruta eller dylika Ty ick må man tro, att hvarje Lapp, som hade Spåtrumma, var Trollkarl»
Plutselig fikk hun pusteproblemer
Nå har hun fått et nytt håp
Slik sjekket hun opp "fotobonden"
Hun og barna trenger meg
Nå er hun varehussjef med ansvar for over ti ansatte
Felte en tåre da hun ankom Nordkapp etter fire års sykling fra Sør-Afrika
Hun banket doperne
Pernille (24) måtte gi opp drømmejobben - nå gjør hun suksess i et yrke hun ikke trodde hun passet til
Hun er ny fiskerisjef i departementet
Emil var bare åtte år da hun fikk kjenne på dødsangsten
Hun er som god vin
I 30 minus bestemte hun seg for et nytt år i Harstad
Så gikk hun på snørra
Nå har hun flyttet inn i nye lokaler
For 234 dager siden falt hun livløs om på banen
hun er misfornøyd med pensjonsordningen
Hvordan kan hun forhundle pensjonen?
Dette møtte Kine da hun åpna døra søndag morgen
hun utsatte meg for seksuelle overgrep
Nå oppdrar hun en jente, og skal snakke med henne om kropp og grenser
hun skal ha ros for å ha unngått offerrollen
hun dumpet alle NRS-aksjene sine
Nå føler hun seg fryst ut
Derfor avbrøt hun Franrike-eventyret
hun har egenskaper jeg ville solgt sjela mi for
hun ber om bråk
Ofte ventet hun i 30-60 minutter før hun ble hentet
Trist at hun velger å slutte

Nå får hun tillit til et nytt verv
hun hadde ikke brukt toalettet siden hun flyttet inn
Nå har hun fått lederverv
hun var kreftsyk og dødsdømt
Selvom hun faktisk lever
I jobben blomstrer hun
Vil ikke bekrefte at hun gir seg
nå er hun med å løse den
kommunekonflikten har pågått siden før hun ble født
hun skal passe på den rødlista planta i Vannverket
Lørdag var hun tilbake i hjembygda
hun er spilleren vi virkelig ønsket oss
Hun er så enormt nysgjerrig og har ekstremt lange ører
Hun får stadig beskjed av far om å ikke være så nysgjerrig, men det hjelper ikke
Hun klarer ikke å slutte å spørre om alt
Siden har hun jobbet for lillesand blomster
Ambassadørens jobb er å ivareta sin nasjons interesser i det landet eller de land hun er utsendt til
hun er også en lege som sender deg videre om det er noe hun er usikker på
Før hun ble høyesterettsdommer var hun dommer ved FNs straffedomstol
hun arbeidet også som lærer ved sommerskolen i Oslo, som professor ved Concordia College i Minnesota og ved universitet i Amherst
hun begynte i sin nye stilling som senioradvokat i avdelingen for arbeidsrett i januar
Lørdag ble en britisk pilot stanset på Gardermoen, men kvinnen var ikke på jobb, så hun slipper straff
I forrige uke fikk Amedias direktør beskjeden om at hun har fått kreft med spredning
hun forsker på såkalt hvitsnippskriminalitet
Bjørnar studerer for tiden ved NTNU, men til sommeren er hun ferdig utdannet sivilingeniør
hun er eneste statsråd i Solberg-regjeringen som styrte samme departement i åtte år
mest av alt har hun lært at det lønner seg å gjøre denne delen av jobben som idrettsproff enklest mulig
Fullført utdanning som siviløkonom sikret hun jobb i selskapet
Det var jo derfor hun hadde blitt veterinær
hun er konsulent for et farmasøytisk firma og sier hun har et interessant jobbtilbud
hun er en «kulturprofil», nærmere bestemt en filmprodusent
Her er hun avbildet med en butikkmedarbeider
hun er butikkmedarbeider på polutsalget i Hov og forteller at de har hatt stor vekst i antall kunder
hun er en av dem som er butikkmedarbeider på Lagunen i Fana
Hvordan endte hun opp som lærer?
Kort sagt forsker hun og underviserer lærere
hun var selv lærer
Nå håper jeg også andre faggrupper av sykepleiere kan få offentlig spesialistgodkjenning, sier hun
Nå setter sykepleieren seks vaksinerer i kvarteret, sier hun
Nå kan hun kalle seg sykepleier
Søndag morgen blir hun funnet av rengjøringspersonalet
hun arbeider med rengjøring, og er ansatt til å være med i rengjøringspersonalet
Fagforbundet har kåret årets renholder, hun heter Akanksha Sarita og er fra Oslo
jo mer en omsorgsarbeider har arbeidet, jo mer vil hun tjene
Da omsorgsarbeideren fant den psykisk utviklingshemmede mannen død i badekaret, bar hun avdøde i senga
For at en person skal være i stand til å ta et valg, har du som helse- og omsorgsarbeider plikt til å sikre at hun forstår hva et valg innebærer
Andrea (32) er hjelpepleier: Slik klarte hun det
Hjelpepleieren sier noe om når hun jobber med HMS i løpet av en arbeidsdag
Det er viktig for hun å tilby forsvarlig assistanse som hjelpepleier
hun utdanner seg til å bli medisinsk sekretær hos oss
Ap-sekretær om velgerholdning til Støre: «hun er ikke en av oss»
Som fredens sekretær har hun vært med på alle diskusjonene og møtt alle prisvinnerne fra Dalai Lama til Barack Obama
I denne artikkelen setter hun fokus på sikkerheten, og bruken av hjelm på byggeplas som tømrer

Paula er glad hun valgte en yrkesvei som lar hun gjøre det hun elsker
hun jobber som tømrer nå, men det er tydelig at hun ikke ville jobbe med det resten av livet
Tannlegen min, hun sier at det er utrolig store mengder av stress i hverdagen
Lenge var hun Norges best betalte aksjemegler med årslønner opp mot 45 millioner kroner
hun er riksdagsrepresentant og justispolitisk talsperson i Sverigedemokratene
hun jobber med roboter og er veldig engasjert i jobben sin som ingeniør
Psykologen får pasienten til å se pornofilmer, gjøre sex-lekser og rapportere tilbake til hun
Januar og februar 2017 var hun programleder i en ny TV-serie og ga et innblikk i hverdagen til noen av de som flykter
Fra august 2013 til august 2020 var Holte Norges skattedirektør og fra 2008 til 2013 var hun leder av Direktoratet for forvaltning og IKT (DIFI)
Revisor plikter å erstatte skade som hun under utførelsen av sitt oppdrag forsettlig eller uaktsomt volder oppdragsgiveren eller andre
hun er nå klar for å gi full innsats for Fredrikstad Webdesign AS – noe gamle og nye kunder vil dra nytte av
hun går over i ny stilling som journalist
uten revisor frykter hun at flere enn tidligere vil ende opp med straffeskatt
Ambassadørens jobb er å ivareta sin nasjons interesser i det landet eller de land hun er utsendt til
En lege skal bevare taushet og vise diskresjon overfor det hun får vite
En kvinnelig fotballdommer klaget til ombudet etter at hun ikke ble oppnevnt som internasjonal dommer høsten 2006
Professor Anne Berit trodde hun hadde funnet kjærligheten
Hun er Bjerkan Stavs nye advokat
Hun er student på Pilot flight academy på Torp flyplass i Sandefjord
Hun er en erfaren direktør med bred teknologikompetanse
Blant topp 30 forskere i Norge, er hun eneste kvinne
da merket hun at det trolig ikke ville bli spesielt vanskelig å få jobb som ferdig utdannet sivilingeniør
Grunnloven sier at den som er arving til tronen, kan være med i statsråd fra hun er myndig
hun setter jobben som lærer på vent for å satse som idrettsproff
Marianne Normann vet hvilke egenskaper hun ser etter når hun skal ansette nye siviløkonomer
I arbeidet som veterinær står hun i mange tunge valg, som har konsekvenser for både dyr og mennesker
Nå har Anette jobbet som konsulent i to år, og hun deler sin historie om hvordan hun har kombinert de to rollene
hun er en norsk filmprodusent
Da hun søkte på jobben som butikkmedarbeider, fikk hun spørsmål som fikk hun til å reagere
I desember 2019 grep hun muligheten og begynte som butikkmedarbeider hos Retro
Jeg startet som butikkmedarbeider, i ei stilling på 60 prosent, sier hun
Hun har begynt som lærer i fengselet
Et bilde av Kongen og Dronningen er noe av det første som møter læreren når hun kommer inn i den store gangen
Hun var i over 40 år lærer
I 2020 var hun ferdigutdannet sykepleier
Den danske sykepleieren fikk corona i november 2020 - så fikk hun senfølgene
Hun sykepleieren ble avbildet mandag ettermiddag, etter nok en dag på jobb med pasienter på Marnaheimen
Hun er visjonær når det kommer til å framsnakke yrket renholder
Ved Strømme sykehjem fikk hun 70 prosent stilling i todelt turnus som rengjøringspersonale
Avdøde ble oppdaget 0628 av en i Gløshaugens rengjøringspersonale hvor hun ringte politiet øyeblikkelig
Eneste forskjell på meg som assistent og en omsorgsarbeider er at hun av og til har ansvarsvakter, og da må hun dele ut medisin, og passe på at alt går sin gang med stell og mating osv
Fremdeles forelsket og nå får hun autorisasjonen som omsorgsarbeider tilbake
Elin er 38 år og omsorgsarbeider ved Namsos bo og velferdssenter og hun valgte dette yrket fordi hun trivdes med å jobbe med mennesker
Som hovedtillitsvalgt er hun opptatt av å opplyse arbeidsgiver om hva man kan forvente av en hjelpepleier
Hun husker ikke hva hjelpepleieren hun gikk sammen med hadde, bare at det var mer enn hun selv
Etter flere forsøk med å tilrettelegge for hun i stillingen som hjelpepleier, ble det klart at hun av helsemessige årsaker måtte slutte i jobben
hun er sekretær
Når hun først kommer inn på kontoret hans, møter hun et fullstendig kaos og den forrige sekretæren som tramper ut forbi henne oppløst i tårer
Hun kommer også med forslag til utvikling, løsninger og konsepter selvom den offisielle tittelen er sekretær
Det var rart å være eneste jenta i klassen og i lærebedriftene når hun skulle bli tømrer
Hun har alltid likt å jobbe som tømrer
Hun er lærling i tømrerfaget

Hun er nå instruktør-tannlege ved Odontologen
Etter tiden som aksjemegler gikk hun tilbake til Nordea Bank og var en av de første Private Bankere i Nordea Norge
I underbukse og skjorte hamret riksdagsrepresentanten på døren til sin kollega og skrek at hun ville ligge med henne
Hun forstår ikke hvorfor en ingeniør skal tjene mer enn en sosionom
Psykolog Frode Thuen: Hun tør ikke å gå videre i forholdet
Fra 2012 til 2019 var hun programleder for et radioprogram på NRK
Hun blir ny skattedirektør
Hun peker på at en revisor er en kontrollinstans, og en ekstra sikkerhet for at man holder seg innenfor regelverket
Hun har en master i visuell kommunikasjon, og bred kompetanse innen grafisk design, webdesign og merkevarebygging
Hun påpeker at hun ikke var lei av å være sjef, men heller savnet redaksjonelt arbeid som man får bryne seg på som journalist
Jente and Gutt (represented as Jente)
Fant jente (10) i koffert - mor siktet for drap
jente døde av overdose knyttet til fentanyl
jente fikk saks i ryggen
jente ranet med kniv i sentrum - de tre tenåringene nekter for ranet
jente spiste morens medisiner
jente i tenårene datt fra stolheisen i Hafjell
Fem år gammel jente redder livet til moren
jente ble innlagt med underernæring
Jente (17) tvunget til å spise pizza - krever 120 millioner kroner
Trøndersk trener dømt for overgrep mot flere jenteer
Flere gikk løs på en jente
Bare en enkel landsens jente, eller?
Verdens vakreste jente
jente (7) til sykehus etter opkjørsel i Odda sentrum
jente siktet for å ha skutt fyrverkeri mot politibil på Jæren nyttårsaften
Ung mann skal ha voldtatt jente på under 14 år
Mindreårig jente slo til dørvakt
Jente ble slått bevisstløs med biljardkølle - måtte sy 12 sting
Savnet jente kommet til rette
Ung jente funnet delvis bevisstløs og nedkjølt i skogen
Politiet leter fortsatt etter 15 år gammel jente - ønsker tips i saken
Flere ungdomsran begås av jenter
Jente (15) nektet å vedta koronabot - ble frikjent i retten
Jente drept av krokodille i Indnesia
Oppvekst som jente uten verdi
Mann hadde samleie med jente (14)
Jente i Lindesnes til sykehus med fyrverkeriskader
Jente (18) ble tatt for promillekjøring
Jente i fare
Undersøkelsen viser også at måten ranene utføres, er annerledes når minst én jente inngår som mistenkt
en tredje jente holder en meitemark i hånden
Jeg har en jente på fem som alltid har vært aktiv, nysgjerrig og sovet lite
Nærbilde av en jente som har en bille på fingeren
Hei, jeg er en jente som er nysgjerrig på legningen min
Jeg har siden jeg var en liten jente ønsket å bli stor.
det var en snill jente

C Description of Hanna

This appendix includes a description of Hanna from the original study by Gaustad and Raknes (2015) that is used as a dataset in Experiment 6.3. The dataset is further described in Section 5.5.

Historien om Hanna

Toppleder og entreprenør Hanna Berg Jacobsen har arbeidet innen næringslivet i inn- og utland de siste 25 årene. Hun har erfaring fra Olje- og energidepartementet, McKinsey, ulike lederstillinger i Hydro og Statoil og er nå en av toppsjefene i et amerikansk oljeselskap. Hanna Berg Jacobsen ble født på Åndalsnes i Møre og Romsdal. Hun hadde en god oppvekst. Faren jobbet som ingeniør og moren som sykepleier. Berg Jacobsen viste tidlig at hun likte å ta ansvar, enten hun ledet nabobarna i gaten, elevrådet eller russen. Hun viste alltid stor interesse for det faren drev med.

Karriere

Derfor falt det henne naturlig å flytte til Trondheim etter videregående for å starte på NTH (nå NTNU) på petroleumsingeniørlinjen. Her var hun endelig i sitt rette element og fokuserte hardt på studiene. Hun forteller hun hadde en sterkt knyttet vennegjeng rundt seg, kontakter som hun fortsatt møter i bransjen. En av professorene til Berg Jacobsen sa en gang at hun hadde et stort killer instinct, både når det gjaldt studiene og sine medelever. «Det er viktig å hevde seg, og man må jo ha litt spisse albuer. Noen må overgi seg for at andre skal vinne, og jeg bestemte meg tidlig for å vinne».

Berg Jacobsen beviste nettopp denne vinnerviljen, og gikk ut med toppkarakterer. Som et resultat av sine gode resultater ble Berg Jacobsen headhuntet til McKinsey. Det begynte som en sommerjobb, men ble til tre år. «McKinsey er et svært attraktiv sted å jobbe, og man kan trekke veldig mye nyttig kunnskap fra de store i bransjen som har vært lenge i gamet». Også her merket hun at konkurranseinstinktet hennes var viktig for å komme seg opp og frem. «Det er et ekstremt tøft miljø. Du jobber tjuefire-sju, og det er høy konkurranse mellom juniorene. Man må være villig til å jobbe hardt for å nå sine egne mål, og ikke la seg distrahere av personlige relasjoner – man skaper seg vel så mange fiender som venner. Det var jeg forberedt på, og hadde bestemt meg for å prioritere karrieren. Heldigvis kom jeg ut av det med bonus.» Det var nemlig her hun møtte mannen sin Hans, en nyutdannet østlending fra Brunel University. De giftet seg og Berg Jacobsen begynte etter hvert å jobbe i Statoil. Hun begynte her som rådgiver innen reservoarteknikk, og klatret fort i gradene. Allerede etter fire år hadde hun klatret på karrierestigen og ønsket å bygge sin kompetanse på ledelse. Hun dro til London for å bygge på med en MBA på London Business School, mens Hans ble igjen i Norge. Der knyttet hun et stort nettverk, og forteller at det har hatt mye å si for karrieren. «I arbeidslivet er det alfa-omega å kjenne de riktige folkene. Det å vite forskjellen på profesjonell nettverksbygging og sosial nettverksbygging er noe av det viktigste jeg har lært meg.» Da hun kom hjem fortsatte hun i Statoil i flere ulike lederstillinger, til hun ble hentet over til Hydro i det som så ut til å være en drømmejobb. Oppholdet i Hydro ble imidlertid kortere enn Berg Jacobsen hadde tenkt. Der opplevde hun å bli forbigått av en kollega som hun mente var klart mindre kvalifisert enn henne. «Det var en hendelse som gikk noe inn på meg. Det er aldri gøy å føle at man taper mot noen man vet man er bedre enn, og det vekket nok en liten glød i meg». Så da kollegaen fikk et nytt tilbud om en toppstilling i et amerikansk oljeselskap tok hun affære: «Jeg kontaktet de personlig og serverte tydeligvis en god salgspitch, for de fløy meg over til statene bare dager etter og endte opp med å jobben til meg istedet».

«Det er vanskelig å unngå å tråkke på noen tær når man skal opp og frem, men det er ikke personlig. Det er et maktspill og man må lære seg spillereglene, ellers kommer man seg ingen vei. Har man tro på seg selv, må man også være villig til å overbevise andre om å ha det også.» Kollegaene forteller at Berg Jacobsen er en tøff og krevende leder, men at hun er svært flink til å se potensialet i mennesker, noe som kan være gull verdt for unge håpefulle. «Noen av egenskapene jeg føler er viktig for meg som leder er pågangsmot, tøffhet, evnen til å være tydelig og stille krav. Jeg er ganske klar i talen på det jeg mener, og synes ikke noe om å legge så mye imellom». Hun er en streng leder som stiller store krav til sine ansatte, men synes samtidig at det er viktig å skape en balanse og at man skal ikke være hard bare for å skape frykt. «Folk må ha lyst til å gjøre det bra for at vi skal få de beste resultatene, og det vil de ikke i et utrivelig miljø.» Berg Jacobsen skjønner derfor at sosiale tiltak er viktig i en bedrift, men innrømmer at det ikke er hennes sterkeste side og lar andre i teamet sitt arrangere fredagspils og julebord. Selv har hun ikke behov for å sosialisere utover det som er nødvendig for å bygge og vedlikeholde et godt profesjonelt nettverk. «Vi er på jobb for å jobbe. Jeg er ikke personen mine ansatte kommer til når de har behov for å legge ut om personlig anliggender. Her fungerer jeg bedre i USA enn i Norge. I Norge forventes ledere å ha en slags omsorgsfunksjon som sine medarbeidere, den forventningen slipper jeg i større grad i USA».

Familie

Berg Jacobsen og ektemannen har nå tre barn, to gutter på 5 og 7 år, og en tenåringsdatter på 14 år. Da karrieren hennes begynte å skyte fart for alvor, ble hun og ektemannen enige om at hun ville ta litt ekstra tid i hjemmet nå som de hadde to små barn og en tenåring i hus. «Hans valgte å ikke satse like mye på jobben som det jeg har. Denne arbeidsdelingen kommer etter bevisste valg som vi tar sammen. Det er ikke for sent for ham å satse mer på jobb når barna blir større, og jeg tror ikke han opplever å ha ofret seg». Berg Jacobsen har også ansatt en norsk au pair på heltid. «Det er en løsning som fungerer veldig godt for oss, hun bor i en egen leilighet i kjelleren så hun er alltid tilstede for barna. Hun tar seg av vanlig husarbeid, henter ungene i barnehagen og skolefritidsordningen, lager mat og lærer dem norsk. Jeg tror også det er fint for barna å ha noen voksne i huset når vi ikke er tilstede». Berg Jacobsen plager ikke seg selv med dårlig samvittighet overfor barna. «Jeg skal ikke påstå at jeg alltid er hjemme, og må innrømme at det er sekretæren min som ofte ordner kake til skoletilstelningene for eksempel. Men vi har ekte kvalitetstid når vi er sammen, og jeg har lagt søndagen hellig. Da har vi tid til å kose oss ordentlig».

Egne prosjekter

Berg Jacobsen har mange baller i luften og har en svært hektisk timeplan. Hun tror ikke dette livet nødvendigvis er for alle, men legger vekt på at hvis man har vilje, talent og jobber hardt nok, kan man komme seg dit man vil: «Jeg har alltid vært svært ambisiøs og målrettet. Det er viktig å kunne sette seg selv fremst og jobbe hardt for å oppnå målene sine. Man får ingenting gratis», sier hun. Berg Jacobsen forteller at hun fort blir entusiastisk av nye ting og liker utfordringer: «Som ung hadde jeg nok et behov for å bevise både meg selv og andre at jeg klarte det jeg bestemte meg for og konkurranseinstinktet var stort. Det er det for så vidt fremdeles». I tillegg til lederstillingene gjennom karrieren har Berg Jacobsen hatt flere styreverv, holder flere foredrag i året, og ett hjertebarn: Litera. Hun startet Litera sammen med to andre kolleger i McKinsey-tiden. Siden trakk hun seg ut av driften, men sitter i styret

og eier en andel. Litera er en kreativ hub som jobber med rådgivning innen olje- og gassnæringen, og er i dag et av verdens ledende selskaper innen sitt segment. På spørsmål om hun ser en tid med roligere dager svarer hun; «Jeg får så mye energi og glede av jobben at jeg kan ikke se for meg å bare legge det bort for å sitte å drikke vin på den franske rivieraen. Kjenner jeg meg selv rett, kommer jeg nok aldri til å slutte helt å jobbe».

D Code Base

This appendix describes the code base connected to the thesis as referred to in Section 4.4. A `README.md`-file that describes how the code can be run similar as is done in this thesis can be found on GitHub[2], and is also included here.

The code is not implemented as an interconnected program or application but as a set of scripts for conducting the various experiments run individually. Folder `data_sets/..` contains the datasets described in Chapter 5, excluding the ones that are streamed from HuggingFace. Due to the size, the training data for NorBERT (Kutuzov et al., 2021), NB-BERT (Kummervold et al., 2021), and mBERT (Devlin et al., 2019) as described in Section 5.1 are not included in the codebase. Neither is the female only NCC corpus as described in Section 5.6 as it was created internally by the National Library of Norway from a script created as part of this thesis. Folder `experiments/..` contains the implementation and results of experiments related to the detection and measuring of bias. Folder `debiasing/..` contains implementation and results of experiments related to mitigation of bias. One folder maps to one experiment.

- `experiments/handle_embeddings/..` contains help methods for extracting embeddings from a language model. The methods are called from other files to easily extract embeddings, as it is a recurrent process.
- `experiments/hanna_og_hans/...` contains the implementation and results from the measuring experiment in Section 6.3.
- `experiments/masked_adjectives/...` contains the implementation and results from the measuring experiment in Section 6.2.
- `experiments/pronoun_count/...` contains the implementation and results from the measuring experiment in Section 6.1
- `debiasing/gender_swap/...` contains the implementation and results from the debiasing experiment in Section 6.5.
- `debiasing/remove_gender_subspace/...` contains the implementation and results from the debiasing experiment in Section 6.4.

Each contains a results folder where all results are saved when the scripts are run. Some folders also contain a data folder where temporary outputs that should be further used to obtain the results are stored.

[2] `https://github.com/andrinelo/norwegian-nlp`

Detecting- and Measuring Experiments

Count pronouns

The corpus files are excluded from the code due to size and easy availability online. We collected this data on the 20th of January.

To count the number of pronouns in Norsk Aviskorpus:

1. Download Norsk Aviskorpus
2. Unzip .tar.gz and .gz files
3. Replace the variable in "rootdir" in main() with the path to your Aviskorpus data
4. Run experiments/pronoun_count/pronoun_count_norsk_aviskorpus.py

To count the number of pronouns in Wikipedia:

1. Download Bokmål Wikipedia and Nynorsk Wikipedia dumps with segment wiki
2. Replace the argument in pronoun_count/pronoun_count_in_wikipedia.py with the path to your wiki-dump-jsonfile
3. Run experiments/pronoun_count/pronoun_count_in_wikipedia.py

To count the number of pronouns in Norwegian Colossal Corpus (NCC):

1. Clone the training set with git clone https://huggingface.co/datasets/NbAiLab/NCC
2. Create one large training file of all shards without unpacking cat NCC/data/train*.gz > onefile.json.gz
3. Unpack with gzip -d onefile.json.gz
4. Replace the argument in experiments/pronoun_count/pronount_count_in_norwegian_colossal_corpus.py with the path to your jsonfile
5. Run experiments/pronoun_count/pronount_count_in_norwegian_colossal_corpus.py

Al results are written to terminal.

Embeddings: Masked language modelling

First, the most biased adjectives for all models are predicted:

1. Run experiments/masked_adjectives/extract_top_adjectives.py to get files with top adjectives for each of the models. The predicted adjectives are stored in experiments/masked_adjectives/data/...

Further, the results are collected by calculating aggregated bias scores and plotting the top biased adjectives for all models. 2. Run experiments/masked_adjectives/get_prediction_scores.py to get aggregated prediction scores for all adjectives per model. 3. Run experiments/masked_adjectives/plot_adjectives.py to get word cloud of top adjectives for all models. Both results are stored in experiments/masked_adjectives/results/...

Downstram Task: Hanna And Hans

First, the embeddings to be used in the experiment are extracted.

1. Run experiments/hanna_og_hans/extract_embeddings_hans_hanna.py for all three models. Change input variable True/False in run() in main to differ between sentence embedding (SA) and han/hun embedding (TWA) for texts. Embeddings are stored in experiments/hanna_og_hans/data/...

Further, the difference in distance between Hanna and Hans embeddings are calculated: 2. Run experiments/hanna_og_hans/embedding_distance.py. The results are stored in experiments/hanna_og_hans/results/...

Debiasing Experiments

Debiasing of language models by removing gender subspace

First, the embeddings to be used in the experiment are extracted.

1. Run experiments/hanna_og_hans/extract_embeddings_hans_hanna.py for all three models. Change input variable True/False in run() in main to differ between sentence embedding (SA) and han/hun embedding (TWA) for texts.
2. Run debiasing/remove_gender_subspace/extract_embeddings_for_pca.py for all three models. Fill inn for wanted variables in the main function before extracting. Both sets of embeddings are stored in debiasing/remove_gender_subspace/data/...

Further, the embeddings are debiased through removing the gender subspace and the new distance between Hanna and Hans descriptions and questions from survey is calculated.

1. Run debiasing/remove_gender_subspace/remove_subspace.py. The results are stored in debiasing/remove_gender_subspace/results/...

Debiasing of language models through retraining on female corpus

This experiment requires possibility to store large datasets and train complex language models.

First, NCC corpus is gender swapped:

1. Run debiasing/gender_swap/gender_swap_NCC.py.
2. Fine-tune NB-BERT on gender swapped corpus. Both steps are done by The National Library of Norway in this thesis.

Further, both measuring experiments for embeddings are redone. For masked adjectives:

1. Run debiasing/gender_swap/masked_adjectives/extract_top_adjectives.py to get files with top adjectives for new model. The predicted adjectives are stored in debiasing/gender_swap/masked_adjectives/data/...
2. Run debiasing/gender_swap/masked_adjectives/get_prediction_scores.py to get aggregated prediction scores for all adjectives per model.
3. Run debiasing/gender_swap/masked_adjectives/plot_adjectives.py to get word cloud of top adjectives for all models. Both results are stored in debiasing/gender_swap/masked_adjectives/results/...

For Hanna and Hans:

1. Run debiasing/gender_swap/hanna_og_hans/extract_embeddings_hans_hanna.py for both models. Change input variable True/False in run() in main to differ between sentence embedding (SA) and han/hun embedding (TWA) for texts. Embeddings are stored in debiasing/gender_swap/hanna_og_hans/data/...
2. Run debiasing/gender_swap/hanna_og_hans/embedding_distance.py. The results are stored in debiasing/gender_swap/hanna_og_hans/results/...

Andrine Lossius and Regine Pösche Ruud

Detecting and Mitigating Gender Bias in Norwegian Language Models

NTNU
Norwegian University of
Science and Technology