MAY 15, 2019

INVESTIGATING GENDER BIAS IN JOB ADVERTISEMENTS WITH WORD EMBEDDINGS

MASTER THESIS

KRISTA VÁGSHEYG, 93615
CHARLOTTE SOPHIE WILHELMSEN, 93173
COPENHAGEN BUSINESS SCHOOL
CAND.MERC.IT

SUPERVISOR: DANIEL HARDT
NUMBER OF CHARACTERS: 178.672
NUMBER OF PAGES: 78,5

ABSTRACT

This thesis is an explorative study that investigates gender bias in Danish job advertisements from the platform Jobindex that is the largest Danish database of job advertisements. The methods used in this thesis combines the fields of computer science with linguistics to utilize natural language processing and specifically word embeddings. The technology used to investigate gender bias in the advertisement is provided by fastText, which was created by an AI research team from Facebook.

This thesis attempts to automatically calculate a gender bias score from the words in a job advertisement, by comparing the similarity of the advertisement to the male and female identifiers 'han' and 'hun'. This is performed by vectorising all terms in the advertisement and averaging the score of the advertisement using the cosine angle of each vector. The scores range from -1 being extremely female bias, and 1 being extremely male bias, scores that are close to 0 are interpreted as neutral.

The empirical data was collected from Jobindex and consist of four years of job advertisements: 2008, 2014, 2017 and 2019, which we compare to the statistics of gender distribution of the Danish industries using the data provided by Statistics Denmark. Our approach in this thesis was to manually annotate 100 advertisements to uncover subconscious bias and whether the advertisements were directed towards a specific gender, which was compared to the automatic scores.

Our results show that the word embeddings can be used to uncover bias, however, there are several questionable aspects to the word embeddings. We found that the level of bias in function words is high, and therefore impacts the overall score. Furthermore, we found that certain occupations related to both teaching and especially public services were very male bias in the pretrained Wikipedia model.

For future work of this project we would like to further develop our approach by discounting function words and building an automatic classifier by extensively increasing the manual annotation.

Table of Contents

1 INTRODUCTION

The emerging use of technology has changed how humans live, behave and interact. Its area of application is wide with a great difference in scope and scale, ranging from automated ordering systems (Chen et al., 2018), to voice assistant services such as Amazon Alexa (Hoy, 2018), to advanced ERP systems (Stadtler, 2002). The possibilities are large and the limitations low resulting in continuous growth and development within these technologies.

The universal adoption of technologies like social media and the world wide web has generated an immense amount of data. The vast textual data has increased the use of text analysis methods for processing and understanding natural human language through computational activities like natural language processing (Deng and Liu, 2018).

The applications of natural language processing (NLP) are wide and powerful. Machine translation is possible through NLP that understands the meaning of a sentence rather than translating each individual word making language barriers smaller (Kim, 2014). Automatic summarisation helps summarise the meaning and information of a text as well as giving insight into the emotional meaning (Tarasov, 2015). Sentiment analysis captures the opinion of an expression and can be applied by companies who seek to understand their customers opinion of their products.

Recently there has been a larger focus on computational linguistic and the semantic and syntactic meaning of words generated from word embeddings by distributing the representation of a text in an n-dimensional vector space (Mikolov et al., 2013). Word embeddings are an extended activity of deep learning through neural nets which are essential for solving many natural language processing problems (Zhang et al., 2014).

The vectors of the words are used to calculate the meaning of each word and can be compared to other vectors to discover similarities or differences (Liu et al., 2018). For this paper we will explore how a word embedding model performs on Danish job advertisements. We want to investigate whether it is possible to identify gender bias within Danish job advertisements drawing from the theoretical framework about the importance of diversity in businesses.

Our interest in gender bias and job advertisements stems from the increase in popularity of diversity practises in businesses. We therefore want to bridge our technical learnings with the business learnings from our study at Copenhagen Business School. Both authors of this thesis have participated in courses in big data analytics and conducted projects on text analytics. Machine learning and text classification has been very interesting to study for us as a lot of information and great results are generated from it. Based on this both of us wanted to continue our studies within natural language processing, but this time for a study at a larger scope and scale hopefully giving even more detailed results.

Diversity has become a popular topic in the news picture, and generally a great focus in society. Companies and institutions have seen the beneficial impacts of diversity, and are striving to implement diversity practices by, among others, working on obtaining gender equality for their employees (lederne.dk, 2017). As a consequence of continuously being exposed to this topic, as well as us being genuinely interested in it, we have chosen to include the topic in our study. Both of us are studying business and IT, which is a male-dominated study program, and we are therefore very familiar with being in minority. This has also triggered our curiosity to further explore diversity and its impact.

We want to expand our knowledge within natural language processing with this thesis. We have chosen to explore job advertisements because little studies have been done within this field. Moreover, it is a topic that has been important for both of us, as we have gone through many job advertisements in our application phase when applying for full-time positions. We chose to look at Danish job advertisements mainly because it interested us most as we both have applied for jobs within Scandinavia. Additionally, there were no similar studies done on Danish job advertisements making it even more compelling. This leaves us with the following research questions:

Is it possible to identify gender bias in job advertisements using word embeddings?

How do the results from the word embeddings compare to the gender distribution in the Danish work force?

We will attempt to answer these questions by collecting Danish job advertisements from Jobindex and calculating the similarities of an advertisement to the Danish terms ‘han’ and ‘hun’. This allows us to discover if an advertisement is closer to the term ‘han’ making it male bias, or ‘hun’ making it female bias. From the scores that we generate we will compare these to the gender distribution of the industries from the statistics of Statistics Denmark. As a remark, we both believe that gender is not binary, however for the limitations of this thesis and because of the limitations of the Danish language we will only be focusing on male vs female.

This paper continues with the following layout: first we will cover the theoretical framework of computational linguistics and natural language processing which will give an insight to the technologies and methods used to develop word embedding technologies. Then we cover the business aspects of diversity and here we will primarily focus on gender diversity and its’ impact on firm performance.

We move on to the literature review which was conducted to examine which similar studies have been researched. Then follows the methodology where we present how we collected our data, and how we processed it and calculated the gender bias of the advertisements. Afterwards we present our results and compare them to the data provided from Statistics Denmark. At last we discuss the applicability of our approach and the future work.

All python codes used in the project are attached as a .zip file in the appendix and can be run with a jupyter notebook[1].

[1] https://jupyter.readthedocs.io/en/latest/install.html

2 THEORETICAL FRAMEWORK

In this section we cover the theoretical aspects of this paper. We will describe natural language processing and computational linguistics and other fields such as artificial intelligence (AI) and machine learning (ML) which are linked to NLP. We will also cover theories about bias and at last we cover the business theories about diversity and inclusion practises.

2.1 Artificial intelligence and language

Artificial intelligence can be rooted back to 1930s and has since then been explained and defined in various ways (Ertel, 2017). In 1955, John McCarthy described artificial intelligence as "The goal of AI is to develop machines that behave as though they were intelligent" (Newell et al., 1957). This early definition is partly right but lacks the aspect of artificial intelligence solving more practical tasks. Looking into the newer literature, Elaine Rich has a broader definition that goes as follows '*Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better*' (Rich, 1983). Here the point that humans are still smarter and more intelligent is underlined, which is important to keep in mind. Even though machines have become able to make decisions for different challenges and problems, they are far from superior to humans on all areas of applications.

Artificial intelligence differs from other types of science, due to its interdisciplinary (Ertel, 2017). It is built upon findings from a variety of different fields such as, linguistics, philosophy, neurobiology and statistics. It is an approach used in modelling, cognitive processes and to replicate and recreate intelligence based on logical, mathematical and computational principles (Frankish and Ramsey, 2014). This makes it applicable for many areas and involves contributions from a variety of scientists, which makes it complex, yet highly important.

The combination of artificial intelligence and natural language processes have increased its importance over the last years as a result of emerging methods for processing text. Artificial intelligence has especially become important recently due to its beneficial ways of understanding language based on repetitive patterns of phrases (Forbes, 2017). Compared to traditional learning methods within language, including learning through repeating grammar and vocabulary, artificial intelligence understands the relation between words through repeating phrases. This also allows for a larger amount of data to be learned in shorter time. Additionally, artificial intelligence systems have become able to learn languages themselves by implementing neural networks, making them very convenient and time saving (Artetxe et al., 2017). Rapid learning has become one of the drivers for the fast-developing technology, and automated language through artificial intelligence is no exception.

2.2 Machine learning

Machine learning is as old as artificial intelligence but has become more and more important as the artificial intelligence systems evolve and become more complex (Frankish and Ramsey, 2014). Like artificial intelligence, machine learning is applicable in multiple fields, from philosophy to mathematics and sociology (Frankish and Ramsey, 2014). What is unique for machine learning is the ability to train models to think and solve human problems without being exactly programmed to do so. Models learn through experiences over time and make it possible to predict the outcome of a future problem (Bell and Jason, 2014). In addition to being able to predict the solution of a problem, called predictive learning, machine learning can also gain knowledge from data, called descriptive knowledge (Alaydin, 2014). The models learn from experiences in the past and knowledge from the data, generated through the machine learning, which makes it possible to perform tasks and recognize patterns in a new dataset (Frankish and Ramsey, 2014). This is convenient as some datasets are challenging or impossible to know the solution for.

The performance of a machine learning program is seen as successful if the performance evolves and improves with more experiences (Frankish and Ramsey, 2014). The main purpose is to find patterns and useful information from complex data, which will be useful and affective for decision making and improvements. Machine learning has become very important for understanding and taking use of text data. Text data is classified as unstructured data, underlining the challenges related to analyse it, due to its structure. Machine learning has made it possible to implement various natural language processing methods, including speech recognition, topic modelling and automatic translation. All methods generate beneficial outputs based on the underlying structures in the text. To be able to get a useful output, a set of actions need to be performed, called the machine learning cycle (Bell and Jason, 2014).

FIGURE 1. THE MACHINE LEARNING CYCLE (BELL AND JANSON 2014).

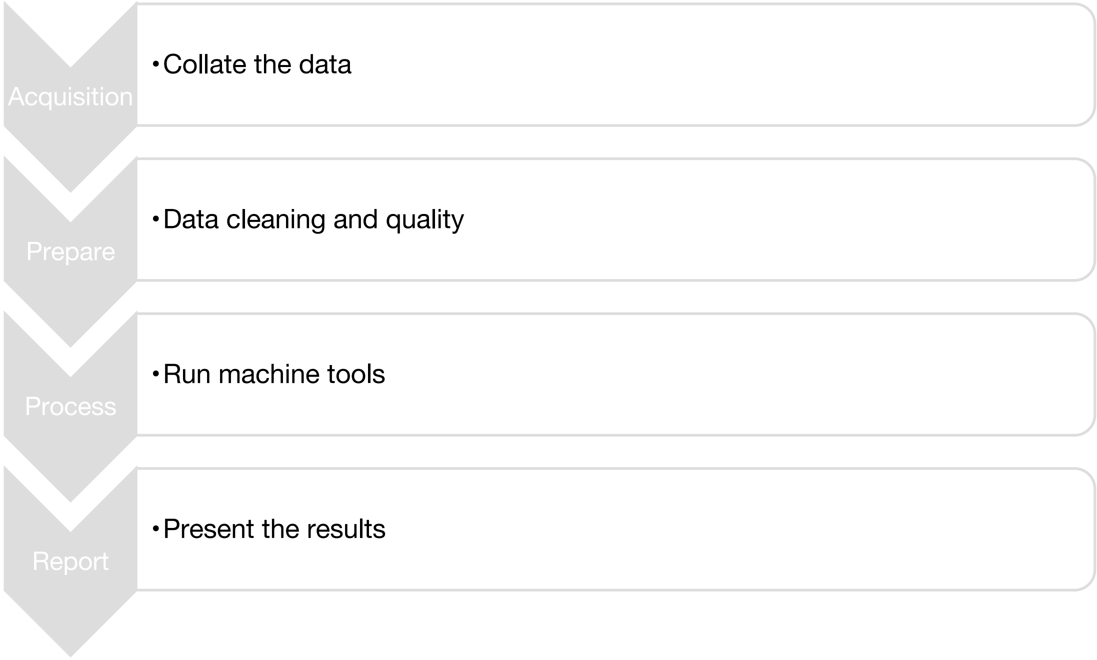


Figure 1 shows the different activities involved to generate useful output from the machine. Collating and cleaning the data are necessary to ensure quality, and to ensure that the data can be used to train the machine. Applying the ML tools is done in the process phase, and the results are shown in the following report phase. The algorithms used in the process phase are chosen based on the desired output (Bell and Jason, 2014). That said, two types of learning define which group the algorithm belongs to, namely supervised or unsupervised learning. These approaches will be further described in the section about text analytics.

2.3 Computational linguistic

Computational linguistics aims to understand language, both written and spoken, from a computational perspective (Schubert and Lenhart, 2019). The discipline combines both science and engineering to build models that process and produce language in a useful dialogue setting. Computational linguistics has its roots in mechanical translation and became a phenomenon in the mid 1960s (Mitkov, 2005). The goal was to create and implement systems which complemented the core purposes and functionalities of linguistic theories. Additionally, language is shaped and processed in human minds, meaning that language reflects and gives insight into intelligence and thinking processes (Schubert and Lenhart, 2019). Computers that are linguistically intelligent therefore meet a crucial need among humans as language is one of the most vital ways of communication.

Linguistic theories include frameworks of grammatical, syntactic, content, and semantic formulations made to characterise languages with computational methods (Schubert and Lenhart, 2019). Additionally, there are cognitive and neuroscience models focusing on how the processes and learning within language take place in the human brain.

There has been a paradigm shift in natural language research from a rationalist approach which depended on a lot of domain specific knowledge and large databases that used hand-coded rules. This turned out to be a very difficult and time-consuming approach and inspired a resurgence in empirical methods. The empirical method is corpus-based and applies learning techniques to extract linguistics knowledge from large corpora of text (Brill & Mooney 1997).

The ultimate goals of applying computational linguistics are broad, and vary from answering questions, both simple and discoverable, and summarization of text including analysis, sentiment and discovering psychological aspects. Moreover, dialogue might be discovered to accomplish tasks or finding answers, and lastly computational systems similar to human competency within language, dialogue and knowledge are highly desirable (Schubert and Lenhart, 2019). Many different methods have been used within computational linguistics since its occurrence in the 60s. The shift that took place in the 1980s was a result of the growing volume of machine-readable text and speech data available. This led to the use of the statistically based techniques in addition to the use of meaning-based techniques.

Semantic method is a sub-category of computational linguistics and focus on the linguistic meaning (Mitkov, 2005). It is an important study as it analyses the meaning of the linguistic through a computational approach to natural language. The meaning is modelled through looking at the individual meaning of the words in the phrase or sentence based on their appearance, and then adding it up together to get the overall meaning. Finding the meaning and thus understanding the text is heavily dependent on background knowledge. This includes large data on similar occasions and situations that the model can be trained on (Schubert and Lenhart, 2019).

Moreover, the corpus-based statistical method focuses on the distributional properties of language (Schubert and Lenhart, 2019). This method was created to cope with the scalability issues and is able to process a large scale of data (Jurafsky and Martin, 2012). Among the areas where this method is applicable is within text recognizers, which have become very accurate and comprehensive (Schubert and Lenhart, 2019). This has been accomplished by building models trained on a corpus of text, to process and explore new text based on the already known corpus. Resultingly languages and texts are processed and compared to other texts and languages to find similarities and patterns.

2.4 Natural Language Processing

Natural language processing focus on understanding human language to perform tasks through computer processes (Deng and Liu, 2018). Natural language processing started out as machine translation in the late 1940s (Chipman, 2017). The aim of machine translation was to use software to build practical methods to translate text from one language to another. However, it was early on understood that translating text word for word was not the way to achieve successful translations. Humans rephrase and express themselves with different words depending on the context they are in. Consequently, instead of word-for-word translation, translating text in relation to its context became vital to generate useful translation. This became the starting point for computational linguistics and later on natural language processing. Moreover, it has affected other areas of application, such as theoretical linguistics and artificial intelligence (Chipman, 2017).

The development of natural language processing builds upon cybernetics and the use of logical reasoning (Chipman, 2017). Cybernetics inspired to the connecting approach of modelling language together with cognition. Furthermore, McCulloch used logical reasoning through artificial neurons, which advanced formal logic (Chipman, 2017). Formal logic included studies of syntax and semantics of language, which finally contributed to the creation of natural language processing systems. Additionally, the principles of successful cryptography used during World War II contributed to information theory (Chipman, 2017). Information theory has played an important role for approaches used for language processing, including statistical and machine learning.

Natural language processing aims to build algorithms that generate and understand text found in natural language (Maynard and Bontcheva, 2019). When creating a natural language processing model, it is often done to be applicable in various languages. The composition does often include a pipeline of processing resources which are replicable to different languages. This include resources that can be removed, adjusted and added depending on the language being analysed (Maynard and Bontcheva, 2019). When building a model, a lot of pre-processing is done on the text to get it to fit in the computational model. The pre-processing include cleaning the data such as tokenization and removing prefixes, stopwords etc. Different pre-processing methods exist for the different languages and this has to be adjusted to take use of the processing resources.

Natural language processing intends to bridge and simplify the interaction between a computer and a human (Deng and Liu, 2018). The fundamental and general concept of natural language is to provide the semantics. Humans take use of context to understand the meaning of words included in the text, and that is the major concern for a computer. Giving meaning to a single word rises many questions as a single word can have multiple meanings based on the context it is written in. Consequently, the use of different machine learning techniques, such as deep learning and word embeddings, are important to create useful results of the processed language (Deng and Liu, 2018). Both techniques will be elaborated further later on in the thesis.

2.5 Text analytics

Text analytics are concerned with analysing text to discover patterns, structures, similarities and other useful information (Provost and Fawcett, 2013). Text data is complex due to its raw and unstructured form it appears in and it therefore requires work to convert it to be suitable for modelling (Dean, 2014). That said, when the pre-processing is done, text data can contribute with highly crucial information. Consequently, it can impact the company's competitive ability in the market, which makes it a topic companies are striving to take advantage of.

The internet is often described as the "new media", even though the content is similar to the old media types (Provost and Fawcett, 2013). This has generated a lot of online available text data, which has become an important source for analysis. Text is mainly categorized as unstructured data, which means that it is not structured as other sorts of data, such as numbers (Dean, 2014). Text is intended for humans, and this often complicates computers' ability to understand it. Challenges include, among others, abbreviations, misspelling, random punctuating, word order and synonyms. Additionally, the context is crucial as, for example, negative topics can be explained by using positive words, and vice versa, which easily leads to misperception. To avoid all these pitfalls, a lot of pre-processing needs to be performed on text data (Provost and Fawcett, 2013).

2.5.1 Text classification

Text classification aims to classify new documents in classes that are pre-defined (Mironczuk and Protasiewicz, 2018). It includes training models, as well as, among others, pre-processing of data and transformation. The main problem with text classification is to acquire enough labelled data to train the model to perform accurately (Nigam et al., 2000). Data acquired from the internet is often unlabelled therefore labelling tends to be done manually which is a very time-consuming process (Nigam et al., 2000). This time process often limits the number of labelled data that the model can be trained on.

That said, this unlabelled text data provides useful information based on the distribution of the words within the text (Nigam et al., 2000). Increased accuracy is to be found in some cases when the unlabelled data is used together with a selection of labelled data. This is due to labelled data that needs to determine some instances for the different classes. Resulting, as far as enough labelled data is available to determine which class the different instances belongs to, the unlabelled data can be used to estimate the parameters (Nigam et al., 2000).

These elements are applicable in this thesis as all the job announcements from Jobindex are unlabelled and labelling them manually would be very time consuming. Word representations through word embeddings are used for generating gender bias scores, and thereby classification. Ultimately this combination should provide useful results about the bias within the advertisements.

2.5.2 Supervised learning

Supervised learning is the process where a machine learns from data where the output is known (Kashyap, 2017). This approach is done for a specific purpose, namely predicting the target variable (Provost and Fawcett, 2013). To be able to predict the target, there needs to be enough data available on the target. The training process takes place until a certain level of accurate classification is achieved (Kashyap, 2017).

Regression, classification and modelling are generally done by using supervised learning methods. However, this will only be achieved if there is adequate historical data on the target. Acquiring enough data is time consuming and a big investment but provides a much more precise learning model (Provost and Fawcett, 2013).

2.5.3 Unsupervised learning

For unsupervised learning the outcome is not known (Kashyap, 2017). Within this method there is no information on the learning, nor the purpose of it, and the conclusions are made from similarities found in the data (Provost and Fawcett, 2013). The outcome is a result of exploration of the algorithms, as well as finding structures in the available dataset (Kashyap, 2017). Clusters are one of the commonly used methods of analysing the unstructured data, done by uncovering similarities in the dataset (Kubat, 2017). Additionally, nearest neighbour and self-organizing maps are other techniques applicable for unstructured data (Kashyap, 2017). The unsupervised learning method is explorative and has no guarantee that the outcome will be useful for a particular purpose. That said, it is one of the most crucial methods used for analysing text data which has become an important task within data science, as this data type has evolved with the increased use of online interaction.

2.6 Neural networks

In a biological definition, neural networks consist of nerve cells networks and is located in the human brain (Ertel, 2017). The networks are created by roughly 100 billion nerve cells and are strongly related to the intelligence of humans (Ertel, 2017). The connections of nerve cells are the reasons for thoughts, associations, consciousness and humans' ability to learn. These networks of connections where previously seen as the fundament to create useful and smart artificial intelligence programs that meet the demands of humans (Frankish and Ramsey, 2014). Consequently, the science side of artificial intelligence has consisted of understanding human intelligence, to create intelligent systems.

Neural networks used today within machine learning are defined as computational tools for processing language (Jurafsky and Martin, 2012). Originally the networks were called neural as it was considered a simplified illustration of the human neuron network. This point of view has evolved and changed with time and todays language processing do not draw direct similarities from biological inspirations. One of the main reasons for this change is better understanding of the human brain and its complexity (Ertel, 2017). The structure of the human brain is changing continuously due to its adaptation of environmental influences. What is to be observed in today's machine learning is the modern neural networks created from computational units (Jurafsky and Martin, 2012). In detail, all computed units produce an output value based on a vector of input values. That means that an output is produced when there is performed computation on some of the valuable inputs.

As many of today's neural networks have numerous layers, they are considered as deep and named deep learning (Jurafsky and Martin, 2012). This is one of the key parameters for why neural networks are powerful. An early layer has the ability to learn representations which can be useful for later layers within the network. A detailed description of a deep learning neural network will not be described here as it is highly complex involving many mathematic formulas.

Deep learning is a major part of natural language processing due to its powerfulness (Deng and Liu, 2018). The previously challenges with processing huge amounts of training data in natural language processing techniques disappeared with deep learning because of the structural composition. Additionally, earlier natural language process techniques involved a lot of manual work such as feature processes. This hit a huge problem due to lack of prominent computer engineers. Both problems were solved with neural networks containing deep layers able to solve general tasks within machine learning by distributing the feature engineering (Deng and Liu, 2018). In neural networks feature extraction is possible through learning representations from the available data using the multiple processing layers of units. Resulting, a hierarchy of concepts is formed as lower level features determine features to higher level of features.

Lastly, the simplicity of deep learning is also to be found in the design of the models (Deng and Liu, 2018). Deep learning can be performed together at the same time for all parts of the model, including tasks related to feature extraction as well as prediction. Moreover, the model is constructed by the same building blocks used generally in various application, making it easy to use the same data in more than one model, as well as replicate a specific task (Deng and Liu, 2018). Resulting, deep neural networks have become a unified method for different machine learning and artificial intelligence techniques used on large datasets, including natural language processing.

2.7 Word embeddings

Word embeddings are distributed word representations in form of vectors that are trained on deep neural networks (Zhang et al., 2014). To better understand the meaning of a word, the company, referred to as the additional words in the sentence, needs to be understood (Senel, 2018). Word embeddings is a form of understanding this company of words through word representations, which is a subcategory of natural language processing (Liu et al.,2018). The embeddings can represent a large amount of words as vectors within a semantic space, and works great in finding distributional similarities, both syntactic and semantic, between words (Jurafsky and Martin, 2012 and Zhang et al., 2014). Word embeddings learn to transform each single word in the raw data into an n-dimensional vector space, where each vector can be compared to each other (Liu et al., 2018).

Semantic and syntactic similarities between words are often the goal in word embeddings (Levy and Goldberg, 2014). To achieve this goal, the hypothesis of Harris 1954 is often used, it aims to show that words occurring in similar contexts have similar meanings (Levy and Goldberg, 2014; Zelling, 1954). Word representations can be done in a variety of methods, but most of them are determined within natural language processing. The range is stretching from clusters based on the context the words are within to high dimensional vectors measured by the association between the given context and the word (Levy and Goldberg, 2014). Moreover, neural network language modelling has also affected word embeddings by introducing dense vectors which are vectors derived by training models. This has further improved the results generated in word embeddings which in general have performed well.

Word embeddings are considered easy to use as low-dimensional matrixes created by efficient computation are used to determine word similarities (Levy and Goldberg, 2014). The skip-gram model implemented in the word2vec software[2] made by Google has become one of the most used word embedding method. The method is easy to train, produces valuable word representations, and is scalable for large corpora and huge vocabularies, both for words and context (Levy and Goldberg, 2014). As context plays an important role for the word similarities, different context means different similarities among the words. The skip-gram model generates broad topical similarities, whereas context based on dependency generates more functional similarities. Neural word-embeddings on the other side is considered more challenging to assign useful similarities

[2] code.google.com/p/word2vec/

as the dimensions of the representations are more undistinguishable (Levy and Goldber, 2014).

However, there are difficulties to be found within the word embeddings' approach. That include the size of the lexicon and challenges regarding recognizing the word. A too large lexicon is problematic as it is difficult to represent every word as an embedding. Moreover, words can be difficult to recognize as new words are included in the language, words from other languages are used, and misspelling occurs (Jurafsky and Martin, 2012). Challenges related to categorizing the words right do also arise (Liu et al., 2018). Even though the context is addressed through word contexts, there are challenges related to task-specific features. A good classification is achieved if the word distribution has clear boundaries for each class (Liu et al., 2018). This is challenging and often hard to reach as some word embeddings methods only focus on the similarities between words. Resulting, many words are classified as closely related based on the given context, even though they have different meanings and should have been divided by a clear boundary. To cope with both aspects, task-specific features and the semantics of words should be processed and found in the word embeddings methods (Liu et al., 2018).

In this thesis we apply Facebooks' word embedding software fastText to calculate the cosine similarity between each word vector in the job advertisement and the male and female identifiers 'han' and 'hun'.

$$similarity = \cos(0) = \frac{A * B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i \, B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

The formula above shows the two vectors *A* and *B* where the cosine similarity is represented as the dot product and magnitude. The similarity is measured with the cosine of the angel between each vector with the python library gensim[3].

[3] https://radimrehurek.com/gensim/models/keyedvectors.html

2.8 Diversity and Inclusion

Diversity and inclusion are two highly related terms but differ in the way they are obtained. Diversity is all about people based on their individual differences (Kreitz, 2008). Inclusion on the other side aims to increase diversity through inclusive practises (Schmidt et al., 2017). The benefits of a diverse firm will not be achieved without inclusive practises and is therefore a crucial part of improving diversity in a business setting. Diversity management is seen as a strategic way of utilizing the benefits of diversity (Richard et al., 2013). The organization should effectively strive for employee diversity and create opportunities based on their unique abilities and skills (Luu et al., 2008).

Diversity should be seen as the variety of perspectives and approaches to working tasks brought together by all identity groups (Thomas and Ely, 2002). When this takes place, companies are better positioned to grow and improve simply by challenging primary activities and assumptions. Several approaches concerning diversity are stated in the literature. Among the assumptions are strategies, practices procedures and approaches. Starting with the study by Thomas and Ely 2002, initiatives for diversity can be divided into two perspectives; a discrimination and fairness paradigm, and an access and legitimacy paradigm.

As understood by their respective names the paradigms concern fairness, discrimination, access and legitimacy. Companies operating within the discrimination and fairness paradigm tend to focus on equal opportunities for all their employees and fairness within recruitment, as well as treatment. These factors are commonly reached by investing in career-developing programs, as well as monitoring solutions. However, there is seen to be a drawback from this solution, even though the diversity among the employees increase, namely if there is no change for the way working tasks are executed (Thomas and Ely, 2002). On the other side, the access and legitimacy paradigm focus on promoting acceptance of differences. The paradigm has fostered possibilities for everyone, including women and people of colour (Thomas and Ely, 2002). Additionally, companies operating within this paradigm tend to have diversity among their customers, making it a good match.

As a consequence of the above-mentioned factors, a learning and effectiveness paradigm has been proposed by Thomas and Ely as the new efficient paradigm focusing on diversity to further improve firm performance (Thomas and Ely, 2002). Different perspectives and approaches to working tasks are proposed to generate valuable insights and opinions. This will lead to a great understanding of that good results are achieved as a consequence of different working approaches. Continuing, excellent performance must be demanded from all employees. This contributes to higher intensity from all employees in their working tasks. Furthermore, there must be room for personal improvement, openness to all employees and leaders must make the employees feel valued. All these factors stimulate a range of knowledge and opportunities, decreased numbers of conflicts and a feeling of inclusion and belonging. Inclusion and belonging will provoke an instinct to apply skills in new areas and work for the firm's best interest (Thomas and Ely, 2002). Lastly, the firm's mission must be clear, and the structure should be non-bureaucratised. Leading to goals which are better understood, and the work is clearly centred to reach it. A flat structure invites all employees to exchange ideas and enlighten new possibilities.

Continuing with the study by Kreitz 2008, different elements are elaborated in relation to diversity. The study proposes that factors that generate differences are often divided into four categories; organizational characteristics, internal characteristics, external characteristics and personality. Organizational characteristics cover department, position, union etc., while personality characteristics are skills, abilities, knowledge etc. External and internal characteristics deal with culture, parent status, nationality, gender, ethnicity, intelligence, race etc. (Kreitz, 2008).

Companies that want to manage diversity successfully need to prioritize diversity practices (Roosevelt, 1991). Moreover, diversity is not a one-time problem that needs to be solved, but rather an ongoing process (Kreitz, 2008). This means that challenges and difficulties occur and need to be solved continuously. Among the normal problems are the employees desire to work in a homogenous group and resistance to change. Both challenges need to be managed by the human resource department but concerns the whole company. The keys are interaction and change management (Kreitz, 2008).

Successful handling of diversity includes changing thoughts and behaviour among the employees. Small steps every day done by the organization as a whole are vital for successful diversity practices. Moreover, changing regulations, norms and procedures internally in the organization affect how people respond and act. Change needs to be done in a broad and inclusive manner so that the workplace benefits all employees. Resulting, this will also contribute to beneficial outcomes of diversity (Kreitz, 2008).

Diversity among employees is a crucial resource for organizations (Tuan et al., 2019). It is a highly important resource for organizations as it, when valued and used effectively, can be seen as one of the firm's capabilities to achieve success. Human resource, (HR), practices that are diversity oriented contribute to an environment where mutual respect among and for all employees takes place, regardless of their diversity (Tuan et al., 2019).

Moreover, practices within HR that focus on diversity can decrease bias, and in best case exclude it, for decision-making processes among the employees. A diversity climate within the working environment is important to eliminate discrimination sources influencing the perceptions and experiences among the employees. Diversity climate is described by Chung et al., 2015, as shared perception among the employees, and that everybody is equally treated and included in the working environment regardless of their background. When treated equally, positive behaviour will be fostered which will affect attitudes and happiness for the employees. This will consequently lead to better satisfaction at work and higher work engagement (Tuan et al., 2019).

In this section diversity has been described in general, but the main focus in this thesis is on gender diversity.

2.9 Bias

Bias is prejudice and attitudes, often in a negative way, for or against a group or a specific person (Oxford Dictionary). As a consequence, behaviour and attitude in the working environment are highly affected by cognitive biases that contribute to the decision-making process (McKinsey, 2011). The human brain is divided into two systems, where decisions are primarily influenced by the emotional and instinctive system 1, rather than system 2 which is deliberated and rational. The outcome of the decisions does therefore tend to be biased, even though humans not intend to do it. The most common reasons for bias affecting diversity management are subconscious bias, favouritism and homogeneity bias (McKinsey, 2011).

Subconscious bias affects the human brain by having stereotypes (McKinsey, 2011). Humans have established associations of stereotypes within their minds, such as men and leading positions, and women and receptionist. Stereotypes are created without the human mind being aware of it. Consequently, the decision-making process is influenced by this underlying stereotype subconsciousness, and candidates suffer from ending up in the wrong position simply by being categorized within a stereotype group (McKinsey, 2011). Continuing, favouritism impacts the human mind to prioritize people that are similar to our self. Factors impacting who a human favour are gender, nationality, religion etc., and might negatively impact diversity. Lastly, homogeneity bias strive to believe that people that belong to a group view their group as more diverse than individuals not belonging to the group (McKinsey, 2011). Meaning that individuals get biased to think that their group is diverse enough and will not actively search for individuals with other backgrounds.

2.10 Diversity and inclusion and its' impacts on the firm performance

Diversity practices can generate valuable outcomes for organizations, including increased profitability, flexibility, creativity, dynamic capabilities and general organizational growth (Thomas and Ely, 2002). Several studies have been conducted related to firm performance as a consequence of diversity. This section will elaborate upon some of the most interesting impacts.

Thomas and Ely did in 2002 publish a study on the beneficial impacts on firm performance related to diversity (Thomas and Ely, 2002). They underline increased productivity, better access to new markets, customer segments and improved morale. They base their result on studies of firms within the banking and law sector. Here the procedures and approaches were seen very streamlined with little room for deviations. Changes in the internal processes showed differentiated results.

Companies aiming for better diversity do best to achieve it through dedicated programmes focusing on specific goals (McKinsey, 2011). Every part of the company needs to be engaged in the goals and programmes, including all department levels as well as the top management team. Diversity is not achieved overnight, the process is ongoing, incremental and slow. The slow ongoing process and lack of rapid noticeable clear results are two of the main reasons for why diversity programmes fail. The biggest contributor to failure is that management and employees do not believe in the programme and therefore fail to prioritize it. It can be solved through support from the top management team, by showing great examples of which advantages can be obtained from successful diversity programmes. Furthermore, it is necessary that all employees get sufficient training which in turn influences their motivation and willingness to prioritize the practises and in long run will pave the way to their set goals. (McKinsey, 2011).

Several studies from leading consultancy companies, including Deloitte in 2018, Pwc in 2017 and McKinsey in 2015, made studies on diversity and the beneficial impacts on the firm. The studies concluded with six key impacts; impact on the talent pool, attracting new talents, boost in reputation, improved bottom line, new economic potential and diversity is the right thing to do (Pwc, 2017; Deloitte, 2018). McKinsey additionally adds that employee satisfaction and improved customer orientation were generated from firms with diversity practices. Lastly, McKinsey found that diversity engaged leaderships are more successful as they attract the most talented employees and therefore are able to take better decisions (McKinsey, 2015).

Impact on the talent pool creates new possibilities. More than 3 out of 4 of the business leaders asked in the study answered that one of their biggest concern regarding growth possibilities is limited availability of talent and skills (Pwc, 2017). Diversity will broaden the talent pool and increase the skills and talent available in the firm. This result is backed by Deloitte, which found that companies which include diversity on their agenda had 22% higher productivity (Deloitte, 2018). A diversity-driven firm attracts new talented people, as people are more aware of diversity today. 80% of the 10.000 millennials asked in the survey answered that diversity and inclusion in the work place are important factors for them when applying to a firm (Pwc, 2017). This means that people are concerned with their own values being aligned with the firm they are working for.

A firm's reputation has become more influenced by the evolving transparency. As many as 60% in the survey answered that the diversity among the leaders within the firm played a crucial role for whether to accept the job offer or not (Pwc, 2017). Low diversity in the firm will negatively affect the reputation and put the firm in a bad spotlight. Meaning that diversity both has a large impact on the people applying for a position within the firm, but also the clients and suppliers.

Business performance is crucial for improving the bottom line. 80% of the CEOs asked in the survey answered that they believed inclusion and diversity heavily impacted the business performance, as fresh ideas and innovation were generated from a diverse work force (Pwc, 2017). Moreover, as other firms engage in the diversity movement, your firm does also need to invest in it, as it will increase the ability to answer and find solutions for other diversity driven firms in an easier and more efficient way. Additionally, Deloitte uncovered 83% better reported firm performance due to better opportunity to innovate as a cause of diversity (Deloitte, 2018).

Striving for inclusion and a diverse working force will increase the possibilities to find new economic potential. Pwc states that letting more females work and participate in the firm will boost GDP (Pwc, 2017). Moreover, letting women return to work after a career break gives increased economic potential. The study uncovered that 2 out of 3 professional women come back to lower-skilled, lower-paid working positions in addition to fewer working hours after a career break. Changing this will potentially give new beneficial opportunities for the firm (Pwc, 2017). Deloitte did also agree on this. They found in their study that diverse companies focusing on being inclusive had 27% higher profitability, and 39% higher satisfaction among their customers (Deloitte, 2018).

Continuing, diversity has become an important topic in the society in general. Fairness and inclusion have become the norms that most people are striving to achieve. Focusing on this within the firm is expected by people and society, and therefore the only right thing to do according to Pwc (Pwc, 2017).

To sum up, the well-being among the employees is affected by diversity (Downey et al., 2015). Job satisfaction and positive mental well-being have been seen to be correlated to diversity as employees feel more included and equally treated at work. This has decreased the number of employees suffering from stress and burn outs. Consequently, the engagement for meeting goals and solving tasks in the working environment have been improved and increased as a consequence of a diversity-oriented climate (Downey et al., 2015). Additionally, Cropanzano and Mitchell 2005, found that relationships emerge over time, which lead to trustful and loyal partnership (Cropanzano and Mitchell, 2005). When organizations support their employees, a relationship will evolve over time stimulating better engagement among the employees for their respective tasks leading to tasks which are executed more efficiently and improving the results.

3 LITERATURE REVIEW

This part of the thesis describes the basis and fundaments for the methodology and scope. The review focuses on several chosen existing research about text analytics using word embeddings, as well as research done within firm performance and diversity. In addition, the machine learning techniques and analysis tools available for python will be described. Lastly, the primary sources for our data are described.

3.1 Word embeddings

Literature related to text analytics is centred around word embeddings as this is the technology used to investigate gender bias in this thesis. Literature based on word embeddings was found through Libsearch and Google Scholar. We used our CBS accounts on Libsearch and selected the articles which were most relevant using the 'highest relevance' sorting function, as well as looking into newest studies. The result showed articles, books, journals and conference documents which included most of the words presented in the search. Words used in the search were, among others*, Word Embeddings, Embeddings*, *Word Representations, Neural Networks* and *Vectorizing*. Additionally, the searches on Google Scholar were chosen based on the rating system the site operates with. The same words were used in this search, as for the Libsearch. Both sorting functions were used to ensure high quality and relevance among the materials.

Bolukbasi et al. 2016, was used throughout this thesis for inspiration on similar approaches. The article has a slightly different focus by analysing News articles written by different authors, with the aim to reduce stereotypes within the embeddings by using debiased algorithms, whereas this thesis aims to find the biased job advertisements through embeddings. The focus on finding gender specific as well as gender neutral words based on the similarity of the vector space is applicable in this thesis. Their results showed that gender stereotypes were found in 19% out of the top 150 analogies. This means that bias was found among the majority of the authors.

Furthermore, Mikolov et al. 2013 provided a study on techniques for ensuring quality for word vectors based on large datasets containing billions of words. Neural network techniques were used with parallel training, which allowed for multiple tasks running at the same time. All parameters were saved at a centralized server synchronizing all parameters. They managed to reach a score of 60% accuracy on a skip-gram model that was trained on 783 million words with a dimensionality of 300. The score indicates that the model provides better results than by chance, but that there is room for improvements. A higher dimension or a larger dataset were both proposed as factors most likely improving the accuracy score.

Lastly, Grave et al. 2018, conducted a study on how fastText was trained on text data from Wikipedia and the CommonCrawl corpus, as well as three analogy datasets, to generate high quality word representation. The training corpora was collected for 157 languages. The study is highly relatable to this thesis as fastText is the model used for processing text data in both cases. Their performance was measured based on the average accuracy of the corpus as a whole, and analogy tasks were performed on the 200.000 most frequent words. The study tested the pre-trained vectors on the analogy tasks for 10 different languages and managed to get an average score of 66,7% performance, a higher score to previously comparable models. Moreover, the languages with the largest training dataset were seen to have the highest average score, whereas smaller training datasets limited the average score. The 10 different languages create the basis for the model to recognize the rest of the languages. This average score does also indicate that word embedding models perform better than chance and that less formal training datasets also contribute to improve the accuracy.

3.2 Diversity, Inclusion and Firm Performance

The literature used to discover firm performance in relation to diversity was similarly found through CBS Libsearch as well as Google Scholar. The searching process included broadly searching through journals, studies, articles, conference as well as books. The same relevance and rating function applied for word embeddings, were used in this review as well. As the newest results are of highest interest for this topic, journals, studies and articles were the most valuable sources of information. We wanted to explore the literature on gender diversity and inclusion because we are analysing job advertisements, and the advertisements are a large factor in who applies for a job and how it shapes the firm.

Results generated in a number of the materials showed that a diverse and including company stimulates and improve the talent pool (Pwc, 2017; McKinsey, 2011). A great talent pool proved to be crucial to generate new ideas and possibilities affecting the bottom line. Additionally, diversity showed to simulate inclusion and better satisfaction among the employees. What was found to be problematic was attracting both women and men candidates for a job position if the company was little diverse in the first place (Pwc, 2017). Studies revealed that candidates were more attracted to companies where it was a mix of men and women among the employees and the leaders. Lastly, the firm's reputation was seen to be positively affected by diversity, which also underlined the importance of diversity.

Studies of job advertisements and how they appeal to both men and women have remained scarce. Previously job advertisements in newspapers were often specified to a specific gender and a study done in 1973 by Bem and Bem showed that this discouraged the opposite gender to apply (Bem and Bem, 1973). Continuing, the study did also show that women found job advertisements which were intended for both genders more interesting to apply for. Male-related jobs were found most attractive to women when the advertisements were not classified for a specific gender.

As time has passed and society has evolved, this classification of job advertisements based on gender is no longer to be found. In 1964 when the U.S. civil rights legislation was in the spotlight, gender classification was seen as an unconstitutional practice (Gaucher and Friesen, 2011). The discrimination of advertisements was ended in 1973 by the Equal Employment Opportunity Commission. In 1983 was this topic in the public attention in Denmark, and the Equal Opportunity Committee in Denmark revised "Teknisk

Landsforbund" statement saying that job advertisements should be gender neutral and should provide both men and women equal rights to choose their work. The Equal Opportunity Committee in Denmark revised this statement by further elaborating about possibilities, claims and what the statement included such as salary and working time[4].

Job advertisements are still concerned to be biased; despite that they no longer are classified based on gender. Gaucher and Friesen found in a study that women preferred more social and emotional words, whereas men preferred associations to agency and leadership (Gaucher and Friesen, 2011). Moreover, advertisements within areas dominated by men contained more masculine wording than advertisements in areas dominated by women. For areas occupied by both men and women the job advertisements appealed in the same way. Participants found advertisements attractive when there were similarities between the gendered wording used and their own gender. Resulting, women were identified to not be attracted to masculine worded job advertisements due to the lack of belonginess. This affect and perpetuates gender inequality, especially within areas that are heavily dominated by men (Gaucher and Friesen, 2011).

[4] https://tl.dk/om-os/teknikeren/artikler/familie-og-ligestillingspolitik-anno-1983/

4 DATA

All the data used in this project had to be pre-processed and analysed. The literature used for this was mainly found in manuals, articles, journals and books. All the materials were found through CBS Libsearch, Google Scholar and GitHub. Moreover, secondary data has been the primary source for the data used in the thesis. In the following sections we will describe our data sources more in-depth.

4.1 Jobindex

Jobindex is the biggest job marked in Denmark, and the most complete online source to get a total overview of the available jobs in Denmark (Jobindex.dk). The webpage includes more than 20.000 job advertisements from 2019, including a large archive dating back to 2005, 130.000 CV's and more than 800.000 users each month (Jobindex.dk).

Jobindex was launched in 1996 by Kaare Danielsen when he lived in the U.S. In the U.S they operated with so called stock exchanges for jobs where advertisements were posted rapidly. Kaare understood that key to success was to provide as many job advertisements as possible in the same place. For that reason, all types of people would stick to the same webpage and that page would be the preferred one due to its many possibilities. Jobindex was created and companies were able to post their job advertisements at the page for free, in comparison to the previously used newspapers which companies had to purchase a spot (Jobindex.dk).

In Jobindex's earliest years the number of advertisements were quite scarce as many companies had not taken the advantage of the World Wide Web (WWW). The first advertisements posted at Jobindex were mainly announcements from 'Ingeniørens og Computerworlds' homepage and 10 other companies. In 2003 came the time when people really started to use Jobindex. Many factors played a role for the increased traffic to the page, but primarily the fact that more people started to use WWW (Jobindex.dk).

Today Jobindex has 271 employees and earnings of 89 million dkk before interest and taxes (Jobindex.dk). Their market share has reached 2/3 of the Danish online job market, making it to the biggest player. Jobindex.dk provides links to 20-25.000 job advertisements, which is almost 90% of all Danish job advertisements online. Their search engine is similar to Google, only within job advertisements. Their main income is from additional products that companies can buy for their job advertisements such as paid spots at the top of the webpage, CV-matching etc. (Jobindex annual report, 2018).

The combination of being a job market targeting people applying for jobs, and a recruitment partner for firms aiming to employ people, make Jobindex a valuable and unified webpage used by many people. The large amounts of users each month strengthens the reliability and trustfulness for the webpage. These factors added together is the reasons for why this company become the source for the data used in this thesis.

4.2 Statistics Denmark

Statistics Denmark was founded in 1850 and have the central authority on statistics about Denmark (dst.dk). The goal of the organization is to be a knowledge generator, focusing on improving the common understanding of Denmark, including social phenomena. Today they deliver knowledge, especially within debates, decisions and research. Their aim is to produce reliable, trusted and accurate statistics which are coherent and comparable to other statistics, which is the reason for why we chose this source for our thesis.

To be allowed the authority to produce statistics from Danish data a strict policy needs to be followed. Statistics Denmark operates with employees specified in law to solve instant statistical needs (dst.dk). Furthermore, they cooperate on an international level with European Statistical System to solve their demands and requests. Additionally, their processes are well-defined, and all outputs are quality checked. Outputs not living up to the standards will be redone to improve the results until they are satisfying. The quality is ensured by the 15 principles provided by the European Statistics Code of Practice (CoP) issued by Eurostat and include, among others, commitment to quality, coherence, relevance and accuracy and reliability (dst.dk).

Statistics Denmark operates with a strategy called ‘Strategy 2020’ where they focus on five main categories; processes, services, data sources, data security and statistical cooperation. The goal is to become a knowledge generator that improve the understanding of social context and circumstances (dst.dk). Increased globalization, growing technological usage, changes in climate and financial crises are all factors that heavily impact the current statistics.

The statistics provided from Statistics Denmark will be used as secondary data to compare the results generated in the fastText model based on the data acquired from Jobindex.dk. To get any comparisons the statistics need to be reliable. As Statistics Denmark is authorised by the Danish government and quality checked by the European Statistics Code of Practice this source of statistics is used in this thesis. fastText will be further explained in the next section.

4.3 fastText

fastText library for learning word embeddings that aims to provide users with the tools to classify and represent text through an open-source lightweight library (fastText.cc). For this thesis fastText was chosen as it can process natural language in Danish, as well as it provides models trained on two different data-sources (Nielsen, 2019). This makes us able to process our job advertisements for different pre-trained models, which is beneficial for comparing the scores.

The tools are provided in a software that is implementable on standard hardware. fastText is broadly used for natural language processing (NLP) tasks as it provides distributed word representations through word vectors. What differs fastText from other standard tools for NLP tasks, is the ability to apply it to other languages than English (Grave et al., 2018). Most NLP techniques are relying on the distributional hypothesis, meaning that the context the word appears in captures the meaning of the word. Consequently, the data the model is trained on heavily impact the quality of the vectors produced in the model. The model therefore needs to be trained on representative qualitative data in the language that the processed data is written in.

fastText is trained on Wikipedia, CommonCrawl and three analogy datasets (Grave et al., 2018). Wikipedia is chosen by fastText due to its existence in many languages and the fact that it is an online encyclopaedia, which means that it provides quality data and can be compared between the languages. Grave et al. do also underline the fact that the articles at Wikipedia are curated, makes them ideal for natural language processing as quality is ensured. The downside with this online encyclopaedia is varying size of it across the languages. Consequently, another large-scale text data provider named CommonCrawl is used in addition as a second pre-trained model. This dataset is less formal than Wikipedia and is seen as noisier but provide larger amounts of data which positively contribute with a wide coverage. The three analogy datasets are an addition to the aforementioned datasets.

Skipgram and CBOW are both extensions available for the fastText model (Grave et al., 2018). The skipgram model provide word representations using character ngrams. To each character ngram a vector representation is applied. This vector is a result of the sum of all vectors per character ngrams occurring in the word. It is important that the model learn one vector per word, which is why it is important to include the full word in the character ngrams section.

Secondly, the CBOW model is quite similar to the skipgram model, except that this model generates character ngrams based on bags of words. (Grave et al., 2018). Moreover, this model provides better information of the position of the word by generating dependent weights. The vector representation is found by taking the average of the word vectors that corresponds to each other.

In this thesis we use fastText because it performs well on the Danish language. We considered using Googles word2vec model but decided not to include it because it does not provide a pre-trained Danish model. The aim for this paper is to use natural language processing tasks to provide linguistic information about the job advertisements from Jobindex. We want to investigate how word embeddings identify gender bias within a job advertisement. For this task we use the similarity function which uses cosine to calculate the distance between word vectors.

5 METHODOLOGY

For this thesis an exploratory study has been conducted. The aim of an exploratory study is to explore new areas of applications that have been little or not discovered before (Matthews et al., 2010). For our study, text data was acquired from Jobindex.dk and the results were compared with statistics provided by Statistics Denmark. The purpose of our study is to explore job advertisements through word embeddings calculating the similarity between the Danish words ‘han’ and ‘hun’ and the advertisements.

5.1 Collecting the data

Before doing the data collection we researched which platforms provided job advertisements and if it would be possible for us to collect them. We found that Jobindex had a large set of current job advertisements in addition to an archive of job advertisements dating back to 2005. They were not able to provide us with a dataset, and we therefore developed scrapers to collect the data.

As a starting point we only collected the current data available, and thereafter we collected the data from the archive. All data was collected using the python library ‘BeautifulSoup4’, which is a library used for parsing HTML and XML documents[5]. To get access to the web we used the requests library[6]. We used the ‘Pandas’ library[7] to create the data structures, which output was a DataFrame. Moreover, we used the python time library[8] to set the scraper to ‘sleep’, meaning a 3 seconds break for each html page scraped, which is to lower the frequency of the request to the website.

[5] https://pypi.org/project/beautifulsoup4/

[6] https://pypi.org/project/requests/

[7] https://pandas.pydata.org/

[8] https://docs.python.org/3/library/time.html

5.1.1 Scraping the data from Jobindex

In order to analyse the job advertisements from Jobindex we constructed scrapers for the individual categories, which collected all of the job advertisements from each category. The categories represent the types of occupations available on Jobindex. 0describes the individual category and its' type of job. All advertisements were scraped on the 25th of February, and the historical advertisements were scraped in the following days. The table below represents the categories of Jobindex and their corresponding number of instances, i.e. the number of job advertisements that were scraped.

TABLE 1. CATEGORIES AND THE NUMBER OF INSTANCES FROM 25.02.2019

Category	No. of instances
Informationsteknologi	963
Ingenør og teknik	1215
Ledelse og personale	1571
Handel og service	686
Industri og håndværk	812
Salg og kommunikation	1084
Undervisning	391
Kontor og økonomi	1841
Social og sundhed	359
Øvrige stillinger	662
Total	9.584

We continued to scrape data from 2017, 2014 and 2008 to allow us to compare the gender bias for each year.

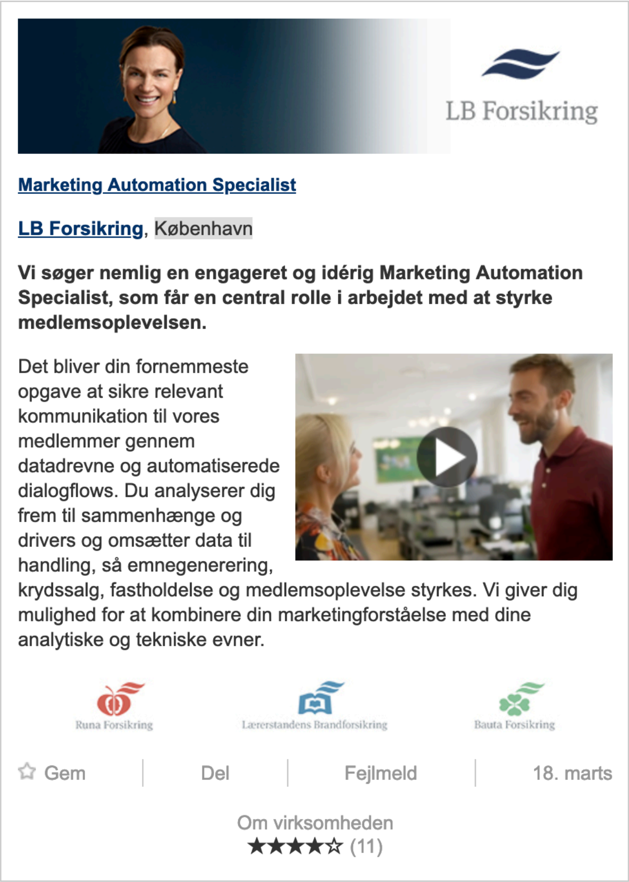
TABLE 2. TOTAL NUMBER OF INSTANCES SCRAPED FROM JOBINDEX

Year	No. of instances
2019	9.584

2017	129.344
2014	84.421
2008	76.912
Total	300.261

Before constructing the scraper, we had to analyse the html code of Jobindex to ensure that we collected the right data. We found that overall the job advertisements were very much alike making them relatively easy to scrape. The figure below is an example of an advertisement.

FIGURE 2. ADVERTISEMENT EXAMPLE



LB Forsikring

Marketing Automation Specialist

LB Forsikring, København

Vi søger nemlig en engageret og idérig Marketing Automation Specialist, som får en central rolle i arbejdet med at styrke medlemsoplevelsen.

Det bliver din fornemmeste opgave at sikre relevant kommunikation til vores medlemmer gennem datadrevne og automatiserede dialogflows. Du analyserer dig frem til sammenhænge og drivers og omsætter data til handling, så emnegenerering, krydssalg, fastholdelse og medlemsoplevelse styrkes. Vi giver dig mulighed for at kombinere din marketingforståelse med dine analytiske og tekniske evner.

Runa Forsikring   Lærerstandens Brandforsikring   Bauta Forsikring

☆ Gem | Del | Fejlmeld | 18. marts

Om virksomheden
★★★★☆ (11)

The scraper works in the way that it identifies a <div> tag which is of the class “PaidJob”, this is the entire advertisement as seen in the figure above. First it scrapes the company name which is a <b> tag within the second <a href> tag of the advertisement, then it scrapes the actual job advertisement which is either a <p> tag or a <li> tag. The job advertisement also includes a rating option on the company for the user, which we also scrape along with the date of the creation of the advertisement, and the URL to the company's job advertisement. We set two indexes to the job advertisements, the first is the overall index, and the second is the index of the current page. We set the index of the page to make it easier to see how many pages we were at in the early stages when we were exploring the data. At last we set the category of the job which corresponds to the category of jobs the scraper is collecting.

The figure below shows the code which was used to scrape the data.

FIGURE 3.   CODE FOR SCRAPING THE ADVERTISEMENTS

```
import pandas as pd
import time
import requests
from bs4 import BeautifulSoup

index = []
list_companies = []
beskrivelse_liste = []
antal_stjerner = []
month_list = []
year_list = []
day_list = []
url_liste = []
category_job = []

antal_sider = 1


for j in range(1,antal_sider + 1):
    url = "https://www.jobindex.dk/jobsoegning/oevrige&page={}".format(j)

    result = requests.get(url)
    src = result.content
    soup = BeautifulSoup(src, 'html.parser')
    divs = soup.find_all("div", {"class": 'PaidJob jix_job_archived'})

    # Code

    for i, article in enumerate(divs):
        # Setting the category name
        category_job.append('ØvrigeArchive')
        # index
        index.append(i)

        # Company name
        try:
            company = article.find_all('a')[2].find('b').text
            list_companies.append(company)
        except:
            list_companies.append('NaN')
        #print(company)

        # text from ad - body text
        text = ''
        for p in article.find_all('p')[1:]:
            text += p.text

        # text from ad - lists i.e bullets and numbered
        opgaver_liste = ''
        opgaver = article.find_all('li')

        for opgave in opgaver:
            if '\n' not in opgave.text:

                text += opgave.text

        beskrivelse_liste.append(text)

        try:
            stjerner = article.find('span', {'class': 'sr-only'}).text.split()[0]
            antal_stjerner.append(stjerner)
        except:
            antal_stjerner.append('NaN')

        # date
        dato = str(article.find('time'))
        year = dato[16:20]
        month = dato[21:23]
        day = dato[24:26]

        month_list.append(month)
        year_list.append(year)
        day_list.append(day)

        # url list
        link = article.find_all('a')[1]['href']
        url_liste.append(link)

    time.sleep(3)
```

5.1.2 Creating the DataFrames

We created individual DataFrames for each category of jobs using the ‘Pandas’ library in python, thereafter we used the ‘concat’ function to merge them together into one large DataFrame that includes the archived job advertisements. The DataFrames were created by setting each html tag from Jobindex to a column in pandas as seen in the figures below.

FIGURE 4. CREATING THE DATAFRAME

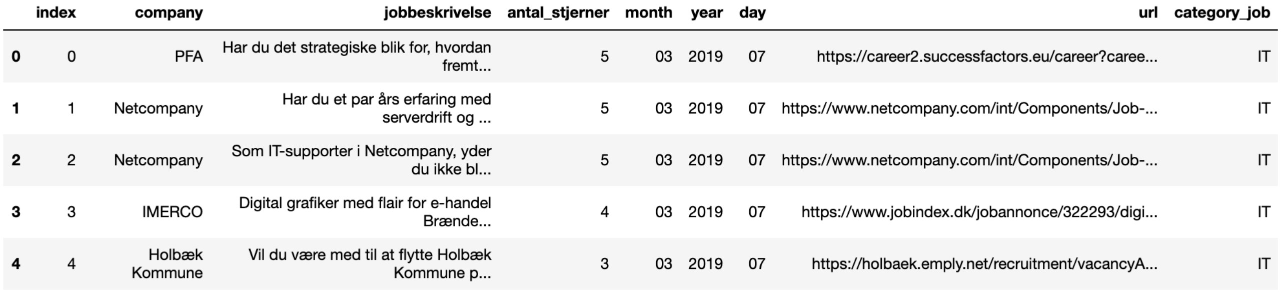
```
# Create the DataFrame
df = pd.DataFrame()
```

FIGURE 5. CREATING THE COLUMNS

```
# Create the columns
df['index'] = pd.Series(index)
df['company'] = pd.Series(list_companies)
df['jobbeskrivelse'] = pd.Series(beskrivelse_liste)
df['antal_stjerner'] = pd.Series(antal_stjerner)
df['month'] = pd.Series(month_list)
df['year'] = pd.Series(year_list)
df['day'] = pd.Series(day_list)
df['url'] = pd.Series(url_liste)
df['category_job'] = pd.Series(category_job)
```

The figure below is an example of the output we got when having created the DataFrames.

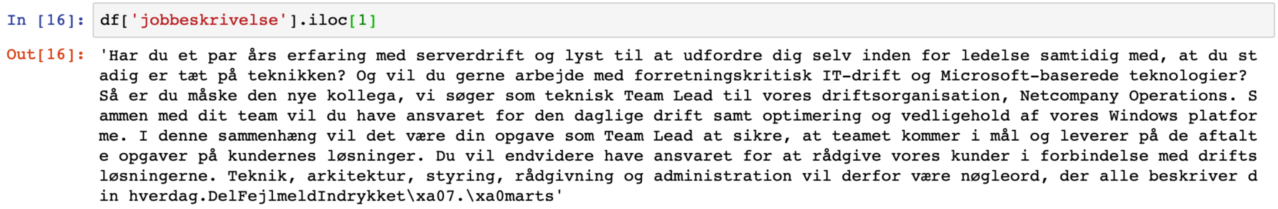
FIGURE 6. EXAMPLE OF A PANDAS DATAFRAME



	index	company	jobbeskrivelse	antal_stjerner	month	year	day	url	category_job
0	0	PFA	Har du det strategiske blik for, hvordan fremt...	5	03	2019	07	https://career2.successfactors.eu/career?caree...	IT
1	1	Netcompany	Har du et par års erfaring med serverdrift og ...	5	03	2019	07	https://www.netcompany.com/int/Components/Job-...	IT
2	2	Netcompany	Som IT-supporter i Netcompany, yder du ikke bl...	5	03	2019	07	https://www.netcompany.com/int/Components/Job-...	IT
3	3	IMERCO	Digital grafiker med flair for e-handel Brænde...	4	03	2019	07	https://www.jobindex.dk/jobannonce/322293/digi...	IT
4	4	Holbæk Kommune	Vil du være med til at flytte Holbæk Kommune p...	3	03	2019	07	https://holbaek.emply.net/recruitment/vacancyA...	IT

The column 'jobbeskrivelse' holds the job advertisement from the company, and an example of an advertisement can be seen in the figure below.

FIGURE 7. EXAMPLE OF A JOB ADVERTISEMENT



```
In [16]: df['jobbeskrivelse'].iloc[1]

Out[16]: 'Har du et par års erfaring med serverdrift og lyst til at udfordre dig selv inden for ledelse samtidig med, at du st
         adig er tæt på teknikken? Og vil du gerne arbejde med forretningskritisk IT-drift og Microsoft-baserede teknologier?
         Så er du måske den nye kollega, vi søger som teknisk Team Lead til vores driftsorganisation, Netcompany Operations. S
         ammen med dit team vil du have ansvaret for den daglige drift samt optimering og vedligehold af vores Windows platfor
         me. I denne sammenhæng vil det være din opgave som Team Lead at sikre, at teamet kommer i mål og leverer på de aftalt
         e opgaver på kundernes løsninger. Du vil endvidere have ansvaret for at rådgive vores kunder i forbindelse med drifts
         løsningerne. Teknik, arkitektur, styring, rådgivning og administration vil derfor være nøgleord, der alle beskriver d
         in hverdag.DelFejlmeldIndrykket\xa07.\xa0marts'
```

Each DataFrame was saved using the ‘Pickle’ library, making it available for pre-processing.

5.1.3 Pre-processing of the collected data

Before doing any pre-processing of the data we explored the data we had collected to familiarise ourselves with it. We found that there was much noise in the data including symbols, error messages, duplicate advertisements and English advertisements.

5.1.3.1 Cleaning the data

From our actual data we created a DataFrame to test our pre-processing methods on. We did the data cleaning in iterations, and for each time we discovered unwanted information in the job advertisements we added it to our cleaning function.

For cleaning the data, we use the library ‘Re’ which is used to match regular expressions[9]. It essentially is a very strong ‘find’ function. We first normalise the data by removing all punctuations, removing all numbers, and then we explored the job advertisements to identify other noise. We found multiple other types of symbols that we removed in addition to an error message which occurred in every job advertisement.

Furthermore, we found that there were many occasions where there was no space after a punctuation. We replace the punctuations with a space to ensure we tokenise the words correctly. We also found that our first cleaning function split words like ‘it-udvikler’ and ‘e-handel’ into two individual words which was incorrect and would have given imprecisely results. Therefore, we remove the dash symbol and insert nothing, to ensure that we have the words ‘itudvikler’ and ‘ehandel’ in our corpus. We created a new column in our

[9] https://docs.python.org/3/library/re.html

DataFrame called ‘cleaned_jobbeskrivelse’ to be able to compare the original advertisement with the cleaned advertisement. In the results section when we present our results, we always present the original advertisement to make it more readable.

Many of the job advertisements are written in English and we therefore had to identify them all and remove them from the DataFrame. We wanted to use a python library that identifies the English advertisements, but we had several issues with running it. Therefore, our initial approach was to go through all of the DataFrames and remove the English ones manually. This turned out to be quite time consuming, so we changed the approach by creating a function that identified all of the job advertisements that did not include the letters ‘å’, ‘æ’ and ‘ø’, due to the high likelihood of the letters occurring in a Danish advertisement. We created a new column called ‘english’ and assigned the number ‘2’ for every job advertisement that did not include the Danish letters. We validated our approach by checking the indexes that were assigned as English, in addition to checking many of the ones that were not assigned as English.

FIGURE 8. FINDING ENGLISH ADVERTISEMENTS

```
# Finding all english/danish
df_processed['english'] = [0 if 'æ' in x or 'ø' in x or 'å' in x else 2 for x in df_processed['cleaned_jobbeskrivelse']]
```

At last we found that as a part of the job advertisements there were many courses provided by Jobindex that we decided to remove. As they are courses, they have nothing to do with job advertisements, and is therefore seen as irrelevant for our study.

We decided to stop cleaning the data after having executed all of our cleaning functions, as well as manually going through parts of the DataFrames to especially identify the remaining English advertisements. This gave us a total number of instances of 257.427 and 28.300.501.

TABLE 3. TOTAL NUMBER OF INSTANCES

Year	**No. of instances**
2019	7.568
2017	108.534

2014	69.664
2008	70.661
Total	257.427

5.1.3.2 Bag of Words, Tokenization, StopWords and Stemming

We first create a Bag of Words which allows the machine to interpret the words as numeric values in a vector space. We tokenize all words in the cleaned_jobbeskrivelse by creating a function that takes all of the words from each instance and joins them. We then identify all unique words and assign them as a word in the corpus.

We tested removing the StopWords with the NLTK library which includes 88 StopWords from the Danish language, but we decided to keep the StopWords in the analysis of gender bias because the words might be of importance for the results. Furthermore, we are not using any Stemmers for the pre-processing because the Word Embeddings of fastText were not stemmed, which therefore would add no value.

5.2 Pretrained models

fastText provides two pretrained models for the Danish language[10]. The first model is trained on the Danish Wikipedia website, and has 300.000 tokens. The other model is trained on CommonCrawl which scrapes data from Danish websites and has 2.000.000 tokens. We searched through the documentation for fastText to gain insight about how large the training datasets were, but we were not able to find information about it.

For comparison we tested the gender bias of the job advertisements on both trained models. The large difference in the number of tokens means that the performance of the gender bias score will most likely be different when we calculate it from the two models. Furthermore, the difference in the language on each source will most likely also give us different results.

[10] https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md

5.3 Calculating the gender bias

To calculate the gender bias, we constructed two functions that we used to calculate the gender bias of an advertisement using the ‘Gensim’[11] library and the similarity function. We first created the dictionary for the unique words in the advertisements and set them up to measure the similarity between each word and the female identifier (hun) and the male identifier (han).

FIGURE 9. CREATING THE GENDER BIAS DICTIONARY

```
unique_words_dic = {}
def calculate_gender_bias_dictionary(df_kolonne, w2vmodel,  word1, word2):

    """
        df_kolonne: Column that needs to be calculated a gender bias on
        w2vmodel: Word2Vec model that is used for calculating gender bias
        Word1: Male "identifier" word
        Word2: Female "identifier" word
    """
    model = w2vmodel
    male_word = word1
    female_word = word2

    # Join all jobannoncer into one big "word"
    all_words = ' '.join(df_kolonne)

    # Finds all unique words in the "big word"
    unique_words = set(all_words.split(' '))

    # Create a dictionary with all unique words with gender bias values
    for word in unique_words:

        if word not in model.vocab.keys():
            unique_words_dic[word] = float(-1000.0)
        else:
            male_sim = float(w2vmodel.similarity(word, word1))
            female_sim = float(w2vmodel.similarity(word, word2))
            difference = male_sim - female_sim
            unique_words_dic[word] = float(difference)
    return unique_words_dic
```

[11] https://pypi.org/project/gensim/

We calculate the bias by looking at each word in the advertisement and the difference between the male identifier and the female identifier giving us an average of the gender bias score for the total document.

FIGURE 10. CALCULATING THE GENDER BIAS

```
def calculate_gender_bias(annonce, gender_bias_dict):
    gender_bias_total = 0
    avg_gender_bias = 0
    count = 0
    list_words = annonce.split()
    for word in list_words:
        bias = gender_bias_dict[word]
        if bias != -1000.0:
            gender_bias_total += bias
            count += 1
    return float((gender_bias_total / count))
```

5.4 Sources of error

In data cleaning we decided to remove the dash between words like ‘it-supporter’ which we should not have done because it is very unlikely that the pre-processing of the trained models and word embeddings had done the same. Furthermore, we found several English advertisements that our function had not removed typically because the advertisement included the location of the offered role which included one of the letters ‘æ’, ‘ø’ or ‘å’. Therefore, we had to manually identify many of the advertisements which included one of the letters and we are unsure if we managed to identify all of them. We noticed this when we applied the gender bias score to the advertisements and sorted them into the lowest scores, where we found many of the English advertisements.

5.5 Intercoder agreement

A Cohens Kappa study has been conducted in this thesis based on the study of Di Eugenio 2004. This study uses manual annotation, which was done carefully to ensure validity as well as identifying the level of agreement among the coders. Both coders in this thesis categorized 100 random samples of the job announcements from 2019. Then the results were compared and the necessary values for solving Cohen’s Kappa were found through manual work in Excel. Finally, the Kappa score was compared to the Kappa Score Table by Landis and Koch 1977 to get an overview of the coherence and validity. The full study is found in the ‘Intercoder agreement’ section.

6 INTERCODER AGREEMENT

An intercoder agreement has been conducted for this thesis to ensure compliance among the coders. This decision was taken to manually annotate advertisements to be able to compare the computed gender bias scores to the manually identified bias by the coders. We assess the computed scores by calculating the accuracy of the models, compared to the manually identified bias in the results section.

Continuing, an intercoder agreement is important for hand-coded data as items are being self-labelled in categories, to either test a computed model or to support an empirical application. Hand-coded data allows for detailed and precise analysis of large amounts of text data, which would have been challenging with automated methods (Fuoli and Hommerberg, 2015). Additionally, the study provides scores of agreement and disagreement for the advertisements being classified, which gives a good overview of how individual-driven bias classification is.

Hand-coded data has risen the question of reliability and how replicable it is as the transparency often decrease compared to automatic processes (Fuoli and Hommerberg, 2015). To be able to use and conclude any results from the hand-coded data the data needs to be reliable (Artesein and Poesio, 2008). Reliability and trust do not only impact the reputation of the coders and the usage of the coded data, it also plays a crucial role for the firm and stakeholders using the results (Fuoli and Hommerberg, 2015).

The fundamental concept behind an intercoder agreement is that coders agree on the categories being used for labelling the data, which increase the reliability and transparency (Fuoli and Hommerberg, 2015; Artesein and Poesio, 2008). When the results are consistently similar among the coders the internal understanding for the study is seen as common and the output is performed equally within the guidelines. This activity for discovering the reliability of the study is only seen as a prerequisite for how valid it is. Even though there is an agreement about what is being studied, there is no direct conclusion ensuring validity (Artesein and Poesio, 2008). Compliance among the coders can still include sharing the same prejudices for the objectives being studied making the study invalid. Many different methods for measuring intercoder agreement and reliability exist, but no specific method is categorized as the best one (Lombard et al., 2004).

In this thesis the focus has been on methods that can measure agreement and reliability manually by hand without software programs. This was decided to be the best approaches as the classification task was manually annotated by both coders. We first considered using a simple percentage method, where the percentage of agreeing answers are divided by the total percentage. However, this method is limited in that it favours tasks with few categories (Scott, 1955). Taking this into consideration, more complex methods were examined. Scott's pi (p) and Cohens Kappa (k) were considered the most valuable methods to be used, as they are two widely used methods within communication studies and computational linguistics (Neuendorf, 2002). Both methods are further elaborated upon below.

We used the intercoder agreement principles to individually categorise 100 advertisements into three categories: male, female and neutral to examine if we believed and agreed that there was gender bias in the Danish advertisements.

6.1.1 Scott's Pi

Scott's pi is measuring agreement among the coders by looking at the joint distribution between the coders (Neuendorf, 2002). This is considered as an informative method as it does better than simple agreement by providing results better than chance. The number of categories and how they are used by the coders are included in the measurement, ensuring indicators better than guessing, which is to be found in a simple agreement method. The values given in the results range from .00 which is agreement at a chance level, to 1.00 which indicate perfect agreement between the coders. If the value is less than .00, then the agreement is less than chance (Neuendorf, 2002).

6.1.2 Cohen's Kappa

Cohen's kappa is measuring differences within the coders' distribution by using a multiplicative method (Neuendorf, 2002). It provides a calculation of the overall agreement of the coders for any classification task (Kvålseth, 1989). It is seen as an improvement of

Scott's pi as the coders' distribution of evaluation is taken into consideration. Cohen's Kappa only works for two coders which can limit larger studies (Kvålseth, 1989). That said, similar to pi, it is doing better than simple agreement as it gives beyond-chance indicators which is highly usable. Moreover, Cohen's Kappa has become favourable based on especially three factors; its reasonability, methods for creating statistical visualizations based on Kappa have been developed, and agreement due to chance is taken into account (Kvålseth, 1989). Lastly, the results given in Cohen's Kappa have the same scale as Scott's pi, .00 agreement at chance, 1.00 perfect agreement and less than chance for indicators under .00.

Both methods have the same formula as seen below. The difference between the methods lies in how PA_E is found.

$$Pi\ and\ Kappa = \frac{PA_o - PA_E}{1 - PA_E}$$

PA_o is the agreement observed, and PA_E is the agreement expected by chance (Neuendorf, 2002). For Scott's pi, PA_E is found by calculating the likelihood of both coders accidentally agreeing on assigning the given item to the same category. This is done by adding the likelihood of both coders agreeing on category 1 by chance together with the likelihood of both coders agreeing on category 2 by chance (Di Eugenio and Glass, 2004). For Cohen's Kappa this calculation is slightly different. The likelihood of both coders accidentally agreeing on assigning the given item to the same category is done similar, but the likelihood of both coders agreeing on a category by chance is differently calculated. This calculation is done by multiplying the number of correctly classified items for each coder in the same category. The calculation is done for both classes. The scores are then added together to get the PA_E value (Di Eugenio and Glass, 2004). The formula for PA_E is as seen below. The calculations are showed in section 6.1.4.

$$\sum_j Pj, 1 * Pj, 2$$

Discussions regarding the conservativeness of the results have arisen with the above-mentioned methods. As both aim to give indicators better than chance, the results under chance have had little attention and credit. For extreme distributions, getting indicators above chance have been challenging and the methods have therefore been of little value (Neuendorf, 2002). In this thesis the distribution is not seen to be extreme and the chance to get indicators higher than ‘by chance’ are seen as achievable. Resulting, using one of the methods are seen as appropriate for this thesis.

As discussed above, the two methods are quite similar and valuable within computational linguistic, but only Cohen’s Kappa is applied in this thesis. The decision is taken based on the number of coders in this paper as well as the usability and possibilities of the methods. As Cohen’s Kappa is taking the distribution of the answers into consideration, this is seen as a better method that often generates more informative results. Additionally, the Kappa method is best usable for indicating agreement and reliability between two coders, which goes perfect with the number of coders in this thesis, namely two.

The Kappa method can also be used for several coders, but then the method is called Fleiss Kappa, and the formula is different (Fleiss, 1971). Scott’s pi can also be applied for multiple coders, but here there will be challenges regarding the distribution (Gwet, 2008). This will not be discussed any further here, as this thesis only has two coders. Lastly, as Feng 2012 describe in the article *“Underlying determinants driving agreement among coders”* the method for measuring reliability should be determined by the codes’ level of difficulty, as well as the scores generated should take it into account. Drawing upon this, Cohen’s Kappa makes most sense to use.

Landis and Koch 1977 determined a Kappa score table which has been used in this thesis as a benchmark. As mentioned above, 0.00 is classified as chance. Any score below this result is classified as poor in the Kappa score table, as it is considered as worse than chance. The table has values between 0.00 and 1, where 0.00-0.20 is seen as slight and 0.81-1 is classified as almost perfect. 1 is seen as perfect but is not included as an own class in the table.

For this paper the highest score as possible is desired, but anything over poor is seen as acceptable. High agreement signalizes coherence and agreement among the coders, meaning high reliability of the results (Gwet, 2008).

TABLE 4. KAPPA SCORE TABLE (LANDIS AND KOCH, 1977).

Kappa Statistic	Degree of Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

6.1.3 Intercoder Agreement Process

The study conducted in this paper consisted of firstly finding 100 random job advertisements in the dataframe of current job advertisements at Jobindex.dk. The job advertisements represented all of the 10 job categories that the job advertisements are classified within at the webpage.

The coders individually classified 100 job advertisements in a spreadsheet in 'Numbers' which can be found in the appendix under 'allsamples'. All advertisements were classified according to our assumptions of the bias in the advertisements, and categorised in 'han', 'hun' and 'neutral. The distribution of the answers of the coders can be seen below.

TABLE 5. DISTRIBUTION OF ANSWERS FOR BOTH CODERS IN THE DIFFERENT CLASSES; *HAN, NEUTRAL* AND *HUN.*

	Coder 1	Coder 2
Han	23	22
Neutral	56	67
Hun	21	11

The table above shows that the distribution is quite similar among the coders for each category. The *Hun* class is the only class which is slightly skewed, which will be an interesting factor for the total score. After this procedure was done a table showing the number of job advertisements which the coders had classified in the same classes was created. We compared the classifications and transferred it into a matrix as seen below:

TABLE 6. THE CLASSES THE CODERS AGREED ON/CLASSIFIED THE SAME

		Coder 1			**Total**
		Han	**Neutral**	**Hun**	
Coder 2	**Han**	10	11	1	22
	Neutral	11	41	15	67
	Hun	2	4	5	11
Total		23	56	21	200

The number of advertisements the coders classified the same are the diagonal dark grey cells. From this table the we move on to the calculations.

6.1.4 Intercoder agreement calculations

We use table 6 above to calculate PA$_E$. The calculations are drawing upon the calculations done by Di Eugenio 2004. The first step to the final score is to calculate Cohen's Kappa PA$_E$, seen in the formula below.

$$\sum_j Pj, 1 * Pj, 2$$

P is the proportion and j is the category, namely Han, Hun or Neutral. 1 and 2 indicate the coder. The proportion for a coder corresponds to the total number of each row and column, divided by the total of all answers seen in the bottom right corner of table 6, 200.

Step 1:

For each coder, the number of advertisements in each class is divided by the total numbers of answers. This gives a percentage value seen in columns Pj,1 and Pj,2. No decimals are removed at this point to get the final score as precise as possible.

TABLE 7. OVERALL PROPORTION OF PJ,1

Step 1.1	Calculation based on table 6 values	*Pj,1*
Han	=22/200	0,1100000
Neutral	=67/200	0,3350000
Hun	=11/200	0,0550000

TABLE 8. OVERALL PROPORTION OF PJ,2

Step 1.2	Calculation based on table 6 values	*Pj,2*
Han	=23/200	0,1150000
Neutral	=56/200	0,2800000
Hun	=21/200	0,1050000

Step 2:

In this step the percentage values for each coder in the same class need to be multiplied together. This means;

$Step\ 2.1 = Pj,1\ Han * Pj,2\ Han$

$Step\ 2.2 = Pj,1\ Neutral * Pj,2\ Neutral$

$Step\ 2.3 = Pj,1\ Hun * Pj,2\ Hun$

The calculations are shown in table 9 below.

TABLE 9. THE LIKELIHOOD OF BOTH CODERS AGREEING UPON CATEGORY J

Step 2.1	Calculations based on step 1.1 and 1.2	*Pj,1 * Pj,2*
Han*Han	=0,1100000*0,1150000	0,0126500
Step 2.2		
Neutral*Neutral	=0,3350000*0,2800000	0,0938000
Step 2.3		
Hun*hun	=0,0550000*0,1050000	0,0057750

Step 3:

The multiplied values calculated in step 3 are added together in this step.

$$Step\ 3 = Han + Neutral + Hun$$

$$0{,}1122250 = 0{,}0126500 + 0{,}0938000 + 0{,}0057750$$

Step 4:

The agreement is calculated in this step. This is done by dividing the number of the same classified advertisements by the number of advertisements.

$$0{,}56 = \frac{56}{100}$$

Step 5:

Finally, the Kappa value can be calculated. The formula is as follow:

$$Kappa = \frac{PA_o - PA_E}{1 - PA_E}$$

$$50{,}4\% = \ 0{,}504378925\ = \frac{(0{,}56 - 0{,}1122250)}{(1 - 0{,}1122250)}$$

The Kappa score is calculated to 50,4%.

6.1.5 Intercoder agreement result

The aim of this intercoder agreement study was to reach a level of agreement better than chance. Based on Landis and Koch's score table, a score better than .00 was the benchmark to exceed.

The generated results for the Kappa calculation showed a percentage of 50,4%. In the Kappa score table, this is classified as a moderate degree of agreement. Thereby the objective to exceed a result by chance is met. A percentage of 50,4% is slightly in the middle of the scale and a higher result would have been favourable. Compared to other classification studies this result is generally a little lower than other scores (Landis and Koch, 1977; Di Eugenio, 2004).

This classification task is based on the coders’ intuition, perception of the advertisement and gut feeling. The advertisements have no clear correct or wrong answer which makes the classification more challenging than right/wrong classification. Therefore, the level of agreement is not that bad compared to the level of difficulty of the task.

To conclude, the benchmark aimed to exceed was exceeded. Even though the percentage score only hit the middle of the score table, the result reached a moderate degree of agreement and the codes are concluded to be reliable in regard to the level of difficulty of the task. Based on all the above-mentioned factors, the manually annotated advertisements are seen as decisive and reliable and the analysis of the job advertisements continues in the next section.

7 RESULTS

In this part of the thesis we present and discuss the results of the gender bias found in the advertisements from Jobindex generated with the fastText word embeddings and the data from Statistics Denmark.

The results are divided into four parts to provide a tidy and meaningful presentation. Firstly, the results from the pretrained models of Wikipedia and CommonCrawl are presented and described. This presentation informs and describe the findings for male- and female biased advertisements, as well as gender neutral advertisements.

Throughout the results we compare the computed bias from the job advertisements to the gender distribution of the branches provided by Statistics Denmark[12]. This table fitted best with the parameters we wanted to explore and compare our results to and is therefore the reason why it was chosen.

Secondly, the results from our job advertisements are compared to our classification results from the intercoder agreement. These scores and comparisons give an understanding of the similarities and differences of manual and computed classification scores. Following, the scores from the fastText model is compared to statistics provided by Statistics Denmark. The first and third sections examine the values from different years to explore the progression. Additionally, a section with our results will be discussed in regard to diversity. Lastly, all results are compared and discussed to provide the clear tendency among our results.

Bias can be identified in different ways as some words are biased in their natural form, such as ‘Købmand’ and ‘Jordmord’, whereas other words are identified as bias based on the relation to ‘han’ and ‘hun’. In our thesis we want to explore subconscious bias and if word embeddings capture it in the Danish job advertisements. Our main focus is on the bias that is not immediately identified but is uncovered by the similarities found in the

[12] https://www.statistikbanken.dk/ligeai3

word vectors. This focus is mainly because there might be biased words that we not manually identify, which have large impact on the scores. This insight is difficult to obtain, and therefore an area we want to explore.

7.1 Results from the Wikipedia and CommonCrawl pretraining models

We have divided our results into the years 2019, 2017, 2014 and 2008, and then we take out the top 10 maximum and minimum gender bias scores for each year in addition to 10 neutral scores. 2019, 2014 and 2008 was the intended years to study, but as statistics provided by Statistics Denmark only were available for years from 2017 and older, 2017 was included too. The years are chosen to be able to compare the results from our job advertisements to different time periods and examine if there is any progression in the scores. A full table of the scores is available in the appendix as excel sheets.

The gender bias scores are interpreted as follows, if an advertisement has a score greater than 0, it is male bias. If an advertisement has a gender bias score lower than 0 it is female bias. The scores range from -1 being extremely female bias, and 1 being extremely male bias, scores that are close to 0 are interpreted as neutral.

7.1.1 Results from 2017

Looking at the advertisements that are most similar to ‘hun’ the results from 2019 show that only the first five advertisements have a negative score with Wikipedia, meaning that the words in the advertisements are according to our calculations female oriented in these five advertisements. The CommonCrawl model performs differently as it provides ten advertisements which are female oriented.

TABLE 10. GENDER BIAS TOP 10 MINIMUM SCORE FOR 2017

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Industri og håndværk	-0,01543696	Social og sundhed	-0,022529054
IT	-0,002432571	Social og sundhed	-0,018186143

Social og sundhed	-0,001062486	Social og sundhed	-0,017882837
Social og sundhed	-0,000281777	Social og sundhed	-0,016806622
Social og sundhed	-0,000055771	Social og sundhed	-0,015902662
Social og sundhed	0,000934367	Social og sundhed	-0,013732674
Social og sundhed	0,001050047	Social og sundhed	-0,013419929
Social og sundhed	0,001095862	Social og sundhed	-0,013113313
Handel og service	0,001508598	Social og sundhed	-0,012142734
Social og sundhed	0,001588307	Social og sundhed	-0,012111802

The scores identify some female gender bias in the advertisements, but it is not clear bias because the scores are relatively far from -1. The lowest score from the Wikipedia model was -0,01543696 and the CommonCrawl model was -0,022529054.

It is interesting to observe the difference of the Wikipedia and the CommonCrawl results. CommonCrawl only has the category 'Social og sundhed' whereas Wikipedia also includes three other categories. As seen in the figure below, in the first advertisement they are looking for an 'elektriker' and the other for a 'sygeplejerske/social- og sundhedsassistent', which are two job titles which are very different from one another.

FIGURE 11. ADVERTISEMENT FOR THE LOWEST SCORE OF 2017

Category: Industri og håndværk - Wikipedia
Dahl A/S søger elektrikere til vores serviceafdeling. Vi forventer at du er: Mødestabil Serviceminded Smilende, glad, udadvendt Vi tilbyder: Godt arbejdsmiljø Uformel tone Efteruddannelse
Category: Social og Sundhed - CommonCrawl
Røde Kors Hjemmet er en selvejende institution med i alt 61 plejeboliger, heraf 6 skærmede pladser til borgere med demens. Udover de 61 plejeboliger rummer Røde Kors Hjemmet botilbuddet Birkebakken, et dagcenter samt modtagekøkken. Vi søger: En sygeplejerske/social- og sundhedsassistent 31 timer pr. uge fortrinsvis i dagvagt, med vagt hver 2. weekend. En sygeplejerske/social- og sundhedsassistent med interesse og erfaring indenfor ældrepleje og demens.

Comparing these jobs to Statistics Denmark, we see that ’elektiker’ has the highest employment of men in 2017, whereas ’sygeplejerske/social- og sundhedsassistent’ is mostly employed by women.

When examining the advertisements which are most similar to ‘han’ for the year 2017, we first noticed that there is a difference in the scores compared to the minimum scores. The highest scores here are for Wikipedia: 0,059438802 and for CommonCrawl: 0,041759014. It identifies some male gender bias and that the male gender bias for 2017 is higher than the female gender bias for 2017.

TABLE 11. GENDER BIAS TOP 10 MAXIMUM SCORE FOR 2017

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Kontor og økonomi	0,059438802	Industri og håndværk	0,041759014
Kontor og økonomi	0,057530651	Industri og håndværk	0,040381328
Handel og service	0,055483315	Industri og håndværk	0,039025455
Ingeniør og teknik	0,055355562	Industri og håndværk	0,038867576
Ingeniør og teknik	0,055198803	Industri og håndværk	0,038678006
Industri og håndværk	0,055089233	Industri og håndværk	0,038578731
Ingeniør og teknik	0,054713316	Industri og håndværk	0,038536753
Handel og service	0,054506484	Industri og håndværk	0,038488882
Ingeniør og teknik	0,05416016	Ingeniør og teknik	0,038429484
Kontor og økonomi	0,054086185	Industri og håndværk	0,038115972

The results show that there is a difference in the gender bias score identified for the two models, as well as the categories. CommonCrawl’s top ten list only includes ‘Ingeniør og teknik’ and ‘Industri og håndværk’, whereas Wikipedia has four different categories in the top ten list.

The figure below shows the two advertisements, the first advertisement is for a ‘forskningssekretær, and the second is looking for a ‘entreprenør-, landbrugsmaskinmekaniker eller kranreparatør’

FIGURE 12. ADVERTISEMENT FOR THE HIGEST SCORE OF 2017

Category: Kontor og økonomi - Wikipedia
Stillingen som forskningssekretær i forskningsenheden for klinisk mikrobiologi er ledig til besættelse d. 1/4 2017. Konkrete arbejdsopgaver: Personaleadministration i forskningsenheden Koordinering og planlægning af undervisningsopgaver Administration af forskningsudgifter og forskningskonti for OUH og SDU
Category: Industri og håndværk - CommonCrawl
Vi har travlt og søger derfor to nye kollegaer Reparatør Er du entreprenør-, landbrugsmaskinmekaniker eller kranreparatør? Vi mangler dig til reparation af kraner, hejselad, mekanisk montage, hydraulik og pneumatik. Dine arbejdsopgaver bliver reparation og fejlsøgning af el og hydraulik på lastvognsopbygninger, herunder PLC el-styring samt fejlfinding med PC.OpbyggerEr du klejnsmed, alsidig lastvognsmekaniker, landbrugsmaskinmekaniker eller lignende? Vi søger smede til opbygning på lastvogne samt til opgaver indenfor svejsning, mekanisk montage, hydraulik og pneumatik. Gerne med erfaring indenfor lastbilbranchen.

The distribution from Statistics Denmark for 2017 shows that positions within ’forsikring’ are in general slightly more employed by women, whereas positions within ’reperatør’, ‘mekaniker’ and ‘entrepenør’ are mostly employed by men.

7.1.2 Results from 2014

We performed the same analysis on data from 2014 to examine if there were any differences in the scores and categories. We first observe that the difference between all ten gender bias scores of Wikipedia are bigger than all ten gender bias scores of CommonCrawl. Moreover, the Wikipedia scores of 2014 are closer to -1 than the scores for 2017, meaning that our models perceive the advertisements from 2014 to be more female biased than the advertisements for 2017. However, the scores from CommonCrawl are very similar to the ones of 2017.

TABLE 12. GENDER BIAS TOP 10 MINIMUM SCORE FOR 2014

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Handel og service	-0,02811	Social og sundhed	-0,01791
Handel og service	-0,0111	Social og sundhed	-0,01653
Ledelse og personale	-0,00563	Handel og service	-0,01601
Industri og håndværk	-0,00511	Social og sundhed	-0,01523
Social og sundhed	-0,00484	Social og sundhed	-0,0152
Undervisning	-0,0029	Undervisning	-0,01454
Øvrige stillinger	-0,0023	Social og sundhed	-0,01438
Social og sundhed	-0,00137	Social og sundhed	-0,01403
Social og sundhed	-0,00101	Social og sundhed	-0,01397
Kontor og økonomi	-0,00081	Social og sundhed	-0,01387

To better understand the advertisements which are most female biased, we present the two top scorers in the figure below. From the Wikipedia model we saw that there was a relative difference between the top advertisement compared to the rest, this is most likely because of how short the advertisement is.

CommonCrawl's most female bias advertisement is for a 'social- og sundhedsassistent' and similarly to 2017 the category of the job is 'Social og sundhed'. For 'social- og sundhedsassistent' there is a higher employment of women in 2014.

FIGURE 13. ADVERTISEMENT FOR THE LOWEST SCORE OF 2014

Category: Handel og service - Wikipedia
Vi tilbyder: Godt arbejdsmiljø Weekendarbejde hver 4. uge. Opgaver: Ledelsesansvar Administrative opgaver Vareopfyldning Ledelse.
Category: Social og sundhed - CommonCrawl
Røde Kors Hjemmet i Sorø søger social- og sundhedsassistent 30 timer ugentligt dag/aften. Røde Kors Hjemmet er en selvejende institution med i alt 61 plejeboliger, heraf 6 skærmede pladser til borgere med demens. Udover de 61 plejeboliger rummer Røde Kors Hjemmet botilbuddet Birkebakken, et dagcenter samt modtager køkken. Vi søger en social- og sundhedsassistent med interesse og erfaring indenfor ældrepleje, gerontopsykiatri og demens.

For the male bias advertisements of 2014, we again observe a Wikipedia model which gives advertisements a higher score than the CommonCrawl model does. There is also a greater fluctuation in the Wikipedia model. The Wikipedia model has advertisements for five different categories whereas CommonCrawl has four. Moreover, there is a larger male bias in the advertisements from the Wikipedia scores of 2014 compared to Wikipedia scores of 2017, whereas the CommonCrawl scores are very similar.

TABLE 13. GENDER BIAS TOP 10 MAXIMUM SCORE FOR 2014

category_wiki	**gender_bias_wiki**	**category_CC**	**gender_bias_CC**
Kontor og økonomi	0,07492	Ingeniør og teknik	0,041647
Ingeniør og teknik	0,067901	Industri og håndværk	0,041427
Industri og håndværk	0,067901	Ingeniør og teknik	0,040892
Undervisning	0,066398	Øvrige stillinger	0,039176
Undervisning	0,065008	Ingeniør og teknik	0,03839
Undervisning	0,062682	Industri og håndværk	0,038179
Industri og håndværk	0,061284	Industri og håndværk	0,037547
Kontor og økonomi	0,061108	Industri og håndværk	0,03734
Kontor og økonomi	0,057229	Ledelse og personale	0,03708
Ledelse og personale	0,057217	Industri og håndværk	0,03708

When observing the two advertisements we see that similarly for the 2014 female bias ad, the male bias advertisement is short. CommonCrawl also here has a longer advertisement. Branches related to ‘maskinkonstruktør’ shows a higher employment rate of men in 2014.

FIGURE 14. ADVERTISEMENT FOR THE HIGHEST SCORE OF 2014

Category: Kontor og økonomi - Wikipedia
Ansøgningsfrist 15. juli og tiltrædelse 1. september 2014

Category: Ingeniør og teknik - CommonCrawl
Technicon er en ung og dynamisk ingeniørvirksomhed, som leverer automationsløsninger, rådgivning, produktion og totalløsninger til industrien. Technicon søger en maskinkonstruktør, som kan være med til at udvikle, producere og implementere nye automationsløsninger og bidrage til videreudviklingen af svejsesoftwaren Technicon Welding. Dine primære arbejdsopgaver vil bestå i: Udvikling af specialmaskiner, herunder konstruktion i CAD programmer Projektstyring af mindre projekter, indkøb og opfølgning på komponenter Dokumentation af løsninger Deltage i montage og indkøring af automationsløsninger

7.1.3 Results from 2008

To round off the results we now present the female bias scores from 2008. Here we observe scores that are lower than 2014 and 2017 meaning that the advertisements are more female bias here than they are in the newer advertisements according to the word embeddings. Furthermore, Wikipedia here only has three different categories, which is the same for CommonCrawl.

TABLE 14. GENDER BIAS TOP 10 MINIMUM SCORE FOR 2008

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Handel og service	-0,07583	Social og sundhed	-0,05743
Handel og service	-0,05846	Social og sundhed	-0,03175
Industri og håndværk	-0,05846	Social og sundhed	-0,02202
Handel og service	-0,03788	Social og sundhed	-0,0212
Social og sundhed	-0,02274	Social og sundhed	-0,02033
Social og sundhed	-0,0205	Social og sundhed	-0,01966
Social og sundhed	-0,01793	Social og sundhed	-0,01942
Social og sundhed	-0,01645	Ingeniør og teknik	-0,01906
Handel og service	-0,01463	Social og sundhed	-0,0173
Social og sundhed	-0,01278	Handel og service	-0,01643

Both models select a short advertisement as the most female bias. The lowest scoring advertisement for Wikipedia has only four words, here we also included the second lowest in the row below, which is a bit longer.

FIGURE 15. ADVERTISEMENT FOR THE LOWEST SCORE OF 2008

Category: Handel og service - Wikipedia
Vi tilbyder: Du får:
Vi søger serviceminded butiksassistenter. JOBBET: Vareopfyldning. Kassebetjening. Kundebetjening. Kvalitetskontrol.
Category: Social og sundhed - CommonCrawl
Multicare vikarservice søger: Social- og sundhedsassistenter Social- og sundhedshjælpere Sygehjælpere Plejehjemsassistenter Sygeplejestuderende

Service-related positions, not related to 'maskin', 'teknik' etc., are mostly employed by women based on statistics from 2008 provided by Statistics Denmark. Within 'Social og sundhed' the majority of the employees are women.

The male bias scores from 2008 are similar to the scores from 2014 however they are a bit higher in 2008. This means that there is a slight increase in the male bias for both models in 2008 compared to 2014. This was the same tendency discovered for the female scores for 2008 as well. Moreover, Wikipedia and CommonCrawl have five different categories for the most male biased job advertisements for 2008.

TABLE 15. GENDER BIAS TOP 10 MAXIMUM SCORE FOR 2008

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Ledelse og personale	0,084004	Kontor og økonomi	0,043697
Kontor og økonomi	0,077078	Salg og kommunikation	0,043367
Handel og service	0,075337	Industri og håndværk	0,043004
Industri og håndværk	0,074134	Ingeniør og teknik	0,042143
Kontor og økonomi	0,073064	Salg og kommunikation	0,041972
Kontor og økonomi	0,072439	IT	0,0419

Kontor og økonomi	0,070256	Ingeniør og teknik	0,041141
Ledelse og personale	0,06986	Ingeniør og teknik	0,041018
Salg og kommunikation	0,068579	Salg og kommunikation	0,040959
Kontor og økonomi	0,065817	Industri og håndværk	0,040611

For the male bias advertisement, we again observe that the ones with the highest scores are very short, therefore we include an additional advertisement in the row below.

FIGURE 16. ADVERTISEMENT FOR THE HIGHEST SCORE OF 2008

Category: Ledelse og personale - Wikipedia
Primære opgaver og ansvar:
Engageret og ansvarsbevidst debitorbogholder / regnskabsmedarbejder søges til nyoprettet stilling i dynamisk virksomhed i Farum. Ansvarsområde: Debitorbogholderi Agent-afregning Opfølgning på ordrestatus
Category: Kontor og økonomi - CommonCrawl
Som valutarådgiver skal du:
Motiveres du af salgsmæssige udfordringer … og har du teknisk flair?

For ‘regnskabsmedarbejder’ there is a slightly higher employment rate of women in 2008, regarding the statistics from Statistics Denmark. ‘Valutarådgiver’ is not a specific branch in the tables from Statistics Denmark, but for ‘Banker, sparekasser og andelskasser’ there is slightly higher employment of women in 2008, compared to ‘centralbanker’ which had a slightly higher employment of men in 2008.

7.1.4 Results from 2019

At last we gathered data from the first two months of 2019 to give an indication of where the gender bias scores are headed according to the calculations from the word embeddings.

The first two months of 2019 are similar to the scores of 2017, but we are not able to compare these to the results from 2017 as we do not have data from the whole year. Looking at the advertisements that are most similar to 'hun' the results from 2019 identify that only the first two advertisements have a negative score with Wikipedia, meaning that the words in the advertisements are according to our calculations female oriented in these two advertisements. The CommonCrawl model performs differently as it provides ten advertisements which are female oriented.

TABLE 16. GENDER BIAS TOP 10 MINIMUM SCORE FOR 2019

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Social og sundhed	-0,00447	Social og sundhed	-0,00965
Social og sundhed	-0,00331	Social og sundhed	-0,00697
Handel og service	0,000584	Øvrige stillinger	-0,00516
Handel og service	0,003945	Undervisning	-0,00502
Øvrige stillinger	0,004715	Social og sundhed	-0,00467
Øvrige stillinger	0,007996	Social og sundhed	-0,00443
Social og sundhed	0,008207	Social og sundhed	-0,00407
Ledelse og personale	0,008844	Handel og service	-0,00327
Social og sundhed	0,008975	Social og sundhed	-0,00308
Undervisning	0,009021	Undervisning	-0,00297

What's interesting to see is that for both models in 2019 the advertisement with the lowest score is from 'Social og sundhed'. As seen in the figure below, in the first advertisement they are looking for a 'lægesekretær' and the other for a 'fysioterapeut'. Most of the positions within 'Social og sundhed' have a majority of female employees regarding statistics from 2017 provided by Statistics Denmark. This include 'lægesekretær' and 'fysioterapeut'.

FIGURE 17. ADVERTISEMENT FOR THE LOWEST SCORE OF 2019

Category: Social og sundhed - Wikipedia
Lægerne Postvænget 1, Aabybro, søger uddannet lægesekretær, gerne med praksiserfaring. Huset er en kompagniskabspraksis med 4 læger, 2 sygeplejersker, 1 bioanalytiker samt 4 sekretærer. Stillingen er på 30 t/ugentligt. Vi forventer en lægesekretær, der har: Gode samarbejdsevner, Bred erfaring med IT, Fleksibilitet og stort overblik.
Category: Social og sundhed - CommonCrawl
Ivaaraq er en landsdækkende døgninstitution for børn, unge og voksne med fysiske/psykiske handicaps. De fleste af beboerne er kørestolsbrugere. Vi søger 1 fysioterapeut med arbejdstid 40 timer ugentligt. Dit arbejdsområde: Vedligeholdelse og optræning af børn og unges fysiske tilstand. Udarbejdelse af træningsprogrammer og behandlingsplaner for beboerne. Deltagelse i tværfaglige behandlingsmøder.

When examining the advertisements which are most similar to ‘han’ for the year 2019, we first noticed that there is a difference in the scores compared to the minimum scores. The highest scores here are for Wikipedia: 0,064149 and for CommonCrawl: 0,03659. It shows that there is some male gender bias and that the male gender bias for 2019 is higher than the female gender bias for 2019.

TABLE 17. GENDER BIAS TOP 10 MAXIMUM SCORE FOR 2019

category_wiki	gender_bias_wiki	category_CC	gender_bias_CC
Salg og kommunikation	0,064149	Ingeniør og teknik	0,03659
Undervisning	0,063523	Ingeniør og teknik	0,035822
Øvrige stillinger	0,059088	Ingeniør og teknik	0,035759
Ingeniør og teknik	0,051404	Industri og håndværk	0,034927
Ledelse og personale	0,050598	Industri og håndværk	0,034908
Industri og håndværk	0,049561	Industri og håndværk	0,033735
Ingeniør og teknik	0,049148	Industri og håndværk	0,033609
Kontor og økonomi	0,049097	Ingeniør og teknik	0,033588
Ingeniør og teknik	0,04845	Ingeniør og teknik	0,033252
Salg og kommunikation	0,048061	Ingeniør og teknik	0,033097

The results show that there is a difference in the gender bias score for the two models, as well as the categories. CommonCrawl's top ten list only includes 'Ingeniør og teknik' and 'Industri og håndværk', whereas Wikipedia has six different categories in the top ten list.

The figure below shows the two advertisements, the first advertisement is for a 'præstestilling', and the second is looking for a 'CTS-tekniker'.

FIGURE 18. ADVERTISEMENT FOR THE HIGEST SCORE OF 2019

Category: Salg og kommunikation - Wikipedia
Under Kalaallit Nunaanni Ilagiit er 1 præstestilling ledig til besættelse 01. april 2019 eller efter aftale. Der kan til stillingen anvises umøbleret personalebolig efter de til enhver tid gældende regler. Husleje og depositum betales efter de til enhver tid gældende regler for den anviste bolig. Personaleboligen er knyttet op på ansættelsesforholdet, og skal fraflyttes ved ansættelsesforholdets ophør.
Category: Ingenør og teknik - CommonCrawl
Insight Building Automation er en af Danmarks førende leverandører af avancerede CTS-systemer. Vi tilbyder kvalitetsløsninger til styring og overvågning af bygningstekniske anlæg som ventilationsanlæg, varmeanlæg, brugsvandsstyringer og lysstyringer herunder systemintegration af fx adgangskontrol og brandalarmering. Vi kombinere innovative idéer med de nyeste tekniske muligheder og søger CTS-teknikere, der er specialister inden for området eller ser det som en mulighed, at udvikle sig til specialist. Arbejdsopgaverne er primært i Jylland. Om jobbet: Opstart, indregulering og servicering af CTS-anlæg. Punktafprøvning og funktionsafprøvning Programmering

'Præstestilling' is not represented in the distribution tables provided by Statistics Denmark used in this thesis. 'Religiøse institutioner og foreninger' is represented with a slightly more employment by women for 2017. For 'CTS-tekniker' the closest relatable branches are 'tekniker'-related ones, which are mostly employed by men in 2017.

7.1.5 Neutral advertisements

We see that for each year there are many advertisements which have a gender bias score that is close to 0. We have taken out 10 of the advertisements which have a score closest to zero to identify the categories that they belong to.

We see that all categories are represented. Furthermore, we see that CommonCrawl has many advertisements under the category 'Social og sundhed', whereas Wikipedia does not really favour a specific category. What is interesting to note from the scores, is that ‘Social og sundhed’ does more or less only occur from female bias scores which are closest to 0. For scores above 0 ‘Social og sundhed’ is quite absent. The ten categories are more random than what we have seen in the male and female tables.

TABLE 18. NEUTRAL ADVERTISEMENTS AND JOB CATEGORIES

2017 category comparision		**2014 category comparision**		**2008 category comparision**	
category_wiki	**category_CC**	**category_wiki**	**category_CC**	**category_wiki**	**category_CC**
Social og sundhed	Undervisning	Kontor og økonomi	Social og sundhed	Handel og service	Handel og service
Social og sundhed	Handel og service	Social og sundhed	Social og sundhed	Kontor og økonomi	Handel og service
Social og sundhed	Undervisning	Handel og service	Social og sundhed	Social og sundhed	Social og sundhed
Social og sundhed	IT	Handel og service	Social og sundhed	Handel og service	Social og sundhed
Social og sundhed	Ledelse og personale	Social og sundhed	Social og sundhed	Handel og service	Social og sundhed
Social og sundhed	Social og sundhed	Ledelse og personale	Social og sundhed	Social og sundhed	Handel og service
Handel og service	Ledelse og personale	Kontor og økonomi	Undervisning	Industri og håndværk	Ingeniør og teknik
Social og sundhed	Salg og kommunikation	Industri og håndværk	Social og sundhed	Industri og håndværk	Social og sundhed
Social og sundhed	Øvrige stillinger	Industri og håndværk	Undervisning	Industri og håndværk	Industri og håndværk
Undervisning	Handel og service	Industri og håndværk	Social og sundhed	Handel og service	Ingeniør og teknik
Social og sundhed = 8	**No highest scorer**	**Industri og håndværk = 3**	**Social og sundhed = 8**	**Handel og service = 4**	**Social og sundhed = 4**

7.1.6 The key takeaways from the results

Having results from both the fastText model trained on CommonCrawl and the model trained on Wikipedia allow us to compare the results of the gender bias. We will start by analysing the results of the female dominant advertisements.

The first interesting factor of the results is that for each year in both the model trained on CommonCrawl and Wikipedia the job category 'Social og sundhed' is the highest scoring category in the top 10 list of the female bias job advertisements. It occurs 41 times out of 60 possible. The second highest category is 'Handel og Service' which occurs 9 times.

For each year every category except for 'Salg og kommunikation' is represented at least once in the female dominant advertisements.

TABLE 19. CATEGORY COMPARISION OF THE TOP 10 MINIMUM SCORE

2017 category comparision		**2014 category comparision**		**2008 category comparision**	
category_wiki	**category_CC**	**category_wiki**	**category_CC**	**category_wiki**	**category_CC**
Industri og håndværk	Social og sundhed	Handel og service	Social og sundhed	Handel og service	Social og sundhed
IT	Social og sundhed	Handel og service	Social og sundhed	Handel og service	Social og sundhed
Social og sundhed	Social og sundhed	Ledelse og personale	Handel og service	Industri og håndværk	Social og sundhed
Social og sundhed	Social og sundhed	Industri og håndværk	Social og sundhed	Handel og service	Social og sundhed
Social og sundhed	Social og sundhed	Social og sundhed	Social og sundhed	Social og sundhed	Social og sundhed
Social og sundhed	Social og sundhed	Undervisning	Undervisning	Social og sundhed	Social og sundhed
Social og sundhed	Social og sundhed	Øvrige stillinger	Social og sundhed	Social og sundhed	Social og sundhed
Social og sundhed	Social og sundhed	Social og sundhed	Social og sundhed	Social og sundhed	Ingeniør og teknik
Handel og service	Social og sundhed	Social og sundhed	Social og sundhed	Handel og service	Social og sundhed

Social og sundhed	Social og sundhed	Kontor og økonomi	Social og sundhed	Social og sundhed	Handel og service
Social og sundhed = 7	**Social og sundhed = 10**	**Social og sundhed = 3**	**Social og sundhed = 8**	**Social og sundhed = 5**	**Social og sundhed = 8**

In the male bias advertisements, the category ‘Industri og Håndværk’ ‘occurs 19 times, followed by ‘Kontor og økonomi’ which occurs 12 times making them accountable for just over half of the advertisements. The only category that does not occur in the most male dominant advertisements is ‘Social og sundhed’.

TABLE 20. CATEGORY COMPARISION OF THE TOP 10 MAXIMUM SCORE

2017 category comparision		**2014 category comparision**		**2008 category comparision**	
category_wiki	**category_CC**	**category_wiki**	**category_CC**	**category_wiki**	**category_CC**
Salg og kommunikation	Ingeniør og teknik	Kontor og økonomi	Ingeniør og teknik	Ledelse og personale	Kontor og økonomi
Undervisning	Ingeniør og teknik	Ingeniør og teknik	Industri og håndværk	Kontor og økonomi	Salg og kommunikation
Øvrige stillinger	Ingeniør og teknik	Industri og håndværk	Ingeniør og teknik	Handel og service	Industri og håndværk
Ingeniør og teknik	Industri og håndværk	Undervisning	Øvrige stillinger	Industri og håndværk	Ingeniør og teknik
Ledelse og personale	Industri og håndværk	Undervisning	Ingeniør og teknik	Kontor og økonomi	Salg og kommunikation
Industri og håndværk	Industri og håndværk	Undervisning	Industri og håndværk	Kontor og økonomi	IT
Ingeniør og teknik	Industri og håndværk	Industri og håndværk	Industri og håndværk	Kontor og økonomi	Ingeniør og teknik
Kontor og økonomi	Ingeniør og teknik	Kontor og økonomi	Industri og håndværk	Ledelse og personale	Ingeniør og teknik
Ingeniør og teknik	Ingeniør og teknik	Kontor og økonomi	Ledelse og personale	Salg og kommunikation	Salg og kommunikation

Salg og kommunikati on	Ingeniør og teknik	Ledelse og personale	Industri og håndværk	Kontor og økonomi	Industri og håndværk
Ingeniør og teknik = 3	**Ingeniør og teknik = 6**	**Undervisnin g = 3**	**Industri og håndværk = 5**	**Kontor og økonomi = 5**	**Ingeniør og teknik = 3**

For the analysis of the results we only observed the top ten extremes of the advertisements. When we examined the full tables of the advertisements and their corresponding bias, we did also see that there were some very clear tendencies in the categories for the female list and the male list which are very similar to the top ten of the extremes. Therefore, we only selected the top ten of the results as it represents the tendencies that we observed in the full tables.

Comparing the results from the different years and the different sources, the scores are generally higher for Wikipedia than CommonCrawl. This can be due to the larger number of tokens in CommonCrawl. There is also a possibility that there is a difference in the language of the two models compared to the language of the job advertisements.

7.2 Comparing the results to the intercoder agreement

In this section we want to analyse our intercoder agreement scores in relation to the scores processed in the fastText model. This comparison is done to identify linkages and disagreement between manually classifications and computational classification. Additionally, this analysis gives informative insight in how one person can identify bias which another person disagrees on, and vice versa.

TABLE 21. MANUAL ANNOTATION LINKED TO JOB CATEGORIES

Category	**Coder 1**			**Coder 2**		
	Han	Neutral	Hun	Han	Neutral	Hun
Informationsteknologi	1	6	3	1	9	0
Ingenør og teknik	3	7	0	5	4	0

Ledelse og personale	4	5	1	2	5	3
Handel og service	3	3	4	2	8	0
Industri og håndværk	4	5	1	6	4	0
Salg og kommunikation	3	7	0	2	7	1
Undervisning	1	6	3	0	9	1
Kontor og økonomi	1	5	4	1	7	2
Social og sundhed	0	6	4	1	5	4
Øvrige stillinger	3	6	1	2	8	0

The table above shows how the two coders have classified bias among the different job categories. In total 100 randomly selected advertisements are classified, and our results show both agreement and disagreement. In the following section different advertisements that we have classified similarly and differently are presented.

FIGURE 19. AGREEMENT OF MALE BIAS

Category: Salg og kommunikation
Til en nyoprettet stilling som Product Manager for vores havemaskiner søger vi nu til vores hovedkontor i København en person, der med passion, viden og drive har lyst til at kaste sig over videreudviklingen af det nordiske marked for klassens frække dreng. I samarbejde med TTI-teamet fastlægger du strategien og sætter retningen for markedsudviklingen. Du sætter den strategiske dagsorden for havemaskiner til det nordiske marked og sikrer helheden i vores produktlanceringer. I samarbejde med vores europæiske produktspecialister og det nordiske team får du ansvaret for at drive udviklingen af vores roadmap. Du tager udgangspunkt i dine dybdegående markeds- og konkurrenceanalyser og en klar forståelse for vores kunder og deres behov.

The advertisement above was classified as male bias because of the following terms ‘havemaskiner’, ‘klassens frække dreng’, ‘strategisk dagsorden for have maskiner’.

FIGURE 20. DIAGREEMENT FOR MALE BIAS

Category: Handel og service
Brænder du for detailhandel, har du erfaring med samme, og vil du udvikle dig hos en succesrig byggemarkedskæde i solid vækst? Som butiksassistent arbejder du tæt sammen med dine kolleger, og du får ansvar for dine egne vareområder. Samtidig er du også med til at sørge for, at butikken fremstår pæn og indbydende, så kunden møder en ordentlig butik og nemt kan finde varerne. Du vil derudover skulle fylde varer op og betjene kunder ved kassen, men før vi kaster dig ud i det, får du grundig oplæring og introduktion til jobbet. Vi tilbyder dig løn efter kvalifikationer, bonusordning, pensionsordning, sundhedssikring fra dag et, rabatkort med 20 % på køb i alle kæder og butikker i Harald Nyborg-koncernen samt sociale arrangementer både lokalt og sammen med dine 1700 jem & fix-kolleger

Coder 1 classified as female because of the terms: ‘pæn og indbydende’, ‘fylder varer op’, ‘grundig oplæring’, ‘bonusordning’, ‘pensionsordning’**.** Coder 2 classified it as male because of the terms ‘succesrig byggemarkedskæde’, ‘harald nyborg-koncernen’**.**

FIGURE 21. AGREEMENT OF FEMALE BIAS

Category: Social og sundhed
Har du lyst til at være en del af teamet ved Bøgegruppen på Møllehøj, har vi en ledig stilling. Møllehøj er og skal være et godt sted at leve, bo og arbejde. Vi tror på livsglæde, hvorfor livslyst, humor og venskab er en del af livet på Møllehøj. Vi tror på værdighed og selvbestemmelse og tager altid udgangspunkt i den enkeltes værdier og behov. Vi arbejder empatisk, respektfuldt, fleksibelt, lyttende, nærværende og professionelt. På Møllehøj skaber vi mulighed for at være aktiv, og der er også plads til fred og ro. På Møllehøj ønsker vi, at borgeren har en oplevelse af at være i eget hjem, omgivet af mennesker der gerne hjælper, når man har brug for det. Vi tilbyder vore borgere professionel pleje og rehabilitering og giver omsorg og tryghed.

Both coders classified the advertisement above as female bias because of the terms ‘livsglæde’, ‘venskab’, ‘værdighed’, ‘værdier og behov’, ‘empatisk, resepktfuld, fleksibelt, lyttende, nærværende og professionelt’, ‘fred og ro’ and ‘omsorg og tryghed’.

FIGURE 22. DISAGREEMENT FOR FEMALE BIAS: (HUN)

Category: Kontor og økonomi
Serviceminded hotelreceptionist, med appetit på store oplevelser i Grønland, søges til fuldtidsstilling på det 4-stjernede Hotel Hans Egede, der også omfatter et 5-stjernet kursus- og konferencecenter, i Grønlands hovedstad Nuuk. Jobbet stiller store krav til at du kan bevare overblikket samt den indre ro i stressede situationer. Løn efter kvalifikationer og erfaring. Møbleret bolig stilles til rådighed, hvor der betales husleje efter gældende regler, ligesom der sørges for rejseomkostninger i forbindelse med tiltrædelsen.

Coder 2 classified as 'hun' because of the terms 'overblik', 'indre ro', 'møbeleret bolig', whereas coder 1 classified as 'neutral' because of 'serviceminded', 'appetit på store oplevelse', 'bevare overblik', 'løn efter kvalifikationer'. This classification clearly shows the disagreement between the coders as some terms are perceived as bias by coder 2 when coder 1 does not perceive the same term as bias.

FIGURE 23. AGREEMENT OF GENDER NEUTRALITY

Category: Salg og kommunikation
Vil du arbejde for, at vores stadig flere udenlandske forskere og studerende oplever en international arbejdsplads og et inkluderende studiemiljø? På Københavns Universitet foregår den interne kommunikation parallelt på dansk og engelsk for at sikre, at alle har lige adgang til information, viden og indflydelse. Til at hjælpe os med den opgave har vi brug for en ny kollega, der både kan fordybe sig i den sproglige detalje og gå forrest i arbejdet med at styrke engelsk som administrativt sprog. Dine opgaver bliver bl.a. at: rådgive om brug af engelsk og parallelsproglighed på tværs af universitetet være en del af et strategisk projekt om at udarbejde en sprogpolitik inkl. en plan for implementering stå for den løbende udvikling af universitetets terminologidatabase og den daglige drift af oversættelsesværktøjet.

Both coders classified as 'neutral' because of the terms 'international arbejdsplads', 'lige adgang til information, viden og indflydelse', 'rådgive', 'løbende udvikling'.

FIGURE 24. DIASAGREEMENT OF GENDER NEUTRALITY

Category: Handel og service
Vi søger en ny, frisk kollega, som kan tage ansvar. Du bliver en del af et ungt team, hvor vi alle sammen sætter en stolthed i det vi laver, og hvor kundeservice kommer i første række. Du vil få grundig oplæring i at opfylde vores høje kvalitetskrav overfor vores kunder, da det er dem vi lever af. Du vil få tildelt et ansvarsområde, det vil primært være afhentning af forvogne og sættevognstræk i området til vask. Vi forventer, at du er i fysisk god form og ikke er bange for at tage fat, når der bliver travlt. På sigt er der gode udviklingsmuligheder internt i virksomheden. Du får: Attraktiv løn plus tillæg for evt. overarbejde Kantineordning Et godt arbejdsmiljø blandt dygtige kollegaer.

Coder 2 classified as 'neutral' because of the terms 'ungt team', 'stolthed', 'kundeservice' 'ansvarsområde', whereas coder 1 classified as 'hun' because of 'frisk kollega', 'ansvar', 'ansvarsområde', 'vask', 'udviklingsmuligheder', 'kantineordning'.

Drawing from the classifications seen in TABLE 21. (Manual annotation linked to job categories) and the examples above, our Kappa score of 50,4% agreement can be identified. Both coders agree on several job advertisements, but many of them are also classified differently. The biggest difference in classification is for 'hun' biased job advertisements where coder 1 classified 21 out of 100 advertisements as 'hun', compared to coder 2 that only classified 11 out of 100 in the same category. 5 similar job advertisements were classified 'hun' by both coders.

For the 'neutral' category, the scores do not differ much, as coder 1 classified 56 advertisements out of 100 and coder 2 classified 66 advertisements out of 100 within this category. This shows that both coders agree that there are more advertisements targeting both males and females, than gender biased advertisements, among the 100 studied advertisements. Additionally, this shows that neutral advertisements are easier to identify than biased ones, which are positive as gender neutral advertisements should be the ultimate goal for companies.

The most similar classified category is 'han' biased job advertisements. Coder 1 classified 23 out of 100 advertisement within this class, compared to coder 2 that classified 22 advertisements out of 100. What is interesting to note in this class is that even though the number of advertisements classified as 'han' is quite similar, the advertisements classified are differing. Only 10 out of the respective 22 and 23 ‘han’ classified advertisements were the same advertisements. The agreed ‘han’ biased advertisements were found in ‘Øvrige stillinger’, ‘IT’, ‘Ingeniør og teknik’, ‘Ledelse og personale’, ‘Salg og kommunikation’ and ‘Industri og håndværk’. This shows that male bias advertisements occur for many categories and are not cantered around particular job branches.

For the rest of the advertisements that coder 1 classified as ‘han’, coder 2 mostly classifies them as ‘neutral’, in addition to two advertisements that were classified as ‘hun’. For the advertisements that coder 2 classified as ‘han’ which coder 1 not agreed on, where most of them also classified as ‘neutral’, only 1 advertisement was classified as ‘hun’. This clearly shows that there are divided opinions for advertisements seen as ‘han’ biased or ‘neutral’, but no doubt that the advertisements are not being ‘hun’ biased.

7.2.1 Assessing the pretrained models

In this section we assess the gender bias scores of the pretrained models in comparison our manual annotation. The gender bias scores generated from the Wikipedia model show that there are no advertisements which have a negative score except from the English and the Faroese advertisement. CommonCrawl also only has two negative scoring advertisements, the Faroese and a Danish advertisement.

We now present the advertisements where both coders agreed that there was male and female bias in the advertisement which matched with the scores generated from the Wikipedia and CommonCrawl training data.

Since we only have a single female bias advertisement that matches with automatic scores, we assess it. The advertisement in the figure below was classified as ’hun bias’ by both coders, because of the terms ‘tryghed’, ‘nærvær’ ‘sygeplejerske’ ‘sosu-assistant’ ‘pleje’ ‘omsorg’.

FIGURE 25.   HIGHEST SCORING FEMALE ADVERTISMENT

Category: Social og sundhed
Louise Mariehjemmet er et lille friplejehjem ”et hjem for ældre” i Brønshøj. Vi søger en faglig dygtig nattevagt til vores hus. På Louise Mariehjemmet bor 39 beboere i egen bolig, fordelt over 3 etager. Du arbejder sammen med en fast makker, hvor I sammen sikre tryghed og nærvær for husets beboere. Vi forventer du har følgende kvalifikationer: Uddannet sygeplejerske eller sosu-assistent. Du har lyst, og erfaring med pleje og omsorg af ældre borgere. Du har viden omkring demens.

Looking at the automatic annotation both coders classified the highest scoring male advertisement for both Wikipedia and CommonCrawl as ‘han bias’**.** The advertisement from CommonCrawl is seen in the figure below and was classified as ‘han bias’ because of the terms: ‘anlægsteknisk’, ‘renseanlægsområdet’, ‘bygningsingeniør’, ‘byggeledelse’, ‘byggemøder’

FIGURE 26.   HIGHEST SCORING MALE ADVERTISMENT FOR COMMONCRAWL

Category: Ingeniør og teknik
Er du en ingeniør med bygge- og anlægsteknisk baggrund, som brænder for at projektere inden for renseanlægsområdet? Så er du den vi søger til vores kontor i Kastrup eller Lyngby.Vi søger en bygningsingeniør med gode projektlederevner, som har fagligt fokus på projektering, tilsyn, byggeledelse etc. Du vil blandt andet få mulighed for at arbejde med: Projektering og udarbejdelse af udbud Projektledelse ved gennemførelse af projekter inden for renseanlægsområdet Lede og afholde byggemøder

The advertisement from CommonCrawl is also under the category ’Ingeniør og teknik’, and was classified as ‘han bias’ for the terms: ‘lede’, ‘udvikle’, ‘ledelsesteam’, ‘laboratoriechefen’ ‘spidsen’

FIGURE 27. HIGHEST SCORING MALE ADVERTISMENT FOR WIKIPEDIA

Category: Ingeniør og teknik
Har du lyst og evne til at lede og udvikle, så send en ansøgning til stillingen som sektionsleder ved Sektion for Plantesundhed og Fodersammensætning hos Fødevarestyrelsens Laboratorie i Ringsted.Du skal: Lede Plantesundhed og Fodersammensætning med 3 akademikere og 12 laboranter. Indgå i laboratoriets ledelsesteam sammen med de 3 øvrige sektionsledere og laboratoriechefen. Stå i spidsen for sektionens faglige prioriteringer og udvikling i overensstemmelse med laboratoriets strategiske handlingsplan og den overordnede politiske prioritering.

To further evaluate the automatically generated scores we defined thresholds for the male bias, neutral and female bias advertisements to calculate the accuracy performance. From the 100 manually annotated advertisements we sorted them from the lowest to highest scores and divided them into the three groups ‘hun bias’, ‘neutral’ and ‘han bias’. The ‘hun bias’ and ‘neutral’ each contain 33 advertisements whereas ‘han bias’ contained 34 to add up to 100.

For this exercise of calculating the accuracy we are classifying an instance as correctly classified if both coders agree on the class that the automatic tool classified. The table below shows the number of correctly classified instances and shows an accuracy of 21% for CommonCrawl and 19% for Wikipedia.

TABLE 22. COMPARING THE CLASSIFICATION OF THE CODERS TO THE TOOL

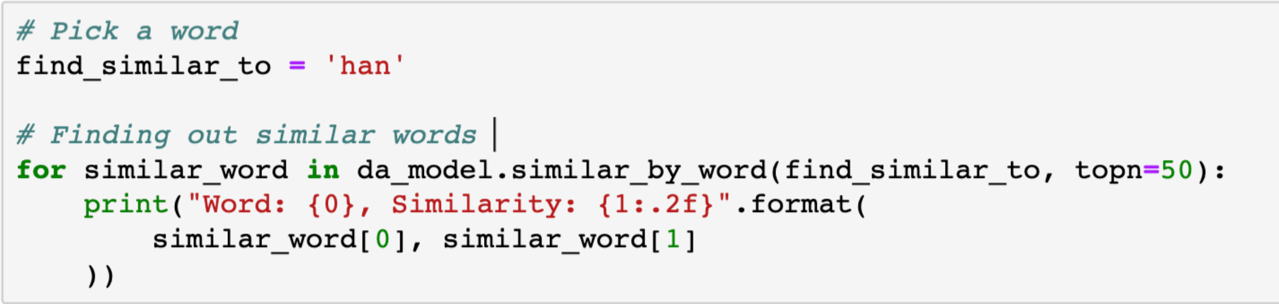
Han bias CC	Neutral CC	Hun bias CC	Han bias Wiki	Neutral Wiki	Hun bias Wiki
4	14	3	4	14	1
21			**19**		

Giving the fact that we are only looking at 100 random advertisements we are not expecting a high accuracy. This is because we do not necessarily know if the random sample includes a good representation of the three groups. For future work on this tool we would take a larger sample of the advertisements and manually annotate them to ensure that the annotation consists of a solid distribution of the three groups. This would give us a better understanding of the performance of the tool.

7.2.1.1 Similarities

We constructed a function that queries the 50 most female and male words for the CommonCrawl and Wikipedia models to get an understanding of how the model handles the semantics of ‘han’ and ‘hun’. We do this because we cannot get the list of words that determine the bias of an advertisement. The function below was used to get the outputs, and the terms are attached in the appendix.

FIGURE 28. QUERING THE 50 MOST SIMILAR WORDS



```
# Pick a word
find_similar_to = 'han'

# Finding out similar words
for similar_word in da_model.similar_by_word(find_similar_to, topn=50):
    print("Word: {0}, Similarity: {1:.2f}".format(
        similar_word[0], similar_word[1]
    ))
```

The list of words shows that for Wikipedia the words that are most similar to the female identifier ‘hun’ and the male identifier ‘han’ do make sense because many of the words are describing a male or female. However, we are seeing some very clear bias in the terms that are most similar to ‘han’ which are problematic because they are all related to occupations. The seventh term on the list is ‘embedskarriere’ and has a similarity of 0.55, which indicates a very problematic interpretation of the word and the occupation. Moreover, there are other terms in relation to teaching ‘lærergerning’ which has a similarity of 0.52, and ‘manuduktør’ and ‘døvstummelærer’ which have similarities of 0.51. We do also see that the category ‘Undervisning’ where ‘lærer’ belongs to is very male oriented when we sort our gender bias scores.

Furthermore, in the 50 words from Wikipedia there are function words like ‘og, i, hvor’ that have no relation to ‘han’. Therefore, we suspect that there are advertisements that are not necessarily male biased but will be classified as such because of the vectors for the function words.

For CommonCrawl the list of the 50 words is primarily functions words for both the male and female identifier. Therefore, we do also expect that this has an impact on the gender bias scores generated from the CommonCrawl data.

7.3 Results in comparison to Statistics Denmark

Studying the automatically identified results for the top 10 maximum and minimum gender biased advertisements for all three years, the job categories are quite similar. Among the most female biased categories are ‘Social og sundhed’, as well as ‘Undervisning’ and ‘Handel og Service’. ‘Social og sundhed’ occurs 41 times out of 60 and is the most represented category among the top female bias job advertisements. Interestingly, ‘Social og sundhed’ is not found among the most male bias categories, whereas ‘Undervisning’ occurs for top male biased categories for 2017 and 2014, and ‘Handel og Service’ occurs in the 2008 table for maximum biased categories.

Continuing, the most male biased categories include ‘Ingeniør og Teknik’, ‘Industri og Håndværk’ and ‘Kontor og Økonomi’. The most male oriented category is ‘Ingeniør og Teknik’ occurring 17 times, followed by ‘Industri og Håndværk’ occurring 15 times. ‘Ingeniør og Teknik’ occurs one time in 2008 for the top female dominated categories, as well as ’Industri og Håndværk’ occurs once in 2008 and “Kontor og Økonomi” once in 2014.

As seen above, even though some job categories occur as male or female dominated, they also occur for the opposite gender. This both-side-occurrence is a good sign as it indicates that advertisements within this category are targeting both males and females. That said, as the categories are reaching the maximum and minimum tables, it indicates bias which preferably should have been discarded. What is seen as surprising is the very high occurrence of ‘Social og sundhed’ among the most female dominated categories, and the fact that it is not occurring among the top male dominated categories. The percentage score is quite low and therefore not very problematic, but the number of occurrences is noticeable.

Drawing from the above-mentioned findings, ‘Social og sundhed’ for female dominated job advertisements, and ‘Ingeniør og teknik’ for male dominated job advertisements, will be further analysed and compared to statics from Statistics Denmark. Adjustments have been necessary to be able to compare our results to the statistics available. The adjustments performed are as following;

1. Statistics Denmark do not operate with the same categories for jobs as Jobindex.dk. Statistics Denmark provide very detailed statistics for the different branches existing in the Danish work market which has made it challenging to directly compare our scores to the statistics. We have therefore chosen to create a representative section of branches from Statistics Denmark for both ‘Social og sundhed’ og ‘Ingeniør og teknik’. The selection is done based on the sub-categories available at Jobindex.dk and in relation to the paper ‘Dansk Branchekode’[13] published by Statistics Denmark. All job categories, with subcategories, from Jobindex can be seen in appendix 1. ‘Dansk Branchekode’ describe and show the industrial classification for economic activities in Denmark and has been a good guideline to choose the right selection.
2. Statistics Denmark operates with 2017 as the newest year for their statistics within employment and gender. Consequently, our comparisons with statistics from Statistics Denmark have been as following: 2017 to 2017 scores on Jobindex, 2014 to 2014 scores on Jobindex, and 2008 to 2008 scores on Jobindex. The yearly statistics are found through Statistics Denmark’s paper ‘Dokumentation af statistikbanktabellerne vedrørende ligestilling’ (Statistics Denmark, 2018).

In the following sections we present our job advertisements in relation to the statistics from Statistics Denmark. Three tables are provided for each gender category, one table for each year. All values in the tables are in %.

7.3.1 Female bias

Some of the most interesting numbers from Statistics Denmark 2017 show that branch ‘Administration af sundhedsvæsen, undervisning, kultur og sociale forhold undtagen

[13] http://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=11119&sid=helepubl

social sikring' is 73,6% employed by women, and only 26,4% men. This is a difference of -47,2% indicating a majority of women in this branch. Moreover, the branch 'Lovpligtig socialsikring mv.' is employed by 74,4% women and only 25,5% men, also showing an imbalanced distribution of -49% towards women.

Both distributions from Statistics Denmark can be compared to the distribution of our job advertisements, which also shows an imbalanced distribution of male and female oriented advertisements. Tables for 2008, 2014 and 2017 are seen below with a selection of branches within 'Social og sundhed', including the abovementioned examples, and their respective distributions among men and women.

TABLE 23. SOURCE: STATISTICS DENMARK - 2017

		2017		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
841200	**Administration af sundhedsvæsen, undervisning, kultur og sociale forhold undtagen social sikring**	26,4	73,6	-47,2
869030	**Psykologisk rådgivning**	20,3	79,7	-59,4
889990	**Andre sociale foranstaltninger uden institutionsophold i.a.n.**	27,2	72,8	-45,6
869010	**Sunhedspleje, hjemmesygepleje og jordmødre mv.**	4,6	95,4	-90,8
869090	**Sundhedsvæsen i øvrigt i.a.n.**	22,5	77,5	-55
871010	**Plejehjem**	7,6	92,4	-84,8
889110	**Dagplejemødre**	2,4	97,6	-95,2
862100	**Alment praktiserende læger**	27,4	72,6	-45,2

Table 23 above shows that the distribution for men and women is differing a lot. Both 'Dagplejemødre' and 'Sundhedspleje, hjemmesygepleje og jordmødre mv.', have a difference of 90% and above, which indicates a large contrast of men and women within these branches. Looking at the percentage scores, only 2,4% and 4,6% of the employed people within these branches are men which is very low. This concretely underlines the differences within “Social og sundhed” and shows a clear correlation to the scores generated for our job advertisements from Jobindex, which also identified female bias within this category.

TABLE 24. SOURCE: STATISTICS DENMARK – 2014

		2014		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
841200	**Administration af sundhedsvæsen, undervisning, kultur og sociale forhold undtagen social sikring**	27,1	72,9	-45,8
869030	**Psykologisk rådgivning**	21,5	78,5	-57,0
889990	**Andre sociale foranstaltninger uden institutionsophold i.a.n.**	28,5	71,5	-43
869010	**Sunhedspleje, hjemmesygepleje og jordmødre mv.**	4,8	95,2	-90,4
869090	**Sundhedsvæsen i øvrigt i.a.n.**	20,2	79,8	-59,6
871010	**Plejehjem**	7,5	92,5	-85
889110	**Dagplejemødre**	2,2	97,8	-95,6
862100	**Alment praktiserende læger**	26,9	73,1	-46,2

As seen in table 23 for 2017, the differences for 2014, seen in table 24 above, are quite similar. Only a few percentage points are changed, mostly to the higher, indicating slightly bigger differences in distribution of men and women in 2014. This can also be compared to the scores for our job advertisements in 2014, which also showed slightly higher identification of bias.

For ‘Sunhedsvæsen i øvrigt i.a.n.’ there is seen a higher employment score of women in 2014, with a score of -59,6%, compared to 2017, which had a score of -55%. Other smaller changes are seen between 2014 and 2017, such as ‘Dagplejemødre’ with a score of -95,6% in 2014 compared to -95,2% in 2017. Furthermore, ‘Psykologisk rådgivning’ has a score of -57,0% in 2014 compared to -49% in 2017, indicating a higher imbalance in 2014. The two last mentioned scores have only minor changes but are still noticeable factors for the imbalanced distribution within 'Social og sundhed'.

TABLE 25. SOURCE: STATISTICS DENMARK 2008

		2008		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
841200	**Administration af sundhedsvæsen, undervisning, kultur og sociale forhold undtagen social sikring**	30	70	-40
869030	**Psykologisk rådgivning**	21,7	78,3	-56,6
889990	**Andre sociale foranstaltninger uden institutionsophold i.a.n.**	27,5	72,5	-45
869010	**Sunhedspleje, hjemmesygepleje og jordmødre mv.**	5	95	-90
869090	**Sundhedsvæsen i øvrigt i.a.n.**	18,9	81,1	-62,2
871010	**Plejehjem**	6,2	93,8	-87,6
889110	**Dagplejemødre**	1,4	98,6	-97,2
862100	**Alment praktiserende læger**	30,2	69,8	-39,6

Table 25 above providing the employment distribution from 2008 shows the biggest imbalance between men and women for the chosen branches within 'Social og sundhed'. For 'Dagplejemødre', only 1,4% of the employed people are men, whereas 98,6% are women. This indicates a branch more or less only employed by women. This is an additionally interesting finding to compare to the computed scores, as 'Social og sundhed' only occurred in the table for female dominated categories, not in the top male dominated table. Resulting, this indicate a good correlation.

Lastly, the percentage scores in the table of 2008 are showing the highest difference score, -97,2% for 'Dagplejemødre', compared to the other years, which is comparable to the scores for our job announcements from 2008, showing overall remarkably higher scores there too.

Comparing the percentage scores from the three years, there is a big difference within the employment rate for men and women within the presented branches of ‘Social og sundhed’. The scores are varying slightly for all years, but the highest percentage point of difference was reached in 2008, which is the same as for the score table for our job advertisements showing female oriented categories for 2008. All the tables above indicate and show a correlation to the gender bias results identified in fastTexts’ word embeddings.

7.3.2 Male bias

Similar tendencies are found for the male oriented job advertisements, as for the female oriented job advertisements. ‘Ingeniør og teknik’ occurred most often among the male biased categories identified in the word embeddings and is therefore compared to statistics for all three years in this section. ‘Ingeniør og teknik’ was also found in the computed table of female oriented job announcements for 2008, making it an interesting category to compare to Statistics Denmark.

‘Ingeniør og teknik’ is wide, spending from veterinarians to architects to chemists. A representative selection is shown in the table below, chosen from the sub-categories found on Jobindex.dk.

TABLE 26. SOURCE: STATISTICS DENMARK 2017

		2017		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
712010	**Kontrol af levnedsmidler**	38,1	61,9	-23,8
711100	**Arkitektvirksomhed**	63,1	36,9	26,2
712020	**Teknisk afprøvning og kontrol**	79,8	20,2	59,6
721100	**Forskning og eksperimentel udvikling indenfor bioteknologi**	44,0	56,0	-12,0
711220	**Rådgivende ingeniørvirksomhed inden for produktions- og maskinteknik**	72,2	27,8	44,4
712090	**Anden måling og teknisk analyse**	51,0	49,0	2,0
750000	**Dyrlæger**	21,9	78,1	-56,2

712010	**Kontrol af levnedsmidler**	38,1	61,9	-23,8

The results in the table above show that the differences in employment of men and women are mixed. This is comparable to the mixed occurrences of the categories in the tables for female- and male oriented job categories identified in the embeddings. Table 26 above shows that just over half of the branches have a dominance in employment of men. What is worth noticing is the very small dominance of men in branch ‘Anden måling og teknisk analyse’, where the distribution among men and women only differs with 2%, indicating a quite balanced distribution.

Moving on to the statistics from 2014, the table below is very similar to the table from 2017 in regard to the percentage scores.

TABLE 27. SOURCE: STATISTICS DENMARK 2014

		2014		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
712010	**Kontrol af levnedsmidler**	39,7	60,3	-20,6
711100	**Arkitektvirksomhed**	64,3	35,7	28,6
712020	**Teknisk afprøvning og kontrol**	77,4	22,6	54,8
721100	**Forskning og eksperimentel udvikling indenfor bioteknologi**	43,1	56,9	-13,8
711220	**Rådgivende ingeniørvirksomhed inden for produktions- og maskinteknik**	73,9	26,1	47,8
712090	**Anden måling og teknisk analyse**	55,9	44,1	11,8
750000	**Dyrlæger**	25,6	74,4	-48,8
712010	**Kontrol af levnedsmidler**	39,7	60,3	-20,6

Both 2017 and 2014 show a high female employment score for ‘Dyrlæger’, with scores of -56,2% for 2017 and -48,8% for 2014. Moreover, ‘Forskning og eksperimentel udvikling indenfor bioteknologi’ is also more female than male employed, with scores of -12% for 2017 and -13,8% in 2014. Lastly, ‘Anden måling og teknisk analyse’ has higher orientation of men in 2014 with a score of 11,8%, compared to 2% in 2017.

At last we present the table from 2008 which has many similarities to the two previous tables.

TABLE 28. SOURCE: STATISTICS DENMARK 2008

		2008		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
712010	**Kontrol af levnedsmidler**	42,0	58,0	-16,0
711100	**Arkitektvirksomhed**	63,5	36,5	27,0
712020	**Teknisk afprøvning og kontrol**	78,7	21,3	57,4
721100	**Forskning og eksperimentel udvikling indenfor bioteknologi**	44,0	56,0	-12,0
711220	**Rådgivende ingeniørvirksomhed inden for produktions- og maskinteknik**	74,0	26,0	48,0
712090	**Anden måling og teknisk analyse**	54,9	45,1	9,8
750000	**Dyrlæger**	27,6	72,4	-44,8
712010	**Kontrol af levnedsmidler**	42,0	58,0	-16,0

2008 shows mixed scores for the distribution of men- and women employed in the selected branches. Here, as well as for 2017 and 2014, the results show 4 out of 7 branches that are employed by more men than women. The minimal difference found in 2017 for 'Anden måling og teknisk analyse' is larger for 2008, with a score of 9,8%. However, it is still a low dominated score compared to other branches.

The most male dominated score for 2008 is branch 'Teknisk afprøving og kontrol' with a score of 57,4%, which is higher than the most female dominated score for 2008, namely 'Dyrlæger' with a score of -44,8%. This indicates that the most male dominated branches within 'Ingeniør og teknik' has higher scores than the most female dominated branches within the same category. This can be related to the scores for our job advertisements indicating more male bias advertisements than female biased within 'Ingeniør og teknik'.

Overall for the male dominated scores, the highest percentage score was reached in 2017 for 'Teknisk afprøvning og kontrol', with a score of 59,6%. Interestingly, the most female dominated branch 'Dyrlæger' has the lowest percentage score in 2008 compared to the other two years, with a score of -44,8% in 2008, -48,8% for 2014 and -56,2% in 2017.

All tables for 'Ingeniør og teknik' are showing very similar results for the three years with a distribution of 4 out of 7 branches mostly dominated by male employed workers. 'Dyrlæger' was the branch with most women employed for all the years, compared to 'Teknisk afprøvning og kontrol' that was the most male dominated branch. 'Dyrlæger' reached scores of -56,2% for 2017, -48,8% for 2014 and -44,8% for 2008. 'Teknisk afprøvning og kontrol' at the other side, reached scores of 59,6% for 2017, 54,8% for 2014 and 57,4% for 2008.

The mix of both male and female employed branches underline the results identified for our job advertisements in fastText. This can indicate that there is a coherence between gender bias in the job advertisements and the distribution of employment. However, these scores are only a selection of many branches, and the differences are not that large, potentially making the job advertisements a factor for the differences, not a concluding reason.

7.3.3 Neutral job advertisements

The scores for our job advertisements showed that some of the advertisements had scores close to zero indicating little bias and a gender-neutral approach. The top ten most neutral job categories are listed in table 18 (Neutral advertisements and job categories) and shows a quite broad spectre of job categories. The most neutral scores for 2014 are found within 'Kontor og økonomi' and 'Social og sundhed', whereas for 2008 the most neutral scores lie within 'Handel og service'.

Comparing the findings from our results to statistics provided by Statistics Denmark there are job branches that are 100% gender equal based on the distribution of men and women within that branch. A list of the equal employed branches can be seen below.

TABLE 29. GENDER EQUAL JOB BRANCHES - SOURCE: STATISTICS DENMARK

		2008		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
110300	Fremstilling af cider og anden frugtvin	50	50	0
132000	Vævning af tekstiler	50	50	0
591400	Biografer	50	50	0
662200	Forsikringsagenters og forsikringsmægleres virksomhed	50	50	0
		2014		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
110300	Fremstilling af cider og anden frugtvin	50	50	0
131000	Forbehandling og spinding af tekstilfibre	50	50	0
265200	Fremstilling af ure	50	50	0
281200	Indsamling af farligt affald	50	50	0
851000	Førskoleundervisning	50	50	0
981000	Private husholdningers produktion af varer til eget brug, i.a.n.	50	50	0
		2017		
BRANCH NR.	**Title**	**Men**	**Women**	**Difference**
649240	FVC-selskaber	50	50	0

Drawing direct linkages between our results and the gender-neutral branches from Statistics Denmark is not possible. Some similarities are to be found, such as branch "Forsikringsagenters og forsikringsmægleres virksomhed" which goes under "Kontor og økonomi", as well as "Biografer" for 2008 and "Førskoleundervisning" for 2014, which go under "Handel og service". However, it is not representative enough to conclude on. Moreover, the statistics provide percentage scores, which give us little information about how many people that are employed within that branch. Number of employees within each branch could be interesting to compare, to get a better picture of how affecting it is for the total number of people employed in the Danish working force. Additionally, many other branches than the ones presented in table 25 showed percentage scores close to 0, indicating relatively gender-neutral distribution. Those scores do also play a significant role to better understand the gender-neutral job categories and can be seen in appendix 4.

A tendency among the almost gender-equal branches, seen in appendix 4, is identified. For the previously years of differences in employment close to 0, the list is much longer for 2014, whereas 2017 and 2008 have the same number of branches. This indicates that more branches were gender neutral employed in 2014, than they are today. This is conflicting with our scores for the job advertisements, which showed that job advertisements are more gender neutral today, holding less bias than years ago. The reason behind this mismatch is difficult to state but can have a relation to the more equal employment of men and women in 2014, as seen in table 25. Moreover, no clear similarities are seen for the scores of branches with employment differences close to 0, and the list of most gender-neutral job categories from our job advertisements, seen in table 18. Again, this is most likely due to very high preciseness of branches provided by Statistics Denmark, making it difficult to directly compare it to the bigger and more general categories that Jobindex operates with. Resulting, the scores show no direct coherence with our scores, rather more similarities to the employment distribution for the respective years.

Adding the values of complete equal employment to the branches with close to 0 differences in employment, 2017 has the shortest list of gender neutrality. This is very surprisingly as 2017 occurred to be the year with lowest bias for the scores on our job advertisements. This can indicate that gender-equal job advertisement have little effect on the distribution of employment, and that men and women apply for given positions regardless of the content in the job advertisement.

7.4 Diversity and Inclusion

Drawing from the theoretical framework about diversity and inclusion we now reflect on what the word embeddings technology means for the gender diversity in the Danish work force.

The scores generated from our gender bias calculations show that there are no advertisements that have a very high, or very low score, indicating that the gender bias within the advertisements is small. However, this does not mean that the scores are insignificant.

'Social og sundhed' only occurred in the table for the most female dominated job categories, and the larger excel sheets of the advertisements, are giving us an indication that the job advertisements are more female oriented. Comparing this to Statistics Denmark, there is a clear tendency that women are employed in positions within 'Social og sundhed'. This shows a very imbalanced employment rate, which eliminates the advantages and possibilities of a gender diverse work-force.

The table below shows that the changes in the Danish work force are fluctuating.

TABLE 30. DANISH WORK FORCE DISTRIBUTION - SOURCE: STATISTICS DENMARK

Year	Men	Women	Difference	Total
2017	1.500.292	1.348.041	152.251	2.848.333
2014	1.426.324	1.293.516	132.808	2.719.840
2008	1.493.769	1.351.045	142.724	2.844.814

2014 is the year of the most balanced employment distribution among men and women, with 1.426.324 men and 1.293.516 women in work, only differing with 132.808 more men than women working. Additionally, this year is also the year with the lowest total employment rate. 2008 and 2017 have a more similar total employment rate, only differing with 3.519 employees more in 2017.

Interestingly, 2017 shows the biggest difference between employed men and women, with a gap of 152.251 more men than women working. This distributional decrease from 2014 to 2017 is inconsistent with our findings for the job advertisements, where the male and female orientation decreased with the years. A concrete reason for the difference is difficult to state. The increased difference for employed men and women is undesired as it negatively impacts the utilization of a gender diverse firm. As discussed earlier in the thesis, firm performance is positively affected by having a diverse firm and therefore what firms should aim for.

Diversity has become an important factor for being an attractive firm, not just for people considering applying for a job, but also for stakeholders and potential investors. Studies conducted by Pwc, McKinsey and Deloitte, elaborated upon earlier in the thesis, all showed that a diverse firm is the key to higher creativity, a broader knowledge base, better decision-making and resulting in higher firm performance.

In society our language plays a crucial role in the way gender discrimination and sexism are perpetrated and it highly affects the gender distribution in the work industries. In the Danish society and in most other societies in the world we observe that women are constantly being declined or neglected when it comes to the top management and the board of directors' positions. This is in part due to the stereotypes associated with our language and culture.

Our results show that the gender bias has decreased over time, which potentially can be related to the process of how the language develops with how our society develops. Over the past years there has been a significant increase in how companies perceive the diversity practices. The media has had a high focus on the importance of gender diverse companies and equality for all genders which has yielded a positive outcome. Moreover, as diversity is an ongoing process that needs to be processed and adapted continuously, it clearly gives meaning that gender neutrality approaches within firms improve with the years.

However, with deeply rooted bias and gender discriminating social constructs it is difficult to not feed these into decision-making algorithms including Natural Language Processing technologies. Therefore, we wanted to examine if we could find the gender bias in the advertisements and if there was any correlation between the gender distribution of the work force and our findings.

From the lens of diversity and inclusion job advertisements have progressed a lot, the advertisements are no longer split into 'female jobs' and 'male jobs' and the language of the advertisements is generally gender neutral today. Moreover, our results show that the bias calculated from the similarity between 'han' and 'hun' has decreased over the years, which could indicate an improvement in the language in relation to continually neutralising gendered wording.

Industries that have traditionally been dominated by a specific gender face a great task in eliminating the imbalance of the gender distribution. We saw that advertisements from ‘Social og sundhed’ were predominantly female oriented which corresponded to the gender distribution from Statistics Denmark, likewise we saw similar tendencies for ‘Ingeniør og teknik’. The underlying bias of an industry is a factor related to how people choose their careers and having a large imbalance in the gender distribution will in some cases hold people back when selecting a career path.

Job advertisements are not the only factor to consider when measuring diversity and inclusion and its’ impact on the firm. However, it is one important factor as it is the first impression that people get when considering a role, therefore, having neutral job advertisements will improve the chances of any person of any gender applying for a role.

7.5 Discussing the results

Our results showed that there is not significant bias to identify in the job advertisement collected from Jobindex. We identify a decrease in bias over the years, and this can indicate that there has been a change in the language over the years which has been inspired by the popularity of themes like gender diversity. However, we cannot say for certain that there is a correlation. It is interesting to observe how the category ‘Social og sundhed’ was the most female oriented, and the categories ‘Industri og håndværk’ and ‘Ingeniør og teknik’ were male oriented, and how this matched with the statistics from Statistics Denmark.

An important note to this study is that the gender bias scores were calculated on pretrained models build on corpuses from Wikipedia and CommonCrawl, which means that the models learn to associate words from these corpuses to ‘han’ and ‘hun’. If there is bias in these corpuses the models learn from the bias and continues to maintain this bias. We saw that the Wikipedia model associated the term ‘embedskarriere’ to be male oriented, which is problematic because it should be a neutral word. However, it is not unlikely that Wikipedia which is a platform that describes a lot of historical events primarily has articles about men that had a career within the public services given the fact that women in Denmark were only allowed to have such a career from 1921[14]. Other debatable associations are the ones of function words which are interpreted as gender bias by the word embedding tool. These should not influence the overall score of a job advertisement. Bolukbasi et al. 2016 research for de-biasing word embeddings could be a great implementation in the fastText models to avoid some of the false bias that we identified.

[14] http://denstoredanske.dk/Samfund,_jura_og_politik/Jura/Retshistorie/kvinders_retlige_stilling

8 APPLICABILITY

The study conducted in this thesis explores if word embeddings can identify gender bias in job advertisements. Our results show valuable outputs that can be used as building blocks for other studies. Moreover, our study has inspired us to apply the word embedding tool to other areas of application. Below are some of our thoughts for future work using word embeddings.

Topic modelling is an interesting part of natural language processes as it can uncover the content or topic for the text being analysed. Word embeddings can contribute to this as it finds the topic through the vector distances. This process is highly valuable as it can give insight in popular topics online, such as for social media platforms without reading it all. The insight can be beneficial to understand what people focus on and are interested in.

Moreover, topic modelling can provide useful information for companies on how their customers perceive and discuss their services and products with other people online, based on the vector spaces. This can be highly useful for, among others, online stores which want to know how their customers review their products. Additionally, topic modelling is very useful for companies wanting to draw highlights from large amount of data, such as surveys and market research as topics are found in the vectors. This can positively contribute to companies' ability to evolve and meet the demands in the market.

Analogy is also a process that can be generated from word embeddings. Analogies are comparisons of two words, which give great understanding of the relations among the words. The vectors created in word embeddings can be used to find words with the same vector distance between each other, creating comparable analogies. This can be interesting to apply to gender-related words, to see what she is to X, as he is to Y.

Lastly, as word embeddings are built on a neural network, they play a crucial role in the development of natural language processes. Neural networks are starting to outperform traditional algorithms due to its performance and ability to process large amounts of data. Therefore, word embeddings in general are seen as a brick in the big picture of the improvement of natural language processes, which aim to bridge the human understanding of language to computers. Resulting, word embeddings that for now is quite unexplored will contribute to important explorations together with other methods, which finally will be important for various areas of application.

9 DISCUSSION

The two pretrained models used in this thesis have not been researched as extensively as other models. Both the English model by Google's Word2Vec team and other models from fastText have been researched by their respective teams (Grave et al., 2018) (Mikolov et al., 2013). Grave et al., 2018 argued that they only did extensive research on the models with the ten largest vocabularies for their research. Therefore, it would be interesting to see if there are any performance differences of our models compared to the ones from the study by Grave et al., 2018 and if this would affect the gender bias scores.

It is difficult to detect gender bias in word embeddings when it is an integrated part of our language. This was also seen in the intercoder agreement scores where the biased classified advertisements differed among the coders. Furthermore, when dealing with subconscious bias it is impossible to assume that individuals have the same bias.

Our comparisons to Statistics Denmark could have been even more accurate if we had managed to go through all the different branches and select every related one to the different categories. Our section with results compared to the statistics from Statistics Denmark only provides a selection of branches, which can affect the comparison outcome. However, the comparisons obtained in this thesis have been valuable for discussing the bias scores identified in fastText.

Our results showed that many short advertisements were on the extreme ends of the bias score tables, which makes sense given these advertisements primarily had function words which we saw were among the most biased words according to the pretrained models. Some of these function words could have been removed from the DataFrames using StopWords, which could have changed the scores. However, the function words should not be biased, therefore we argue that de-biasing the algorithm is the right approach here. Moreover, it would have been interesting to only examine advertisements of a similar length to avoid the short advertisements, as they do not represent how real job advertisements look like.

Moreover, our intercoder agreement results are built upon the classification of 100 job advertisements, whereas the total number of advertisements classified with fastText is 257.427. If we had manually annotated a larger number of advertisements, we would have had a clearer assessment of the performance of the gender bias scores. However, this is a very time-consuming task and would therefore not be appropriate for a master's thesis. Therefore, for future work it would have been interesting to investigate and annotate more advertisements to build a classifier, where we use the annotated advertisement as thresholds for 'female bias', 'neutral' and 'male bias'.

At last, our approach to calculating the gender bias of an advertisement was to get the average gender bias score of the advertisement. This is not necessarily the right approach because as we observed function words that tend to be biased and therefore could determine the classification of an advertisement. Secondly, some content words that appear in an advertisement could be linked to a gender, 'jordmor, købmand, landmand', and would therefore make the advertisement bias. This is an issue related to language not being gender neutral. Third, when we take the average of the overall score and get a neutral score, there could be words which are highly biased in either direction, but even each other out by averaging the score. Therefore, it would be interesting to see which terms of an advertisement are most bias and evaluate them to investigate if the model actually captures bias or if it is more random.

10 CONCLUSION

The aim of this explorative study was to investigate if Danish job advertisements on Jobindex were gender bias using the word embedding technology provided by fastText. We conducted this study based on an intercoder agreement which aim was to manually annotate 100 job advertisement to examine if we found any subconscious bias in the advertisements. The manual annotation was done to classify the advertisements in the classes ‘male bias’, ‘neutral’ or ‘female bias’ to be able to compare the computed scores to the intercoder agreement.

Our results showed that both the intercoder agreement and the computed scores identified bias in the job advertisements. Our approach was to test our dataset on two pretrained models trained on data from Wikipedia and CommonCrawl. These models were provided by fastText and were therefore selected. It can therefore be concluded that gender bias can be identified with word embeddings.

We used a relatively simple approach to identify gender bias, we averaged the score of the advertisements and used this as the gender bias score. When assessing the pretrained models we found that the models implemented bias into function words which should not be bias. Therefore, we concluded that it is problematic to use a pretrained model trained on a dataset that is not adapted to job advertisement because of their understanding of the terms ‘han’ and ‘hun’.

For future work we would like to continue to build upon this explorative study by removing the scores from function words, to avoid their impact on the overall scores. We would like to show which terms on an advertisement are gender bias to provide an overview of what the model interprets as bias and evaluate if it is valid. At last, we would like to build a classifier that uses de-biased algorithms to classify job advertisements and their bias. It should be a tool that utilizes NLP and word embeddings for a more inclusive language and could be adopted by companies that aim to exploit diversity.

<u>**References:**</u>

Alpaydin, E., 2014. Introduction to machine learning Third., Cambridge, Massachusetts London, England: The MIT Press.

A. Newell, J. C. Shaw, and H. A. Simon. Empirical explorations with the logic theory machine: A case study in heuristics. In J. Siekmann and G. Wrightson, editors, Automation of Reasoning 1: Classical Papers on Computational Logic 1957-1966, pages 49–73. Springer, Berlin, Heidelberg, 1983. Erstpublikation: 1957. CHAPTER

Artstein, R. & Poesio, M., 2008. Intercoder Agreement for Computational Linguistics. Computational Linguistics, 34(4), pp.555–596.

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041.

Bell, Jason, 2014. Machine Learning: Hands-On for Developers and Technical Professionals, Polity Press.

Bem, S. L., & Bem, D. J. (1973). Does Sex-biased Job Advertising “Aid and Abet” Sex Discrimination? 1. Journal of Applied Social Psychology, 3(1), 6-18

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Brighton & Gigerenzer, 2015. The bias bias. Journal of Business Research, 68(8), pp.1772–1784.

Brill, E. & Mooney, R.J., 1997. Overview of empirical natural language processing. AI Magazine, 18(4), pp.13–24.

Chen, B. T., Lin, A., & Tseng, B. (2018, June). The Application of Automated, Point of Sale System Customer Service Robots to AttractT. In 8th Advances in Hospitality and Tourism marketing and management (AHTMM) Conference (p. 812).

Chipman, S.E.F., Nirenburg, S. & McShane, M.J., 2017. Natural Language Processing. In The Oxford Handbook of Cognitive Science. Oxford University Press, pp. The Oxford Handbook of Cognitive Science, Chapter 13.

Cropanzano, R., & Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. Journal of management, 31(6), 874-900.

Deng, Li, Liu, Yang & SpringerLink, 2018. Deep Learning in Natural Language Processing, Singapore: Springer Singapore Imprint: Springer.

Dean, J., 2014. Text Analytics. In Wiley & SAS Business Series. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 175–191.

Di Eugenio, B. and M. Glass. 2004. 'The kappa statistic: a second look', Computational Linguistics 30 (1), pp. 95–101.

Downey, S.N. et al., 2015. The role of diversity practices and inclusion in promoting trust and employee engagement. Journal of Applied Social Psychology, 45(1), pp.35–44.

E. Rich. Artificial Intelligence. McGraw-Hill, 1983. CHAPTER

Ely, R. J., & Thomas, D. A. (1996). Making differences matter: A new paradigm for managing diversity. Harvard Business Review, 74(5), 79-90.

Ertel, W. & SpringerLink, 2017. Introduction to Artificial Intelligence 2. ed. 2017., Cham: Springer International Publishing Imprint: Springer. Chapter 1, 9.

Feng, G., 2013. Underlying determinants driving agreement among coders. Quality & Quantity, 47(5), pp.2983–2997.

Fleiss, J.L. & Deese, James, 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), pp.378–382.

Fuoli, M. & Hommerberg, C., 2015. Optimising transparency, reliability and replicability: Annotation principles and intercoder agreement in the quantification of evaluative expressions. Corpora, 10(3), pp.315–349.

Gaucher, D. et al., 2011. Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. Journal of Personality and Social Psychology, 101(1), pp.109–128.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48.

Herbert, Claire (2013). Unconscious Bias and Higher Education. Equality Challenge Unit.

Hoy, M. B. (2018). Alexa, siri, cortana, and more: An introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88

Jurafsky, D., & Martin, J. H. (2012). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Braille Jymico. Chapter 7

Kashyap, P., 2018. Machine learning for decision makers: Cognitive computing fundamentals for better decision making, Apress Media LLC.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kubat, M. & SpringerLink, 2017. An Introduction to Machine Learning 2. ed. 2017., Cham: Springer International Publishing Imprint: Springer. Chapter 14

Landis, J., & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159-174. doi:10.2307/2529310

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 302-308).

Liu, P., Joty, S., & Meng, H. (2015). Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. doi:10.18653/v1/d15-1168

Liu, Q., Huang, H., Gao, Y., Wei, X., Tian, Y., & Liu, L. (2018, August). Task-oriented word embedding for text classification. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2023-2032)

L.M. Shore, A.E. Randel, B.G. Chung, M.A. Dean, K. Holcombe Ehrhart, G. Singh Inclusion and diversity in work groups: A review and model for future research. Journal of Management, 37 (4) (2011), pp. 1262-1289

Luu, Rowley & Vo, 2019. Addressing employee diversity to foster their work engagement. Journal of Business Research, 95, pp.303–315.

Matthews, Bob & Ross, Liz, 2010. Research Methods 1. ed., Harlow: Pearson Education UK.

Maynard, D. & Bontcheva, K., 2014. Natural language processing. In Perspectives on Ontology Learning. IOS Press, pp. 51–67.

Mikolov et al.2013 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Mitkov, R., 2005. The Oxford Handbook of Computational Linguistics 1st ed., Oxford University Press. - Part 2nr.5 semantics

Nielsen, Finn Årup, 2019. Danish Resources. DTU.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2-3), 103-134.

Nishii, L.H.H., 2013. The benefits of climate for inclusion for gender-diverse groups. Academy of Management Journal, 56(6), pp.1754–1774.

P.F. McKay, D.R. Avery, S. Tonidandel, M.A. Morris, M. Hernandez, M.R. Hebl

Racial differences in employee retention: Are diversity climate perceptions the key? Personnel Psychology, 60 (1) (2007), pp. 35-62

Provost, Foster & Fawcett, Tom, 2013. Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly.

Rajput, Dharmendra Singh et al., 2019. Sentiment analysis and knowledge discovery in contemporary business, Hershey, Pennsylvania (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA): IGI Global. (chapter 1).

Richard et al., 2013. O.C. Richard, H. Roh, J.R. Pieper "The link between diversity and equality management practice bundles and racial diversity in the managerial ranks: Does firm size matter?" Human Resource Management, 52 (2) (2013), pp. 215-242

R. Roosevelt Thomas, Jr., Beyond Race and Gender: Unleashing the Power of your Total Work Force by Managing Diversity (Saranac Lake, NY: AMACOM, 1991) xv

Richard and Johnson, 2001. O.C. Richard, N.B. Johnson "Understanding the impact of human resource diversity practices on firm performance". Journal of Managerial Issues, 13 (2) (2001), pp. 177-195

Schmidt, Macwilliams & Neal-Boylan, 2017. Becoming Inclusive: A Code of Conduct for Inclusion and Diversity. Journal of Professional Nursing, 33(2), pp.102–107.

Schubert, Lenhart, "Computational Linguistics", The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.). (Chapter 1).

Schwarz, N (2000) 'Emotion, cognition, and decision making'. *Cognition and Emotion* 14(4): 433–440.

Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding". Public Opinion Quarterly. 17 (3): 321–325

Senel, L. et al., 2018. Semantic Structure and Interpretability of Word Embeddings. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 26(10), pp.1769–1779.

Shore, L. M., Randel, A. E., Chung, B. G., Dean, M. A., Holcombe Ehrhart, K., & Singh, G. (2011). Inclusion and Diversity in Work Groups: A Review and Model for Future Research. Journal of Management, 37(4), 1262–1289.

Soares, Carlos, Ghani, Rayid & ProQuest, 2010. Data mining for business applications, Amsterdam: IOS Press. (Chapter 1)

Stadtler, H., & Kilger, C. (2002). Supply chain management and advanced planning (Vol. 4). Springer-Verlag

Sun, Maosong et al., 2017. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings, Cham: Springer International Publishing Imprint: Springer.

Tarasov, D. S. (2015). Natural language generation, paraphrasing and summarization of user reviews with recurrent neural networks. In Materials of international conference" Dialog.

Thomas, D. A., & Ely, R. J. (2002). Making differences matter: A new paradigm for managing diversity Harvard business review on managing diversity.

Zellig Harris. 1954. Distributional structure. Word, 10(23):146–162

Zhang L., Liu B. (2017) Sentiment Analysis and Opinion Mining. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA (p.1152)

Zhang, Y. et al., 2014. Ontology matching with word embeddings. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8801, pp.34–45.

Webpages:

https://en.oxforddictionaries.com/definition/bias (06.03.19, 17:53).

https://www.pwc.co.uk/human-resource-services/assets/documents/real-diversity-2017-no-holding-back.pdf (07.03.19, 09:32)

https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Tax/us-tax-inclusive-mobility-mobilize-diverse-workforce-drive-business-performance.pdf (09.04.19, 13:30)

https://www.forbes.com/sites/jenniferhicks/2017/12/30/the-role-of-artificial-intelligence-and-language/#38858c58530b (30.04.19, 15:06)

Links to documentation from Statistics Denmark:

https://ml-eu.globenewswire.com/Resource/Download/f2ea11cf-b8cf-49ab-941a-8a701a9f0fd4 (Jobindex's yearly report for 2018).

https://webcache.googleusercontent.com/search?q=cache:DB67JleFYBcJ:https://www.dst.dk/ext/5861453677/0/befolkning/Dokumentation-af-statistikbanktabellerne-vedroerende-ligestilling--pdf+&cd=1&hl=no&ct=clnk&gl=dk (documentation for the tables provided by Statistics Denmark, 2018).

APPENDIX 1 DESCRIPTION OF JOBINDEX CATEGORIES

TABLE 31. JOBINDEX CATEGORIES

Category	Description of employment
Informationsteknologi	Database, IT drift og support, IT kurser for ledige, IT ledelse, internet og WWW, systemudvikling og programmering, tele- og data kommunikation, økonomi- og virksomhedssystemer.
Ingenør og teknik	Bygge- og anlægsteknik, elektroteknik, kemi og bioteknik, ledelse indenfor ingenør og teknik, maskinteknik, medicinal og levnedsmiddel, produktions- og procesteknik
Ledelse og personale	Detailledelse, freelancekonsulent, HR- og ledelseskurser for ledige, IT ledelse, institutions- og skoleledelse, ledelse, ledelse inden for ingenør og teknik, personale og HR, projektledelse, salgsledelse, topledelse og bestyrelse, virksomhedsudvikling, økonomiledelse
Handel og service	Bud og udbringning, børnepasning, detailhandel, detailledelse, ejendomsservice, frisør og personlig pleje, hotel, restaurant og køkken, rengøring, service, sikkerhed
Industri og håndværk	Blik og rør, bygge og anlæg, elektriker, industriel produktion, jern og metal, lager, landbrug skov og fiskeri, maling og overfladebehandling, mekanik og auto, nærings- og nydelsesmiddel, tekstil og kunsthåndværk, transport, træ- og møbelindustri, tømrer og snedker.
Salg og kommunikation	Design og formgivning, ejendomsmægler, grafisk, kommunikation og journalistik, kultur og kirke, marketing, salg, salgs- og kommunikationskurser for ledige, salgsledelse, selvstændig virksomhedsdrift, telemarketing
Undervisning	Bibliotek, forskning, institutions- og skoleledelse, lærer, pædagog, voksenuddannelse
Kontor og økonomi	Akademisk og politisk arbejde, ejendomsmægler, ejendomsservice, finans og forskning, indkøb, jura, kontor, kontor- og økonomi kurser for ledige, kontorelev, logistik spedition, offentlig administration, oversættelse og sprog, sekretær og reception, økonomi og regnskab, økonomiledelse
Social og sundhed	Læge, lægesekretær, offentlig administration, pleje og omsorg, psykologi og psykiatri, socialrådgivning, sygeplejerske og jordemorder, tandlæge og klinikpersonale, teknisk sundhedsarbejde, terapi og genoptræning
Øvrige stillinger	Elevpladser, forsvar og efterretning, frivilligt arbejde, kontorelev, studiejob og fritidsjob, studiepraktik, øvrige, øvrige kurser for ledige

APPENDIX 2 LIST OF STOP WORDS

TABLE 32. LIST OF STOPWORDS FROM NLTK.CORPUS

'og',	'til',	'af',	'mig',	'min',	'fra',	'man',	'her',
'i',	'er',	'for',	'sig',	'havde',	'du',	'hans',	'alle',
'jeg',	'som',	'ikke',	'men',	'ham',	'ud',	'hvor',	'vil',
'det',	'på',	'der',	'et',	'hun',	'sin',	'eller',	'blev',
'at',	'de',	'var',	'har',	'nu',	'dem',	'hvad',	'kunne',
'en',	'med',	'mig',	'om',	'over',	'os',	'skal',	'ind',
'den',	'han',	'sig',	'vi',	'da',	'op',	'selv',	'når',
'være',	'efter',	'mit',	'hende'	'vor',	'hos',	'været',	'hendes',
'dog',	'ned',	'også',	'mine',	'mod',	'blive',	'thi',	'nogle',
'noget',	'skulle'	'under',	'alt',	'disse',	'mange',	'jer',	'sine',
'ville',	'denne',	'have',	'meget',	'hvis',	'ad',	'sådan'	'anden',
'jo',	'end',	'dig',	'sit',	'din',	'bliver',	'deres',	'dette',

APPENDIX 3 WORDS MOST SIMILAR TO HAN/HUN

MODEL WIKIPEDIA: Top 50 words most similar to ’han’	
Word: hun	Similarity: 0.71
Word: ham	Similarity: 0.62
Word: sin	Similarity: 0.61
Word: hans	Similarity: 0.60
Word: senere	Similarity: 0.55
Word: enelærer	Similarity: 0.55
Word: embedskarriere	Similarity: 0.55
Word: drengeårene	Similarity: 0.54
Word: faderen	Similarity: 0.54
Word: elleveårig	Similarity: 0.54
Word: og	Similarity: 0.53
Word: lærergerningen	Similarity: 0.52
Word: årig	Similarity: 0.52
Word: zølck	Similarity: 0.52
Word: i	Similarity: 0.52
Word: hjemkomsten	Similarity: 0.52
Word: hvor	Similarity: 0.52
Word: pröck	Similarity: 0.52
Word: sine	Similarity: 0.52
Word: faderens	Similarity: 0.52
Word: faders	Similarity: 0.52
Word: »vor	Similarity: 0.51
Word: tolvårig	Similarity: 0.51
Word: vikarierede	Similarity: 0.51
Word: manuduktør	Similarity: 0.51
Word: døvstummelærer	Similarity: 0.51
Word: stedfaderen	Similarity: 0.51
Word: studieårene	Similarity: 0.51
Word: nittenårig	Similarity: 0.51
Word: privatmand	Similarity: 0.51
Word: attenårig	Similarity: 0.51
Word: hjembys	Similarity: 0.51
Word: tyveårig	Similarity: 0.51
Word: morbroderen	Similarity: 0.51
Word: eftermand	Similarity: 0.51
Word: livsopholdet	Similarity: 0.51
Word: hvor	Similarity: 0.50
Word: lawætz	Similarity: 0.50
Word: agreèret	Similarity: 0.50
Word: hvermand	Similarity: 0.50
Word: fødebys	Similarity: 0.50
Word: professortitelen	Similarity: 0.50
Word: karrierekvinde	Similarity: 0.50
Word: udenlandsophold	Similarity: 0.50
Word: dèr	Similarity: 0.50
Word: derefter	Similarity: 0.50
Word: broderen	Similarity: 0.50

Word: han	Similarity: 0.50
Word: trettenårig	Similarity: 0.50
Word: succederede	Similarity: 0.50

MODEL WIKIPEDIA: Top 50 words most similar to ’hun’	
Word: han	Similarity: 0.71
Word: hendes	Similarity: 0.71
Word: hende	Similarity: 0.70
Word: hun	Similarity: 0.62
Word: karrierekvinde	Similarity: 0.61
Word: storesøster	Similarity: 0.58
Word: mor	Similarity: 0.57
Word: hendes	Similarity: 0.57
Word: forkvinde	Similarity: 0.57
Word: kvinde	Similarity: 0.57
Word: storesøsteren	Similarity: 0.57
Word: teenagedatter	Similarity: 0.56
Word: plejesøster	Similarity: 0.56
Word: talskvinde	Similarity: 0.56
Word: danselærerinde	Similarity: 0.56
Word: drømmepige	Similarity: 0.55
Word: storesøstre	Similarity: 0.55
Word: stedmoderen	Similarity: 0.55
Word: stedsøster	Similarity: 0.55
Word: gudmoderen	Similarity: 0.54
Word: barnepige	Similarity: 0.54
Word: foregangskvinde	Similarity: 0.54
Word: stedsøstre	Similarity: 0.54
Word: politikvinde	Similarity: 0.54
Word: stedmor	Similarity: 0.54
Word: barnepigen	Similarity: 0.54
Word: attenårig	Similarity: 0.54
Word: bedstemoderen	Similarity: 0.54
Word: næstforkvinde	Similarity: 0.53
Word: højgravid	Similarity: 0.53
Word: stedmoren	Similarity: 0.53
Word: teenagedreng	Similarity: 0.53
Word: pigesjov	Similarity: 0.53
Word: surrogatmor	Similarity: 0.53
Word: lillesøster	Similarity: 0.53
Word: teenagepige	Similarity: 0.53
Word: elleveårig	Similarity: 0.53
Word: piget	Similarity: 0.53
Word: gravid	Similarity: 0.53
Word: paige	Similarity: 0.53
Word: danselærer	Similarity: 0.52
Word: teenageidol	Similarity: 0.52
Word: skuespillerkollegaen	Similarity: 0.52
Word: piger/kvinder	Similarity: 0.52
Word: teenagerne	Similarity: 0.52
Word: veninden	Similarity: 0.52
Word: skønhedsdronning	Similarity: 0.52

Word: brudepige	Similarity: 0.52
Word: teenageårene	Similarity: 0.52
Word: kærestesorger	Similarity: 0.52

MODEL CommonCrawl: Top 50 words most similar to ‘han’	
Word: hun	Similarity: 0.79
Word: Han	Similarity: 0.75
Word: ham	Similarity: 0.71
Word: jeg	Similarity: 0.69
Word: man	Similarity: 0.67
Word: der	Similarity: 0.63
Word: selv	Similarity: 0.63
Word: havde	Similarity: 0.62
Word: faderen	Similarity: 0.61
Word: ham.Han	Similarity: 0.61
Word: også	Similarity: 0.61
Word: alligevel	Similarity: 0.60
Word: dog	Similarity: 0.60
Word: var	Similarity: 0.60
Word: imidlertid	Similarity: 0.60
Word: desuden	Similarity: 0.60
Word: jeg.Han	Similarity: 0.60
Word: fik	Similarity: 0.60
Word: derimod	Similarity: 0.59
Word: aldrig	Similarity: 0.59
Word: det.Han	Similarity: 0.59
Word: samtidig	Similarity: 0.59
Word: ellers	Similarity: 0.59
Word: manden	Similarity: 0.59
Word: men	Similarity: 0.59
Word: ikke	Similarity: 0.59
Word: ham.Hun	Similarity: 0.58
Word: ikke.Han	Similarity: 0.58
Word: Hun	Similarity: 0.58
Word: hans	Similarity: 0.58
Word: Fikumdikmanden	Similarity: 0.58
Word: engang	Similarity: 0.58
Word: blev	Similarity: 0.58
Word: tillige	Similarity: 0.58
Word: godt.Han	Similarity: 0.57
Word: har	Similarity: 0.57
Word: kom	Similarity: 0.57
Word: derfor	Similarity: 0.57
Word: som	Similarity: 0.57
Word: hvor	Similarity: 0.57
Word: snart	Similarity: 0.56
Word: DFeren	Similarity: 0.56
Word: allerede	Similarity: 0.56
Word: ligesom	Similarity: 0.56
Word: ham.Nu	Similarity: 0.56

Word: nok	Similarity: 0.56
Word: selv.Han	Similarity: 0.56
Word: ligeledes	Similarity: 0.56
Word: retten.Han	Similarity: 0.56
Word: gjorde	Similarity: 0.56

MODEL CommonCrawl: Top 50 words most similar to 'hun'	
Word: Hun	Similarity: 0.81
Word: han	Similarity: 0.79
Word: hende	Similarity: 0.74
Word: jeg	Similarity: 0.68
Word: hendes	Similarity: 0.68
Word: hende.Hun	Similarity: 0.67
Word: ham.Hun	Similarity: 0.66
Word: af.Hun	Similarity: 0.65
Word: det.Hun	Similarity: 0.65
Word: ikke.Hun	Similarity: 0.63
Word: ud.Hun	Similarity: 0.63
Word: mig.Hun	Similarity: 0.62
Word: pigen	Similarity: 0.62
Word: den.Hun	Similarity: 0.62
Word: man	Similarity: 0.61
Word: selv.Hun	Similarity: 0.61
Word: Han	Similarity: 0.60
Word: alligevel	Similarity: 0.59
Word: kvinden	Similarity: 0.59
Word: dem.Hun	Similarity: 0.59
Word: selv	Similarity: 0.59
Word: hende.Da	Similarity: 0.59
Word: mormoren	Similarity: 0.59
Word: der	Similarity: 0.58
Word: farmoren	Similarity: 0.58
Word: bedstemoren	Similarity: 0.58
Word: hende.Han	Similarity: 0.57
Word: år.Hun	Similarity: 0.57
Word: henden	Similarity: 0.57
Word: om.Hun	Similarity: 0.57
Word: moderen	Similarity: 0.57
Word: igen.Hun	Similarity: 0.57
Word: også	Similarity: 0.57
Word: datteren	Similarity: 0.56
Word: for.Hun	Similarity: 0.56
Word: bare	Similarity: 0.56
Word: eksen	Similarity: 0.56
Word: gammel.Hun	Similarity: 0.56
Word: jordemoren	Similarity: 0.56
Word: mor	Similarity: 0.56
Word: samtidig	Similarity: 0.56
Word: til.Hun	Similarity: 0.56
Word: kæmpekvinden	Similarity: 0.56
Word: kvinde	Similarity: 0.56

Word: frøkenen	Similarity: 0.56
Word: hende.Jeg	Similarity: 0.56
Word: sig.Hun	Similarity: 0.56
Word: mandet	Similarity: 0.56
Word: med.Hun	Similarity: 0.55
Word: hende.I	Similarity: 0.55

APPENDIX 4 BRANCHES WITH CLOSE TO 0 DIFFERENCES

		2017		
BRANCH NR.	Title	Men	Women	Difference
102020	Forarbejdning og konservering af fisk, krebsedyr og bløddyr, undtagen fiskemel	49,5	50,5	-1
105200	Fremstilling af konsumis	50,8	49,2	1,6
120000	Fremstilling af tobaksprodukter	50,5	49,5	1
139600	Fremstilling af andre tekniske og industrielle tekstiler	50,9	49,1	1,8
205300	Fremstilling af æteriske olier	50,5	49,5	1
329100	Fremstilling af koste og børster	50,2	49,8	0,4
479119	Detailhandel med andre varer i.a.n. via internet	49,7	50,3	-0,6
561010	Restauranter	50,6	49,4	1,2
649230	Andre kreditselskaber	50,4	49,6	0,8
853200	Tekniske skoler og fagskoler	49,2	50,8	-1,6
854200	Videregående uddannelser på universitetsniveau	49,4	50,6	-1,2
932100	Forlystelsesparker og lignende	50,7	49,3	1,4
949200	Politiske partier	49,5	50,5	-1

		2014		
BRANCH NR.	Title	Men	Women	Difference
102020	Forarbejdning og konservering af fisk, krebsedyr og bløddyr, undtagen fiskemel	49,6	50,4	-0,8
139600	Fremstilling af andre tekniske og industrielle tekstiler	50,9	49,1	1,8
463890	Specialiseret engroshandel med fødevarer i.a.n.	49,9	50,1	-0,2
471130	Discountforretninger	50,6	49,4	1,2
473000	Servicestationer	49,8	50,2	-0,4
479114	Detailhandel med bøger, kontorartikler, musik eller film via internet	49,7	50,3	-0,6
479119	Detailhandel med andre varer i.a.n. via internet	49,7	50,3	-0,6
561010	Restauranter	50,9	49,1	1,8
649230	Andre kreditselskaber	50,4	49,6	0,8
641900	Banker, sparekasser og andelskasser	49,5	50,5	-1
649230	Andre kreditselskaber	49,8	50,2	-0,4
662900	Andre hjælpetjenester i forbindelse med forsikring og pensionsforsikring	49,4	50,6	-1,2
702100	Public relations og kommunikation	50,9	49,1	1,8
732000	Markedsanalyse og offentlig meningsmåling	50,7	49,3	1,4
773300	Udlejning af kontormaskiner og -udstyr, computere og it-udstyr	49,1	50,9	-1,8
829900	Anden forretningsservice i.a.n.	49,9	50,1	-0,2
854200	Videregående uddannelser på universitetsniveau	49,7	50,3	-0,6

889160	Fritids- og ungdomsklubber	50,6	49,4	1,2
910300	Historiske monumenter og bygninger og lignende attraktioner	49,5	50,5	-1
932100	Forlystelsesparker og lignende	49,3	50,7	-1,4

		2008		
BRANCH NR.	Title	Men	Women	Difference
464410	Engroshandel med porcelæns- og glasvarer	50,1	49,9	0,2
553000	Campingpladser	49,8	50,2	-0,4
561020	Pizzeriaer, grillbarer, isbarer mv.	49,7	50,3	-0,6
139600	Fremstilling af andre tekniske og industrielle tekstiler	50,9	49,1	1,8
591300	Distribution af film, video- og tv-programmer	49,6	50,4	-0,8
329100	Fremstilling af koste og børster	50,2	49,8	0,4
822000	Call centres virksomhed	50,3	49,7	0,6
823000	Organisering af kongresser, messer og udstillinger	50,8	49,2	1,6
853200	Tekniske skoler og fagskoler	49,1	50,9	-1,8
854100	Videregående uddannelser på universitetsniveau	49,2	50,8	-1,6
854200	Videregående uddannelser på universitetsniveau	50,1	49,9	0,2
932100	Forlystelsesparker og lignende	50,7	49,3	1,4
910300	Historiske monumenter og bygninger og lignende attraktioner	49,4	50,6	-1,2