Master's thesis

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

Gunvor Huso and Ingvild Løver Thon

From Binary to Inclusive

Mitigating Gender Bias in Scandinavian Language Models Using Data Augmentation

Master's thesis in Informatics
Supervisor: Björn Gambäck
June 2023

NTNU
Norwegian University of
Science and Technology

Gunvor Huso and Ingvild Løver Thon

From Binary to Inclusive

Mitigating Gender Bias in Scandinavian Language Models Using Data Augmentation

Master's thesis in Informatics
Supervisor: Björn Gambäck
June 2023

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

The Transformer revolutionised the field of natural language processing, including inspiring the Bidirectional Encoder Representations from Transformers (BERT) and the Generative Pre-trained Transformers (GPT). However, the word representations created by these architectures capture more information about the words than the semantics of the word. It has been shown that language models can exhibit social biases such as gender bias. These social biases can appear as the language models are trained using data from, among others, newspapers, books and web crawling. In this thesis, this is seen from a pronoun count performed on a traditional Norwegian dataset based on newspapers and a Scandinavian dataset based on social media data. The count shows that the traditional dataset contains three times more male than female pronouns and 1305 times more gendered pronouns than gender-neutral pronouns. The dataset gathered from social media is in this sense less biased and contains an almost equal representation of male and female pronouns. There are, however, 55 times more gendered pronouns compared to gender-neutral pronouns.

Gender bias has been detected in Norwegian language models published by the National Library of Norway and the University of Oslo. This requires that research is to be made regarding the mitigation of gender bias in Norwegian language technology. Through an experimental approach, this Master's Thesis mitigates gender bias in Norwegian language models using data augmentation techniques such as gender-swapping and gender-balancing. The results show that neither gender bias nor performance is significantly affected by most of these approaches. This is promising for future debiasing with data augmentation. Furthermore, it was attempted to use transfer learning from Norwegian datasets to Scandinavian language models to mitigate gender bias. The results are promising for decreasing gender bias using transfer learning. Additionally, the performance of the Scandinavian models is unaffected by the transfer learning.

Gender is viewed as a fluid attribute. Yet, research in the field of natural language processing often uses a binary definition of gender when investigating gender bias in language models. This thesis broadens the definition of gender bias by including gender-neutral pronouns when debiasing the Norwegian language models.

Samandrag

Prosessering av naturleg språk vart revolusjonert då Transformaren (the Transformer) vart introdusert og inspirerte både tovegs-omkodar-representasjonar frå Transformarar (BERT) og generative førehandstrena Transformarar (GPT). Ordrepresentasjonane som vert laga av desse arkitekturane fangar likevel opp meir informasjon om orda enn semantikken av ordet. Det har vorte vist at språkmodellar kan utvise sosiale skeivheiter slik som kjønnsskeivheit. Dei sosiale skeivheitene kan kome til syne i språkmodellane etter som dei vert trena på data frå blant anna aviser, bøker og nettgjennomgang. I denne masteroppgåva kan ein sjå dette av pronomenteljinga utført på eit tradisjonell norsk datasett basert på aviser og eit skandinavisk datasett basert på data frå sosiale medium. Teljinga viser at tradisjonelle datasett inneheld tre gonger fleire mannlege pronomen enn kvinnelege pronomen og 1305 gonger meir kjønna pronomen enn kjønnsnøytrale pronomen. Datasettet henta frå sosiale medium er såleis mindre kjønnsskeiv og inneheld ein nærast lik representasjon av kvinnelege og mannlege pronomen. Likevel inneheld datasettet 55 gonger meir kjønna pronomen samanlikna med kjønnsnøytrale pronomen.

Kjønnsskeivheit vart påvist i norske språkmodellar publisert av Nasjonalbiblioteket og Universitet i Oslo. Dette krev at ein forskar meir på måtar å redusere eller fjerne kjønnsskeivheiten frå norsk språkteknologi. I masteroppgåva vert ei eksperimentell tilnærming nytta til å redusere kjønnsskeivheiter i norske språkmodellar gjennom teknikkar der ein gjer endringar på datasetta. For å modifisere datasetta vart det nytta ulike teknikkar som å bytte om på kjønna eller å balansere ut kjønna. Resultatet viser at korkje kjønnskeivheitene eller ytinga vert nemneverdig påverka i dei fleste av desse tilnærmingane. Dette er lovande for framtidig redusering av kjønnskeivheit ved bruk av datamanipulasjon. Vidare vart det forsøkt å bruke overføringslæring frå norske datasett til skandinaviske språkmodellar for å redusere kjønnsskeivheit. Resultatet av denne tilnærminga verkar lovande og ytinga til modellane er nærast upåverka.

Kjønn vert sett på som eit flytande attributt. Likevel vert ofte den binære definisjonen nytta i forsking når ein undersøkjer kjønnsskeivheit i språkmodellar. I denne masteroppgåva vert definisjonen av kjønnsskeivheit utvida ved å inkludere kjønnsnøytrale pronomen når ein reduserer og fjernar kjønnsskeivheiter i norske språkmodellar.

Preface

This Master's Thesis concludes our studies for a Master of Science in Informatics at the Department of Computer Science at the Norwegian University of Science and Technology in Trondheim.

Chapter 2 and Chapter 4 contain text that is reproduced from a delivery in the course *IT3915 - Master in Informatics, Preparatory Project*. In addition, in Chapter 2 parts of the text are reproduced from Ingvild's written assignment in TDT05. IT3915 is a preparation for the Master's Thesis and is a course worth 15 points. Our preparatory project was a literature review of gender bias in natural language processing. TDT05 is a theory module called *Self-Supervised Machine Learning* worth 3.75 points. Course content consists of different topics within self-supervised learning such as learning from co-appearance, masked learning, contrastive learning and applications of self-supervised learning.

We would like to thank our supervisor, Björn Gambäck, for guiding us through this thesis and giving us valuable feedback throughout the entire process of writing it. In addition, thank you to Andrine Lossius and Regine Pösche Ruud for meeting with us to discuss the direction of the thesis. Further, we want to thank AI Sweden, RISE and WASP WARA Media & Language for access to GPT-SW3. Lastly, we want to thank Audun Asdal for proofreading the thesis and Kjell Huso for proofreading the Norwegian abstract.

Gunvor Huso and Ingvild Løver Thon
Trondheim, 1st June 2023

Contents

List of Figures

List of Tables

1. Introduction

Bolukbasi et al. found that gender bias was present in English word embeddings in 2016. Moreover, Lossius and Ruud (2022) found that the Norwegian language models NorBERT (Kutuzov et al., 2021), NB-BERT (Kummervold et al., 2021) and mBERT (Devlin et al., 2019) exhibit gender bias. Debiasing techniques are used to reduce or remove bias from language models. However, there is no standardised way to perform debiasing. This thesis explores data augmentation as an approach to mitigate bias from Scandinavian language models. This chapter will introduce the motivation and background for the research done in this thesis, including human rights, United Nations Sustainable Development Goals, and the Norwegian AI (artificial intelligence) Strategy. Furthermore, research questions and goals are presented based on knowledge gaps found in related work. The experimental research method is presented and the contributions of the Master's Thesis are listed. Lastly, an overview of the thesis structure is presented.

1.1. Background and Motivation

Gender equality is a human right and in the Universal Declaration of Human Rights Article 2 it is stated that: "Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status". Furthermore, in 2015 the United Nations (UN) presented 17 Sustainable Development Goals (SDGs) to be reached by 2030 and gender equality is one of these goals (United Nations, 2015). The goal consists of several smaller targets, one of them being: "End all forms of discrimination against all women and girls everywhere", showing the importance of research regarding gender equality. Moreover, Norway is one of the highest-ranking countries in the world in the category of gender equality and was ranked third in the World Economic Forum's global gender gap report in 2022 (Zahidi, 2022). However, Fjeld (2015) investigated the presence of gender bias in Nordic dictionaries and found that the dictionaries still present an outdated and stereotypical view of women and men and their roles. Furthermore, Lossius and Ruud (2022) investigated two Norwegian language models and one multilingual model and found gender bias present in all of them. This shows that one of the highest-ranking countries still has a lot of

work to do regarding gender equality.

The introduction of the Transformer (Vaswani et al., 2017) led to new research on deep neural networks for natural language processing, making it a "hot" field for both researchers and the media. With the introduction of the competing architecture GPT (Brown et al., 2020), further development in the field ensued. Chatbots and other AI tools based on natural language processing have become increasingly popular. Many instances of models using the Transformer architecture and the GPT architecture have followed, both open- and closed-source. To sum it up; the development of language technology has accelerated and the widespread use of language models among the world population makes it an important field of research. Technology can become harmful if it is used in the wrong way. In this thesis, the focus is gender bias in language technology. This is an important field of study as gender bias in NLP could lead to decisions made on wrongful terms in downstream tasks. An example of this is Amazon's AI-based recruiting tool from 2018 which favoured men over women. The recruiting tool gave a lower score to resumes including the word "women's", meaning resumes containing "women's chess club captain" would be penalized[1]. In addition, two all-women's universities were also penalized, as these also included "women's". When publishing new language models very few of the researchers investigate if gender bias is present in their model, the focus is performance and performing better than state-of-the-art. This leaves the problem with the end-user, who might not have competencies regarding gender bias in natural language processing. Thus unintended discrimination might appear due to a lack of knowledge.

Previous research in natural language processing has mostly been conducted on the English language and language models. In 2021, two Norwegian language models (Kummervold et al., 2021; Kutuzov et al., 2021) were published. However, ethics was not investigated when these models were published. In 2020 the Norwegian government published a national strategy for artificial intelligence (AI) which included seven ethical principles (Norwegian Ministry of Local Government and Modernisation, 2020). One of these principles stated that AI systems should facilitate inclusion, diversity and equal treatment. It is further stated that discriminating bias therefore should be removed from datasets when the data is gathered. As previously mentioned, Lossius and Ruud detected gender bias in Norwegian language models in their Master's Thesis in 2022 and further investigation of gender bias in Norwegian language models is therefore necessary as the Norwegian government stated in their AI strategy. Thus, in this thesis, variations of data augmentation will be investigated as an approach to mitigate gender bias in Scandinavian language models as Norwegian, Swedish and Danish are quite similar

[1] https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

languages. An interesting topic is achieving a sound ratio between gender bias in a model and the performance of a model. Is debiasing the best option if the performance of the model is negatively impacted in a significant way? It is also important to state that today, gender is viewed as a spectrum and not as a binary state, male or female. Furthermore, people identifying as gender-fluid or another gender than assigned at birth are often vulnerable to discrimination which further leads to poorer mental health (Tabaac et al., 2018). This further drives the need for more research on fair language technology as most research on gender bias is done with a binary view on gender.

The motivation of this thesis is thus to contribute to fair language technology given any gender. Simultaneously, it is crucial to investigate how bias can be mitigated in other languages than English, in this thesis specifically in Norwegian, but also in Swedish and Danish. Due to the increasing popularity of artificial intelligence, it is important to make this technology ethical and fair.

1.2. Goals and Research Questions

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

Goal *Mitigate gender bias in Scandinavian language models through data augmentation and broaden the definition of gender in Norwegian language technology.*

Moreover, the goal will be reached by investigating the following research questions:

Research question 1 *How do current mitigating strategies for gender bias affect the performance of Norwegian language models?*

Lossius and Ruud (2022) found that the Norwegian BERT-models, NorBERT (Kutuzov et al., 2021), NB-BERT (Kummervold et al., 2021) and mBERT (Devlin et al., 2019), are gender biased. They also tested out two mitigation strategies but did not investigate the impact this had on the performance of the models. It could be interesting to see how much debiasing affects the performance of a model as this might affect the need for mitigation.

Research question 2 *How do different datasets affect the presence of gender bias in Norwegian language models?*

Training data is a crucial part of why language models exhibit gender bias in the first place. Experimenting with different datasets to see the increase or decrease in both bias and performance related to each other could be of interest. Furthermore, is there a difference in training a model on data retrieved from books compared

to data retrieved from for instance social media? Books could contain attitudes that are no longer present in society, while social media could be a more realistic picture of today's society.

Research question 3 *Do Scandinavian language models exhibit the same gender bias as the monolingual Norwegian models, and can the same mitigation techniques be used on Scandinavian language models and Norwegian language models?*

Sahlgren and Olsson (2019) found that the mitigation techniques proposed by Bolukbasi et al. (2016) amplified gender bias in Swedish word embeddings. Thus, showing that impressive results in one language might not transfer to other languages. It is therefore important to research the field of gender bias in NLP in more than English. As Norwegian, Swedish and Danish are quite similar languages syntactically it could be interesting to see if mitigation techniques that are successful on Norwegian language models show the same success for Scandinavian language models.

Research question 4 *How is gender bias in Norwegian language models affected by the introduction of the gender-neutral pronoun "hen" in the Norwegian language compared to Scandinavian language models and languages?*

"Hen" as a pronoun is a newly introduced word in Norwegian dictionaries and is therefore not used considerably. Thus, the word is sparsely presented in the training data of Norwegian language models. This further introduces a new form of gender bias in the context of natural language processing, as gender bias against people who identify as non-binary. The gender-neutral pronoun "hen" has been in Swedish dictionaries since 2015. This might affect the Swedish language models in a positive direction regarding bias against this group of people, as "hen" might be more common in training data.

1.3. Research Method

For this Master's Thesis, an experimental research strategy has been used to achieve the research goal. An experimental strategy entails performing experiments to prove or disprove a cause-and-effect relationship (Oates, 2006). This is the research method best suited for this Master's Thesis because it can best answer the research questions. For instance, the presence of gender bias in Norwegian language models can be found by conducting experiments and interpreting the results. An experimental strategy may have some challenges, for instance, it might be difficult to control all variables used. After performing the experiments, a quantitative data

analysis is performed. Lastly, the results are discussed in relation to the research goal and research questions.

The Master's Thesis builds upon a preparatory literature review of gender bias in natural language processing. Parts of the work in this thesis are reused or inspired by the text in this previous work. The literature review is the pre-study of this thesis and it will be clearly stated when text from the pre-study is reused.

1.4. Contributions

The main contributions of this thesis can be summed up as follows:

- Contributed to acceptance and equality in society by broadening the definition of gender bias in Norwegian language technology. This was done by including the gender-neutral pronoun "hen" in the mitigation of gender bias.
- Demonstrated that data augmentation techniques such as gender-swapping and gender-balancing maintain the performance of Scandinavian language models.
- Proved that transfer learning between Scandinavian language models can be efficient when mitigating gender bias.
- Provided proof that datasets from social media can be less gender biased compared to traditional datasets from newspapers, books and web crawling.
- Established gender bias as an evaluation metric alongside performance when evaluating language models. This was done by comparing gender bias and performance of the Scandinavian language models after debiasing.

1.5. Thesis Structure

The rest of the thesis is structured in the following manner:

- **Chapter 2 Background Theory**: introduces theory about gender and concepts in natural language processing as word embeddings, neural networks, language models and downstream tasks. In addition, tools and evaluation metrics commonly used in NLP are presented.
- **Chapter 3 Datasets**: introduces Norwegian and Scandinavian datasets used in this thesis. In addition, commonly used large datasets are briefly described. For descriptions of the processing and generation of datasets used in this thesis, see Chapter 5.

- **Chapter 4 Related Work**: presents how gender bias is defined in the context of information technology and NLP. Furthermore, this chapter includes a discussion regarding state-of-the-art techniques to detect and mitigate gender bias in natural language processing. Moreover, the quantification of bias and gender bias with gender-neutral pronouns is debated. Lastly, gender bias in Scandinavian language models is presented.
- **Chapter 5 Experiments and Results**: includes a presentation of the experimental plan for each of the experiments, continued by both the experimental setup and experimental results for each of the experiments performed.
- **Chapter 6 Discussion**: includes an evaluation and discussion of the methods and choices made throughout working with the thesis. Likewise, the results are discussed.
- **Chapter 7 Conclusion and Future Work**: sums up the work and points to ways it can be improved or extended in the future.
- **Appendix A Description of the Code Base**: describes the code base related to the Master's Thesis.

The code base used for the experiments, presented in Chapter 5, can be found on GitHub[2].

[2] https://github.com/ingvlt/master-project

2. Background Theory

In this chapter, relevant background theory will be introduced. Firstly, this chapter introduces gender definitions and gender in the Scandinavian languages as these are important topics for this thesis. For understanding gender bias, both a technical and a social understanding are needed. For the technical understanding; relevant foundational topics in machine learning (ML) and natural language processing (NLP) will be presented. These include neural networks, the Transformer, transfer learning, word embeddings and language models. Building on these foundations, bidirectional encoder representation from Transformers (BERT) and generative pre-trained Transformers (GPT) will be introduced followed by Scandinavian language models. Further, relevant downstream tasks for natural language processing such as named entity recognition and part-of-speech tagging will be introduced. Lastly, tools and libraries that are useful for the project and some evaluation metrics are made familiar. Section 2.1 is reproduced from the preparatory project in IT3915 - Master in Informatics, Preparatory Project and is similar to the final delivery. Section 2.2, 2.3.1, 2.3.2, 2.4, 2.5, 2.7.1, 2.9 and 2.10 are reproduced from the preparatory project but have been modified for the Master's Thesis.

2.1. Gender

The World Health Organization defines gender as "the socially constructed characteristics of women and men" (World Health Organization, 2022). This entails norms, behaviours and roles, in addition to relationships. The definition of gender varies from society to society and can change over time. Sex, on the other hand, refers to the biological and physiological characteristics of females and males.

The gender definition varies between society and linguistics. Stańczak and Augenstein (2021) and Cao and Daumé III (2020) presented respectively four and three distinct categories of gender in linguistics. The three gender categories in common are grammatical gender, referential gender and lexical gender. In addition, Stańczak and Augenstein presented a fourth gender category (bio-) social gender. Grammatical gender is the classification of nouns into different categories. The number of such classes varies between languages and can range from two to twenty (Stańczak and Augenstein, 2021; Theil, 2022). Referential gender is the classification of referents as female, male or neuter. Moreover, lexical gender is the property of

gender that lexical units carry. An example is a father, carrying male property, or waitress carrying a female property. Lastly, (bio-)social gender classifies gender roles based on an individual's characteristics, norms and identity.

One usual way to determine the gender of an individual is by looking at their name. Many languages have gendered name dictionaries which makes it possible to decide the gender. However, most languages have some gender-neutral names and thus this method does not transfer well to all languages. In addition, this method neglects that gender can be fluid (Stańczak and Augenstein, 2021).

2.2. Gender in the Scandinavian Languages

The Scandinavian languages consist of Norwegian, Swedish, Danish, Icelandic and Faroese. In this thesis, the focus is on Norwegian, Swedish and Danish. All of these languages are based on Norse, and the languages are thus quite similar. The focus is Norwegian, Swedish and Danish because these languages have the most native speakers and the most available resources.

In Norway, there are two official written Norwegian languages. These are Nynorsk and Bokmål. Nynorsk is written by a minority of Norwegians and in 2022 only 11.6% of pupils in Norwegian schools used Nynorsk as their primary language (Foss, 2022). As the two written languages are equated in law, Språklova (the language law) states that government agencies should use at least 25% of Nynorsk and Bokmål in commonly available documents to maintain both languages (Språklova, 2022). Nevertheless, many agencies struggle with this, especially writing 25% of the documents in Nynorsk.

There are three grammatical genders in the Norwegian language: masculine, feminine and neuter. These grammatical genders refer to "en/ein", "ei", "et/eit" for Bokmål/Nynorsk. In the later years, there has become a division between Nynorsk and Bokmål, where Bokmål now can operate with only two grammatical genders, common gender ("en") and neuter ("et") (Theil, 2022). The common gender turns feminine words into masculine words.

In Sweden, the standardised way to talk and write is called "Rikssvenska". Swedish only includes two grammatical genders, common gender ("en") and neuter ("ett"). The same is applicable to the Danish standard language where the grammatical gender includes common gender ("en") and neuter ("et").

The personal pronouns used in Swedish, Danish, and Norwegian can be seen in Table 2.1, in addition to English. As can be seen from the table there are some variations between Norwegian Nynorsk and Norwegian Bokmål; however, overall there are similarities between all three languages.

In June 2022 Språkrådet (The Language Council of Norway) agreed upon introducing the gender-neutral pronoun "hen" in the official norms of both Nynorsk and

Table 2.1.: Personal pronouns in Norwegian Nynorsk, Norwegian Bokmål, Swedish, Danish and English.

	Feminine	Masculine	Gender-Neutral
Norwegian Nynorsk	Ho, Ho, Hennar	Han, Han, Hans	Hen
Norwegian Bokmål	Hun, Henne, Hennes	Han, Ham, Hans	Hen
Swedish	Hon, Henne, Hennes	Han, Honom, Hans	Hen
Danish	Hun, Hende, Hendes	Han, Ham, Hans	De
English	She, Her, Hers	He, Him, His	They

Bokmål and the word is now found in both the Nynorsk dictionary and the Bokmål dictionary (Aasmundsen, 2022a). The word is a loan word from Sweden and has been used in the Swedish dictionary since 2015. The inspiration for the use of the word in Sweden came from Finnish where "hän" is gender neutral. In Danish, the singular gender-neutral pronoun, "de", is more similar to the English "they". The different gender-neutral personal pronouns in the Scandinavian languages can also be seen in Table 2.1.

In Norway there have also been discussions about introducing a third juridical gender (Aasmundsen, 2022b), this could influence today's language models as it would appear more often in texts and thus in the training corpora of language models.

2.3. Machine Learning

Machine learning is a sub-field of artificial intelligence where the goal is to make machines learn. This is done by developing techniques and algorithms where machines perform better at certain tasks. The idea with machine learning is that a program or model can solve a task without being programmed for that task specifically. By learning, the program or model can make predictions on new tasks. Training data is given to learn and gather knowledge from, and from there, the model can make predictions and inferences on tasks never seen before. In this section, important topics within machine learning for this thesis will be described. This includes neural networks, the Transformer and transfer learning.

2.3.1. Neural Networks

Artificial Neural Network An artificial neural network is a computing system that tries to mimic a biological neural network. For example, when you see a red

light at a pedestrian crossing, neurons in your brain get signals from the eyes, process them, and send signals to your legs to stop you from walking into traffic. This network of neurons in the brain is the inspiration for the artificial neural network. An artificial neural network consists of three types of layers, an input layer, hidden layers and an output layer, which can be seen in Figure 2.1. A hidden layer can sound mysterious but only means that the layer is not an input or output layer. The input layer consists of input neurons where data is encoded, and the output layer consists of only one output neuron which outputs a value. For hidden layers, on the other hand, there are numerous distinctive design heuristics. Furthermore, in this network, the output of one layer is used as input in the next layer, which means there are no loops. It is called a feedforward neural network.

Recurrent Neural Network The recurrent neural network (RNN) is a neural network where backpropagation is introduced. Backpropagation allows cycles in the network, meaning that the network can learn from mistakes by sending output back into the same neuron as input. The state of a hidden layer is dependent on the value of the hidden layer from a preceding point in time. This makes the architecture sequential and suitable for natural language processing where input can be long sequences of words. When using RNNs as a language model the input sequence is processed one word at a time. To predict the next word using an RNN the current word and the previous hidden state is used. This makes it possible for the RNN to have more context when predicting the next word as it is possible for the hidden state to present information about all the preceding words back to the beginning of the sequence.

However, there are some problems with RNNs, one of them being short-term memory. This means that while RNNs have memory, the network is not able to remember over a more extended period, and thus forgets previous inputs. Another problem with the RNN is exploding and vanishing gradients. This problem arises from backpropagating an error signal back through time. When training, the hidden layers can end up with repeated multiplications which are determined by the length of the sequence. This further results in gradients eventually being driven to zero, and is called the vanishing gradient problem. The exploding gradient problem also arises from backpropagating error signals. However, in this case, the result is an exceptionally large gradient that can make the weights overflow and result in not-a-number(NaN)-values. The explosion happens when multiplying gradients continually through the layers of the network that have values larger than 1. Both problems can result in stopping the learning process. Moreover, the basic RNN is one-directional which means dependency is assumed only one way. This is not always the case, meaning the sequential input makes the RNN slow to train.

The long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is an extension of the network architecture created to address the gradient problems. The

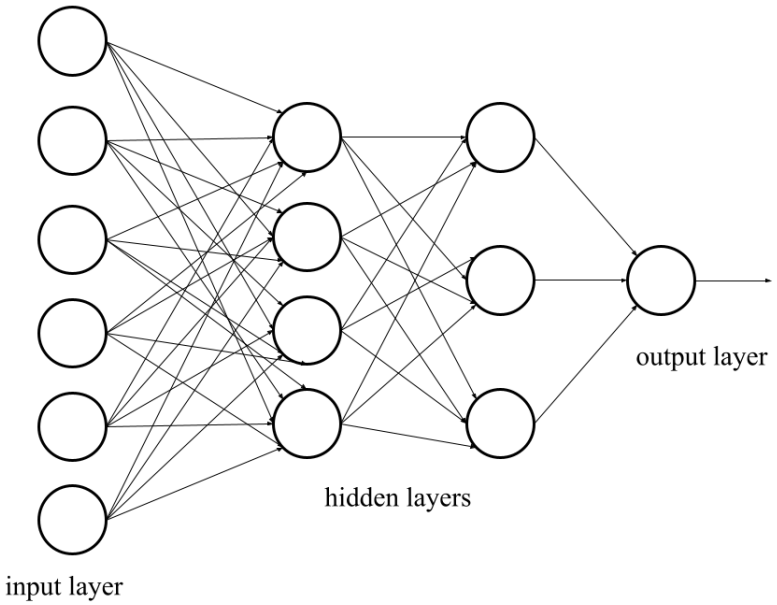


Figure 2.1.: Architecture of artificial neural networks.

goal is to maintain relevant context over time, by learning to forget information that is not useful and remembering information needed for future decisions. Another extension of the RNN is the gated recurrent unit (GRU) (Cho et al., 2014), which is like the LSTM but with a few different parameters.

2.3.2. Transformers

In 2017 the paper "Attention is all you need" was published by Vaswani et al. (2017). It made a significant impact on the natural language processing (NLP) community because of its state-of-the-art performance in machine translation. The Transformer is based on an encoder-decoder architecture and relies only on the attention mechanism, as the paper title suggests. In Figure 2.2 the architecture is shown in more detail. Attention is, simply stated, what to bring focus to. When we read, our attention moves from word to word to form meaningful sentences. In the same way, the Transformer uses attention to weigh the significance of each word or token in an input sequence.

Both the encoder and the decoder blocks are stacked in n layers, Vaswani et al. used n=6. The encoder takes all the input in at once, then the attention matrices are computed. There are numerous operations in the architecture that are overly complex, thus this will be a superficial description. For a deeper understanding, see

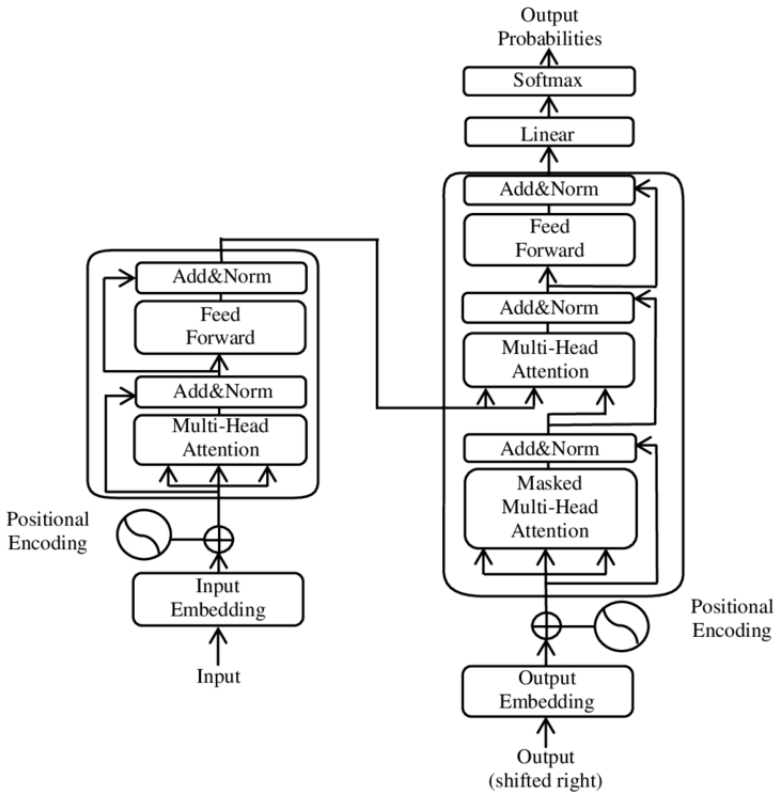


Figure 2.2.: The Transformer architecture by Yuening Jia.
DOI:10.1088/1742-6596/1314/1/012186, CC BY-SA 3.0
URL: https://commons.wikimedia.org/w/index.php?curid=121340680

Vaswani et al. (2017). The most crucial step however is the attention. The output of the encoder is put into the encoder-decoder attention block in the decoder. This attention block is different from the other attention blocks because it is not self-attention, but cross-attention or attention over more than itself. The attention in this architecture is called scaled dot-product attention and each block of attention is built up by multi-head attention, which is several dot-product attention heads. For each token in the sequence, often words, keys, values and a query are calculated. These are used to calculate attention scores and how each token relates to itself or others. The decoder uses the same attention mechanisms as the encoder in addition to a masked multi-head attention block. This block uses masked learning to learn a language by masking out some of the words in the input. The last step in the decoder is the softmax layer, which makes a probability distribution over all the words in the vocabulary in the case of translation. The word with the highest probability is the most likely right answer.

If you were to use the Transformer architecture to translate a sentence you could think that the encoder transforms the input into representations of words and the meaning between them, and the decoder decodes that representation into the other language.

The Transformer became state-of-the-art for numerous NLP tasks and showed promising results in the fields of computer vision and time series forecasting after its publication. By making such an impact, the Transformer is present in countless fields today within machine learning and especially NLP. Since the publication, there have been multiple alterations and models based on the Transformer architecture. A few of these models are presented in this chapter.

2.3.3. Transfer Learning

Transfer learning is a technique used in machine learning to transfer learning from one domain to another domain, also called domain adaptation. This is often done to shorten training times or due to a lack of data in a certain field. If a model is trained for recognising apples and another model is built for recognising oranges, then a lot of the training done on the first model can be reused on the new model. This saves the amount of training and data needed for the orange model because it can build on the part of the other model already trained on apples. The saying that you cannot compare apples with oranges will then no longer be true. The same thought goes for other fields and domains. Even though transfer learning as a concept is a general technique, it will here be presented in the setting of natural language processing.

In transfer learning, there is a **source domain** and a **target domain**. The source domain is the domain you want to transfer knowledge from and the target domain is the domain you want to transfer knowledge to. Considering the above example;

the target domain is pears and the source domain is apples. There could also be multiple source domains, but for simplicity one is used to describe the technique. The **learning task** would be recognising apples and pears, respectively. A domain consists of a feature space X and a marginal probability distribution $P(X)$. A task, on the other hand, consists of a label space Y and a predictive function $f(\cdot)$. Transfer learning can be divided into two main categories; **homogeneous** and **heterogeneous** transfer learning. Each of these has its own solutions and approaches. In homogeneous transfer learning, $Xt = Xs$ and $Yt = Ys$. This means that both the feature space and the label space for the source and target domains are the same. Techniques for solving homogeneous transfer learning include instance-based approaches, feature-based approaches, parameter-based approaches, hybrid-based approaches and relationship-based approaches. In heterogeneous transfer learning, on the other hand, $Xs \neq Xt$. This means that the source domain and the target domain do not have the same feature space. Because of the different feature spaces, only the feature-based approach is used as a solution to heterogeneous transfer learning. Transfer learning between different languages, as done in this thesis, is a heterogeneous transfer learning task as the different Scandinavian languages have different words - feature spaces. Since so many of the words are similar it could be argued that the problem is a homogeneous one.

Notations and definitions are the same as used in Pan and Yang (2010), refer to this for further reading.

2.4. Natural Language Processing

In this section, aspects of natural language processing will be described. Some background theories on word embeddings and language models will be discussed to give a better understanding of the experiments presented in Chapter 5.

2.4.1. Word Embeddings

Word embeddings are used to make words machine-readable. It is a way to represent words through numbers, more specifically vectors. To find the similarity between two words one can calculate the similarity between the vectors of the two words using different similarity measures like cosine similarity, Euclidean distance and Manhattan distance. Words get their embedding by looking at which words they tend to appear next to. Similar words will also have similar vector representations. Bag of Words, Word2vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017) and GloVe (Pennington et al., 2014) are examples of different methods used to create such word embeddings but are however not further discussed in this thesis.

2.4.2. Language Models

A language model is a mathematical model that assigns probabilities to sequences of words (Jurafsky and Martin, 2023). The simplest form of a language model is the n-gram model. An n-gram is a sequence of n words, e.g., "my name is" is a trigram because the sequence consists of three words. The n-gram model is often used for computing the probability of the next word based on the previous n words. The assumption that the next word is dependent on the previous word and only the previous word is called a Markov assumption. This is the assumption the bigram is built upon. The trigram can then be derived from the generalization of the bigram. And thus we can generalize the trigram to the n-gram.

Language models can be multilingual or monolingual. A multilingual model consists of more than one language, while a monolingual model is only trained on one language. mBERT (Devlin et al., 2019) is an example of a multilingual language model, while NorBERT (Kutuzov et al., 2021) is a monolingual Norwegian language model.

Some of the state-of-the-art language models are ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023). ELMo is not further discussed in this thesis.

2.5. Bidirectional Encoder Representation from Transformers (BERT)

BERT (Devlin et al., 2019) is one of the language models that are based on the Transformer. Specifically, BERT is based on the encoder part of the Transformer and is widely used in NLP. BERT was developed by Google and generates context-based embeddings. Furthermore, BERT can be used in text classification, sentiment analysis, machine translation, named entity recognition and question answering. Devlin et al. (2019) stated that BERT made improvements on the state-of-the-art for eleven downstream NLP tasks when published.

BERT has a few different configurations, but the two most common ones are BERT-base and BERT-large. BERT-base has 12 encoder layers, 12 attention heads and 768 hidden units in the feed-forward network. Conversely, BERT-large has 24 encoder layers, 16 attention heads and 1024 hidden units in the feed-forward network. Both BERT-base and BERT-large are trained on 340 million parameters.

Shared for both configurations is how the input data is handled. Firstly, input data is converted into embeddings using the following three layers; token embeddings, segment embeddings and position embeddings. The first token of each sequence is the classification token, [CLS]. A sequence can consist of one or more sentences separated using the token, [SEP]. Token embeddings are used to represent

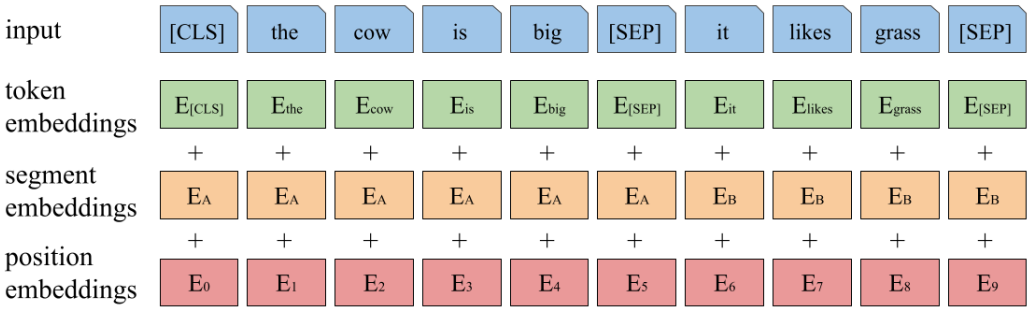


Figure 2.3.: BERT input representation.

individual tokens to vectors. Furthermore, segment embeddings are used to show which sentence a given token belongs to if the sequence consists of more than one sentence. Lastly, position embeddings show the order of the tokens in the input sequence. Figure 2.3 shows an example of the input representation.

BERT is pre-trained and fine-tuned. Pre-training is done with unlabelled data using two different tasks, namely masked language modelling (MLM) and next sentence prediction (NSP). MLM is a widespread technique for training language models. By giving the model a sentence, the model masks some of the words, hiding some of the words from itself. Then the model must guess which words are masked. By doing this the model can learn from its own mistakes by giving feedback on the guess. Next sentence prediction, on the other hand, is when the model gets two sentences and must predict the probability of sentence two coming after sentence one. BERT is pre-trained on the BookCorpus (800M words) and English Wikipedia (2500M words).

Examples of MLM (sentence 1) and NSP (sentence 2 and 3):

Sentence 1: “It is [MASK] today, so I need an umbrella”
Sentence 2: “It is raining today”
Sentence 3: “I need an umbrella”

The first example shows masked learning where a word in the sentence is masked. Here, the masked word can be raining. The next two examples show NSP. We want the model to predict that it is highly likely that sentence 3 follows sentence 2.

Fine-tuning is performed by initializing the model with the pre-trained parameters. Furthermore, these parameters are fine-tuned using labelled data from a given

downstream task.

There are numerous available optimizations and variations of the standard BERT. Some examples include RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019) and various language-specific models such as CamemBERT (Martin et al., 2019), a French BERT-based model.

2.6. Generative Pre-Trained Transformer (GPT)

Generative Pre-trained Transformer (GPT) is a general architecture based on the Transformer architecture. The models are generative, meaning they can produce text, and are pre-trained, to be able to predict the next token in a document. A notable instance of the architecture is the family of language models developed by OpenAI[1]. Further in this section, the special instance of the OpenAI's GPT family will be discussed after a general discussion on training the GPT model with causal language modelling. There are two forms of language modelling; masked language modelling (MLM) and causal language modelling (CLM). As described in the previous section, BERT is an example of a language model which is trained on masked language modelling. This makes BERT adapted for question answering, machine translation and natural language understanding. GPT on the other hand, is trained in causal language modelling which makes it good at text generation. Causal language modelling is similar to masked language modelling where the goal is to predict a masked token. The difference is that CLM can only predict the next token based on the previous tokens, making it unidirectional in contrast to MLM which can be bi-directional. CLM has its name from causality, cause and effect, and reflects the prediction's dependence on the previous tokens. The prediction is an effect of the previous tokens, which is the cause.

GPT-3 (Brown et al., 2020) showed impressive results in several downstream tasks and has gained much attention. The model is trained on 175 billion parameters, which, compared to $BERT_{BASE}$ with 110M parameters, is a lot. GPT-3 is evaluated under three conditions; few-shot learning, one-shot learning and zero-shot learning. Few-shot learning, also called in-context learning, means that demonstrations are only limited by the context window of the model. One-shot learning only allows one demonstration and for zero-shot learning, no demonstration is allowed, only natural language instruction is given. GPT-3 shows the most promising results as a one-shot and zero-shot learner but can compete with the state-of-the-art as a few-shot learner in some tasks. This model was only pre-trained, and fine-tuning was left for future work.

In 2023 OpenAI introduced the newest addition to the family with GPT-4 (OpenAI, 2023). GPT-4 is a multimodal model which takes both image and text

[1] https://openai.com/

inputs and can produce text outputs. The model was first pre-trained using both publicly available data and data from third-party providers and then fine-tuned using Reinforcement Learning from Human Feedback (RLHF). More information on RLHF can be found on Huggingface[2]. GPT-4 has been assessed on different exams originally created for humans and OpenAI (2023) stated that GPT-4 can perform better than most human test takers. Furthermore, OpenAI reported some safety challenges that might become prevalent when publishing a large language model. Among others, these safety challenges include hallucinations, harmful content, harms of representation, allocation and quality of service, disinformation and influence operations, privacy, cybersecurity, the potential for risky emergent behaviours, interactions with other systems, economic impacts, acceleration and overreliance. Measures were taken to mitigate some of these challenges, however, OpenAI still stated that caution should be taken when using the model.

2.7. Scandinavian Language Models

In this section, the different Scandinavian mono- and multilingual models mentioned throughout this thesis will be presented, including mBERT, NB-BERT, NorBERT2, NorBERT3, KB-BERT, SwedishMegatron, GPT-SW3, DanishBERT and DanishRoBERTa.

2.7.1. Norwegian Language Models

In 2021 two Norwegian language models were published, NB-BERT by Kummervold et al. and NorBERT by Kutuzov et al. mBERT (Devlin et al., 2019) was published in 2018 and is a multilingual model based on BERT. The model is trained on Wikipedia articles from 104 languages, including Norwegian Wikipedia. The languages were chosen based on the top one hundred languages with the largest Wikipedia collections. Kummervold et al. estimated the size of the Norwegian Wikipedia used to be around 172 million words. This is not a large amount in relation to NLP and Norwegian might be under-represented in this model.

The National Library of Norway (NLN) published NB-BERT. The model is based on mBERT and is further trained with substantial amounts of historical data, thus being a multilingual language model. NB-BERT is trained on 18.4 billion words, gathered from various sources including books, newspapers, Norwegian Wikipedia, parliament documents and more. Kummervold et al. decided to use the pre-trained mBERT model because this would allow for a working model for both newer texts including loanwords from for example English and the Scandinavian languages, in addition to older texts. Kummervold et al. found that NB-BERT outperformed

[2] https://huggingface.co/blog/rlhf

mBERT in NLP tasks like part-of-speech tagging and named entity recognition. No tests were done regarding the fairness or ethics of the new language model.

NorBERT was published by the Language Technology Group at the University of Oslo (UiO). In the same paper, Kutuzov et al. also introduced NorELMo, however, this model is not discussed further in this thesis. NorBERT was trained on around 1.9 billion words, both Nynorsk and Bokmål from Norsk Aviskorpus and Wikipedia. The model was trained from scratch, different from NB-BERT. Kutuzov et al. stated that the model had much better coverage of Norwegian words than mBERT and NB-BERT, leading to better tokenization. They further compared NorBERT with both NB-BERT and mBERT using different NLP tasks and concluded that both monolingual models performed better than mBERT on most tasks. Kutuzov et al. did not perform any tests of fairness or gender bias on the published model. In 2022 an updated version of this model was published, NorBERT2. NorBERT2 is trained on C4 (Raffel et al., 2020) and the Norwegian Colossal Corpus[3] (NCC) using Whole Word Masking. Whole Word Masking means masking all of the tokens corresponding to a word at once. NorBERT3 (Samuel et al., 2023) was published in 2023. The training data includes Norwegian Wikipedia (both Nynorsk and Bokmål), NBDigital[4], Norsk Aviskorpus, NCC and the Norwegian part of the mC4 corpus (Xue et al., 2020). For NorBERT3-base this means the model was trained using 123 million parameters.

2.7.2. Swedish Language Models

KB-BERT (Malmsten et al., 2020) was created by the KBLab at the National Library of Sweden (KB). It is trained on approximately three billion tokens from various sources, including books, news articles, government publications, Swedish Wikipedia and internet forums. Malmsten et al. (2020) stated that a larger corpus was crucial to increase the performance of the model. Moreover, they found that the model outperformed both mBERT and the Swedish BERT model created by Arbetsförmedlingen. However, the evaluation only included performance tests of specific NLP tasks and social biases were not investigated.

SwedishMegatron is another Swedish BERT model. It was trained using the Megatron-LM[5] library. The training data consisted of 70GB from Swedish newspapers and the OSCAR[6] corpus.

Ekgren et al. (2023) published GPT-SW3. The original model had 3.5 billion parameters and was trained on a Swedish corpus of 100GB. However, the authors continued working on the model and the newest version is trained on a 1.2TB corpus

[3] https://huggingface.co/datasets/NbAiLab/NCC
[4] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-34/
[5] https://github.com/NVIDIA/Megatron-LM
[6] https://oscar-project.org/

called the Nordic Pile (Öhman et al., 2023). The Nordic Pile is further discussed in Section 3.6. Performance was measured in perplexity, and no ethical concerns were considered when publishing the model. In April of this year, the people behind GPT-SW3 released an update to the GPT-SW3 models with instruction-tuned variants of the available models. Instruct tuning was done with inspiration from Ouyang et al. (2022), and concerns fine-tuning models with human feedback. Ouyang et al. used a dataset with prompts and expected behaviour written by human labellers to fine-tune the model. Then they later applied reinforcement learning from human feedback (RLHF) to further fine-tune the model. RLHF was not used to train the GPT-SW3-instruct models. The GPT-SW3 models are not yet open to the public, but researchers can apply for early access[7].

2.7.3. Danish Language Models

DanishBERT[8] was created by the Danish and Spanish technology startup Certainly. It is trained on 1.6 billion Danish words from Danish Wikipedia, Danish OpenSubtitles, and Danish language text from Common Crawl, in addition to data from the two biggest Danish debate forums (dindebat.dk and hestenettet.dk). DanishBERT is a multilingual model trained on Danish, Norwegian and Swedish data.

RøBÆRTa[9] (Danish RoBERTa) is a Danish pre-trained Roberta base model. The model was trained on the Danish part of the mC4 (Xue et al., 2020) dataset and was organized by Dansk Data Science Community.

2.8. Downstream Natural Language Processing Tasks

A downstream task is an application of a language model to solve a task. This means that the actual training of a language model is not a downstream task, however, the application of the model to a problem is a downstream task. There are many such tasks like named entity recognition, part-of-speech tagging, sentiment analysis, coreference resolution, machine translation, text generation and more. In this thesis only named entity recognition and part-of-speech tagging are further discussed.

[7] https://www.ai.se/en/node/81535/gpt-sw3
[8] https://huggingface.co/Maltehb/danish-bert-botxo
[9] https://huggingface.co/DDSC/roberta-base-danish

2.8.1. Named Entity Recognition

Named entity recognition (NER) is a downstream task where the goal is to identify and classify named entities. An entity could be a person, organization, location, product, event or date. In the sentence "Max works in the company Max Fun." you can identify that *Max* is a person and *Max Fun* is a company. It is difficult to write rules to make a computer understand this. Therefore, the solution is to train a language model using labelled data where named entities and types of entities are found in the dataset. The process of NER involves tokenizing the input data and further analysing each token to determine whether it represents a named entity or not. If a token is classified as a named entity, it is further classified into one of the predefined categories.

The AI-lab at the National Library of Norway fine-tuned NB-BERT (Kummervold et al., 2021) on the NorNE[10] (Jørgensen et al., 2020) dataset to create a Norwegian NER-model[11]. NorNE is introduced in Chapter 3. NB-BERT-base-ner can predict the nine different entity types; person, organization, location, geo-political entity, product, event, derived and miscellaneous.

2.8.2. Part-of-Speech Tagging

In part-of-speech (POS) tagging the goal is to mark each word in a text with a tag corresponding to a part of speech, based on the syntactic role of a given word in a sentence. Part of speech is a category of words that have similar grammatical properties. Examples of POS categories are among others nouns, verbs, adjectives, adverbs, pronouns and prepositions. POS tagging can be a challenging task as different words might represent more than one POS at different times. An example of this is the sentences "I went for a walk." and "I like to walk.", where "walk" in the first sentence is a noun, while in the second sentence "walk" is a verb. Thus, the context of the word is important to consider.

Universal Dependencies[12] (UD) is a framework used to annotate data with syntactic tags like part-of-speech (POS) and named entities. In the experiments performed in this thesis, introduced in Chapter 5, the POS tags from UD are used. The framework defines a set of universal POS tags that can be applied to any language, regardless of its specific linguistic features. In Table 2.2 all POS tags in UD are listed.

[10] https://huggingface.co/datasets/NbAiLab/norne
[11] https://huggingface.co/NbAiLab/nb-bert-base-ner
[12] https://universaldependencies.org/

Table 2.2.: Part-of-speech tags in the Universal Dependencies framework.

Tag	Description
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

2.9. Tools and Libraries

Python is used as the programming language for the experiments in this Master's Thesis. Python has many accessible libraries and resources which makes it easy to use, the ones used for this project are further discussed in this section. To gain access to state-of-the-art language models **Huggingface**[13] library *Transformers* (Wolf et al., 2020) is used. The *Datasets* library made loading datasets, both from the Huggingface hub and locally, very easy and efficient. Huggingface is much used in the NLP and machine learning community and almost all of the datasets used for the experiments were accessed through the Huggingface Hub. Most of the models used in the experiments were also accessed through the hub. All libraries in Huggingface are open-source. Among others, Huggingface has a library named *evaluate* which can be used to evaluate both models and datasets. Other important Python libraries used in this thesis are **scikit-learn** (Pedregosa et al., 2011) and **PyTorch** (Paszke et al., 2019). PyTorch is an open-source framework for machine learning and is widely used in the NLP community. PyTorch was used together with the Huggingface libraries. Other alternatives to PyTorch are **Tensorflow**

[13] https://huggingface.co/

(Abadi et al., 2015) and **Keras** (Watson et al., 2022). Huggingface *Trainer* is an API for training models in PyTorch and is part of the Transformer library. Trainer was used in the experiments to simplify the training process.

In addition to these libraries specialized at machine learning, **Pandas** (McKinney, 2010) and **NumPy** (Harris et al., 2020) have been used. Pandas is a Python library that provides data structures and data analysis tools. NumPy makes it easy to compute arrays and matrices in Python and integrates well with Pandas.

2.10. Evaluation Metrics

Evaluation is crucial for finding the effectiveness of a model and measuring how "good" a model performs. It is also useful to compare different models with the same metric. Without evaluation, there is no room for improvement as there is no way to say if the model is bad or good. Evaluation can be divided into two main groups, intrinsic and extrinsic evaluation. Intrinsic evaluation is when the output is evaluated with the criteria for the functionality of the task. This can often be a sub-task in a bigger system. In machine translation, the intrinsic evaluation could for example be how precise the translations were or how understandable it is. Extrinsic evaluation, on the other hand, evaluates the impact the output has on an external task or system where the output of the sub-task may affect another task in the system. Extrinsic evaluation is more complex to measure because it is often tied to a user task. Following, some of the most used intrinsic evaluations will be presented.

2.10.1. Measuring Model Performance

Accuracy is a measure of how accurate the predictions are, meaning how close they are to the true or right values. The formula for accuracy can be seen in Equation 2.1. Accuracy is mostly used in classification, to see how accurate the classifications were. In binary classification, the accuracy formula can be described with the terms true positive, false positive, true negative and false negative, as seen in Table 2.3. Accuracy can be seen as true positives added together with true negatives, then divided by all classifications.

Table 2.3.: Overview of classification terms.

		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \tag{2.1}$$

Precision describes how precise the predictions are, meaning how close the predictions are to each other. If looking at a dart board, high precision is when all darts or bullets are close to each other, regardless of whether they are close to the bullseye or not. If they are close to the bullseye, then that is high accuracy. Precision is often used in information retrieval and classification. The precision formula can be seen in Equation 2.2.

$$\text{Precision} = \frac{\text{relevant elements} \cap \text{retrieved elements}}{\text{retrieved elements}} \tag{2.2}$$

Recall is in information retrieval often used together with precision because they together say something about how good the system is at retrieving the right documents. Recall says something about how many of the predictions were right (true positives) among all the relevant elements. The formula can be seen in Equation 2.3. If a system returns all documents or a program predicts all the elements as relevant then recall is one hundred per cent, which means that recall is not a good measure. Therefore, the F1-score is often used.

$$\text{Recall} = \frac{\text{relevant elements} \cap \text{retrieved elements}}{\text{relevant elements}} \tag{2.3}$$

F1-score, also called the harmonic mean. The formula is shown in Equation 2.4. For a more generic version of the formula where it is possible to weight recall as more or less important than precision, the more general F_β exists, sometimes called weighted F-score. The general formula of F_β can be seen in Equation 2.5. F1 is widely used as an NLP metric, especially when comparing models to each other.

$$\text{F} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{2.4}$$

$$\text{F}_\beta = (1 + \beta^2) \times \frac{\text{recall} \times \text{precision}}{(\beta^2 \times \text{precision}) + \text{recall}} \tag{2.5}$$

Furthermore, F1 can be divided into *macro* F1 score (Equation 2.6) and *micro* F1 score (Equation 2.7). The Macro F1 score is the unweighted mean of the F1 scores calculated per class. Micro F1 score is calculated using the total number

of true positives (TP), false positives (FP) and false negatives (FN), instead of individually for each class.

$$\text{Macro F1 score} = \frac{\text{sum (F1 scores)}}{\text{number of classes}} \tag{2.6}$$

$$\text{Micro F1 score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2} \times (\text{FP} + \text{FN})} \tag{2.7}$$

Another famous evaluation metric is **BLEU** (bilingual evaluation understudy) (Papineni et al., 2002). It is used as a benchmark for machine translation and other tasks such as text summarising and text generation.

Perplexity is a metric that can evaluate how good the predictions of a language model are. It says something about how confused, or perplexed, the model is when predicting the outcome. The lower the perplexity score, the better the language model is.

Training loss and **validation loss** are used in deep learning to describe the fitting of a model, i.e., how well adapted the model is to the data. Together they can indicate underfitting or overfitting in a model. Underfitting is when a model makes big mistakes and is not able to find a pattern in the data. The model is too simple and cannot represent the underlying data structure in a satisfactory manner. Underfitting can occur if a model is not trained for long enough. Another reason for underfitting is a too small training set. In Figure 2.4 the red line represents an underfitted model because it does not represent the data points, it is too simple to capture the trends. Overfitting is when a model is able to model the training data too closely, which can be a sign of copying instead of learning. When an overfitted model is presented with new data it will make many mistakes because it cannot generalize the knowledge it has learnt from the training data. This can happen if the model is trained for too long. Overfitting can occur when the training error decreases and the validation error increases. This can be seen in Figure 2.5. The red line represents the validation error, and the blue line represents the training error. When these two diverge as seen in the figure, it can be a sign of overfitting.

2.10.2. Measuring Gender Bias

Gender bias is challenging to measure due to no collective agreement on definitions or evaluation metrics. This is the same as for many other forms of bias in natural language processing. Bias can also be thought of as subjective, hence it can be challenging to measure objectively. There are no universally effective ways to

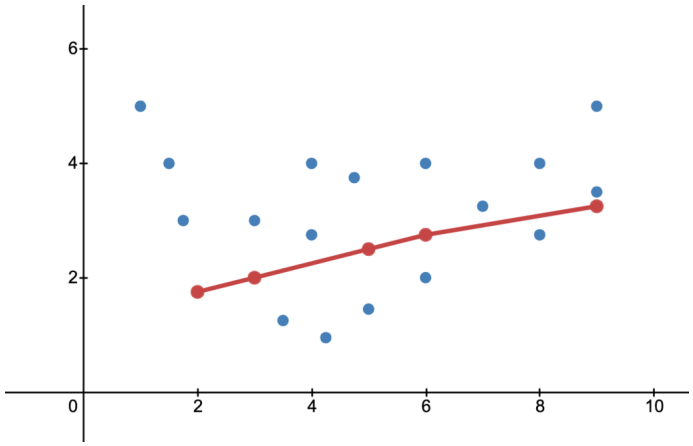


Figure 2.4.: Underfitted model in red trying to represent the blue data points, by AAStein
CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons.

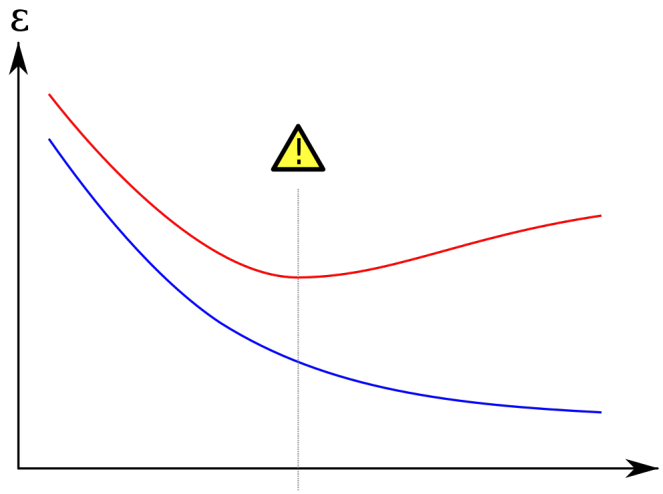


Figure 2.5.: Training error in blue and validation error in red diverge in an overfitted model, by Gringer
CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, via Wikimedia Commons.

measure bias, which is why it is a withstanding challenge in NLP. There are some proposed methods, and these will be introduced in this section.

WinoBias was proposed by Zhao et al. (2018) as a new benchmark for gender bias in coreference resolution. The dataset is comprised of two test sets with pro- and anti-stereotypical sentences. Statistics from the U.S. Bureau of Statistics[14] determine if a sentence is pro- or anti-stereotypical. Test set 1 is on the form: *[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]* and test set 2 is on the form: *[entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances]*.

Webster et al. (2018) introduced a **bias score** in their paper to evaluate the performance of different systems on their new GAP dataset. It is calculated by dividing the feminine F1-score by the masculine F1-score. This score can show the difference in performance on female and male pronouns.

Stereotype and Skew, as presented by de Vassimon Manela et al. (2021), are metrics to quantify gender bias present in pre-trained and fine-tuned language models. Stereotype is the phenomenon that occurs when a model has an unequal preference when assigning pronouns to stereotypical and anti-stereotypical professions. For instance, when a model predicts male pronouns in pronoun resolution for the profession doctor and builder, these are stereotypical male professions. Skew is when a model favour male pronouns overall. This is called a male skew. This is quite common in natural language processing because male pronouns are often over-represented in the training data. de Vassimon Manela et al. suggested a negative correlation between skew and stereotype, implying a trade-off between the two. de Vassimon Manela et al. used WinoBias (Zhao et al., 2018) to compare the prediction of the models with pro- and anti-stereotypical labels. This is done by masking the pronoun in each example sentence and letting the model predict the pronoun. Then the F1 scores of male pronouns for pro- and anti-stereotypical sentences are calculated, and the same with female pronouns. These F1 scores are then used to find skew and stereotype. Skew and stereotype are formally described in Equation 2.8 and 2.9, respectively.

$$\mu_{\text{Skew}} \triangleq \frac{1}{2}(|\text{F1}^{\male}_{\text{pro}} - \text{F1}^{\female}_{\text{pro}}| + |\text{F1}^{\male}_{\text{anti}} - \text{F1}^{\female}_{\text{anti}}|) \tag{2.8}$$

$$\mu_{\text{Stereotype}} \triangleq \frac{1}{2}(|\text{F1}^{\male}_{\text{pro}} - \text{F1}^{\female}_{\text{anti}}| + |\text{F1}^{\male}_{\text{pro}} - \text{F1}^{\female}_{\text{anti}}|) \tag{2.9}$$

In addition to these methods, Bernstein-Bounded Unfairness (BBU) proposed by Ethayarajh (2020) is a method that can show the uncertainty of a predicted bias. This method will not be further discussed in this thesis.

[14] https://www.bls.gov/cps/cpsaat11.htm

3. Datasets

This chapter will present and discuss the different datasets that are commonly used in natural language processing, primarily in Norway. Data is the foundation of natural language processing as it is the source of all knowledge that language models can learn. There are some challenges with data in natural language processing, for instance, too little data, too much data, biased data and more. In addition, both gathering and annotating data can be time-and resource-consuming. Especially data annotation can require vast resources, for instance, the need for many human annotators and the time-consuming task of annotating data manually. Recently there has been research suggesting using language models for automatic annotation, but there is still a need for humans to verify or supply annotations for models to learn from.

3.1. Common Crawl

Common Crawl is a non-profit organisation that provides datasets and metadata to the public for free. The data is gathered by crawling the web. The size of the dataset is per October 2022 380 TiB. The Common Crawl started crawling the web in 2011. The dataset is broadly used as training data for different language models. Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020) is a cleaned version of the Common Crawl corpus for the English language. mC4 (Xue et al., 2020) is a multilingual variation of the C4 dataset and contains natural text from 101 languages gathered from the Common Crawl web crawl.

3.2. The Pile

The Pile (Gao et al., 2020) is an 800GB dataset made for training large language models. The company behind the Pile is EleutherAI[1] which is a non-profit organization within AI-research. The Pile is made up of twenty-two high-quality datasets in English. Data in the Pile is gathered from, among others, Common Crawl, PubMed, Wikipedia, and ArXiv. This means the data ranges between scientific

[1] https://www.eleuther.ai/

text, web pages, mathematics, chat logs and medical texts. The diversity in the data is seen as a great advantage over other huge datasets.

3.3. Norsk Aviskorpus

Norsk Aviskorpus (NAK)[2] is a collection of Norwegian news texts from the period 1998 to 2019 and was created by the National Library of Norway. The corpus contains 1.68 billion words for Norwegian Bokmål and 68 million words for Norwegian Nynorsk, a total of 2.36 billion words. This is a monolingual dataset, meaning that it only contains Norwegian. The data was collected by crawling news websites. Both NorBERT (Kutuzov et al., 2021) and NB-BERT (Kummervold et al., 2021) are trained on this data. In this thesis, NAK is used for data augmentation in several of the experiments introduced in Chapter 5.

3.4. Norwegian Colossal Corpus

The Norwegian Colossal Corpus[3] (NCC) was also created by the National Library of Norway. The corpus is a collection of multiple small Norwegian corpora coming from newspapers, books, the parliament, in addition to many other organisations. The corpus consists of 18.4 billion words; however, only parts of it are available due to the sensitivity of the data. The Norwegian models NB-BERT and NorBERT2 are both trained on NCC, in addition to NAK. The dataset is divided into a training and a validation split, where the training set is sharded in 1GB chunks and the validation set is a file of 1GB.

3.5. Norwegian Dependency Treebank and NorNE

Norwegian Dependency Treebank (NDT)[4] (Solberg et al., 2014) is a manually annotated dataset from the National Library of Norway. The NorNE[5] (Jørgensen et al., 2020) dataset is manually annotated with Norwegian Named Entities extended from NDT. NorNE was created in a collaboration between Schibsted Media Group, Språkbanken, the National Library of Norway and the Language Technology Group at the University of Oslo. The NER- and POS tags make the dataset ideal for

[2] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/
[3] https://huggingface.co/datasets/NbAiLab/NCC
[4] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-10/
[5] https://huggingface.co/datasets/NbAiLab/norne

fine-tuning downstream tasks such as Named Entity Recognition (NER) and Part-Of-Speech tagging (POS). The data was collected from parliament speeches and government reports, as well as from Norwegian blogs and newspapers. The dataset has nine splits: validation, test and train for Bokmål, Nynorsk and combined. In addition, there are nine classes, as seen in Table 3.1. The entity distribution of Bokmål in NorNE can be seen in Table 3.1. In addition, each entry in NorNE has seven data fields, which are idx, lang, text, tokens, lemmas, ner_tags and pos_tags. For further information about the splits, classes and data fields see Jørgensen et al. (2020).

Table 3.1.: Overview of the entity distribution for Bokmål in the NorNE dataset.

Entity	**Train**	**Dev**	**Test**	**Total**
Person (PER)	4033	607	560	5200
Organisation (ORG)	2828	400	283	3511
Geo-political location (GPE_LOC)	2132	258	257	2647
Product (PROD)	671	162	71	904
Location (LOC)	613	109	103	825
Geo-political organisation (GPE_ORG)	388	55	50	493
Derived (DRV)	519	77	48	644
Event (EVT)	131	9	5	145
Miscellaneous (MISC)	8	0	0	0

3.6. The Nordic Pile

The Nordic Pile (Öhman et al., 2023) is a dataset with a focus on the Nordic languages and pre-training large Nordic language models. It consists of texts in Swedish, Danish, Norwegian, English and Icelandic. The dataset is comprised of 1.2TB of text. According to Öhman et al., the goal of creating this dataset is to make a high-quality training set for large language models. The process of collecting data is similar to The Pile, mentioned in Section 3.2, hence the name the Nordic Pile. The data is gathered from various sources divided into nine categories; articles, books, code, conversational, math, miscellaneous, web cc (common crawl), web sources and Wikipedia. To process the data, Öhman et al., created a pipeline which includes normalization, metrics, quality filtering, exact deduplication, language segmentation, fuzzy deduplication and merging. For more information on this pipeline, readers are referred to Öhman et al. (2023)

3.7. Scandi-Reddit

Scandi-reddit[6] is a dataset comprised of thirteen million comments from Reddit containing four different languages. The distribution is as follows: ∼7 million Swedish comments, ∼5 million Danish comments, ∼1,3 million Norwegian comments and ∼200 thousand Icelandic comments. The dataset was created by Dan Saattrup Nielsen from the Alexandra Institute[7]. The Alexandra Institute is a Danish institute owned by Aarhus university research foundation (Aarhus Universitets Forskningsfond)[8]. The data fields are described in Table 3.2. The PushShift API[9] was used to collect the data from 2005 up until 2022. To filter the Reddit data based on languages, the FastText language detection model was used. The dataset was created with the intention of using it to train language models. As Reddit is comprised of many different forums called subreddits, characterised by r/ before the forum name, the source forum collected from can be important. The top five subreddits are r/sweden, r/Denmark, r/norge, r/svenspolitik and r/InfluencergossipDK.

Table 3.2.: The data fields in the Scandi-reddit dataset.

Data field	**Description**
doc	the comment
subreddit	the name of the subreddit the comment is retrieved from
language	a two-letter abbreviation stating the language of the comment, no (Norwegian), da (Danish), se (Swedish) or is (Icelandic)
language_confidence	the confidence with which the FastText language detection model stated the language of the comment

Moreover, Scandi-reddit is licenced under CC BY 4.0 license[10] which allows for free use of the dataset, also commercially, with attributions to the creator.

[6] https://huggingface.co/datasets/alexandrainst/scandi-reddit
[7] https://alexandra.dk/
[8] https://auff.au.dk/en/
[9] https://files.pushshift.io/reddit/
[10] https://creativecommons.org/licenses/by/4.0/

4. Related Work

Gender bias can be defined in different ways, in this section, gender bias is presented in relation to natural language processing. Bolukbasi et al. (2016) reported that gender bias existed in English word embeddings. This started a wave of research in the field of gender bias and NLP. Gender bias can be detected in training data, using psychological tests, using masked language modelling and through downstream tasks. Mitigation techniques that have been tested include data augmentation, gender tagging, fine-tuning for bias, learning gender-neutral embedding, adjusting adversarial discriminators, hard debiasing and prediction constraining. These techniques are mostly well-tested for English language models and word embeddings; however, Scandinavian language models have not been tested to the same degree. Finally, using a fluid definition of gender has been suggested by some researchers (Stańczak and Augenstein, 2021; Cao and Daumé III, 2020; Sun et al., 2019), and Manzini et al. (2019) showed that it was possible to use multiclass settings when detecting other social biases like religion and race. In this chapter, Section 4.1 through 4.3, and Section 4.5 through 4.7 include text reproduced from the final delivery in the course IT3915 - Master in Informatics, Preparatory Project. These sections are mostly similar to the preparatory project, but some alterations have been made by removing or adding information.

The papers researched in this thesis were recommended by our supervisor, found by snowballing and by searching for relevant terms in Google Scholar. Snowballing was performed on Lossius and Ruud (2022), Stańczak and Augenstein (2021), Bolukbasi et al. (2016) and Sun et al. (2019) and thus this chapter is mostly based on these papers.

4.1. Gender Bias in Natural Language Processing

Gender bias is often defined in different ways, Stańczak and Augenstein defined gender bias as "systematic, unequal treatment based on one's gender". Furthermore, Touileb et al. (2022) defined bias as "the cases where automated systems exhibit a systematic discrimination against, and unfairly process, a certain group of individuals". Another definition of bias is the deviation between the distribution of ideal and predicted outcomes of a model (Touileb et al., 2022). Stańczak and Augenstein refer

to the paper on bias in computer systems by Friedman and Nissenbaum (1996) and their categorization of bias in information technology. Friedman and Nissenbaum divide bias into three distinct categories: pre-existing bias, technical bias and reporting bias. Pre-existing bias appears before the computer system is created, and can come from, among others, individuals, organizations, and historical and cultural contexts. This means that bias in the computer system itself might be unintended. Technical bias can appear from the technical design of a model, such as hardware or software. Lastly, emergent bias can appear when the context in which the computer system is used changes. Furthermore, Stańczak and Augenstein also mentioned reporting bias and interpretation bias. Reporting bias appears when there is a difference between the frequency of a specific situation being written in text and a specific situation appearing in the real world. Interpretation bias can appear if researchers assume that gender is relevant. This can lead to the researchers not questioning their results, especially if the results align with common stereotypical perceptions.

In natural language processing, bias can be divided into two types, structural bias and contextual bias. Stańczak and Augenstein stated that structural bias can occur when patterns related to gender bias are seen in the construction of sentences, while contextual bias is about the context of a sentence. To observe contextual bias, background information and human perception are required. Contextual bias can further be divided into societal stereotypes and behavioural stereotypes (Stańczak and Augenstein, 2021). Societal stereotypes show traditional gender roles reflected in social norms, while behavioural stereotypes are different attributes used to describe a specific gender (Stańczak and Augenstein, 2021).

Gender bias in natural language processing can pose harm to the end-user in downstream tasks. These types of harm can be divided into allocational harm and representational harm. Allocational harm is to unfairly allocate resources to certain groups over others, while representational harm is the harm where a social identity or certain group is less represented (Stańczak and Augenstein, 2021; Sun et al., 2019).

Stańczak and Augenstein also mentioned the term gender gap. Gender gap is described as a phenomenon that influences gender bias in texts. This appears as women are underrepresented in different parts of society and therefore most texts discuss and quote men. This can lead to biased datasets that are being used by researchers to, for instance, create a new language model.

4.2. Detection of Gender Bias in Natural Language Processing

Bias is difficult to detect and a standardised way to detect gender bias does not exist. However, the techniques used usually fall into one of two categories: detecting gender bias in specific NLP tasks or detecting gender bias in society.

4.2.1. Detecting Gender Bias in NLP Tasks

In 2016 Bolukbasi et al. published a paper pointing out the presence of gender bias in word embeddings. In 2021, Stańczak and Augenstein performed an extensive literature review of papers published regarding gender bias and natural language processing up until June 2021, in total 304 papers were reviewed. Stańczak and Augenstein further presented a graph showing the exponential increase in published papers on the topic of gender bias in natural language processing since 2015. Thus, an increase in research on gender bias in NLP was seen after the paper by Bolukbasi et al. was published. Among others, gender bias has been detected in coreference resolution (Rudinger et al., 2018; Webster et al., 2018; Zhao et al., 2018; Cao and Daumé III, 2020), machine translation (Stanovsky et al., 2019) and named entity recognition (Mehrabi et al., 2020). Bolukbasi et al. proved the presence of gender bias in word embeddings using analogies, and a lot of researchers followed (Manzini et al., 2019; Mikolov et al., 2013). However, Nissim et al. (2020) presented a problem with this approach. A famous example of an analogy used to prove gender bias is "Man is to doctor as woman is to nurse". If the premise of an analogy, on the form A : B :: C : D (A is to B as C is to D), is that all four terms must be distinct, then what is the expected result? According to Nissim et al. there are two main problems with using the analogy tasks to detect gender bias in natural language processing; propagation and misleading. When detecting gender bias with an analogy, the result can be quite sensational and therefore propagate through science and mainstream media, gaining more attention than it might deserve. *Man is to computer programmer as woman is to homemaker*, Bolukbasi et al. creates a headline which is easy to spread. When readers outside of the field read analogies as such, they have no way to know how sound this analogy is, or how the theory behind it works. The other problem with analogies in this setting is that they can be misleading in the way we search for bias. If the detection of bias is faulty, then the proposed debiasing technique can be faulty too.

Detecting gender bias through downstream tasks is another approach. A downstream task is the application of a word embedding or language model. It is important to see the effect gender bias in NLP can have in real-world applications. Sahlgren and Olsson (2019) described a scenario where limited companies in Sweden

must get their company name approved by the registration office. The decision of the registration office is based on the relationship between the company name and the company description. If someone wanted to register a company that does business with cars this could for instance be done by using pre-trained embeddings, where the similarity between the suggested company name and company description could be quantified. Sahlgren and Olsson found that male names like "Fredrik" and "Magnus" were closer to "cars" than female names like "Maria" and "Anna" in pretrained Swedish ELMo embeddings. Thus, the use of word embeddings in this case could introduce gender bias into the real world.

4.2.2. Detecting Gender Bias in Society

To detect gender bias in society, a commonly used approach is analysing the training data. Zhao et al. (2019) investigated the training data of ELMo for gender bias. Zhao et al. performed an analysis of the One Billion Word Benchmark (Chelba et al., 2013) and found that there was an unequal representation of female and male pronouns in the corpus. The female ("she" and "her") and male ("he", "him" and "his") pronouns were counted, in addition to the co-occurrence of occupations with those pronouns. The results showed that male pronouns occurred three times more often than female pronouns. Male pronouns co-occurred more often with occupations than female pronouns, whether these were stereotypical female or male occupations. Lossius and Ruud (2022) counted pronouns in the training data used by NorBERT, NB-BERT and mBERT, and found that the corpora these models are trained on all include more than three times as many male pronouns as female pronouns. Thus being consistent with the findings from Zhao et al.

In 2020 the 6th Global Media Monitoring Project (Macharia, 2020) was published. This study included research on Norwegian news and showed that 60 per cent of all sources on TV was male, meaning 40 per cent were women. In newspapers, however, only 28 per cent of the cited sources were women. This shows that the gender gap is present in media and thus will be present in training data as many corpora are built on news articles. Asr et al. (2021) also found the same results in Canadian news and stated that men are quoted three times more often than women. News articles often discuss politics, leadership and economics, where men are over-represented. In addition, women often do not want to appear in the media or are more sceptical about being photographed. NRK (The Norwegian Broadcasting Cooperation) performed an unofficial test where five out of five men asked said yes to being interviewed, while they had to ask 15 women to get five of them to be interviewed[1]. Thus, showing a fundamental problem when it comes to

[1] https://www.nrk.no/vestfoldogtelemark/xl/det-er-langt-flere-menn-enn-kvinner-pa-norske-nyhetssider-_-fortsatt-ikke-likestilling-1.15168666

creating fair datasets.

4.3. Mitigation of Gender Bias in Natural Language Processing

In this section, the state-of-the-art methods used to mitigate bias (debiasing) in NLP will be presented. Mitigation methods are techniques used to remove or reduce the bias present in NLP. In NLP there does not exist a standardised way to mitigate bias. However, Sun et al. (2019) divided the mitigation techniques into two categories, retraining of models and inference of models. Data augmentation, gender tagging, fine-tuning for bias, learning gender-neutral embeddings and adjusting adversarial discriminators are all retraining techniques, while hard debiasing and prediction constraining are inference techniques (Sun et al., 2019).

4.3.1. Retraining of Models

Retraining are debiasing methods where gender bias is addressed in the initial stages of modelling or even at the source. The models are retrained on new datasets, which might be both time- and resource-consuming (Sun et al., 2019).

Data augmentation is one of the retraining techniques introduced by Zhao et al. (2018). To perform data augmentation, Zhao et al. gender-swapped the sentences in the dataset. This means that sentences like "He is a computer scientist" would be swapped to "She is a computer scientist" and vice versa. To do this Zhao et al. first anonymised the named entities in the dataset using a named entity finder. Furthermore, a dictionary of gendered terms and their realisation as the opposite gender was built. Then there were created rules to obey this, a rule could for example be "she - he", "Mr. - Mrs." or "mother - father". If the rules had multiple different phrases this was managed by using the most frequent term. Zhao et al. further trained the language model on the union of the gender-swapped dataset and the original dataset. According to Zhao et al. this approach can remove bias.

Bias fine-tuning is another approach used to mitigate gender bias. Park et al. (2018) based the approach on transfer learning from a dataset that is less biased. After training a model with the less biased dataset, the model can be fine-tuned using a dataset with more bias. Park et al. also tested out gender-swapping and found that bias fine-tuning was less effective at removing bias and performance than gender-swapping.

Vanmassenhove et al. (2018) introduced gender tagging as a debiasing technique, where gender information was integrated into neural machine translation systems. When translating "I am happy" from English to French there are two options "Je suis heureux" for a male version and "Je suis heureus" for a female version

(Vanmassenhove et al., 2018). To get the gender right in machine translation Vanmassenhove et al. therefore suggested using gender tagging. The technique is based on adding a tag indicating the gender of the source. The gender tag is added to the beginning of every data point. Vanmassenhove et al. stated that gender tagging could sometimes lead to improvements in machine translations, however, the results were varying. According to Sun et al. the approach is expensive as the meta-information could be costly both in memory and time. In addition, machine translation models would have to be redesigned to parse the gender tags correctly.

Costa-jussà and de Jorge (2020) attempted to create gender-balanced datasets. According to Costa-jussà and de Jorge, an unbalanced dataset would influence the methods built on top. In addition, people using such a system would learn these biases and preserve these biases for the future. To achieve a gender-balanced dataset Costa-jussà and de Jorge removed male-related samples until the dataset had the same amount of male and female instances. The results showed that this balanced dataset had less bias than more massive datasets. Webster et al. (2018) introduced the gender-balanced dataset GAP (gender ambiguous pronouns) for coreference resolution. The dataset contains 8,908 labelled pairs sampled from Wikipedia. The motivation behind the dataset was the gender bias present in corpora, resulting in systems favouring masculine entities.

4.3.2. Inference of Models

Inference are the mitigation techniques where bias is reduced without using the original dataset. Instead, existing models are adjusted to provide testing-time debiasing (Sun et al., 2019).

Bolukbasi et al. (2016) suggested a debiasing technique where the gender subspace is removed from the dataset. This is done by first identifying the gender subspace, followed by neutralising and equalising the dataset. By neutralising the dataset, the gender-neutral words are kept at zero in the gender subspace while equalising makes sets of words outside the subspace equal. This approach is called hard debiasing. Another approach suggested by Bolukbasi et al. is soft debiasing. Soft debiasing reduces the difference between the set of words outside the gender subspace, but at the same time keeps the similarity with the original embedding. Bolukbasi et al. reported remarkable results, however, Gonen and Goldberg (2019) stated that the debiasing techniques suggested by Bolukbasi et al. only hid the gender bias and did not remove it. Gonen and Goldberg were able to show that for example, “nurse” and “receptionist” were closer to each other in the clusters. This was shown by using the k-means algorithm to classify the gender-neutral words into two different classes. After this, the k-nearest neighbours’ algorithm was used. This showed that words in the same cluster could show the bias Bolukbasi et al. claimed to have removed.

4.4. Quantification of Gender Bias

Quantification of bias is not an easy task. As mentioned in Section 2.10.2 there is no standardised way to measure gender bias in natural language processing. Often different tasks will require different measures. Even the definitions of bias can vary from person to person or between fields of study. As presented earlier in this chapter there have been several proposed methods to detect and mitigate gender bias in natural language processing. With several different methods for measuring bias, detecting bias, mitigating bias and even different ways to define bias, it is obvious that quantifying bias can be challenging.

Czarnowska et al. (2021) performed a study of fairness metrics in NLP. A fairness metric quantifies the difference in model behaviour across different social groups. Czarnowska et al. focused on measuring bias in downstream tasks and found that the existing metrics could be divided into three generalized fairness metrics. They proposed a three-step process to help choose which metric to use:

1. Identify which type of question to ask and choose the appropriate generalized metric to answer it.

2. Identify a scoring function that targets the studied type and aspect of bias.

3. Choose the remaining parameters.

Furthermore, Czarnowska et al. suggested considering at least one probability-based metric and one prediction-based metric.

Kurita et al. (2019) suggested quantifying gender bias in contextualised word embeddings. As mentioned in Section 2.5, BERT is trained using masked language modelling (MLM). To measure the bias in the word representation Kurita et al. used the predictions for the [MASK] tokens from MLM. This was done by computing the association between targets and attributes. The generalised procedure as described by Kurita et al. starts with preparing a template sentence e.g. "[TARGET] is a [ATTRIBUTE]". Next, replace [TARGET] with [MASK] and compute $p_{tgt} = P([MASK] = [TARGET]|sentence)$. Following, replace both [TARGET] and [ATTRIBUTE] with [MASK] and compute the prior probability: $p_{prior} = P([MASK] = [TARGET]|sentence)$. Lastly, compute the association: $\log \frac{p_{tgt}}{p_{prior}}$ (Kurita et al., 2019).

Touileb et al. (2022) approached quantification by exploring to which degree language models reflect Norwegian demographics. This was done by using data from Statistics Norway, templates containing occupations and pronouns, and language models. The data from Statistics Norway contains 418 occupations which presents the demographic distribution of men and women in these occupations. Furthermore, Touileb et al. used the gender-to-occupation ratio as the "gold standard" when

investigating the language models. Touileb et al. used five different templates, among others "[pronoun] is a/an [occupation]". As the data from Statistics Norway only contains a binary gender distribution, the pronouns do not include the gender-neutral pronoun "hen". To calculate the bias Touileb et al. generated a probability distribution of masked tokens in each template. The probability distribution is then mapped to percentage. Moreover, the difference between female and male scores is quantified. A positive value indicates occupations associated with females more than males, while a negative value means the opposite. This is also done for the data from Statistics Norway. Furthermore, the macro F1 score is calculated for each model.

Samuel et al. (2023) presented NorBench[2], a collection of Norwegian datasets and evaluation scripts that introduces a standardised way to compare performance between different language models. The benchmark tasks include morpho-syntactic token-level tasks like part-of-speech tagging, named-entity recognition, sentiment analysis, linguistic acceptance, question answering, machine translation and diagnostics of harmful predictions such as gender bias.

4.5. Gender Bias with Gender-Neutral Pronouns

As stated in Section 2.1, gender can be defined in different ways. Stańczak and Augenstein (2021) performed a survey of datasets and papers regarding gender bias and natural language processing. They concluded that further research should be conducted with a more fluid definition of gender in mind. Many researchers consider gender as a binary attribute. However, in our modern society, this is not the case, gender is a fluid attribute. Hence, it is important to consider this in research as well. Cao and Daumé III (2020) and Sun et al. (2019) also stated that it is important to consider gender as a fluid attribute or the research in itself would contribute to bias.

Manzini et al. (2019) showed that it is possible to debias multiclass settings such as race and religion using soft and hard debiasing as presented by Bolukbasi et al.. This shows that using a multiclass setting of gender also should be possible. Manzini et al. further quantified bias removal by using mean average cosine similarity (MAC), and found that the score increased after debiasing, meaning a reduction in bias. However, they also assessed the approach suggested by Gonen and Goldberg (2019) and found that the approach was insufficient at removing multiclass "cluster bias". The formula for MAC can be seen in Equation 4.3. Here, T is the set of target embeddings that contain some form of social bias, A is the set of attributes containing word embeddings not to be associated with the set T, the

[2] https://github.com/ltgoslo/norbench

function S computes mean cosine similarity between T and A. The function S is shown in Equation 4.1 and the cosine distance used in S is shown in Equation 4.2.

$$S(t, A_j) = \frac{1}{N} \sum_{a \epsilon A_j} cos(t, a) \tag{4.1}$$

$$cos(u, v) = 1 - \frac{u \cdot v}{||u||_2 \cdot ||v||_2} \tag{4.2}$$

$$\text{MAC}(T, A) = \frac{1}{|T||A|} \sum_{T_i \epsilon T} \sum_{A_j \epsilon A} S(T_i, A_j) \tag{4.3}$$

A challenge with gender-neutral pronouns is that there sometimes is more than one meaning for the pronoun. In English "they" is used to refer to both third person singular and second person plural. In Norwegian, a similar problem can be found as "hen" can be used as both a personal pronoun, a place and as an adverb. This can confuse language models and increase the need for context.

Brandl et al. (2022) investigated if the performance of language models would be affected by including gender-neutral pronouns in downstream tasks for Swedish, Danish and English language models. They found that the language models had a drop in performance when including gender-neutral pronouns, for the English language model the drop in performance was significant, while the Danish language model only experienced a small reduction in performance. Brandl et al. argued that this could be because of sparse training data, as not many datasets include gender-neutral pronouns. Furthermore, language models are rarely updated after being published so they are not retrained on newer data possibly including more gender-neutral pronouns.

Lindqvist et al. (2018) investigated the use of the gender-neutral pronoun "hen" as a gender-fair strategy. They found that using the gender-neutral "hen" can eliminate the male bias. Interestingly, they did not find the same to be true for the gender-neutral "NN" and the gender-neutral noun "the applicant". The two last are associated with male entities. In addition, they also found that paired pronouns "he/she" had the same effect as "hen" in eliminating male bias. This is promising for the use of "hen" in Norwegian and in Norwegian language models to eliminate the male bias.

4.6. Gender Bias in Scandinavian Language Models

Stańczak and Augenstein (2021), Bender et al. (2021) and Sun et al. (2019) all suggested that further research should be conducted on multilingual language models or at least in low-resource languages (not English). They argued that by focusing on high-resource languages, a limited view of gender bias in NLP is presented. Lossius and Ruud (2022) took this into consideration and investigated gender bias in Norwegian language models. They were able to detect gender bias in the Norwegian BERT models, NorBERT and NB-BERT, by counting pronouns in the training data, and by using analogy tasks as proposed by Bolukbasi et al. (2016) In addition, they found that gender bias was present in downstream tasks. Furthermore, Lossius and Ruud criticised Kummervold et al. (2021) and Kutuzov et al. (2021) for not including research regarding ethics and fairness when publishing language models. Moreover, they suggested that there should be conducted more research regarding Norwegian language models, datasets and debiasing techniques.

Sahlgren and Olsson (2019) investigated the presence of gender bias in pre-trained Swedish embeddings by matching names with occupations and found that gender bias was present. Moreover, they found that the debiasing techniques proposed by Bolukbasi et al. amplified the bias already present in the training data. Sahlgren and Olsson suggested that this might be because Bolukbasi et al. thoroughly cleaned up their data before debiasing, while they themselves did not.

Touileb et al. (2022) performed a similar experiment to Sahlgren and Olsson on Norwegian and Multilingual language models. Touileb et al. investigated if the demographic distribution of occupations was reflected in pre-trained language models. Their results showed that the language models have a biased representation of gender-balanced occupations.

Two different strategies to mitigate bias in NB-BERT and NorBERT were tested by Lossius and Ruud, hard debiasing and fine-tuning on a female-only dataset. Hard debiasing gave better results for NorBERT than NB-BERT regarding a decrease in absolute bias. Fine-tuning on a female-only dataset resulted in gender bias against men, and they suggested looking into actual gender swapping.

González et al. (2020) proposed a new challenge dataset for detecting gender bias. The Anti-reflexive Bias Challenge dataset (ABC) includes four languages: Russian, Swedish, Danish and Chinese. These are all examples of languages with type B reflexivization (where third-person reflexive pronouns are not gendered whereas third-person anti-reflexive is). An example of this type of reflexivization is: "The surgeon put a book on PRON.POSS.REFL.3RD table" versus "The surgeon put a book on PRON.POSS.3RD table". In English, this would have the same meaning, and the pronouns would be the same, but for languages with type B

reflexivization like Swedish, Danish and Norwegian there is a distinction. This distinction can be used to detect gender bias present in systems. An example of Danish can be seen here:

"Teknikeren mistede sin tegnebog ved huset." (Neutral)
"Teknikeren mistede hans tegnebog ved huset." (Male)
"Teknikeren mistede hendes tegnebog ved huset." (Female)
Eng: "The technician lost his/her drawing book by the house."

The first sentence contains the third-person reflexive pronoun "sin" which is gender-neutral. The two following sentences contain male and female pronouns, respectively. To uncover gender bias González et al. present four tasks: language modelling, machine translation (MT), coreference resolution and natural language inference. In the MT task, the systems tested would get a source sentence in English and the task would be to translate to one of the four languages Russian, Swedish, Danish and Chinese. The system then has to choose if the English "his/her" should be translated into a reflexive genderless pronoun or to the corresponding gender in the English sentence. This can uncover bias. Furthermore, they found that almost all systems show a worsening in results in the different tasks using the ABC dataset in contrast to the baseline. This is seen as the presence of gender bias. The paper by González et al. can be seen as a response to the amount of research on gender bias in NLP in English and the lack of research for other languages, especially for languages with different grammatical structures.

4.7. Consequences of Large Language Models

Bender et al. (2021) discussed the size of training data. Datasets used to pre-train models are growing, which often leads to better results in benchmark tasks. However, it also leads to costs regarding the environment, finance, and opportunity, in addition to stereotyping, wrongful arrests, increase in extremist ideology and denigration (Bender et al., 2021). "Size doesn't guarantee diversity" Bender et al. stated. Minorities and marginalized groups are underrepresented in web crawls because they to a lesser degree have access to and engage in the same digital platforms as the majority. Furthermore, Bender et al. questioned if it was right that these communities pay the price for training and deploying large English language models when such models are not being produced in their native language. In the paper, Bender et al. further encouraged researchers to use more resources on processing datasets than on building bigger datasets. Moreover, they stated that it was important to be critical of which data is used in the training process of a language model to avoid bias and other harm.

5. Experiments and Results

In this chapter, four experiments will be presented and discussed. The experimental plan shows an overview of the experiments performed in this thesis. The experiments include retraining different language models on augmented data from both Norsk Aviskorpus[1], Scandi-reddit[2] and NorNE (Jørgensen et al., 2020). The language models investigated include NorBERT2 (Kutuzov et al., 2021), NorBERT3 (Samuel et al., 2023), NB-BERT (Kummervold et al., 2021), KB-BERT (Malmsten et al., 2020), DanishBERT[3] and GPT-SW3 (Ekgren et al., 2023). Furthermore, the experimental setup is presented. This includes code infrastructure, experiment running times, experiment parameters and the setup of each experiment. Lastly, the results from the experiments are presented. Among others, these results show that all models exhibit gender bias both before and after debiasing. In addition, accuracy is not substantially decreased for part-of-speech tagging after debiasing. A discussion of the results can be found in Chapter 6.

5.1. Experimental Plan

The experimental plan is to conduct four different experiments that investigate different approaches to using data augmentation as a mitigation technique for gender bias in Scandinavian language models. With these experiments, it should be possible to answer the research questions as stated in Chapter 1. The mapping between the different research questions and experiments can be seen in Figure 5.1. As seen, each question should be answered by at least one of the experiments. Primarily, research question 1 is answered by experiment 1 and research questions 2 and 3 are answered by experiments 4 and 3, respectively. Research question 4 is investigated using experiments 1, 2 and 4. In the following chapters, the experimental plan for each individual experiment is presented.

[1] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/
[2] https://huggingface.co/datasets/alexandrainst/scandi-reddit
[3] https://github.com/certainlyio/nordic_bert

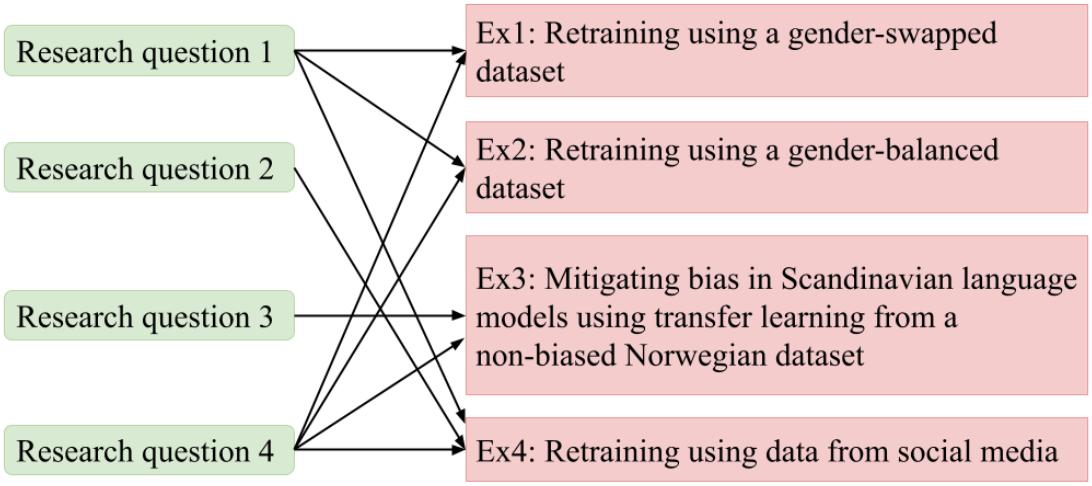


Figure 5.1.: Mapping between the research questions and proposed experiments in the experimental plan.

5.1.1. Retraining Using a Gender-Swapped Dataset

Lossius and Ruud (2022) tried to retrain NB-BERT (Kummervold et al., 2021) using a dataset where each male pronoun was switched to a female pronoun. This approach turned out to result in bias, but this time the bias was against men. They suggested retraining the model using a dataset with three times as many female pronouns as male pronouns to make the original bias neutral or to use gender-swapping as proposed by Zhao et al. (2018). Zhao et al. gender-swapped the dataset by swapping each male pronoun to female and vice versa. Then the model was retrained using the union of the gender-swapped dataset and the original dataset. This approach is previously introduced in Section 4.3.1.

In this experiment, a subset of Norsk Aviskorpus (NAK) will be gender-swapped, before fine-tuning NorBERT (Kutuzov et al., 2021) and NB-BERT (Kummervold et al., 2021) on this dataset. NAK is introduced in Section 3.3 and both NorBERT and NB-BERT are described in Section 2.7.1. In addition to the original gender-swapping, as proposed by Zhao et al., gender-swapping will be performed by swapping both male and female pronouns to gender-neutral pronouns to create another dataset, which will be gender-neutral. Lastly, it will be investigated if the performance of the model is weakened by retraining on a gender-swapped dataset.

5.1.2. Retraining Using a Gender-Balanced Dataset

Costa-jussà and de Jorge (2020) tried to create a gender-balanced dataset by removing male pronouns until the dataset had the same amount of female and

male pronouns. For this experiment, NorBERT and NB-BERT will be retrained using a subset of NAK which has been balanced for male pronouns. As mentioned in Section 4.5, Stańczak and Augenstein (2021), Cao and Daumé III (2020) and Sun et al. (2019) all stated that more research should be conducted using a fluid gender definition. Thus, a gender-balanced neutral dataset will also be created by removing both male and female pronouns until the dataset contains the same amount of female, male and gender-neutral pronouns. Lastly, calculations regarding bias and performance before and after retraining will be carried out.

5.1.3. Transfer Learning from a Non-Biased Norwegian Dataset

Bias fine-tuning is another approach used to mitigate bias. The strategy was proposed by Park et al. (2018) and is based on transfer learning from a less biased dataset. This approach is presented in Section 4.3.1. Using the Norwegian datasets created in the above experiments, the question is if gender bias will be reduced in the Scandinavian language models. This experiment will be performed using two Swedish language models, KB-BERT (Malmsten et al., 2020) and GPT-SW3 (Ekgren et al., 2023), and one Danish language model, DanishBERT[4]. The results will be compared to both previously introduced experiments.

5.1.4. Retraining Using Data from Social Media

The Norwegian language models NB-BERT (Kummervold et al., 2021) and NorBERT2 (Kutuzov et al., 2021) are trained using among others NAK, the Norwegian Colossal Corpus (NCC) and Norwegian Wikipedia. This data might be less up-to-date than other data that exist. Explorations will be done to find out if fine-tuning language models on social media data would create less bias in the models. As social media reflects society at a specific point in time, is that point in time more gender equal than historic data? Social media data could, for instance, be gathered from Twitter, Reddit, or similar platforms. This experiment will use data from Reddit. Moreover, to compare bias in the dataset collected from social media with NAK, pronouns will be counted in each dataset.

5.2. Experimental Setup

This section explains the experimental setup. Firstly, the technical setup details such as code infrastructure, training parameters, running times and model configuration

[4] https://github.com/certainlyio/nordic_bert

will be described. Following, each experiment will be explained in detail. Here the individual parameters or technical configurations for each experiment will be presented. The code base for all the experiments is available on GitHub[5]. For a description of the code base, see Appendix A.

5.2.1. IDUN

The code for all the experiments was run at NTNU's computer cluster IDUN[6] (Själander et al., 2019). IDUN is a platform for running high-intensity computations for its shareholders, which are faculties and departments at NTNU, in addition to the NTNU IT division. The cluster is maintained and operated by the high-performance computing group at NTNU. A high-performance cluster is a network of computers set to work together and be viewed as one system. By combining the computing power of several computers and nodes, the system performs much better than each individual node. In contrast to single supercomputers where the hardware is specialized and often expensive, clusters can use off-the-shelf CPUs and GPUs to obtain the same computing power. This is a shift in the world of supercomputers where distributed clusters with off-the-shelf hardware take over for monolithic single systems with specialized hardware. IDUN is used due to the amount of data that is processed and the size of the models. It would not be feasible to run the code on a single generic laptop as it would be extremely inefficient. By using IDUN, users get access to both GPU and CPU computing power and the high-performance computing infrastructure. IDUN uses the Slurm workload manager[7] to manage resources and schedule jobs on the resources. When starting a job on IDUN the number of nodes and tasks per node has to be specified. The number of nodes used in the experiments varied between 2 to 4, and tasks per node varied between 5 to 10. For generating datasets, 1 node was used with 3 tasks run on this node.

5.2.2. Training Parameters

The training parameters used to fine-tune the models with augmented datasets were chosen with a background in tutorials found on Huggingface. For BERT-based models where masked language modelling is used for training, a tutorial from Huggingface[8] on masked language modelling (MLM) was used. The tutorial

[5] https://github.com/ingvlt/master-project
[6] https://www.hpc.ntnu.no/idun/
[7] https://slurm.schedmd.com/overview.html
[8] https://huggingface.co/docs/transformers/tasks/masked_language_modeling

describes how to fine-tune DistilRoBERTa[9] on a subset of the ELI5[10] dataset called r/askscience. The ELI5 datasets are gathered from Reddit. MLM is previously described in Section 2.5. The tutorial provides code for both PyTorch and TensorFlow. In these experiments, PyTorch was used. Lastly, the tutorial explains how to use the trained model for inference. The dataset and model were changed to the ones created in the experiments.

For GPT-based models, a tutorial on causal language modelling (CLM) from Huggingface[11] was used. GPT-SW3 is the only GPT-based model included in this thesis. CLM is presented in Section 2.6. Similarly to the tutorial on MLM, this tutorial on causal language modelling uses the r/askscience subset of the ELI5 dataset for fine-tuning a model. Here, the DistilGPT[12] model is the model to be fine-tuned. PyTorch was also chosen in this tutorial.

The same training parameters as Kutuzov et al. (2021) used for part-of-speech (POS) tagging were used for POS in this Master's Thesis. The parameters used for both fine-tuning and part-of-speech tagging can be seen in Table 5.1.

Table 5.1.: Parameters used for fine-tuning and part-of-speech (POS) tagging.

	Fine-tuning	**POS tagging**
chunk_size	128	-
batch_size	3	8
evaluation_strategy	epoch	epoch
epochs	3	20
learning_rate	2e-5	2e-5
weight_decay	0.01	0.01

5.2.3. Run Times

Table 5.2 shows the run times for generating the different datasets. The difference in run times between the small and big datasets is evident in this table. The NorNE dataset and Scandi-reddit datasets were generated on a private laptop due to queuing issues at IDUN. Following is a description of the datasets. **Gender-swapped** is where all pronouns are exchanged for opposites. **Gender-neutral swapped** is when all pronouns are swapped to the gender-neutral pronoun ("hen"). **Gender-balanced** describes a dataset where it is a balance between the number

[9] https://huggingface.co/distilroberta-base
[10] https://huggingface.co/datasets/eli5
[11] https://huggingface.co/docs/transformers/tasks/language_modeling
[12] https://huggingface.co/distilgpt2

of male and female pronouns. **Gender-neutral balanced** is a dataset where there is a balance between all three pronouns, female, male and neutral. The process of creating these datasets is further discussed in the next sections.

Table 5.3 shows the run times for fine-tuning different models with the different augmented datasets. Note the significant difference in running times for fine-tuning GPT-SW3 compared to the other language models. The run times for part-of-speech (POS) tagging can be seen in Table 5.4. The dataset Universal Dependencies was used for all POS tagging performed in the experiments. There were also differences in run time for POS tagging, from the shortest at around 14 hours to the longest at around 32 hours.

Table 5.2.: Run times in minutes for generating dataset.

Task	**Run time**	**Dataset**
Gender-swap	2 min	Part of NAK
Gender-swap	∼60 min	NorNE
Gender-swap	5 min	Scandi-reddit
Gender-neutral swap	1 min	Scandi-reddit
Gender-balance	1 min	Scandi-reddit
Gender-neutral balance	1 min	Scandi-reddit
Gender-swap	52 min	NAK
Gender-balance	48 min	NAK
Gender-neutral balance	45 min	NAK

5.2.4. Model Configurations

All the different language models used for the experiments in this thesis have several configurations. In this section, all configurations are presented as well as the chosen configuration with an explanation for why it was chosen.

For NorBERT, there are three main models: NorBERT, NorBERT2 and NorBERT3. The Language Technology Group at the University of Oslo is behind all of them. For this thesis, NorBERT2 is used for all the experiments. NorBERT2 is an updated and improved version of NorBERT and is described more closely in Section 2.7.1. NorBERT3-base is also used for experiments 1 and 2.

The AI Lab at the National Library of Norway has created many different models, including NB-BERT. NB-BERT exists in a few different configurations, among others, NB-BERT-base, NB-BERT-large and NB-BERT-base-samisk. For the experiments in this thesis, only NB-BERT-base is used.

DanishBERT currently includes two models, one DanishBERT uncased model and one DanishBERT model fine-tuned for named entity recognition (NER). In

Table 5.3.: Run times for fine-tuning.

Fine-tuning		
Model	**Dataset**	**Run time**
NorBERT2	Gender-swapped NAK	40.5 min
	Gender-neutral swapped NAK	40.8 min
	Gender-balanced NAK	25.7 min
	Gender-neutral balanced NAK	22.2 min
	Gender-swapped Scandi-reddit	23.4 min
	Gender-neutral swapped Scandi-reddit	41.4 min
	Gender-balanced Scandi-reddit	4.2h
	Gender-neutral balanced Scandi-reddit	2.2h
	Scandi-reddit	23.5 min
NorBERT3	Gender-swapped NAK	37.4 min
	Gender-neutral swapped NAK	49.1 min
	Gender-balanced NAK	32.2 min
	Gender-neutral balanced NAK	21.9 min
NB-BERT	Gender-swapped NAK	3.3h
	Gender-neutral swapped NAK	2.0h
	Gender-balanced NAK	59.8 min
	Gender-neutral balanced NAK	1.3h
	Gender-swapped NorNe	3.3h
KB-BERT	Gender-swapped NAK	1.0h
	Gender-neutral swapped NAK	55.4 min
	Gender-balanced NAK	34.9 min
	Gender-neutral balanced NAK	42.7 min
DanishBERT	Gender-swapped NAK	48.5 min
	Gender-neutral swapped NAK	1.0h
	Gender-balanced NAK	46.6 min
	Gender-neutral balanced NAK	46.6 min
GPT-SW3	Gender-swapped NAK	15h
	Gender-neutral swapped NAK	15.5h
	Gender-balanced NAK	4.2h
	Gender-neutral balanced NAK	16.8h

Table 5.4.: Run times in hours for part-of-speech tagging.

Part-of-speech tagging	
Model	**Run time**
Gender-swapped NorBERT2	17.3h
Gender-neutral swapped NorBERT2	14.1h
Gender-balanced NorBERT2	18.2h
Gender-neutral balanced NorBERT2	16.9h
Gender-swapped NorBERT3	12.2h
Gender-neutral swapped NorBERT3	15.0h
Gender-balanced NorBERT3	14.5h
Gender-neutral balanced NorBERT3	14.3h
Gender-swapped NB-BERT	55.1h
Gender-neutral swapped NB-BERT	21.7h
Gender-balanced NB-BERT	19.7h
Gender-neutral balanced NB-BERT	29.9h
Gender-swapped NB-BERT (NorNE)	17.6h
Gender-swapped KB-BERT	17.7h
Gender-neutral swapped KB-BERT	18.1h
Gender-balanced KB-BERT	20.5h
Gender-neutral balanced KB-BERT	32.1h
Gender-swapped DanishBERT	16.8h
Gender-neutral swapped DanishBERT	16.9h
Gender-balanced DanishBERT	17.0h
Gender-neutral balanced DanishBERT	18.7h
Gender-swapped NorBERT2 (Scandi-reddit)	31.1h
Gender-neutral swapped NorBERT2 (Scandi-reddit)	31.7h
Gender-balanced NorBERT2 (Scandi-reddit)	17.0h
Gender-neutral balanced NorBERT2 (Scandi-reddit)	16.5h
Scandi-reddit NorBERT2	17.7h

the experiments in this thesis, only DanishBERT uncased is used since the other model is fine-tuned for another downstream task.

The Swedish KB-BERT exists in three versions: bert-base-swedish-cased, bert-base-swedish-cased-ner and albert-base-swedish-cased-alpha. For the experiments performed in this thesis, bert-base-swedish-cased is used. The reason for this is that it is the best fit for the experiments in this thesis. The bert-base-swedish-cased are the closest to the other BERT-based Scandinavian models and therefore make comparisons easier. Bert-base-swedish-cased-ner is already fine-tuned for named entity recognition (NER) and is therefore not suitable.

GPT-SW3 also comes in a few different configurations: 126M, 356M, 1,3B, 6,7B, 20B and 40B. In addition to the instruct models mentioned in Section 2.7.2. The number refers to the number of parameters the model is trained on. The GPT-SW3 models used for these experiments are the GPT-SW3-1.3B and the GPT-SW3-126M. The 1.3B model is chosen because it is the middle ground; not too small and prone to repetition, and not too big to manage efficiently. This was seen as a good compromise between performance and time. After finding that run times were slow with this model, the 126M model was used instead.

5.2.5. Retraining Using a Gender-Swapped Dataset

The first experiment is retraining using a gender-swapped dataset. To create the dataset, a subset of the dataset Norsk Aviskorpus (NAK) was swapped. The subset used was the data from AP2019. NAK was downloaded as a zip file containing .tar and .tar.gz files from Språkbanken[13] on the 1st of February 2023. These files are large and time-consuming to unzip. To be able to use the data, some pre-processing had to be done. Thus, all HTML tags were removed from each file for cleaner text to work with. Only the start and end "%hmlsymbol" were left due to the code only weeding out tags encapsulated by "<>". To gender-swap the data, a dictionary containing the words seen in Table 5.5 was used to map both ways. This means that "ho" and "hun" was swapped to "han" and vice versa. As seen from the table, swapping was done in both Norwegian Nynorsk and Norwegian Bokmål. Tokenization was done with AutoTokenizer which is available through the Huggingface Transformer library[14].

Another approach was also investigated, where the words in the dataset were swapped to make it gender-neutral. This was done by switching male and female pronouns to the gender-neutral pronoun "hen". The pronouns swapped to create the gender-neutral dataset can be seen in Table 5.6. In this case, however, it was

[13] https://www.nb.no/sprakbanken/

[14] https://huggingface.co/docs/transformers/main/en/model_doc/auto#transformers.AutoTokenizer

Table 5.5.: Terms used to gender swap datasets.

Female	Male
Ho/Hun	Han
Hun	Ham
Hennes	Hans
Kvinner	Menn
Fru	Herr
Jente	Gut/Gutt
Jenta	Guten/Gutten
Jenter	Gutar/Gutter
Jentene	Gutane/Guttene
Kvinne	Mann
Kvinnene	Mennene
Damene	Herrene
Kvinna/Kvinnen	Mannen
Damer	Herrar/Herrer

decided to try to make a gender-neutral dataset and thus only switch out the gendered pronouns with gender-neutral pronouns, not vice versa. The reason for doing this is the few occurrences of the gender-neutral pronoun "hen".

Table 5.6.: Gender-neutral pronouns used to create a gender-neutral dataset.

Gendered	Gender-neutral
Ho/Hun	Hen
Han	Hen
Henne	Hen
Ham	Hen

After gender-swapping, the data had to be anonymized, or more correctly pseudonymized. This was achieved by using named entity recognition. This way the model could categorize which words were people. *Nb-bert-base-ner* (Kummervold et al., 2021) was used to achieve this. After categorizing each word, the words classified as "person" were swapped out with for example "P1". A person would have the same number throughout the whole text, meaning that if "Erna Solberg" was given P8, she would be recognized as P8 for the entire text.

Furthermore, NB-BERT, NorBERT2 and NorBERT3 were fine-tuned using these newly created datasets. To see if there were any changes to the models after fine-tuning, predicting masked words in sentences was tried. The sentences chosen

are good examples of unbalanced occupations regarding gender. According to the gold data used by Touileb et al. (2022) from Statistics Norway, 90% of nurses in Norway are women and 10% are men. Software development is also an unbalanced occupation in Norway with 18,9% women and 81,1% men. Sentence S3 was chosen to investigate whether the models would suggest a male- or female-dominated occupation or a more balanced occupation for a gender-neutral pronoun. In a study by Lindqvist et al. (2018), as mentioned in Section 4.5, they investigated three gender-fair language strategies. They found that using "hen" and "she/he" were the two gender-neutral strategies with the most balanced perceived gender by the human test subjects. Similarly, it would be interesting to see which occupations the language models will relate to the gender-neutral pronoun "hen". For GPT-SW3 the MASK token was left out as it is trained on causal language modelling and not masked language modelling. Hence, it could generate the rest of the sentence with text generation.

The sentences are as follows:

- **S1:** "[MASK] arbeidar som programvareutviklar." (English: [MASK] works as a software developer.)
- **S2:** "[MASK] arbeidar som sjukepleiar." (English: [MASK] works as a nurse.)
- **S3:** "Hen jobber som [MASK]." (English: They (singular form) works as [MASK].)

Furthermore, the gender-swapped models were fine-tuned on the NLP task part-of-speech (POS) tagging for evaluation. This task was chosen to measure the performance of all the models, making it easy to compare. The three models were trained using the Norwegian part of the Universal Dependencies. To make the results comparable with previous results the training parameters from Kutuzov et al. (2021), seen in Table 5.1 on page 49 were used.

Lastly, measuring bias in the models was done as proposed by Touileb et al. (2022). The code for the method proposed by Touileb et al. can be found on GitHub[15]. As previously described in Section 4.4, an F1 macro score is calculated between the "gold data" from Statistics Norway and the predicted values calculated using the code from Touileb et al. There are problems with this method and these will be further discussed in Section 6.5.

5.2.6. Retraining Using a Gender Balanced Dataset

NorBERT2, NorBERT3 and NB-BERT were retrained using two different gender-balanced datasets. The first dataset was created by counting all pronouns, and

[15] https://github.com/SamiaTouileb/Biases-Norwegian-Multilingual-LMs

removing male pronouns until there was an equal amount of male and female pronouns in the dataset. The second dataset, a gender-neutral balanced dataset was created by removing both female and male pronouns until there was an equal amount of gender-neutral and gendered pronouns. Furthermore, both models were fine-tuned for part-of-speech tagging to measure the model's performance on an NLP task. Measuring bias was performed as described in experiment 1, where the macro F1 score is calculated based on the gold data and the predicted data. The results are presented in Section 5.3.

5.2.7. Transfer Learning from a Non-Biased Norwegian Dataset

In this experiment, transfer learning is used to fine-tune two Swedish and one Danish language model with augmented Norwegian datasets. The Swedish models chosen are KB-BERT (Malmsten et al., 2020) and GPT-SW3 (Ekgren et al., 2023). In addition, DanishBERT[16] is also explored. All models are previously discussed in Section 2.7. The four datasets created during the two previous experiments, **gender-swapped**, **gender-neutral swapped**, **gender-balanced** and **gender-neutral balanced**, are used to perform transfer learning. This makes it possible to compare and discuss the findings across different languages and models. Performance and bias are calculated the same way as in experiment 1.

5.2.8. Retraining Using Data from Social Media

The Scandi-reddit dataset, mentioned in Section 3.7, is used to fine-tune NorBERT2. Data from Reddit might be more or less biased than historical data from newspapers and books. It was therefore decided to count the pronouns of both Scandi-reddit and NAK to compare the datasets before using the dataset for fine-tuning. The code for counting pronouns is inspired by Lossius and Ruud (2022), the code can be found on their GitHub[17]. For the Reddit dataset, the code had to be modified as it is loaded from the Huggingface hub and not from local files. Following the pronoun count, the dataset is gender-swapped and gender-balanced as described in Section 5.2.5 and Section 5.2.6. Lastly, performance is measured using part-of-speech tagging and bias is measured as presented in Section 5.2.5.

[16] https://github.com/certainlyio/nordic_bert
[17] https://github.com/andrinelo/norwegian-nlp

5.3. Experimental Results

In this section, the results from the experiments will be presented. All experiments were conducted as described in the previous section. The results show that bias is reduced when using the gender-balanced dataset for fine-tuning. The performance of the different models in the part-of-speech tagging measured in accuracy is barely decreased. Moreover, it was found that Scandi-reddit had a more equal representation of male and female pronouns than Norsk Aviskorpus (NAK). However, the dataset used for fine-tuning was too small and not representative of the ratio between the genders and thus the models retrained with NAK achieved a higher F1 score, meaning the models included less bias.

5.3.1. Retraining Using Gender-Swapped and Gender-Balanced Datasets

Predicting Masked Words: By fine-tuning NorBERT2 on only 2504kB of gender-swapped data, the probability of "ho" in the sentence "[MASK] arbeidar som programvareutviklar." ("[MASK] works as a software developer.") went from 0.041 to 0.541. And when fine-tuning with a gender-neutral swapped dataset of only 2314 kB, "hen" appears as one of the top 5 words. The results can be seen in Table 5.7 where S1 and S2 refer to the sentences presented in Section 5.2.5. For both S1 and S2 fine-tuning with the gender-swapped dataset makes "ho" the most likely word to be the [MASK]. When fine-tuning with the gender-neutral swapped dataset, "hen" finally appears as a suggestion for the masked word in S1. The probability of the masked word being "hen" in the three different models for S1 was calculated and went from 0.0 for the original NorBERT and gender-swapped NorBERT to 0.049 with the gender-neutral swapped NorBERT.

Table 5.8 shows results from testing different models with sentence S3, as mentioned in Section 5.2.5. Note here that the Original GPT-SW3 changes the gender-neutral pronoun "hen" to "han" (he in English) in the generated follow-up sentence. It is also worth mentioning that GPT-SW3 is a generative model and hence there is no probability calculated. Interestingly, the gender-swapped GPT-SW3 ends the generated sentence with %htmlsymbol. Both original NB-BERT and NorBERT2 give meaningful suggestions for occupations. If you look away from ":" as proposed by the original NB-BERT. Gender-swapped NB-BERT on the other hand gives no good suggestions and even uses English words like "Roads". The gender-swapped NorBERT2 becomes a bit more uncertain, with probabilities going down from the original to the gender-swapped. Occupations such as assistant, consultant and apprentice are all vague and not typically gendered occupations. Interestingly, "regnskapsfører" (Eng: accountant) is an occupation with 74.2% women statistically.

Table 5.7.: Results from masked language modelling for sentences S1 and S2.

Model	S1		S2	
	Masked word	Prob.	Masked word	Prob.
Original NorBERT	"Eg"	0.078	"Eg"	0.131
	"Han"	0.052	"Ho"	0.090
	"Ho"	0.041	"Rektor"	0.024
	"Me"	0.023	"Siv"	0.017
	"Morten"	0.019	"Mor"	0.014
Gender-Swapped NorBERT	"ho"	0.465	"ho"	0.416
	"han"	0.036	"eg"	0.049
	"eg"	0.026	"Eg"	0.029
	"Ho"	0.019	"Ho"	0.027
	"og"	0.017	"som"	0.022
Gender-Neutral Swapped NorBERT	"og"	0.080	"som"	0.064
	"men"	0.056	"eg"	0.049
	"som"	0.055	"men"	0.042
	"hen"	0.049	"ho"	0.036
	"han"	0.048	"og"	0.035
Gender-Balanced NorBERT	"Eg"	0.273	"Eg"	0.281
	"Han"	0.133	"Ho"	0.085
	"Ho"	0.065	"ho"	0.045
	"Dei"	0.042	"som"	0.030
	"Vi"	0.031	"eg"	0.025

Measure of Performance with Part-of-Speech Tagging: Kutuzov et al. (2021) stated that NorBERT was able to achieve a score of 98% accuracy on the part-of-speech (POS) task when NorBERT was published in 2021. After fine-tuning on a gender-swapped dataset and then fine-tuning for POS tagging NorBERT2 achieves 95% accuracy. This is a decrease of only 3%. The results can be seen in Figure 5.2. The figure shows the performance achieved in POS tagging for the three first experiments, retraining using a gender-swapped dataset and retraining using a gender-balanced dataset. As seen from the figure the gender-swapped NB-BERT was able to achieve the highest score in accuracy. The models perform overall well on the POS task, however, as seen from the figure DanishBERT, KB-BERT and NorBERT3 have a lower accuracy score than NB-BERT and NorBERT2. NorBERT3 overall scores lower than all the other models. Compared to itself the performance is not affected much by the fine-tuning.

Table 5.8.: Results from masked language modelling and text generation with sentence S3.

Model	S3	
	Masked word/generated sentence	Prob.
Original NorBERT2	"lærling"	0.210
	"regnskapsfører"	0.042
	"saksbehandler"	0.032
Original NB-BERT	"assistent"	0.187
	":"	0.101
	"lærer"	0.066
Gender-Swapped NorBERT2	"assistent"	0.126
	"konsulent"	0.097
	"lærling"	0.075
Gender-Neutral Balanced NorBERT2	"lærling"	0.363
	"prosjektleder"	0.046
	"frisør"	0.025
Gender-Swapped NB-BERT	"ordet"	0.018
	"##gede"	0.004
	"Roads"	0.004
Original GPT-SW3	"lærer på en skole i Oslo. Han har vært i Norge i 12 år."	-
Gender-swapped GPT-SW3	"1.amanuensis ved høgskolen i innlandet. %htmlsymbol"	-

Measuring Bias: To measure bias, the macro F1-score was calculated between the gold data from the statistical data and the predicted data on the template "[pronoun] is [occupation]". Figure 5.3 shows the F1 macro scores for the different language models and the different generated datasets. The darker the colour the better the scores, meaning the higher the F1 macro score is, the lower the bias in the model. NorBERT2 retrained with a gender-neutral swapped dataset and NB-BERT retrained with a balanced dataset are the best achievers with an F1 score of 0.75 or 75%. However, as seen from the figure, there is no change in the F1 value from the original NB-BERT to NB-BERT fine-tuned using a balanced dataset. NorBERT3 performs seemingly worse than NorBERT2. From these results, data augmentation as a debiasing method is not optimal for NorBERT3.

NB-BERT Retrained Using Gender-Swapped NorNE: NorNE, mentioned in Section 3.5, was also gender-swapped. NB-BERT was retrained using this dataset and it scored an accuracy of 0.94 in part-of-speech tagging. Bias was also measured,

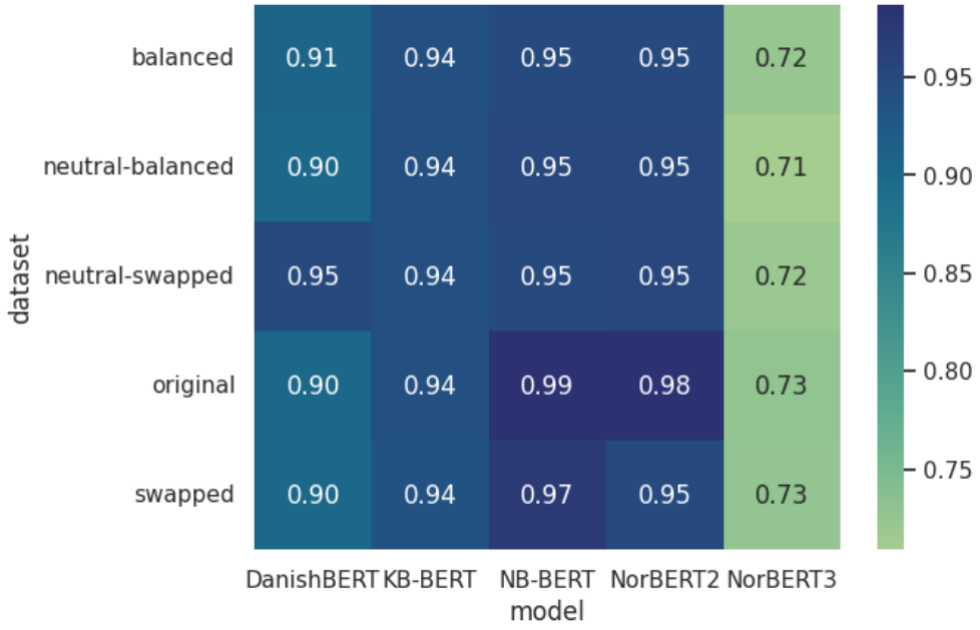


Figure 5.2.: Performance measured in accuracy for the NLP task part-of-speech tagging.

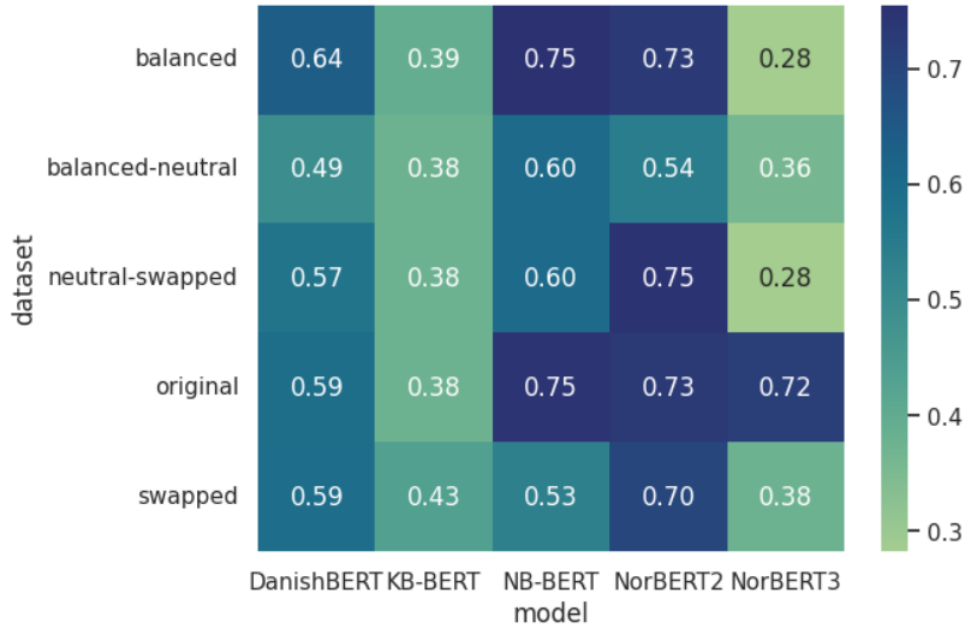


Figure 5.3.: F1-macro score between golden data from Statistics Norway and the predicted data from the templates.

and the F1 macro score was 0.32. As seen from Figure 5.3, this result is evidently worse than the original NB-BERT which scores an F1 macro score of 0.75 and worse than NB-BERT fine-tuned with gender-swapped Norsk Aviskorpus (NAK) which scores 0.53. From these two results, it seems that fine-tuning NB-BERT on gender-swapped NorNE is not the best option to mitigate gender bias.

NorBERT2 with Larger Batch Size: The parameters used when fine-tuning the models are introduced in Table 5.1 on page 49. However, to see if it would be effective to change the parameters, one experiment was run with a batch size of 64. The experiment run with a batch size of 64 was fine-tuning NorBERT2 with the gender-swapped NAK dataset. The model scored a 0.75 F1 macro score, which is an increase of 0.05 from the gender-swapped model with the parameters shown in Table 5.1. The accuracy in part-of-speech tagging was 0.95. It seems the parameters have an impact on the results and a larger batch size could be one way to decrease the amount of bias in the models. However, the model's performance does not seem to be affected to a considerable extent by this change in the parameters.

5.3.2. Transfer Learning from a Non-Biased Norwegian Dataset

The results from transfer learning from a non-biased Norwegian dataset to different Scandinavian language models can be seen in Figure 5.2 and 5.3. As seen from Figure 5.3, the F1 score for KB-BERT fine-tuned on a gender-swapped dataset increased from the original model. For DanishBERT, on the other hand, there are no significant changes; however, retraining using a gender-balanced dataset gives a slightly better F1 macro score. Furthermore, DanishBERT and KB-BERT both achieved lower accuracy in part-of-speech tagging in comparison to NorBERT2 and NB-BERT. However, DanishBERT scores higher when retrained with the gender-neutral swapped dataset. This might be because "hen" is much used as an adverb in Danish. This is further discussed in Section 6.3.

5.3.3. Retraining Using Data from Social Media

Pronoun Counting: To compare Scandi-reddit with Norsk Aviskorpus (NAK), pronouns in both Scandi-reddit and NAK were counted, and the results can be seen in Table 5.9. The original NAK dataset contains 3.3 times more male pronouns than female pronouns and 1305 times more gendered (female and male pronouns) than gender-neutral pronouns. The smaller dataset used in the experiment exhibits the same problem, but to a lesser degree. For the smaller NAK, there are 1.8 times more male than female pronouns and there are no gender-neutral pronouns. In

comparison, the male-to-female ratio in Scandi-reddit is 1.2. Gendered pronouns still occur 55 times more often than gender-neutral. In the subset of Scandi-reddit that was used for retraining, the ratio is different. In this subset, male pronouns occur 3.2 times more often than female pronouns and gendered pronouns occur 24.5 times more often than gender-neutral pronouns. This shows that both Scandi-reddit and NAK exhibit gender bias. Yet, Scandi-reddit does have a more equal representation of male and female pronouns in comparison to NAK.

Table 5.9.: Results from pronoun counting in Norsk Aviskorpus and Scandi-reddit. The results from the smaller dataset used in the experiments are in parentheses.

	Norsk Aviskorpus	**Scandi-reddit**
Female pronouns	2 279 685 (116)	607 230 (200)
Male pronouns	7 463 735 (213)	748 562 (634)
Gender-neutral pronouns	7465 (0)	24 649 (34)

Retraining Using Data From Reddit: The results from fine-tuning and part-of-speech tagging on Reddit data can be seen in Table 5.10. The same datasets were created with Scandi-reddit as was done in experiments 1 and 2. In addition to fine-tuning on these datasets, NorBERT2 was also fine-tuned using the original Scandi-reddit dataset. As seen from Table 5.10, all the augmented datasets except for the gender-neutral swapped dataset increase the F1 macro score, meaning that bias is decreased. The performance of the models on part-of-speech tagging is unchanged. A larger part of Scandi-reddit could have been used to decrease the amount of bias in the model even more. Compared to retraining the original NorBERT, seen in Figure 5.2 and Figure 5.3, the only dataset that does not increase bias in the model is using the gender-swapped Scandi-reddit dataset. Using this dataset maintains the F1 macro score from the original NorBERT2. Retraining using the original Scandi-reddit dataset decreases the F1 macro score by 0.07. Overall NorBERT2 retrained with both the original and the augmented Scandi-reddit datasets achieved the same accuracy score of 0.95 as NorBERT2 retrained with augmented NAK. All the models retrained with augmented data sees a decrease in accuracy from the original NorBERT2, this is expected.

Table 5.10.: Results from fine-tuning NorBERT with Scandi-reddit.

Dataset	**Accuracy**	**Bias**
Gender-swapped Scandi-reddit	0.95	0.73
Gender-neutral swapped Scandi-reddit	0.95	0.60
Gender-balanced Scandi-reddit	0.95	0.70
Gender-neutral balanced Scandi-reddit	0.95	0.69
Original Scandi-reddit	0.95	0.66

6. Discussion

This chapter presents an evaluation and discussion of the experiments and results in this Master's Thesis. This includes how the datasets were made, the amount of data used to retrain the models, using social media as training data and if bias should be mitigated. The methods used and results are reflected upon considering related work previously presented in Chapter 4.

6.1. Creating the Datasets

For the first three experiments, it was decided to use Norsk Aviskorpus (NAK) as the basis. The size of the dataset made it difficult to run code in trial and error mode as it took a long time before finishing. It was therefore decided to test out the functionality with a smaller dataset, and therefore AP2019 from NAK was used. As demonstrated in Section 5.3, the smaller datasets used in experiment 1 appeared to generate meaningful changes in predicting masked words. This made it seem like the changes were larger than they were. Were the experiments to be repeated, a larger dataset could have given more visible results when measuring bias and performance. The same can be seen when looking at the generated sentence by the gender-swapped GPT-SW3. The model ends the sentence with "%htmlsymbol" which all the articles from the NAK dataset start and end with. Originally, each article in the NAK dataset had a full HTML outline with div-tags and paragraph-tags. Most of these tags were removed in the preprocessing, but the start and end were left out. See Section 5.2.5 for a description of the process of creating the dataset. The swapped model clearly learned from the small dataset that a sentence often ends with "%htmlsymbol" and copied that behaviour. This shows how even a small dataset can have a significant impact on the end results.

In the process of gender-swapping and gender-balancing, decisions were required concerning the selection of words to be swapped or balanced. Specifically, it was necessary to determine whether words such as "mor/far"(mother/father) or "datter/sønn"(daughter/son) should be swapped or balanced. Moreover, it was important to consider the question of whether only the gendered part of a word should be altered, such as replacing "brann**mann**" (fireman) with "brann**kvinne**" (firewoman). Given the time constraints a Master's Thesis presents, the decision was made to solely focus on balancing pronouns in the dataset, rather than retrieving an

extensive list of all gender pairs. While this approach aligns with the methodology employed in related work, it may have significant implications for the findings of this thesis, as the inclusion of additional gendered words could have resulted in different outcomes. For gender-swapping, the words swapped are presented in Section 5.2.5. If a word is partly gendered, it is swapped. This could lead to creating sentences which obscure the context of different words. For this thesis, it was decided that it would be better to swap more words as this could give a greater impact on the results.

In addition to gender-swapping and gender-balancing, alternative data augmentation techniques could give different outcomes. Other techniques that could be tested out include counterfactual augmentation and reweighting. Counterfactual augmentation is a similar technique to gender-swapping, where counterfactual sentences are created. An example is switching "The programmer is excellent." to "The nurse is excellent.", thus giving positive associations to an occupation which traditionally is female-dominated. Reweighting, on the other hand, is based on assigning more weight to underrepresented words or sentences in the training data, thus telling the language model that these words or sentences are significant and should be given greater attention. Each of these techniques alone may not give optimal results. Therefore, a potential strategy for mitigating gender bias could involve combining some or all of the aforementioned methods. In future work it could be interesting to assess both counterfactual augmentation and reweighting, in addition to a combination of the four different techniques.

6.2. Anonymising with Named-Entity Recognition

The NB-BERT-base-ner model was used to anonymise the data. It is noted that this approach has some shortcomings, as the named entity recognition (NER) model is not fully dependable. Some names may not be classified as the entity "person", which results in incomplete pseudonymisation. While it may be worthwhile to address this issue in future work, for this Master's Thesis, it was believed it would be feasible to pseudonymise as many names as possible. The reason behind this step is to remove the link between a given name and its associated gender, thereby making it less prone to bias. Moreover, it is important to anonymise the data because of privacy. It is not believed this should have affected the findings in this thesis to a substantial extent.

6.3. Social Media as Training Data

The results seen in Section 5.3 showed that Scandi-reddit dataset[1] did not perform better than the original NorBERT2 or NorBERT2 retrained with augmented Norsk Aviskorpus. Thus, disproving the hypothesis that social media data provides less bias in the language model. This is an interesting finding as it could be seen from the pronoun count that Scandi-reddit includes an almost equal amount of female and male pronouns. For Norsk Aviskorpus (NAK), in comparison, there are 3.3 times more male pronouns than female pronouns. From the count made on the subset of both datasets, it could be seen that the ratio switched between the two datasets. Meaning, if the whole NAK and whole Scandi-reddit had been used, the outcomes might have been switched. In future work, fine-tuning should be performed using the whole datasets or with subsets that maintain the ratio between the pronouns equal to the original datasets.

The pronoun count for gender-neutral pronouns, as seen in Table 5.9 on page 62, may not be representative as the word "hen" can have different meanings according to the context. Counting pronouns was done without consideration of the context of the word. Especially in the Danish portion of Reddit comments, there are more cases of using "hen" in other contexts than as a gender-neutral pronoun. For instance, in these examples from the Reddit dataset: "Jeg tror at så længe du poster godt hen på aftenen [...]" (Eng: I think as long as you post well into the evening) and "Er langt hen af vejen enig med dig [...]" (Eng: Totally agree with you). In these examples "hen" is used as an adverb. In addition, only the Norwegian pronouns were counted.

Social media consists of many different platforms, another platform than Reddit could have given a different outcome. The data from Reddit is also highly dependent on which subreddits the comments are retrieved from. Some forums may be more or less gendered than others. As mentioned in Section 3.7, the top subreddits comments are retrieved from are among others the national subreddits of Norway, Sweden and Denmark. It is difficult to say if these subreddits are more or less gendered than other subreddits. According to a survey by Statista[2], only 36.2% of Reddit users are female. This is not an ideal foundation for creating fairer datasets. It would be interesting to see how gender distribution in a subreddit can affect the language used. In comparison to Reddit's user base, Twitter is a more balanced social media regarding gender. According to Statista[3] 43.6% of Twitter users identify as women. Fine-tuning a language model with data from Twitter

[1] https://huggingface.co/datasets/alexandrainst/scandi-reddit

[2] https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/

[3] https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/

could therefore give a language model with less bias.

Another interesting observation regarding the Scandi-reddit dataset is the language classification done by the FastText language detection model. A skim through the dataset preview on Huggingface shows that the model may not be that accurate in classifying the comment as the right language. For instance, the first comment in the dataset preview, "Bergen er ødelagt. Det er ikke moro mer." (Eng: Bergen is destroyed. It's not fun anymore), is classified as Danish. Even though it was published in the subreddit r/Norway and reads as Norwegian Bokmål.

6.4. Choice of Model Configuration

The model configuration is presented in Section 5.2.4. There was no system when choosing which configuration of the models to use for the experiments. This varies throughout the different models. However, for GPT-SW3 there was a tremendous difference in performance between the different models, where the smaller models performed worse than the larger ones. The smaller models often repeated themselves a lot when generating new text. This is most likely due to the lower data volume the model is trained on. Thus, the findings from this thesis are specific to the model configuration chosen and different outcomes might be present in different configurations. It could therefore be intriguing to compare bias found in the different configurations of the same model.

6.5. Evaluation of Bias

Measuring bias was a challenging task, and in the end it was decided to use the approach suggested by Touileb et al. (2022) as presented in Section 4.4. A similar approach is also used in NorBench (Samuel et al., 2023), as presented in Section 4.4. A problem with this approach is that gender is defined as a binary attribute. The word "hen" was first added to the Norwegian dictionary in 2022, as mentioned in Section 2.2, thus there are little to no statistics regarding the use of this word. In addition, Statistics Norway does not include any statistics regarding people who identify as non-binary and their occupations. It is suggested that future research includes a more inclusive way to measure bias. One suggestion could be to experiment with templates for masked language modelling and the gender-neutral pronoun "hen". In experiment 1, retraining using a gender-swapped dataset, it was attempted to see if the model would associate "hen" with female- or male-dominated occupations or a balanced occupation with a sentence (sentence S3 from Section 5.2.5). Experiment 1 is described in Section 5.2.5. This was inspired by Lindqvist et al. (2018), described in Section 4.5, who found that "hen" and "he/she" were

the most balanced options for a gender-neutral pronoun. Their study concerned the perception of Swedish students, so S3 would be a similar experiment but on language models instead of humans, and with occupations instead of pictures. To investigate gender bias more inclusively, this experiment could be developed further. The more used the word "hen" becomes, the more training data the models have. And the more likely the models are to connect the word "hen" with a gender-neutral pronoun which in turn makes it easier to measure bias more inclusively. There are other, more implicit, ways to measure bias more inclusively without the word "hen", but in this thesis, the focus was on the gender-neutral pronoun.

6.6. Data Augmentation as a Debiasing Technique

Costa-jussà and de Jorge (2020) and Zhao et al. (2018) stated that data augmentation as a debiasing technique was successful in removing gender bias. As seen from the results of the experiments, the only time the F1 score is increased compared to the original Norwegian models, is NorBERT2 retrained with a gender-neutral swapped dataset. This is an interesting outcome as the bias measure is done with a binary definition of gender, while fine-tuning is done with a gender-neutral dataset. Thus, for these Norwegian models and these datasets data augmentation does not seem to reduce bias to an extensive degree. This could be a result of a too small dataset when fine-tuning or due to the way bias is measured as previously discussed in this chapter.

The overall best performing dataset with regards to bias for NorBERT2 and NB-BERT is the balanced dataset where male pronouns were removed until there was an equal amount of male and female pronouns. Still, the results do not remove or mitigate bias, the results are equal with the original models. This way of mitigating bias is not optimal as can be seen from the results of all the different Norwegian language models. From the performance scores, it is seen that the accuracy is decreased from the original models to the fine-tuned models by about 3 to 4 percentage points. This decrease is not that substantial. In the process of removing gender bias from language models, there has to be made a compromise between gender bias and performance, this is further discussed in Section 6.8. Additionally, further training on augmented data or changing the training parameters could increase the accuracy.

6.7. Transfer-Learning as a Debiasing Technique

The results show that KB-BERT scores worse than the other models when it comes to bias. Even the original KB-BERT exhibits a lot more bias compared to the other Scandinavian models. This could be because the gold data is based on Norwegian demographics. Moreover, the bias score is barely increased in the fine-tuned versions of the model. To get a larger increase in the bias score one option would be to increase the size of the dataset the model was fine-tuned on.

DanishBERT had no noteworthy results, as the changes were quite small in bias. Fine-tuning the model with a balanced model still decreased the bias present in the model by 0.5 percentage points. Fine-tuning using a gender-neutral balanced dataset made the model more biased by 1 percentage point.

Transfer learning seems to work better for DanishBERT compared to KB-BERT. However, the results are not that significant and future research should be conducted using a larger dataset, using different approaches to debiasing or changing the training parameters when fine-tuning the models.

6.8. Performance versus Bias

From the results, it is seen that the performance is quite stable with the amount of data used for debiasing. Finding a threshold value between performance and the amount of bias a model exhibits is a critical area to further research. How should this be decided? One suggestion could be a central entity being made responsible for creating the threshold value for publishing a new model. This approach might become too strict, and another suggestion is to require researchers to at least measure bias somehow and publish their scores so that people are aware that bias is present in the model before using it to make substantial choices.

Since Scandinavian languages are all low-resource languages, there are few to no benchmarks regarding gender bias. This makes it difficult to compare the different models as each researcher has to produce their own benchmark. This is also seen in this Master's Thesis as it is difficult to compare the results found in this thesis with related work. However, NorBench was published in May 2023 and is a promising new benchmark for Norwegian language models. This should be investigated further for both Swedish and Danish language models as well.

The development and use of generative AI have accelerated quickly. Even though Brown et al. (2020) and OpenAI (2023) include much research into ethics and fairness, other researchers do not necessarily follow them. A pre-release of GPT-SW3 was introduced in January of 2023. Ekgren et al. (2023) did not include any measures of bias and fairness. Researchers are usually more concerned with beating the performance of the previous state-of-the-art than to contribute to fair language

technology.

6.9. Training Parameters

As seen from the results of experiment 1 in Section 5.3.1, NorBERT2 fine-tuned with a gender-swapped dataset performed better with an increased batch size. This shows that different training parameters could impact the outcome of the experiments performed in this thesis. Other parameters that could be interesting to change are the number of epochs and the size of the learning rate used for the experiments. The learning rate controls how fast the model adapts to a problem, and a small learning rate needs a higher number of epochs, while a large learning rate needs a lower number of epochs. It could be interesting to investigate which approach is best for these experiments.

6.10. Investigating Gender Bias in Generative Pre-Trained Transformers

In Norway, there has not been performed research regarding gender bias and generative pre-trained Transformers (GPT). However, in this thesis it is investigated if GPT-SW3 exhibits gender bias and furthermore how this can be mitigated. As GPT has become increasingly popular compared to Bidirectional Encoder Representation from Transformers (BERT), it is equally important to investigate how gender bias can be mitigated in GPT models.

When predicting masked words for BERT-based models and generating sentences for GPT-SW3, it could be seen that the original GPT-SW3 changes the pronoun from “hen” to “han” (Eng: “they” to “he”). This is interesting because it means that the model may be associating “hen” with male pronouns. The gender-swapped GPT-SW3 suggests “1. amanuensis” (Eng: Associate professor) which is an extremely specific profession. From the gold data used by Touileb et al., which is gathered from Statistics Norway[4], the closest profession could be professor which has an uneven distribution with 31.2% female professors. This can indicate that the model relates “hen” to male pronouns. Interestingly teacher, as the original GPT-SW3 suggests, is as unevenly distributed as professor but the other way around with 74.4% female teachers.

[4] https://www.ssb.no/en

6.11. Should Gender Bias be Mitigated

Mitigating gender bias in language models is generally desirable as it promotes fairness and inclusivity. Gender bias can lead to harmful stereotypes and result in unequal treatment and opportunities as introduced in Chapter 1.1. Gender bias is already found in the training data, as seen from the results of the pronoun count in both Scandi-reddit and Norsk Aviskorpus in Table 5.9 on page 62. This is an example of representational harm and gender bias is carried out in the language models. For example, statistics show that the occupation nurse is female-dominated, and if the model then more intricately connects women with nurses, then the model more closely mirrors the real world. If the goal of the model is to represent real-world connections, then the model has achieved that goal. However, if the goal is to mitigate gender bias, it might still show signs of that if the model for instance penalizes male applicants for a position as a nurse, as seen with Amazon's recruiting tool in Section 1.1.

Sahlgren and Olsson (2019) found that bias was amplified by pre-trained Swedish word embeddings, previously introduced in Section 4.6. Thus, there is a difference between the bias created by the model and the bias found in the training data. If the model is biased, this gender bias should be mitigated as the model should not increase the amount of bias found in the model. When the training data is biased, should it be mitigated? Training data is usually based on newspapers, books and forums, and one might think that the data would reflect society when published. However, training data does not reflect the world at publishing time because newspapers and books are biased, as seen from Macharia (2020) and Asr et al. (2021). Thus, this type of gender bias comes from deeper in society than the actual language models. As seen in Chapter 5.3, when bias is calculated as presented by Touileb et al. (2022) comparisons can be made between the gold data from Statistics Norway and the predicted data by the model. The different occupations in the gold data might be skewed in the real world, thus a skewed representation in the language model is not gender bias.

6.12. Ethics

As discussed earlier in Chapter 1 and 2, ethics is an important part of research in the field of natural language processing and gender bias. Sadly, it is often forgotten when new language models are presented.

In addition to biased datasets and biased models, as previously discussed in this chapter, the research itself can also be biased. From biased researchers to biased grants. New research depends on which researchers and which institutions get the appropriate funding for conducting relevant research.

Another topic for discussion regarding ethics is the vast size of the large language models as presented in Section 4.7. Is it appropriate for big technology companies to use enormous amounts of electricity to fuel clusters for generating large language models and large datasets if the result only benefits a smaller part of the world's population? (Bender et al., 2021) Climate change is a much debated subject in today's politics, social media, news and research. Can the amount of electricity used to produce large language models and large datasets be justified?

The size of the dataset may not reflect the fairness of the data, as seen from the results of the pronoun count of NAK and Scandi-reddit in Section 5.3.3. Bigger does not always equate to better, as stated by Bender et al. (2021), especially when the sources of the data are biased. This is particularly true when gathering data from social media, where there is a majority of white males on both Twitter and Reddit, which are the social media discussed in this thesis. Kummervold et al. (2021) stated that using a larger corpus is beneficial when trying to improve the performance of a Transformer-based model. However, Bender et al. (2021) stated the negative impact these large corpora have on the environment and marginalised countries, and further suggested creating models where the quality of the data is more important than quantity. Quality over quantity could also give the right push against a non-biased language model.

Norwegian is seen as a low-resourced language. Meaning there are fewer available resources and monetary grants for research on Norwegian language technology compared to English. Everyone in the world should have the same access to technology and resources; technology should be available in all languages! There is a long way to go for this to happen, and many languages do not have the same available resources as Norwegian, even though Norwegian is seen as a low-resourced language. By dividing languages into two groups: one where there is research and funding and another where there are no such things, a bias towards languages spoken in richer countries is created (Bender et al., 2021). Therefore, transfer learning can become a valuable tool to achieve good language technology for more languages. Languages in the same language group have similar characteristics, which may be useful in transfer learning. Transfer learning is previously described in Section 2.3.3.

Gender and gender identity are much discussed topics in news and media. As mentioned in Section 1.1, people who identify as gender-fluid or another gender than assigned at birth are often vulnerable to discrimination leading to poorer mental health (Tabaac et al., 2018). Therefore, it is crucial that researchers include a broader definition of gender when investigating gender bias.

6.13. Limitations

This section describes the limitations encountered in this Master's Thesis. This includes external and internal limitations such as queuing problems at the IDUN cluster and knowledge gaps on the author's part. Since this is a Master's Thesis, time has been a significant limitation. The thesis had a time limit of 20 weeks. For a large-scale research project, this is not a long time, therefore some shortcuts will have to be made, and some mistakes may not be sufficiently solved within the time limit.

Another limitation is the available resources. As mentioned in Section 5.2.1, the IDUN cluster was used to run the code for the experiments. IDUN is a widely used platform, thus being subject to heavy user traffic which led to long queuing times, particularly in the latter stages of the experiments. Due to the protracted waiting times, there would often be significantly longer feedback loops, where the code was not immediately available for testing, and the final outcome was delayed. Consequently, the extended delay in obtaining error messages prolonged the time required for error correction.

7. Conclusion and Future Work

This chapter concludes the work done in this Master's Thesis. The research questions from Chapter 1 are answered, followed by the main contributions of this Master's Thesis. Lastly, research gaps for future work are described. This includes creating standards, inclusive evaluation of gender bias, investigating gender bias in Generative Pre-trained Transformers (GPT) and in Sami language models, and creating a gender gap tracker for Norwegian newspapers.

Research question 1 *How do current mitigating strategies for gender bias affect the performance of Norwegian language models?*

Data augmentation was chosen as the mitigation strategy of choice for the experiments performed in this thesis. It can be seen from the results that this approach did not work as expected, and bias was only mitigated in a few cases. In addition, performance was not affected by retraining to a substantial degree. Some of the reasons for this could be the smaller datasets used, how bias is measured and how performance is measured. Nevertheless, it was found that NorBERT3 scored seemingly worse on accuracy in the part-of-speech tagging compared to the two other Norwegian language models, NB-BERT and NorBERT2.

Research question 2 *How do different datasets affect the presence of gender bias in Norwegian language models?*

Based on the findings in this thesis, Norsk Aviskorpus[1] (NAK) seems to be the best performing dataset compared to Scandi-reddit[2] and NorNE (Jørgensen et al., 2020). Data augmentation was only conducted to its fullest potential for NAK. However, as mentioned in Section 6.3, the results from fine-tuning on data from Scandi-reddit could have been different if the whole dataset had been used. The gender-balanced dataset seemed to overall perform best for all language models when it comes to bias. For the performance measure, there is not one winner, however, NB-BERT fine-tuned with a gender-swapped dataset is the model which performs the best in part-of-speech tagging. Both NorBERT2 and NB-BERT

[1] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/
[2] https://huggingface.co/datasets/alexandrainst/scandi-reddit

achieve a similar score for all language models fine-tuned with augmented data. NorBERT3 performs notably worse than both NorBERT2 and NB-BERT when it comes to gender bias.

Research question 3 *Do Scandinavian language models exhibit the same gender bias as the monolingual Norwegian models, and can the same mitigation techniques be used on Scandinavian language models and Norwegian language models?*

Both KB-BERT and DanishBERT show seemingly worse results when measuring bias, compared to both NorBERT2 and NB-BERT. Thus, gender bias is present in these models as well. Neither of the Norwegian language models were checked for bias before publication, and the same goes for KB-BERT and DanishBERT. Transfer learning was tested out to mitigate bias in the Scandinavian language models, showing promising results. An increased F1 macro score could be observed in both DanishBERT and KB-BERT for some of the augmented datasets, however, this increase was small. This could be due to the small size of the training data or because of the chosen training parameters. Both DanishBERT and KB-BERT achieved a lower score overall in the part-of-speech tagging, compared to the Norwegian language models NorBERT2 and NB-BERT. Yet, DanishBERT scores higher on the POS task when retrained with the gender-neutral swapped dataset. Future work should use a larger dataset when retraining and assess if other training arguments can affect both performance and gender bias.

Research question 4 *How is gender bias in Norwegian language models affected by the introduction of the gender-neutral pronoun "hen" in the Norwegian language compared to Scandinavian language models and languages?*

There are none to few statistics as to how many people identify as non-binary in Norway. This makes it harder to find ways to measure gender bias in relation to non-binary people. For women and men, there are many statistics making it easier to calculate bias, as seen from Touileb et al. (2022). The gender-neutral pronoun "hen" therefore introduces a problem for measuring bias in Norwegian language models. In addition, "hen" is not used in that many newspapers or articles yet, as seen from the pronoun count made for both NAK and Scandi-reddit, which makes it harder for the language model to learn which context this word belongs to.

Finally, we can conclude that mitigating bias in Scandinavian language models is a challenging task. It is especially complex to measure gender bias and future research is required for both mitigating gender bias and broadening the terms of gender and gender bias.

7.1. Contributions

This thesis has shown new possibilities when mitigating gender bias from Scandinavian language models. The definition of gender bias is broadened by including the gender-neutral pronoun "hen" when augmenting the datasets in the experiments. Thus, the Master's Thesis contributes to more acceptance and equality in society and research. This approach is inspired by related work for debiasing binary gender bias, and both gender-swapping and gender-balancing are techniques used. Measuring binary gender bias is challenging and measuring non-binary gender bias is even harder as there is not much previous research in this area. In addition, gender bias is mitigated in Scandinavian language models using data augmentation and transfer learning. Data augmentation is approached using both gender-swapping and gender-balancing and it is proven that the performance of the models is maintained when debiasing. Transfer learning is performed using Norwegian augmented datasets for fine-tuning Scandinavian language models and proves that transfer learning can be efficient at mitigating bias in lower-resourced languages. In addition, gender bias is established as a performance metric juxtaposed to performance metrics by comparing the two. Lastly, the thesis provides clear proof that datasets from social media, such as Scandi-reddit, can be less gender-biased compared to traditional datasets gathered from newspapers. In conclusion, the research goal of the Master's Thesis has been reached:

Goal *Mitigate gender bias in Scandinavian language models through data augmentation and broaden the definition of gender in Norwegian language technology.*

7.2. Future Work

In this thesis, many topics discussed still maintain as research gaps that can be filled in future work. These gaps are presented in the following sections.

7.2.1. Create Better Training Data

Training data is the key to creating and fine-tuning language models. From the results it can be seen that Scandi-reddit have a more equal representation of pronouns than NAK, however, NorBERT2 fine-tuned using Scandi-reddit contains more bias than NorBERT2 fine-tuned using NAK. When looking closer into this, it was found that the subset of Scandi-reddit that was used for fine-tuning had almost the same representation as NAK. This shows the importance of training data, and in future work, it is important to curate the datasets to have an equal representation of the genders.

7.2.2. Create and Use Standards

Within the field of natural language processing and gender bias, there are many definitions. All these definitions makes it difficult to find a standardised way to measure bias. Likewise, comparing the results from different models is hard, as bias and performance are measured differently by each researcher. In the future, standard ways to both measure and mitigate bias should be made. In addition, a threshold value for bias could be introduced. Another option could be to require researchers to measure bias before publishing new language models.

NorBench (Samuel et al., 2023), as described in Section 4.4, proposes a standardised benchmark for Norwegian language models. Samuel et al. (2023) also include a gender bias evaluation and a harmfulness score. This is important for comparing results and evaluating models. Future work should therefore include NorBench as a standardised benchmark.

7.2.3. Inclusive Evaluation of Bias

In addition to creating standards, a big focus point of future work should investigate which possibilities there are regarding measuring bias with concern to people identifying as non-binary. As there does not exist much statistics on people identifying as non-binary this is a challenging task, however, important.

7.2.4. Investigate Gender Bias in Generative Pre-Trained Transformers

As generative artificial intelligence has accelerated in development and interest, it is important that gender bias is investigated in these models as well. GPT-SW3 (Ekgren et al., 2023) is one example where it should be investigated if gender bias is present, and if so how to mitigate it. In this thesis, preliminary research has been conducted in this area. However, future research should go more in-depth.

7.2.5. Create a Gender Gap Tracker

Bias in language models comes from the training data it is trained on. Usually, the training data comes from newspapers and books. From the study made by Macharia (2020), it was found that only 28% of the cited sources in Norwegian newspapers are female. Asr et al. (2021) developed a gender gap tracker for Canadian newspapers. Creating a gender gap tracker could motivate journalists to write more inclusive articles, and thus contribute to a more equal representation in the training data of language models. As suggested by Lossius and Ruud (2022) a gender gap tracker

for Norwegian newspapers could be made to contribute to fair Norwegian language technology.

7.2.6. Investigate Gender Bias in the Sami Languages

The Sami languages are protected under the Norwegian language law and are equivalent to the Norwegian language. Future work should therefore include the Sami languages in language models. The National Library of Norway has published a Sami model on Huggingface called NB-BERT-base-samisk[3]. Future work includes examining the model for gender bias and possibly mitigating that bias. This is especially interesting in Sami languages as there are no gendered pronouns.

[3] https://huggingface.co/NbAiLab/nb-bert-base-samisk

Bibliography

Jakob Semb Aasmundsen. Språkrådet har vedtatt «hen». flere kjønnsnøytrale pronomen kan følge etter., 2022a. URL https://www.aftenposten.no/kultur/i/47Pjaq/spraakraadet-har-vedtatt-hen-flere-kjoennsnoeytrale-pronomen-kan-foelge-etter. Accessed at: 16th of September 2022.

Jakob Semb Aasmundsen. Norge kan få et tredje kjønn, 2022b. URL https://www.aftenposten.no/kultur/i/L5dq1V/norge-kan-faa-et-tredje-kjoenn. Accessed at: 16th of September 2022.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

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. *PLoS ONE*, 16(1):e0245533, 2021. doi:http://dx.doi.org/10.1371/journal.pone.0245533.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi:10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

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 *Advances in Neural Information Processing Systems*, volume 29, page 4356–4364, Barcelona, Spain„ 2016. Curran Associates, Inc.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. How conservative are language models? Adapting to the introduction of gender-neutral pronouns. *arXiv preprint arXiv:2204.10281*, 2022.

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. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 33:1877–1901, 2020.

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. Association for Computational Linguistics, 2020. doi:http://dx.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. *arXiv preprint arXiv:1312.3005*, 2013.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

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, 2020.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021.

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, pages 2232–2242. Association for Computational Linguistics, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. GPT-SW3: An autoregressive language model for the nordic languages. *arXiv preprint arXiv:2305.12987*, May 2023.

Kawin Ethayarajh. Is your classifier actually biased? Measuring fairness under uncertainty with Bernstein bounds. *arXiv preprint arXiv:2004.12332*, 2020.

Ruth Vatvedt Fjeld. Om ordbokseksempler og stereotypisering av kjønn i noen nordiske ordbøker. *Perspektiv på lexikografi, grammatik och språkpolitik i Norden. Helsinki: Institutet för de inhemska språken*, pages 35–65, 2015.

Emmie Stolpe Foss. 1 av 9 har nynorsk som hovudmål i skolen, 2022. URL https://www.ssb.no/utdanning/grunnskoler/statistikk/elevar-i-grunnskolen/artikler/1-av-10-har-nynorsk-som-hovudmal-i-skolen. Accessed at: 15th of October 2022.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transaction on Inforamtion Systems (TOIS)*, 14(3):330–347, jul 1996. ISSN 1046-8188. doi:10.1145/230538.230561. URL https://doi.org/10.1145/230538.230561.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. Type b reflexivization as an unambiguous testbed for multilingual multi-task gender bias. *arXiv preprint arXiv:2009.11982*, 2020.

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, Marthen H. van

Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abassi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.559.

Dan Jurafsky and James H. Martin. *Speech and Language Processing*. Stanford University, 3rd edition, January 2023. URL https://web.stanford.edu/~jurafsky/slp3/.

Per E Kummervold, Javier de 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 (Online), 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. *arXiv preprint arXiv:1906.07337*, 2019.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Anna Lindqvist, Emma Aurora Renström, and Marie Gustafsson Sendén. Reducing a male bias in language? Establishing the efficiency of three different gender-fair language strategies. *Sex Roles*, 81:109–117, 2018. doi:https://doi.org/10.1007/s11199-018-0974-9.

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. *arXiv preprint arXiv:1907.11692*, 2019.

Andrine Lossius and Regine Pösche Ruud. You shall know a female word by the company it does not keep. Master's thesis, Norwegian University of Science and Technology, 2022.

Sarah Macharia. *Who makes the news? Global Media Monitoring Project (GMMP)*. World Association for Christian Communication, 2020. ISBN 978-1-7778038-0-3.

Martin Malmsten, Love Börjeson, and Chris Haffenden. Playing with words at the National Library of Sweden – making a Swedish BERT, 2020.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: A tasty French language model. *arXiv preprint arXiv:1911.03894*, 2019.

Wes McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, 2010.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 231–232. Association for Computing Machinery, 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

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.

Norwegian Ministry of Local Government and Modernisation. National strategy for artificial intelligence. 2020.

Briony J Oates. *Researching information systems and computing*. SAGE publications, 2006. ISBN 1-4129-0223-1.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. The Nordic Pile: A 1.2 TB Nordic dataset for language modeling. *arXiv preprint arXiv:2303.17183*, 2023.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, oct 2010. ISSN 1041-4347. doi:10.1109/TKDE.2009.191. URL https://doi.org/10.1109/TKDE.2009.191.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Sooumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Cristopher 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, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, and Matt Gardner. Deep contextualized word representation. *arXiv preprint arXiv:1802.05365*, 2018. doi:https://doi.org/10.48550/arXiv.1802.05365.

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. 21(1), jan 2020. ISSN 1532-4435.

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*. arXiv preprint arXiv:1804.09301, 2018.

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

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Sergeevna Palatkina. Norbench – a benchmark for norwegian language models. In *The 24rd Nordic Conference on Computational Linguistics*, 2023. URL https://openreview.net/forum?id=WgxNONkAbz.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. The Norwegian dependency treebank. 2014.

Språklova. Lov om språk. (lov-2021-05-21-42), 2022. URL https://lovdata.no/dokument/NL/lov/2021-05-21-42. Accessed at: 5th of October 2022.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*, 2019.

Karolina Stańczak and Isabelle Augenstein. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*, 2021.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Ariella Tabaac, Paul B. Perrin, and Eric G. Benotsch. Discrimination, mental health, and body image among transgender and gender-non-binary individuals: Constructing a multiple mediational path model. *Journal of Gay & Lesbian Social Services*, 30(1):1–16, 2018. doi:10.1080/10538720.2017.1408514. URL https://doi.org/10.1080/10538720.2017.1408514. PMID: 30880881.

Rolf Theil. Genus (grammatikk), 2022. URL https://snl.no/genus_-_grammatikk. Accessed at: 22nd of September 2022.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, 2022.

United Nations. Transforming our world: the 2030 agenda for sustainable development. *United Nations: New York, NY, USA*, 2015.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*, 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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Matthew Watson, Chen Qian, Jonathan Bischof, François Chollet, et al. Kerasnlp. https://github.com/keras-team/keras-nlp, 2022.

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.

Thomas Wolf, Lysandre Debu, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. pages 38–45. Association for Computational Linguistics, 2020. URL https://github.com/huggingface/transformers.

World Health Organization. Gender and health, 2022. URL https://www.who.int/health-topics/gender#tab=tab_1. Accessed at: 22nd of September 2022.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Saadia Zahidi. The global gender gap report 2022. World Economic Forum, 2022. ISBN 978-2-940631-36-0.

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, pages 15–20. Association for Computational Linguistics, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1. Association for Computational Linguistics, 2019.

Appendices

A. Description of the Code Base

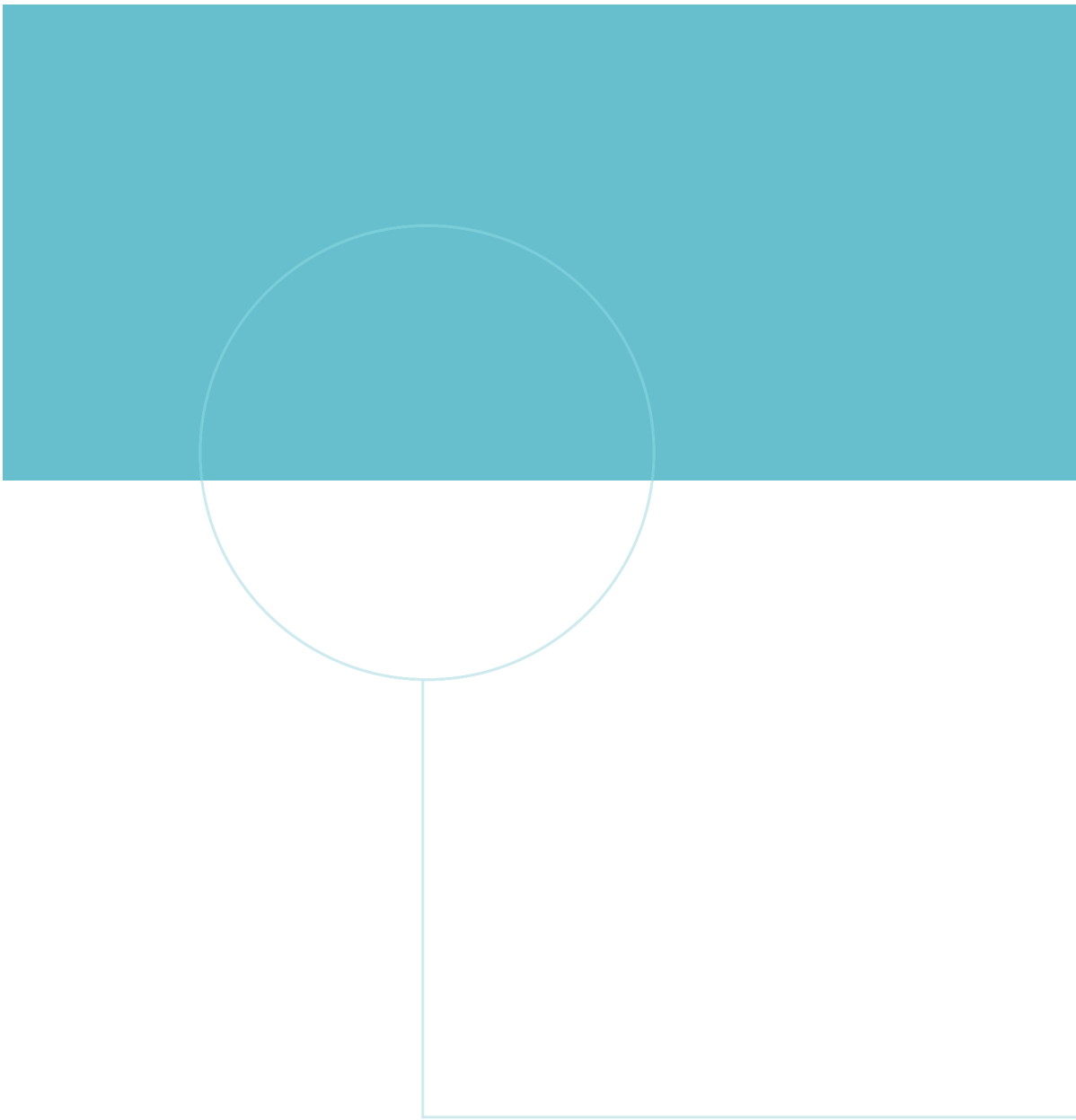
This appendix describes the code base created throughout this Master's Thesis. The code base can be found on GitHub[1], including a `README.md` which explains how the code is run.

The code is structured as scripts that can be run for the experiments. The folder `scripts/..` includes all of the code and following is a description of each of the files:

- `scripts/balance_dataset.py`: code used to generate datasets for experiments 2 and 4 as described in Chapter 5.
- `scripts/data_clean.py`: code used to clean the datasets before swapping or balancing.
- `scripts/evaluation.py`: code used to evaluate the models for bias and accuracy.
- `scripts/fine_tuning.py`: code used to fine-tune the models.
- `scripts/gender_swap.py`: code used to generate datasets for experiments 1 and 4 as described in Chapter 5.
- `scripts/make_dataset.py`: makes all datasets and counts pronouns.
- `scripts/ner.py`: code used to anonymise data in the datasets.
- `scripts/pos.py`: code used to fine-tune the models for part-of-speech tagging.
- `scripts/predict.py`: code used to predict masked words and generate sentences.
- `scripts/pronoun_counting.py`: code used to count pronouns in the datasets. Used in experiment 4 as described in Chapter 5.

In addition, `slurm/job.slurm` includes an example of how a Slurm file is structured to run the code at IDUN.

[1] https://github.com/ingvlt/master-project



NTNU

Norwegian University of
Science and Technology