MASTER THESIS

Identifying Possible Bias Indicators in Job Advertisements

by

R.P.H. Frissen

i6258734

Maastricht University, School of Business and Economics

MSc Business Intelligence & Smart Services

Regular thesis

Munstergeleen, 22 August 2021

Thesis supervisor: Dr. Rohan Nanda (Maastricht University)

Second reader: Dr. Niels Holtrop (Maastricht University)

Abstract

In recent years, the work of organizations in the area of digitization has intensified significantly. This trend is also evident in the field of recruitment where job application tracking systems (ATS) have been developed to allow job advertisements to be published online. However, recent studies have shown that recruiting in most organizations is not inclusive, being subject to human biases and prejudices. Most discrimination activities appear subtly but early in the hiring process, for instance, a non-inclusive choice of words in describing a job advertisement can inadvertently discourage qualified applicants from applying. In this thesis, we present various systems which utilized linguistic and semantic features along with the most recent state-of-the-art contextual word embeddings and transformer language models to identify language choices that explicitly or implicitly reflect bias. These features were fed to supervised machine learning classifiers to identify bias and discriminatory language in job advertisements. Our research focuses on five broad categories of extant biased language in job advertisements. The results indicate that for the semantic text classification models, the Random Forest classifier with FastText word embeddings, achieved the best performance with 10-fold cross validation. Regarding the different named entity recognition models, the results indicate that NER model 4, using lemmatized words and word vectors during model training, leads to the best performance with an accuracy of 99.84, a recall of 99.78 and an F1 score of 99.81.

Table of contents

List of Figures

List of Tables

List of abbreviations

DT - Decision Tree

EMSCAD - Employment Scam Aegean Dataset

LR - Logistic Regression

MLP - Multi-layer Perceptron Classifier

NB - Naive bayes

NER - Named Entity Recognition

NLP - Natural Language Processing

RF - Random Forest

RQ - Research question

SVM - Support Vector Machine

TD-IDF - Term Frequency–Inverse Document Frequency

Acknowledgements

Before you can proceed to read this thesis, I would first like to thank a number of people who made it possible for this result to be presented to you. First of all, I would like to thank Dr. Rohan Nanda as my supervisor who provided the resources whenever they were needed. In addition, I would also like to thank Kolawole John Adebayo for his input and guiding my thinking and methodology in the right direction to achieve this result.

Furthermore, I would like to thank Dr. Rui Jorge De Almeida e Santos Nogueira and Dr. Niels Holtrop who provided a very valuable contribution as course coordinators within my curricula at the University of Maastricht. The skills obtained in the courses Business Analytics and Analysing unstructured data have certainly contributed to the successful completion of my thesis.

I would like to thank my family members for being always there for me.

Richard Frissen

1 Introduction

A conventional approach to think about bias and discrimination is that it confers preference on someone or treating someone differently than another person. In today's society, discrimination and prejudice cuts across a wide area of human endeavors and affects countless number of people, especially those designated as underrepresented or minority. Bias and discrimination serve as a thumbprint of socially constructed stereotypes as they are often a product of extensive cultural and societal learning [22]. For instance, right from an early age, cultural attitudes about gender and race are often learned. In general, bias can be explicit or implicit. The latter of which is more pervasive, ingrained in peoples' minds and hearts and lends itself to an unconscious classification of information (gender, race, age, sexual orientation, disability, etc.) into associations to the disadvantage of the disfavored social groups.

1.1 Bias and discrimination in recruitment

It is well documented how some applicants get an unfair advantage due to specific physical appearance, gender, and ethnicity [3] [10]. For instance, women and Black people are usually at a disadvantage position than their White male peers when applying for jobs. A study also established that Black candidates are historically at a disadvantage compared to Caucasians with similar qualification and experience, and in fact, substituting Black candidates' names with fictitious White names significantly increased interview and job chances [23].

Previous studies have analyzed and categorized bias as it pertains to the recruiting industry [5] [8] [24]. The study carried out in [2] shows the occurrence of bias in the job market and provides insight into the evolution of the number of job advertisement compared to the occurrence of bias. Similarly, studies have shown that bias and discrimination, nonetheless implicitly, can also be reflected in the language in which employment job advertisements are written [24] [21]. For instance, certain masculine-leaning words or phrases in job advertisements have been found to dissuade female job applicants [1] while some racially sensitive language discourages minority/immigrant applicants from considering to apply for the advertised jobs. In particular, the authors in [3] show the appearance of gender bias in IT job descriptions and propose a tool to determine whether a job description is male, or female oriented based on the writing style in the job description. The essence of these studies is to establish the presence of bias and discrimination in job descriptions. This is intuitive given that the first point of the hiring process is the public release of a job advertisement. More importantly, the hiring process consists of the attraction, selection, and retention phases. The attraction phase is the first point where an employer seeks to invite and convince suitable applicants to join the organization. This is usually done through job advertisement. The selection phase involves assessing the candidates that applied for the job, e.g., reviewing candidates' CVs, matching candidates to job positions and shortlisting qualified candidates for interview. While attention is usually focused on the selection phase and machine learning systems have been developed to automatically review candidates' profile and select suitable candidates for recruitment, the attraction phase has largely been overlooked. If potential and well qualified candidates from certain groups feel alienated from applying for jobs solely because the job description portray stereotypes that feel offensive to them then the labour market would be losing significant skills from these groups of people.

In today's changing society, there is an increasing need for inclusiveness in the workplaces. Importantly, studies have established a strong correlation between diversity and inclusiveness in workplaces and its attendant innovation, productivity and profitability for employers [25]. With the advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies and the wide adoption it is enjoying in solving social and humanitarian problems, it is expedient that AI technology is employed to analyze job advertisements and to detect and identify bias or discriminating language choices that could cause qualified candidates to feel excluded from applying for jobs.

1.2 Thesis scope

This thesis presents a system that uses state-of-the-art NLP technology to empower recruiters in writing more inclusive job descriptions or adverts. The proposed system autonomously performs semantic analysis on an input job advert written in English language and allows recruiters to quickly identify any choice of language at the word, phrasal, or sentence level that reflect bias or discrimination to prospective job applicants. The identification of such language is important to ensure that everyone can feel involved in a job when the content appeals to them. A reason why someone may not feel engaged by a job description may be due to the prominence of innocuous-looking words like “challenging” or “dominant”. Studies have shown that these words tend to appeal to male applicants [3] [19]. In addition to the above words, there is a wide range of other words that studies have attributed to unconsciously enticing candidates of a particular gender at the expense of the other, thereby enhancing bias and discrimination through job descriptions. The recruiter using our system can then act on the analysis to make the advert more inclusive.

The significance of our approach stems from the fact that recognizing bias and non-inclusive language is very complex. It involves identifying language choices that implicitly or explicitly convey discrimination. This in return requires carrying out an extensive text analysis at different levels of text granularity – words, phrases, sentences, etc. To achieve this, the underlying system must approximate language semantics and contextual understanding of job descriptions. In addition, the ideal system must provide an interpretable or explainable feedback loop for users to understand the result, which makes the task more complex. Currently available systems exclusively focus on the selection phase of the hiring process and therefore lack the capabilities to eliminate bias at the attraction phase of hiring which in turn have formed the framework on which the proposed system has been developed [26] [12].

This study focuses on five broad categories of bias and discriminatory language in job descriptions. These categories include the following:
a) Masculine-coded language
b) Feminine-coded language
c) Exclusive language
d) Demographic and Racial-coded language and
e) LGBTQ-colored language.

For example, the Masculine-coded language category contains words like “ambitious” and “dominant”, which tend to lean towards the male gender and are therefore more likely to attract men. In contrast, the Feminine-coded language category contains terms like “collaborative” and “honest”, which tend to lean towards the female gender and are therefore more likely to attract women.

Exclusive language is a category built on a number of smaller stand-alone categories, namely: disempowering terms such as “ninja” and “rockstar”; offensive terms that can be related to people with disabilities, illnesses, addictions; or characteristics of people that may be perceived as “not natural” by another person. The Demographic and Racial-coded language category contains words that tend to favour people of a specific ethnicity or historically sensitive words, including words like “African” or “Arabs”. Finally, the category LGBTQ-colored language focuses on people’s relationship status as well as sexual preference. Think of terminology such as “married” or “non-binary”.

1.3 Research questions

To this effect, this thesis seeks to pursue a research objective which is to identify and classify bias indicators in job descriptions utilizing NLP. In order to achieve the desired research objectives, this study aims to provide answers to the following research questions:

RQ1 - “In what way is bias and discrimination currently present during the recruitment process and how does this manifest itself in the context of job descriptions?”

RQ2 - “Which state-of-the-art natural language processing technologies can best be employed to achieve the optimal model to automatically identify and classify possible bias indicators in job descriptions?”

RQ1 focuses on examining the literature for existing work on behavioural science and in particular, bias and discrimination in the recruiting industry. This would in turn aid the curation of bias and discrimination language indicator and terms which we can code into the respective categories earlier highlighted in this research work.

By RQ2, this study seeks to investigate and develop suitable machine learning models for semantic analysis of job adverts and eventual identification of non-inclusive languages in job descriptions. Furthermore, the model will classify the identified information into their respective coded categories. Collectively, answering these research questions will deliver methods and techniques that will help fulfils our research objective.

1.4 Research contribution

The main contribution of this study is the developed system to identify bias and discrimination in job descriptions using the various approaches. Different methods and techniques were utilized to develop various machine learning models to identify bias and discrimination in job descriptions. In addition, a sample of the annotated dataset (up to 3000 job descriptions, published in batches) and the custom trained machine learning models used during this research, will be published on the publicly available github repository https://github.com/RichardFrissen/A-Machine-Learning-Approach-to-Recognize-Bias-and-Discrimination-in-Job-Advertisements. The aforementioned together with the clarified approach, implications and limitations presented in this research, make up the major contributions of this research. In addition to the above, based on the findings and knowledge presented in this thesis, the corporate sector will probably be able to implement such a system more easily in its organizations. This will help stimulate adoption of such a complex technical application, after which organizations and individuals will benefit from a more diverse environment with thanks to a more inclusive recruitment process.

To the best of our knowledge, this is the first study to propose the use of machine learning (ML) and NLP to tackle bias and discrimination at the attraction phase of hiring, by focusing on classifying an extensive set of the five specific categories as listed above. The research work and the reported results are based on a publicly available real world job advertisement dataset, Employment Scam Aegean Dataset (EMSCAD) [16]. A gazetteer-based approach was used to semi-automatically generate an annotated corpus by tagging the biased language terms in the job advertisements.

1.5 Thesis outline

The remaining parts of the thesis is organized as follows. In chapter 2, we examine the literature and discuss the hiring process as well as bias and discrimination in recruitment. In chapter 3, we present our methodology while in chapter 4, we present and analyze the result of the experiments conducted. Chapter 5 constitutes a comprehensive discussion on the findings described in chapter 4. In the last chapter, we give the conclusion and recommendation for future works.

2 Literature review

The previous chapter provided the introduction, context, and relevance of this thesis within its stated scope. In this chapter, the conducted systematic literature review is described. Based on this literature review, an endeavour will be made to obtain results to answer the first research question "In what way is bias and discrimination currently present during the recruitment process and how does this manifest itself in the context of job descriptions?".

Section 2.1 describes the different forms of bias and discrimination and how they may manifest themselves during the recruitment process. Section 2.2 reviews what measures are currently documented to minimize the occurrence of bias and discrimination during the recruitment process. In section 2.3 we focus on commerical applications that contribute to a more inclusive way of recruiting by using applications similar to those presented in this thesis.

2.1 Bias and discrimination in recruitment

Hiring is usually not a single decision but a chain of events that result into a job offer for an applicant. The first step is the talent attraction or sourcing phase where an employer hopes to generate a strong set of applicants. Typically, employers disseminate available job positions, the description of the role as well as ideal profile of candidates. The second step is the selection or screening phase where an employer/recruiter independently or with the aid of some AI algorithms assess and ranks the various applicants in order of their employability. The outcomes of the steps above might be influenced by bias. In the first step, a candidate from a particular group (e.g., female) may feel not enchanted to apply for the advertised job due to the way the job description is written [19]. In the latter step, a human recruiter may unconsciously have an ingroup preference for candidates with similar ethnicity or look [23] and an algorithm may have encoded societal stereotypes found in the data it was trained with [4].

2.1.1 Various types of bias and discrimination

Bias and discrimination exist in different forms in hiring. It is an implicit inclination or prejudice for or against one person or group. According to [19], discrimination can be either implicit (unconscious) or explicit (conscious), and can occur on the basis of gender, ethnicity, sexual orientation, ethnicity, culture, religion, age, etc. The driving force behind these influenced choices is the feeling that one develops during his or her life based on events, but also for example the media [18]. Discrimination is recognized when someone is treated unfairly in the same situation another person is in and preferential behavior occurs. An example of this is: a male and female person both apply for the same job, after which the male is offered the job while the female has better qualifications. Furthermore, the researcher states that ethnic, gender and age discrimination are the three most common forms of discrimination an applicant faces during the application process. In addition to these three forms, "pregnancy discrimination, political views and religious beliefs" are also mentioned as common forms of discrimination. Apart from the above, people with disabilities or other forms of impairments, for example because they are carriers of a disease, experience discrimination during the application process [19].

A study by [5] shows that as employees age, they are more likely to face discrimination because of their age during the application process. This discrimination occurs at the time of assessing which of the applicants will become the chosen new employee. Age is then often considered, despite the fact that an older applicant has often gained more skills and experience during their career.

Similar effects occur when people of color apply for jobs. A study conducted by [6] shows that individuals with dark skin color experience lower job suitability ratings than individuals with light skin color, in a situation in which both applicants master the same qualifications.

Unconscious bias can be divided into a number of different types, for instance: Halo effect, Affinity bias, Conformity bias, Cloven hoof effect, attribution bias, beauty bias and confirmation bias. Characteristics of these different types of bias are presented in Table 1.

Table 1: Common types of bias during recruitment

Bias type	Characteristics
Halo effect	Evaluating a person by one positive trait
Affinity bias	The “like me” bias, evaluating a person by their similarities to oneself
Conformity bias	Caused by peer pressure
Cloven hoof effect	Generalising one negative aspect of a person to all their performance
Attribution bias	Taking credit on one’s own successful work when blaming others for unsuccess, and vice versa when it is about someone else
Beauty bias	Beautiful/handsome people are more successful than others
Confirmation bias	Looking for evidence to support one’s opinions when forming an opinion about someone

Table reprinted from “Preventing discrimination in recruiting through unconscious biases”, by ÖSterlund, 2020, p. 14-15.

A study by [15] shows that gender bias is recognized being a common form of unconscious bias. In one specific example, it is stated that job titles are skewed toward both the male and female gender. A consequence that arises from this fact is that job titles will no longer ensure equality in terms of salary but, on the contrary, will ensure that this equality is undermined.

When it comes to the senior positions within companies, it is evident that there is less equality within many organizations. According to [21] we can state that the higher the position is hierarchically within an organization, the less equity occurs between men and women. This is a possible explaination for the uneven distribution which is evident in many companies these days.

Implicit discrimination is often based on implicit attitudes or certain stereotypes and is often unintentional. In contrast, explicit discrimination is often based on a specific aversion to a particular group of individuals. By using automation in the form of algorithmic decision making, more standardization of procedures takes place. This leads to decisions to become more objective, less biased and there will also be fewer failures since information that is processed by humans is often executed unsystematically, often leading to unwarranted decisions [13].

Unlike unconscious bias, conscious bias refers to perceptions of individuals or groups in society. *"Conscious bias may lead to disparate treatment of coworkers and can derail the process of search to bring new people into an organization"*. An example of this is the preference to work with men rather than women. This may lead to the exclusion of specific people when it comes to opportunities within the labor market, thus ultimately leading to discrimination [18].

Once a person has been confronted with bias and has taken cognizance of it unconsciously, it is difficult to remove it from their thoughts and way of thinking. Therefore, it is important to prevent bias before it occurs and to take the necessary precautions for it to stimulate awareness [19].

2.2 Measures taken to prevent bias or discrimination

Possible measures and solutions that are utilized to prevent bias or discrimination from occurring during the recruitment process, are for example: anonymizing the application process, whereby personal information of the applicant is removed from the resume so that bias cannot or can hardly take place; intensifying the application of proper recruiting management and interviewing techniques, by only conducting competence based structured interviews; or facilitating an unconscious bias training whereby people are made aware of how unconscious bias can be prevented [19].

The work in [3] presents an approach that can be used to measure gender bias in IT job postings. This approach involves the development of a prototype tool based on the training of a machine learning classifier, that allows gender bias to be identified in a job post. In addition, they also developed a user interface in which the job description can be reviewed for possible occurrence of bias and discrimination. In this way recruiters can be informed when gender bias unintentionally occurs in their just written job advertisement and it can be corrected by applying suggestive language proposed by the tool. HTML pages are processed to obtain the title and text of the job post, for further analysis. Next, the title and text are checked for the occurrence of gender biased language. On the grounds of this occurrence, a gender-neutrality score is calculated that indicates how gender-neutral a job post is. Findings of this study include the proposal of a keyword repository which courage or discourages women during the application process. In addition, they also proposed a prototype tool which allows recruiters to evaluate job posts and correct them if biased or discriminatory in nature by using suggested language by the tool. The first tests, based on three samples from different job portals show that the results are plausible but the way of calculating scores still needs improvement.

The research work in [7] presents a model consisting of three steps that explain the underlying causes for biased resume screening. The three steps included in the model are: Application information, impression formation and screening outcomes. Applicant information focuses on the qualifications, but also the non-job-related information that one can determine from these qualifications and characteristics that an applicant has. The second step of the model, impression formation, focuses on how the data retrieved for this purpose is processed by the recruiters. Despite the processes in this step being automatic or unconscious, there can still be a high degree of consciousness present in the decision making. For example, the assessment of a person's qualifications and characteristics based on the data collected for this purpose. The third phase of this model focuses on the outcomes arising from the resume screening process. Here, perceptions of similarity can influence the way applicants are attracted and retained. The three-stage model shows why resume screening is vulnerable to bias, but not why discrimination occurs in ethical terms. However, they identified a number of factors that contribute to biased resume screening. For example, they argue that the lack of extensive personal information of a particular applicant can lead to biased decision making. As a result, people are more likely to be pigeonholed based on stereotypes. In contrast to the previous, they also argue that including non-work-related information of the job applicant, can lead to biased decision-making. Their research recommends that the candidate should aim to provide sufficient information about themselves in the application. However, the non-work related information should be limited to prevent biased decision-making. They also present a number of interventions that can be used to avert biased resume screening. Some of these include: anonymizing resumes, standardizing processes, training recruiters more intensively and holding them accountable for their hiring decisions.

The application of NLP within the recruitment industry is not new, but it is certainly innovative. NLP has been used to analyze verbal video resumes to examine the relationship between verbal content and perceived hireability ratings [17]. Tools based on NLP have also been developed for the automated extraction of relevant information such as skills, work experience and interests from resumes [20].

The authors in [11] developed a tool named RésuMatcher, which utilizes text similarity and machine learning models to find the optimal match between a job applicants' resume and a specific job post. Such tools are used in the recruitment industry to select ideal matching candidates for a particular job post.

A similar work is presented in [14] where the researchers designed a tool to identify the optimal candidate for a job description by using a Deep Siamese Network. This deep Siamese Network consists of a Convolutional Neural Network to effectively capture the underlying semantics. Their approach captures underlying semantics and significantly outperforms six commonly used representations used for similar practices, namely, Word-n-grams, TF-IDF, Bag-of-Words, Bag-of-Means (by using the average Word2Vec embedding of the training data), Doc2Vec and finally a Convolutional Neural Network.

Similarly, an application which matches the optimal candidate's profile considering the job criteria in job descriptions, as introduced in [8]. They propose a framework that makes use of NLP in combination with the hadoop framework to provide a fully scalable solution. The authors use a three-phase algorithm consisting of data gathering, in which resumes are retrieved and pushed to the hadoop distributed file system; Data processing, in which necessary fields are retrieved based on the previous step; and an attribute tagger, by which for example the name, email, phone number, etc. are gathered.

A study by [2] proposed a web application to predict the best fit resumes against given job descriptions posted by recruiters, with the main goal to lower the workload of recruiters to prevent them to go through all applicant's details. A comparison between a given job description and a candidate's resume is facilitated by using TF-IDF cosine similarity scores. An implication during this study occurred due the processing of unstructured data. While calculating the relevant work experience of a job applicant, years in which an applicant had been studying were sometimes counted.

In a recent study by [10], the researchers are referring to the improvements they want to make based on an opinion from The Association of American Medical Colleges. In order to make a positive contribution to diversity within organizations, they argue that many organizations employ targeted staff who promote diversity within their organizations. These so-called diversity officers can provide a helpful perspective to the recruitment processes currently in place. Based on these perspectives, recruiters can adjust their language. For example, to avoid situations like in which recruiters write words like "chairman" and "fellow" that create a bias toward men. This may discourage women from applying for a position described in a job description in which this choice of words has occurred in, for example, the job title or title of the job description.

2.3 Related commercial applications

In addition to research as described above, there are also facilities being organized in practice, whether commercial or scientific, to minimize or prevent the occurrence of bias and discrimination in job advertisements.

Project implicit is a nonprofit organization dedicated to human social cognition. Research is conducted within this body that often serves as the scientific foundation upon which knowledge is based. Project implicit's mission is to educate the public about prejudice. In addition, everyone can take an Implicit Association Test (IAT), which measures the strength of associations between concepts and evaluations or stereotypes. During this test, scores will be measured based on how fast a person answers a certain question. As a result, it can be determined which implicit preference a person has.

Gender decoder a public free tool, available at http://gender-decoder.katmatfield.com/, is supported by scientific research [24]. The tool allows anyone to check a written job description for the presence of gendered language.

From a commercial perspective, there are a number of companies that facilitate solutions in the area of limiting biased language in job descriptions, several of these are discussed. Textio, an augmented hiring platform, supports users through what is called an inclusion guidance. With this guidance it is easier for users to control the possible occurrence of gender, age and ability biased language. The user receives a “Textio-score” based on the extent to which these biases are present, which indicates how inclusive the written job description is. There is also integration with several popular communication information systems such as Microsoft Outlook, Gmail and LinkedIn.

Ongig is a similar tool that offers the possibility to check job descriptions for gendered, racial, disability and age bias. Using a gender neutrality score, the user can easily see how neutral the written text is.

Finally we have Diversely, a tool that facilitates recruiters in the verification of a job description and its release. First of all, a user can review a written job description for the presence of gendered, cultural, experience, impersonal or inclusive and non-inclusive language. Scores are calculated to check how inclusive a written job description is. The tool also checks for structure to determine a score. When the job ad is ready to be posted, the user can easily do this through their portal. Through this portal it is possible to post job advertisements on paid, unpaid and job boards dedicated to a specific target group, like women.

2.4 Discussion

Based on the systematic literature review, we can report some interesting findings. First, we looked at how bias and discriminatory language might occur during the recruitment process. Then we looked at what different types we can identify in this occurrence. Next, we looked at how measures are currently being taken to prevent or minimize this occurrence. Finally, a review of several companies currently offering leading solutions that embrace similar objectives as this thesis took place.

1. Bias occurs on both a conscious and unconscious basis. Where conscious bias refers to the perception of individuals or groups of people in society. Unconscious bias is often based on events an individual has experienced in his or her life.

2. Ethnic, gender and age discrimination are the most common forms of bias and discrimination during the recruitment process. In addition, pregnancy discrimination, political views and religious beliefs are also three common forms that appear during this process.

3. In previous studies where similar applications have been developed, the inclusiveness of an advertisement or text is often indicated by a score.

4. Various applications have already been developed using NLP in the recruitment process. The majority of these applications focused on finding the optimal match of applicant's resumes and job advertisements.

5. In the commercial field, there are a number of leading applications offered that include similar functionalities to the tool developed in this thesis, however, it is not possible to see what these tools are developed on. So the black box principle is in place here, most likely because these companies want to maintain their revenue model.

2.5 Summary

In this chapter we have discussed the background and literature that focuses on the occurrence of biased and discriminatory language in job descriptions. Furthermore, we have looked at the different ways in which this occurrence can exist and which measures are already in place and being used to minimize this occurrence.

The next chapter describes the methodology used to achieve the results presented in this thesis.

3 Methodology

This chapter is used to describe the methodology used to develop the Natural LanguagePprocessing pipeline to identify biased and discriminatory language in job descriptions. Section 3.1 describes the properties of the collected data used to train the different machine learning classifiers. Subsequently, section 3.2 describes the preparatory work that took place before further processing could take place. Section 3.3 explains how the annotations of the five different categories took place and where they were based on. After this, section 3.4 describes how the different named entity recognition models have been designed and how they vary from each other. Section 3.5 describes how the semantic text classification models were developed. In section 3.6 you will find an evaluation on the proposed applied methodology in which we briefly elaborate on the limitations.

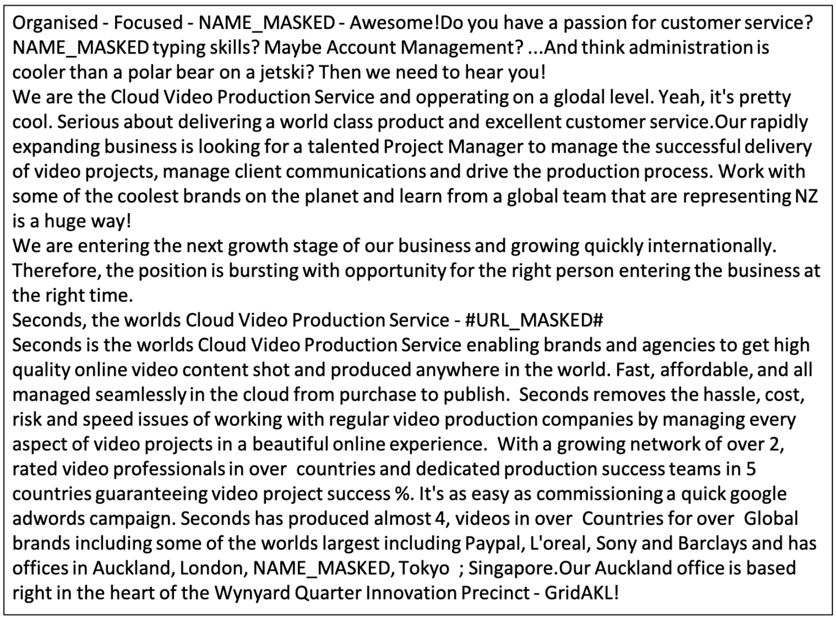
3.1 Data collection

The publicly available Employment Scam Aegean Dataset, EMSCAD, was used for this study [9]. It was selected because it contained real-life job descriptions and has been widely used for research purposes. The dataset consisted of 17014 legitimate and 866 fraudulent job advertisements. It was decided to only use the legitimate job advertisements. The personal information present in the dataset was either anonymized or removed by using regular expressions. Due to the limited computational resources available for training we utilized 3000 random sampled job descriptions for experiments. An example of a job description used during further development is presented in Figure 1.

3.2 Data pre-processing

For the purposes of this study, the focus was solely on the job descriptions of the job advertisements. Therefore, we only utilized the job description column from the EMSCAD dataset. By using regular expressions, HTML code and other noise-causing parts like empty lines and special characters were removed from the job descriptions. Next, the job descriptions were tokenized into sentences, after which invalid sentences were removed, for example when a sentence contained less than 2 words. Lastly, the sentences were then tokenized into words.

Figure 1: Example job description



Organised - Focused - NAME_MASKED - Awesome!Do you have a passion for customer service? NAME_MASKED typing skills? Maybe Account Management? ...And think administration is cooler than a polar bear on a jetski? Then we need to hear you!

We are the Cloud Video Production Service and opperating on a glodal level. Yeah, it's pretty cool. Serious about delivering a world class product and excellent customer service.Our rapidly expanding business is looking for a talented Project Manager to manage the successful delivery of video projects, manage client communications and drive the production process. Work with some of the coolest brands on the planet and learn from a global team that are representing NZ is a huge way!

We are entering the next growth stage of our business and growing quickly internationally. Therefore, the position is bursting with opportunity for the right person entering the business at the right time.

Seconds, the worlds Cloud Video Production Service - #URL_MASKED#

Seconds is the worlds Cloud Video Production Service enabling brands and agencies to get high quality online video content shot and produced anywhere in the world. Fast, affordable, and all managed seamlessly in the cloud from purchase to publish. Seconds removes the hassle, cost, risk and speed issues of working with regular video production companies by managing every aspect of video projects in a beautiful online experience. With a growing network of over 2, rated video professionals in over countries and dedicated production success teams in 5 countries guaranteeing video project success %. It's as easy as commissioning a quick google adwords campaign. Seconds has produced almost 4, videos in over Countries for over Global brands including some of the worlds largest including Paypal, L'oreal, Sony and Barclays and has offices in Auckland, London, NAME_MASKED, Tokyo ; Singapore.Our Auckland office is based right in the heart of the Wynyard Quarter Innovation Precinct - GridAKL!

3.3 Data annotation

The dataset was annotated using 524 unique biased or discriminatory terms divided into the five categories: Masculine-coded language, Feminine-coded language, Exclusive language, LGBTQ-colored language, Demographic and Racial-coded language. The words used for annotation are mostly based on literature, the reputable websites such as universities or logical assumption.

A study by [24] contributes to the dictionary of Masculine- Feminine-coded language by their results. In addition, [29] [32] [34] and [35] contributed a list for the categories of Masculine- and Feminine-coded language. [30] [32] [34] and [35] formed the list for Exclusive language. Reports by [32] [34] and [35] led to a list of terms for the Demographic and Racial-coded language category. And finally, the dictionary used for the annotations related to the category LGBTQ-colored language was based on [30] [32] and [35]. To expand these dictionaries, in some cases words were added manually. For example, when the word “fireman” was mentioned in an article, variations such as “salesman” and “mailman” were also added.

Given the large number of unique words that needed to be annotated, and the labour intensive nature of manual annotation, a semi-automatic method for annotation was chosen. A gazetteer-based approach was used to semi-automatically generate an annotated corpus by tagging the biased language terms in the job advertisements. A thorough inspection was done to ensure that the tagged annotations were correct. The annotated dataset was made available for further processing. Due to limited computational resources, a random sample of 500000 tokens (3000 job descriptions) was used for further processing.

3.4 Named Entity Recognition models

Various Named Entity Recognition (NER) models were developed to discover which combinations of configurations resulted in the optimal performing model. SpaCy's NLP library was utilized to develop and evaluate these different NER models.

3.4.1 Model properties

Utilizing this library brought the advantage that the use of word vectors was supported, this ensured a short processing time during development and model training. However, it was not possible to add custom features to the model. As a result, we were constrained to use the standard features. In Table 2 the various tested configurations are presented. The column "properties model pre-processing" refers to the model used during pre-processing, "Properties model training" refers to the the pre-trained model used to train the different NER models. "Tokens lemmatized" refers to the option of whether tokens are broken down to the root of the word or not.

3.4.2 Data annotation

Due to the fact that different pre-processing configurations are used as can be seen in Table 2, the amounts in annotations vary by type of pre-processing. These various pre-processing steps result in three different data sets, named NER_Dataset_01, NER_Dataset_02 and NER_Dataset_03 for clarity. The number of annotations in each category for every dataset is presented in Table 3. These different in numbers arise based on different configurations such as, for example, the use of different models during the preprocessing of the data.

Table 2: Named Entity Recognition model configurations

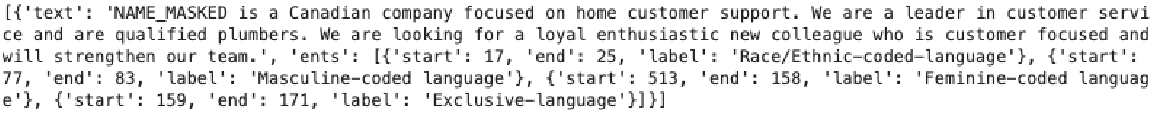
Model ID	Properties model pre-processing	Properties model training	Tokens lemmatized
NER Model 1	no word vectors	no word vectors	No
NER Model 2	no word vectors	large word vector (500k vectors)	No
NER Model 3	no word vectors	no word vectors	Yes
NER Model 4	no word vectors	large word vector (500k vectors)	Yes
NER Model 5	no static word vectors	no word vectors	No
NER Model 6	no static word vectors	large word vector (500k vectors)	No
NER Model 7	no static word vectors	no word vectors	Yes
NER Model 8	no static word vectors	large word vector (500k vectors)	Yes

Table 3: Named Entity Recognition number of annotations per dataset

	NER Model 1 & 2	NER Model 3 & 4	NER Model 5 & 6	NER Model 7 & 8
Category	NER_Dataset_01	NER_Dataset_02	NER_Dataset_03	NER_Dataset_04
Demographic and Racial language	74	290	74	286
Exclusive language	480	643	480	665
Feminine-coded words	5248	8947	5248	9082
LGBTQ-colored language	13	23	13	24
Masculine-coded words	6362	8869	6362	9251
O	487823	481228	487823	480692

Using the aforementioned annotated datasets, the tokens were either lemmatized or not, depending on the model's configuration. Next, several processing steps are required to achieve the final binary input format required to train the NER models. This final input format is created based on a tuple-like annotation format and results in a train- and test set. A proportion of 80% was used for model training, the remaining 20% was used for model evaluation, resulting in 2400 descriptions used for model training and 600 descriptions for evaluation. An example of a tuple-like annotation is presented in figure 2.

Figure 2: Example NER annotation



```
[{'text': 'NAME_MASKED is a Canadian company focused on home customer support. We are a leader in customer servi
ce and are qualified plumbers. We are looking for a loyal enthusiastic new colleague who is customer focused and
will strengthen our team.', 'ents': [{'start': 17, 'end': 25, 'label': 'Race/Ethnic-coded-language'}, {'start':
77, 'end': 83, 'label': 'Masculine-coded language'}, {'start': 513, 'end': 158, 'label': 'Feminine-coded languag
e'}, {'start': 159, 'end': 171, 'label': 'Exclusive-language'}]}]
```

3.5 Semantic Text Classification models

Various machine learning classifiers have been utilized to develop a variety of models in order to recognize biased and discriminatory language in job descriptions. This section is used to describe the configurations on which the resulting semantic text classification models are based.

3.5.1 Data annotation

The amount of annotations per category for the semantic text classification models, unlike the NER models, are the same for each trained semantic text classification model. This is because there is a different method of pre-processing compared to the NER models. Because this different method is applied, and features are only created for each token using the same set of tokens, this resulted in the exact same dataset for each model to be trained. The number of annotations per category are shown in Table 4.

Table 4: Number of annotations per category

Category	Number of annotations
Feminine-coded words	8947
Masculine-coded words	8869
Exclusive language	643
Demographic and Racial language	290
LGBTQ-colored language	23
O	481228

3.5.2 Feature engineering

Before the dataset as described before could be used to train the different models, it was first necessary to extract features. For each semantic text classification model the same linguistic features were extracted as described in section 3.5.2.1 Linguistic features. Semantic features were also created based on different types of word embeddings, as described in section 3.5.2.2 Semactic features. A combination of a type of semantic feature and the standard linguistic features, together with the token and an annotation category formed the dataset.

3.5.2.1 Linguistic features

Several linguistic features have been determined for each token. Each feature represents a characteristic property of the word. Table 5 presents the explanation of each linguistic feature.

Table 5: Linguistic features

Feature	Explanation
token.pos	Coarse-grained part-of-speech from the Universal POS tag set.
token.ent_type	Named entity type.
token.is_alpha	Does the token consist of alphabetic characters?
token.is_ascii	Does the token consist of ASCII characters?
token.is_digit	Does the token consist of digits?
token.is_lower	Is the token in lowercase?
token.is_upper	Is the token in uppercase?
token.is_title	Is the token in titlecase?
token.is_punct	Is the token punctuation?
token.is_space	Does the token consist of whitespace characters?
token.like_num	Does the token represent a number? e.g. “10.9”, “10”, “ten”, etc.
token.is_oov	Is the token out-of-vocabulary (i.e. does it not have a word vector)?
token.is_stop	Is the token part of a “stop list”?
token.lang	Language of the parent document’s vocabulary.
token.sentiment	A scalar value indicating the positivity or negativity of the token.
token.len(word)	The length of the token.

3.5.2.2 Semantic features

In addition to the features described in the previous section, we utilized various state-of-the-art pre-trained word embeddings to extract semantic features for the desired machine learning models. The different word embeddings which were used are: Word2Vec, BERT, ELMo, GloVe, Flair and FastText. Pre-trained word embedding models were used. The reason for choosing pre-trained models to determine word embeddings stems from the fact that word embeddings trained on the EMSCAD dataset would not demonstrate sufficient semantic quality due to the smaller size of the dataset. Word embeddings like BERT, ELMo and Flair also consider the context of the target word while computing word embedding vectors. Other pre-trained word embeddings do not support context and thus create the embedding vector based on a specific token only. Table 6 presents the pre-trained models used by word embedding type and shows whether they are context aware or not.

Table 6: Word embeddings used

Word embedding	Context-aware	Pre-trained model used
Word2VEc	No	"en-glove"
BERT	Yes	"bert-base-cased"
ELMo	Yes	"medium"
GloVe	No	"en-glove"
Flair	Yes	"news-forward-fast"
FastText	No	"cc.en.300.bin"

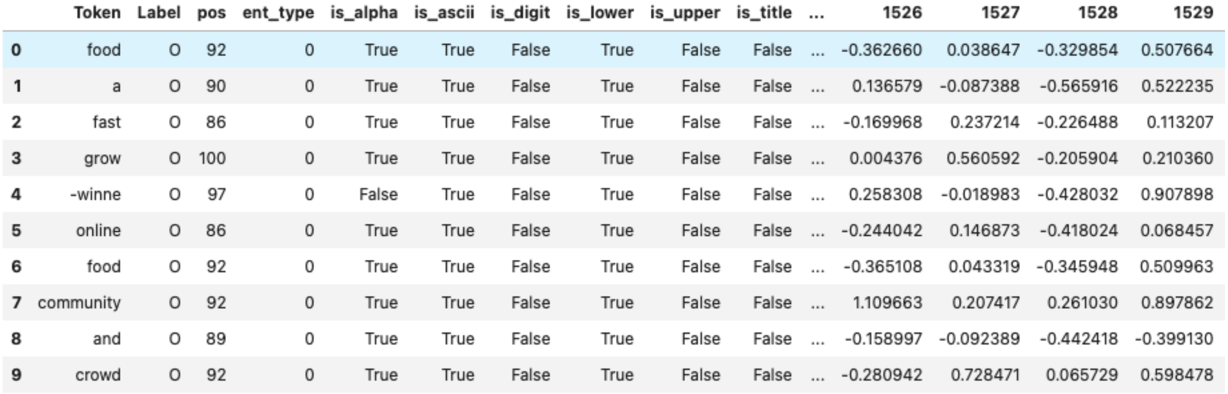
3.5.3 Feature selection

Once all the features were determined for each token, there are a total of 6 different feature sets. Each based on one type of word embedding combined with the linguistic features. The resulting featuresets are presented in tabel 7. A preview of featureset 6 is shown in figure 3.

Table 7: Semantic Text Classification feature sets

Feature set #	Linguistic features		Semantic features
Feature set 1	Linguistic features	—	Word2Vec Word embeddings
Feature set 2	Linguistic features	—	BERT Word embeddings
Feature set 3	Linguistic features	—	ELMo Word embeddings
Feature set 4	Linguistic features	—	GloVe Word embeddings
Feature set 5	Linguistic features	—	Flair Word embeddings
Feature set 6	Linguistic features	—	FastText Word embeddings

Figure 3: Example feature set 6



	Token	Label	pos	ent_type	is_alpha	is_ascii	is_digit	is_lower	is_upper	is_title	...	1526	1527	1528	1529
0	food	O	92	0	True	True	False	True	False	False	...	-0.362660	0.038647	-0.329854	0.507664
1	a	O	90	0	True	True	False	True	False	False	...	0.136579	-0.087388	-0.565916	0.522235
2	fast	O	86	0	True	True	False	True	False	False	...	-0.169968	0.237214	-0.226488	0.113207
3	grow	O	100	0	True	True	False	True	False	False	...	0.004376	0.560592	-0.205904	0.210360
4	-winne	O	97	0	False	True	False	True	False	False	...	0.258308	-0.018983	-0.428032	0.907898
5	online	O	86	0	True	True	False	True	False	False	...	-0.244042	0.146873	-0.418024	0.068457
6	food	O	92	0	True	True	False	True	False	False	...	-0.365108	0.043319	-0.345948	0.509963
7	community	O	92	0	True	True	False	True	False	False	...	1.109663	0.207417	0.261030	0.897862
8	and	O	89	0	True	True	False	True	False	False	...	-0.158997	-0.092389	-0.442418	-0.399130
9	crowd	O	92	0	True	True	False	True	False	False	...	-0.280942	0.728471	0.065729	0.598478

3.5.4 Machine Learning Classifiers

The machine learning classifiers were trained using the six unique feature sets as described in the prvious sections. The following classifiers were selected in order to find out which classifier performs best in combination with each of these six feature sets.

- Baseline classifier
- Support Vector Machine (SVM)
- Random Forest (RF)
- Logistic Regression (LR)
- Decision Tree (DT)
- Naive Bayes (NB)
- Multi-layer Perceptron classifier (MLP)

By performing parameter optimization using GridSearch, it was possible to seek for the optimal parameters for all machine learning classifiers. For all classifiers, the maximum iterations were increased to infinity to ensure that the models converge, this because of the widely varying categories that are difficult to classify. The SVM was trained using the radial basis function (rbf) kernel. The regularization parameter was set to 10. For the LR classifier, we found that the 'newton-cg' solver seems to give the optimal performance. For the remaining models, no better configuration options were found or they were not applicable. For the Baseline classifier, Scikit-learn's Dummy classifier has been utilized.

We used proportions of 80% for training and 20% for evaluating the models, resulting in 400000 tokens used for training and the remaining 100000 tokens for model evaluation. The performance of each model was evaluated using the accuracy, precision, recall and F1 scores. In addition, all machine learning classifiers were validated using a 10-fold cross validation. In evaluating this process, the macro averages of the precision, recall and F1 score were measured.

3.6 Evaluation

3.6.1 Sample dataset

A sample of 3000 job descriptions was used within this study due to the limited availability of computational resources. This is quite sufficient to obtain reliable results and to support the conclusions made. Nevertheless, in a future study it may be valuable to use a larger sample or even the complete 17880 job descriptions that are in the EMSCAD dataset. For example, this can be facilitated by using an architecture such as Maastricht University's Data Science Research Infrastructure (DSRI). This infrastructure was used in the final phase of the study, but due to time constraints it was not possible to use a larger sample.

3.7 Summary

In this chapter, the overall process carried out during this thesis, including the underlying steps, to arrive at the results obtained, has been discussed. First, the data collection process and the necessary pre-processing steps were explained. After this, the activities concerning the development of the various NER and semantic text classification models were explained. It was found that the chosen NER models resulted in four different datasets as different configurations were applied during the NER-specific pre-processing. However, this was not the case for the semantic text classification models, as no different pre-processing was used for these models. Finally, a short evaluation took place on this process and pain points were explained.

The following chapter presents the results that were achieved using the methodologies described in this chapter.

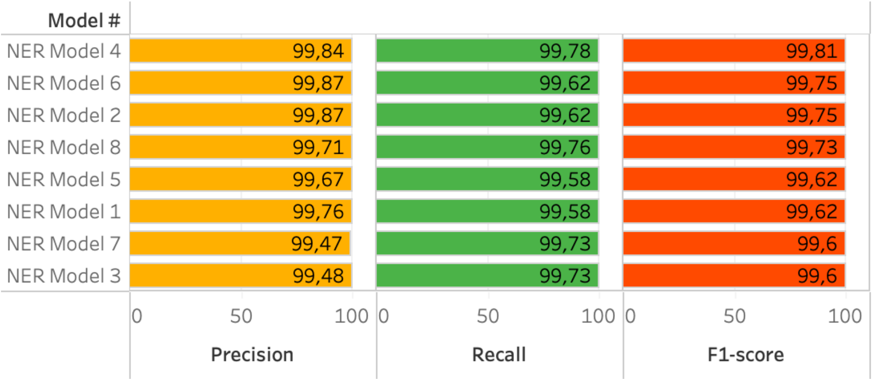
4 Results and evaluation

In this chapter, the results of the different NER and Semantic Text Classification models trained on the EMSCAD dataset are presented. By presenting these results, we will answer the second research question - “Which state-of-the-art natural language processing technologies can best be employed to achieve the optimal model to automatically identify and classify possible bias indicators in job descriptions?” - which is concerned with a thorough evaluation of the performance of the different utilized feature sets to identify and classify possible bias indicators in job advertisements using machine learning classifiers. The performance of each model was evaluated in terms of accuracy, precision, recall, and F1 scores.

4.1 Named Entity Recognition models

As a first attempt to answer the research question, various NER models have been trained according the models' configurations presented in Table 4. By training these various models, we wanted to seek for the optimal configuration and properties in order to get the best performing model. The NER models have been trained using 80% of the dataset whereafter the models have been evaluated using 20% of unseen test data.

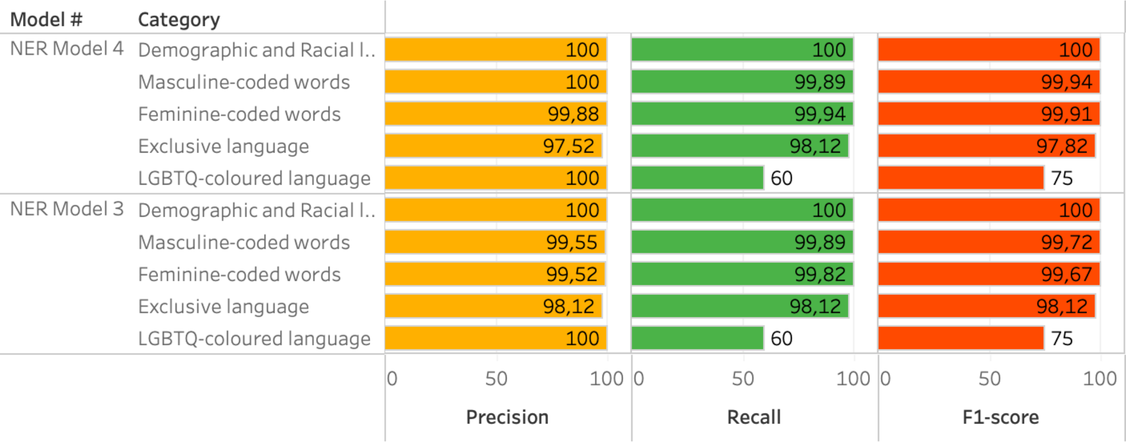
Figure 4: NER models performance



The results of the NER models are presented in Figure 4, in descending order based on performance measured by the F1 score. The best performing model is NER Model 4 with a precision score of 99,84, a recall of 99,78 and a finally a F1 score of 99,81. This model makes use of lemmatized words, no word vectors during pre-processing of the dataset and a large word vector containing 500k vectors during model training.

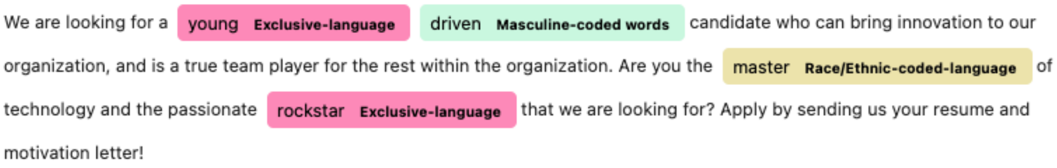
The least performing model is NER Model 3, with a precision score of 99,48, a recall of 99,73 and finally a F1 score of 99,6. This model is training on lemmatized tokens, using no word vectors during model pre-processing and no word vectors during model training. Scores determined for each class individually, are shown in Figure 5.

Figure 5: NER class performance



An example prediction made by using NER Model 4 is presented in Figure 6. A job description which does not originate from the training dataset is used to show predictions made by the model. The trained NER model identifies the biased- and discriminatory language and indicates these using color coding per category.

Figure 6: Example prediction - NER Model 5



We are looking for a young **Exclusive-language** driven **Masculine-coded words** candidate who can bring innovation to our organization, and is a true team player for the rest within the organization. Are you the master **Race/Ethnic-coded-language** of technology and the passionate rockstar **Exclusive-language** that we are looking for? Apply by sending us your resume and motivation letter!

4.2 Semantic Text Classification models

As described in chapter 3 Methodology, multiple machine learning classifiers were trained using six different feature sets to seek for the best performing combination of classier and use of features. The performance of each model was measured and evaluated using the accuracy, precision, recall and F1 scores.

Figure 8 presents the results for every machine learning classifier based on the evaluation data, trained on the various features sets. The results are sorted in descending order by the F1 score. Based on these results we can conclude the combination of BERT word embeddings and the Random Forest (RF) classifier leads to the best performing model, with an accuracy of 0,9992, a precision of 0,99999, a recall of 0,92992 and a respectable F1 score of 0,95660.

The results obtained using a 10-fold cross validation on the train data are presented in Figure 9. The results are again sorted in descending order by the F1 score. In contrast to the results in Figure 8, the classifier Random Forest in combination with the FastText word embeddings performs best with a precision of 0,9976, a recall of 0,9900 and finally a F1 score of 0,9936.

The class specific scores for the two aforementioned best performing models are shown in Figure 7.

Figure 7: Class performance - Random Forest BERT and FastText word embeddings

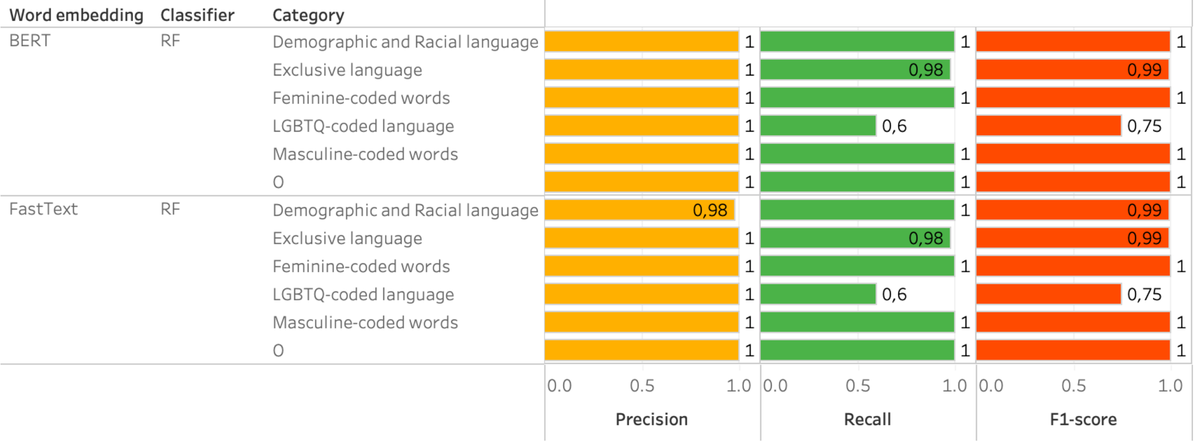
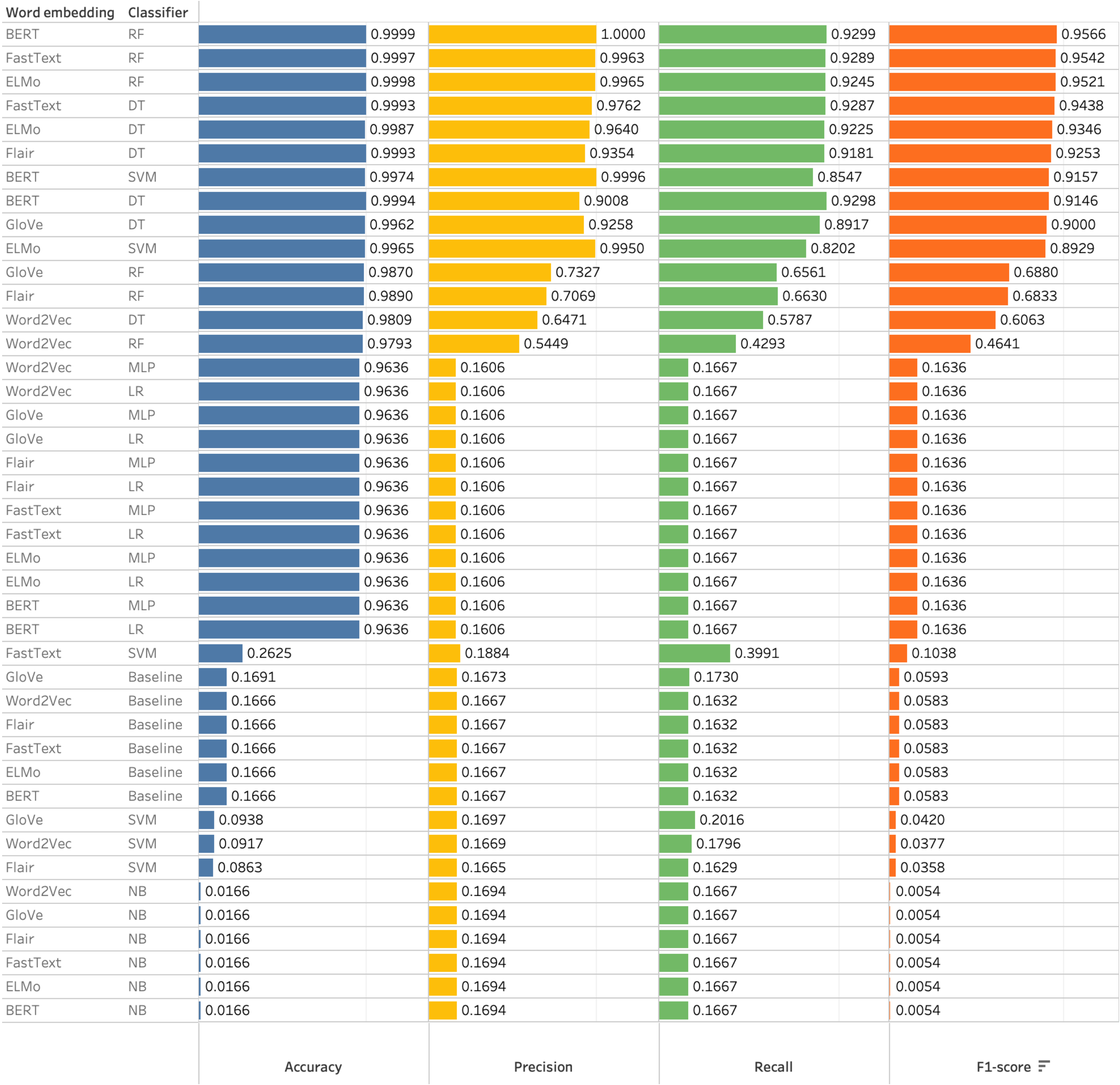
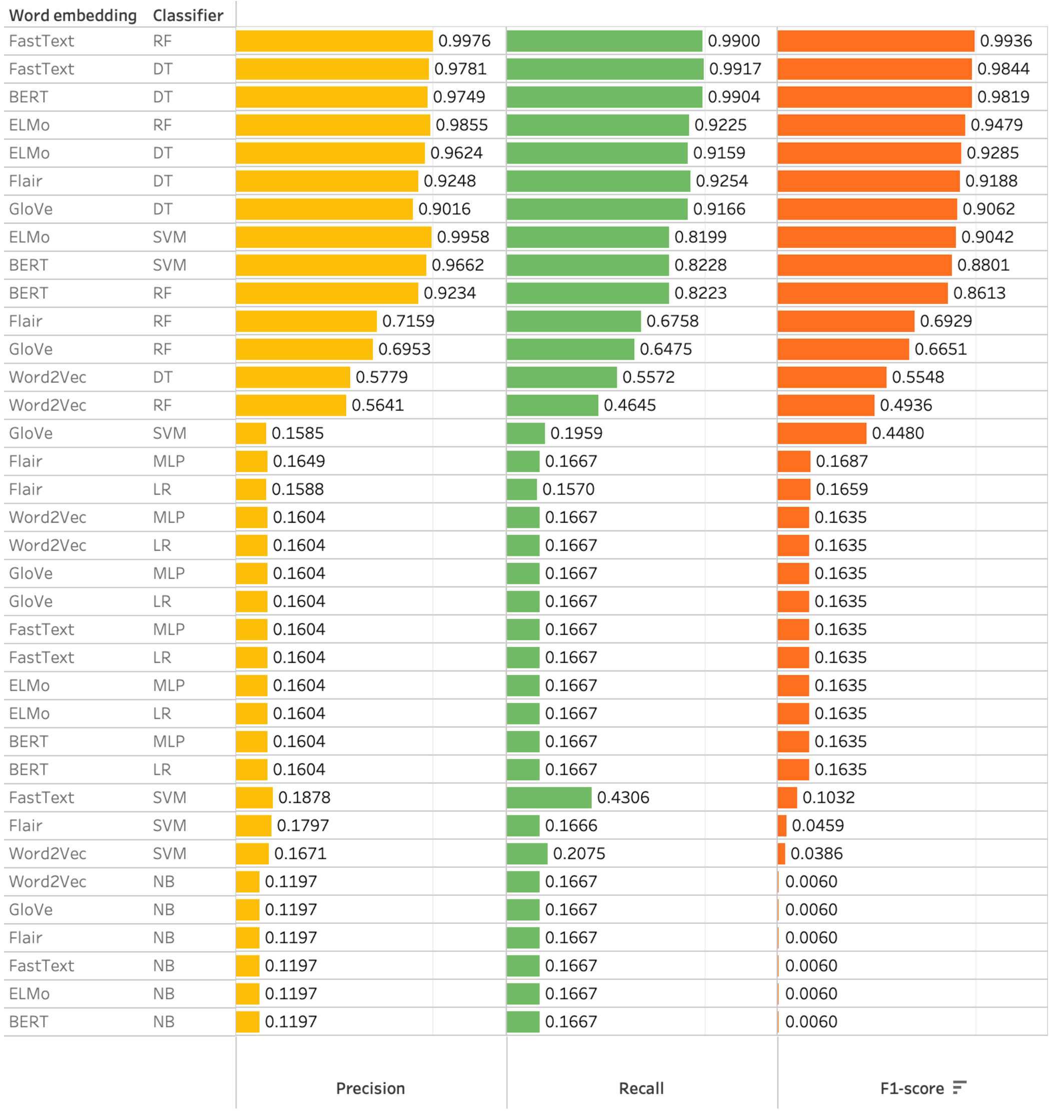


Figure 8: Semantic Text Classification models performance



Word embedding	Classifier	Accuracy	Precision	Recall	F1-score
BERT	RF	0.9999	1.0000	0.9299	0.9566
FastText	RF	0.9997	0.9963	0.9289	0.9542
ELMo	RF	0.9998	0.9965	0.9245	0.9521
FastText	DT	0.9993	0.9762	0.9287	0.9438
ELMo	DT	0.9987	0.9640	0.9225	0.9346
Flair	DT	0.9993	0.9354	0.9181	0.9253
BERT	SVM	0.9974	0.9996	0.8547	0.9157
BERT	DT	0.9994	0.9008	0.9298	0.9146
GloVe	DT	0.9962	0.9258	0.8917	0.9000
ELMo	SVM	0.9965	0.9950	0.8202	0.8929
GloVe	RF	0.9870	0.7327	0.6561	0.6880
Flair	RF	0.9890	0.7069	0.6630	0.6833
Word2Vec	DT	0.9809	0.6471	0.5787	0.6063
Word2Vec	RF	0.9793	0.5449	0.4293	0.4641
Word2Vec	MLP	0.9636	0.1606	0.1667	0.1636
Word2Vec	LR	0.9636	0.1606	0.1667	0.1636
GloVe	MLP	0.9636	0.1606	0.1667	0.1636
GloVe	LR	0.9636	0.1606	0.1667	0.1636
Flair	MLP	0.9636	0.1606	0.1667	0.1636
Flair	LR	0.9636	0.1606	0.1667	0.1636
FastText	MLP	0.9636	0.1606	0.1667	0.1636
FastText	LR	0.9636	0.1606	0.1667	0.1636
ELMo	MLP	0.9636	0.1606	0.1667	0.1636
ELMo	LR	0.9636	0.1606	0.1667	0.1636
BERT	MLP	0.9636	0.1606	0.1667	0.1636
BERT	LR	0.9636	0.1606	0.1667	0.1636
FastText	SVM	0.2625	0.1884	0.3991	0.1038
GloVe	Baseline	0.1691	0.1673	0.1730	0.0593
Word2Vec	Baseline	0.1666	0.1667	0.1632	0.0583
Flair	Baseline	0.1666	0.1667	0.1632	0.0583
FastText	Baseline	0.1666	0.1667	0.1632	0.0583
ELMo	Baseline	0.1666	0.1667	0.1632	0.0583
BERT	Baseline	0.1666	0.1667	0.1632	0.0583
GloVe	SVM	0.0938	0.1697	0.2016	0.0420
Word2Vec	SVM	0.0917	0.1669	0.1796	0.0377
Flair	SVM	0.0863	0.1665	0.1629	0.0358
Word2Vec	NB	0.0166	0.1694	0.1667	0.0054
GloVe	NB	0.0166	0.1694	0.1667	0.0054
Flair	NB	0.0166	0.1694	0.1667	0.0054
FastText	NB	0.0166	0.1694	0.1667	0.0054
ELMo	NB	0.0166	0.1694	0.1667	0.0054
BERT	NB	0.0166	0.1694	0.1667	0.0054

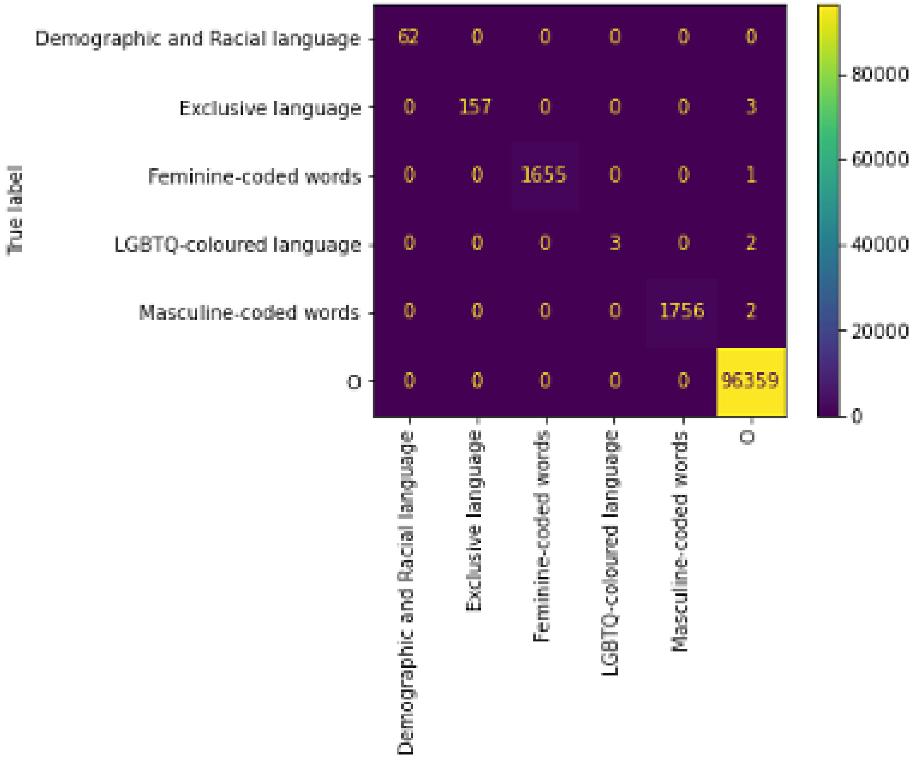
Figure 9: Semantic Text Classification models performance cross-validated



The resulting models were also tested against a Baseline model. Figure 12 presents the accuracy, figure 13 presents the precision, figure 14 presents the recall and finally figure 15 presents the F1 score. Based on the results presented in these figures we can conclude that the baseline classifiers performed poor compared with the other machine learning classifiers. Another observation we can withdraw based on the results shown is that the cross-validated best performing model, Random Forest with FastText word embeddings, outperforms the Random Forest with BERT word embeddings model by only a very small margin.

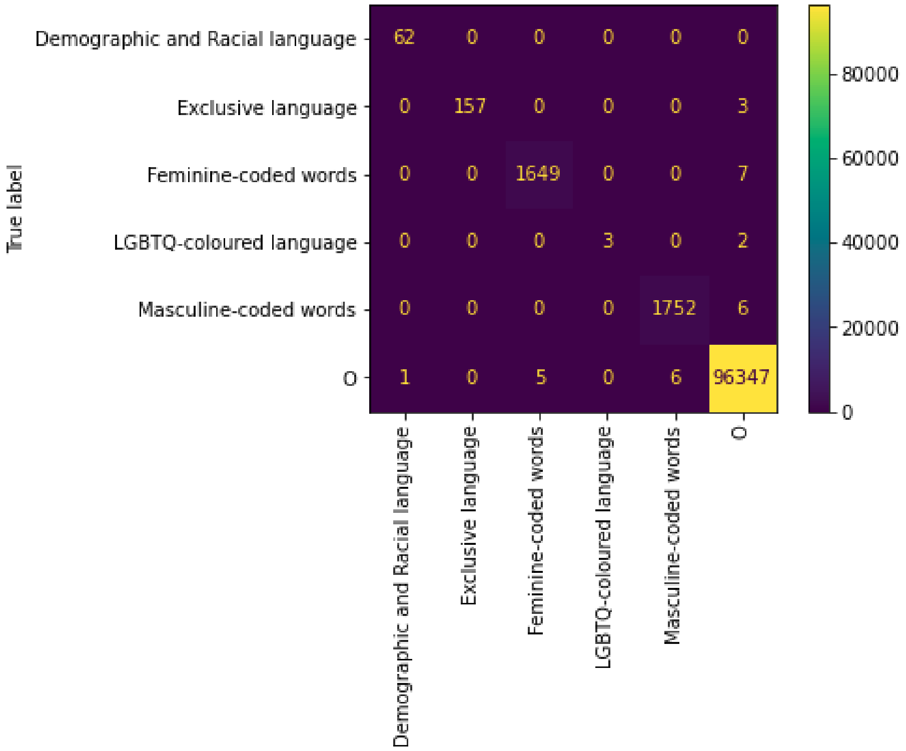
The confusion matrix of the two best performing models are presented in Figure 10 for the Random Forest classifier using BERT word embeddings, and Figure 11 for the cross-validated Random Forest classifier using FastText word embeddings.

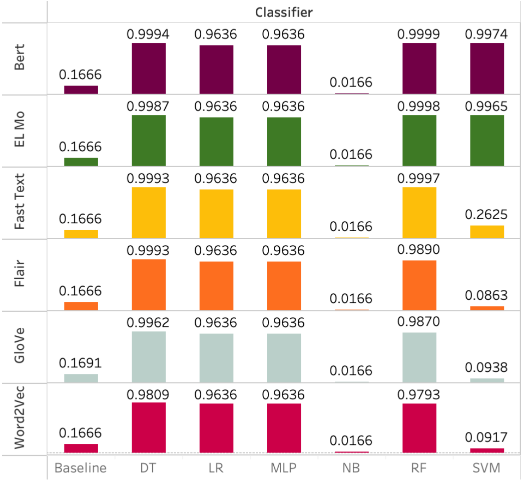
Figure 10: Confusion matrix - Random Forest BERT word embeddings



	Demographic and Racial language	Exclusive language	Feminine-coded words	LGBTQ-coloured language	Masculine-coded words	O
Demographic and Racial language	62	0	0	0	0	0
Exclusive language	0	157	0	0	0	3
Feminine-coded words	0	0	1655	0	0	1
LGBTQ-coloured language	0	0	0	3	0	2
Masculine-coded words	0	0	0	0	1756	2
O	0	0	0	0	0	96359

Figure 11: Confusion matrix - Random Forest FastText word embeddings





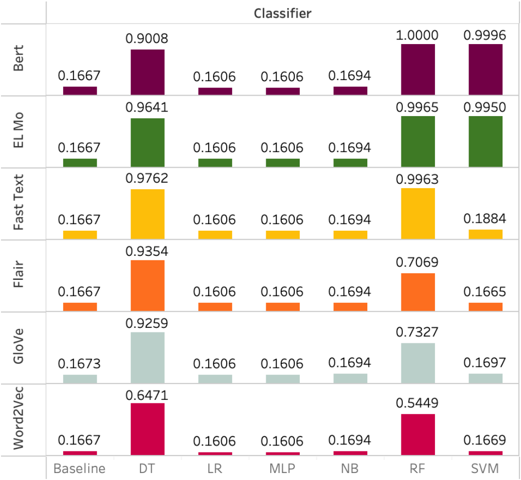


Figure 12: ML model performance - Accuracy

Figure 13: ML model performance - Precision

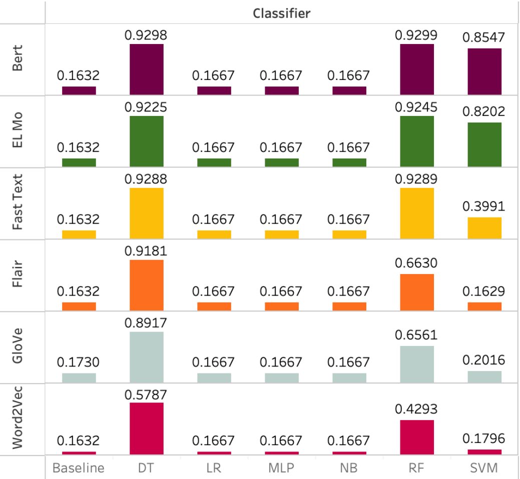


Figure 14: ML model performance - Recall

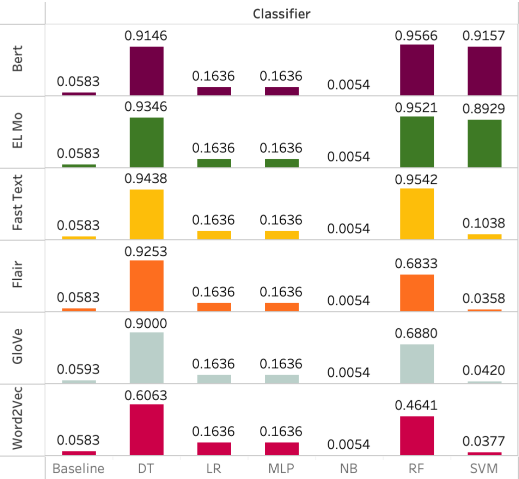


Figure 15: ML model performance - F1

4.2.1 Semantic Text Classification - Attempt to address class imbalance

As presented earlier in Table 2, the classes “LGBTQ-colored language”, “Demographic and Racial language” and “Exclusive language” are significantly less represented in the dataset. In this section an attempt to address this imbalance is described and evaluated.

Several tests have taken place by applying the resampling method. The purpose of this method is to ensure that classes with a large proportion are reduced in size, and smaller classes are increased in size, in order to ensure that there is more balance in the dataset. In all cases, majority classes were reduced in volume by undersampling them and minority classes were oversampled to increase their volume. It is important to note that this over- and undersampling only applies to the trainset. The main reason for this is that the actual test set may not be distorted because it needs to represents the reality.

Table 8 presents the proportions and table 9 shows the difference in number of annotations for the initial dataset and the three different balanced datasets. The impact on the results was examined when the initial proportion of the minority classes was multiplied by, 5 times, 10 times, and 15 times in volume. The minority classes include the categories: ”Masculine-coded words”, ”Feminine-coded words”, ”LGBTQ-colored language”, ”Demographic and Racial-coded language” and ”Exclusive language”. Based on this increase the size of the majority class ”O” decreases resulting in the total number of tokens remaining at 400000 as was initially the case for the original train set.

Resampling the trainset allows the underrepresented classes to probably become more predictable by the models. The three different balanced datasets were tested against the two best performing models, namely Random Forest using BERT word embeddings and Random Forest using FastText word embeddings.

Table 8: Proportions initial and balanced number of annotations per category in trainset

Category	Initial	Balanced_01	Balanced_02	Balanced_03
O	96,22%	81,08%	62,17%	43,26%
Feminine-coded language	1,82%	9,10%	18,20%	27,30%
Masculine-coded language	1,78%	8,90%	17,80%	26,70%
Exclusive language	0,12%	0,60%	1,20%	1,80%
Demographic and Racial-coded language	0,06%	0,29%	0,57%	0,86%
LGBTQ-colored language	0,00%	0,02%	0,05%	0,68%

Table 9: Initial and balanced number of annotations per category in trainset

Category	Initial	Balanced_01	Balanced_02	Balanced_03
O	384869	324345	248690	173035
Feminine-coded language	7291	36455	72910	109365
Masculine-coded language	7111	35555	71110	10665
Exclusive language	483	2415	4830	7245
Demographic and Racial-coded language	228	1140	2280	3420
LGBTQ-colored language	18	90	180	270

Figure 16 presents the performance of the models trained on the balanced dataset, after which they were evaluated on the untouched testset. The class specific scores for the best performing models are presented in Figure 17. Finally, the performance of the cross-validated models using a 10-fold cross validation are presented in Figure 18.

The confusion matrix for the Random Forest BERT model is presented in Figure 19 and Random Forest FastText is presented in Figure 20.

Figure 16: Semantic Text Classification models performance - balanced trainset

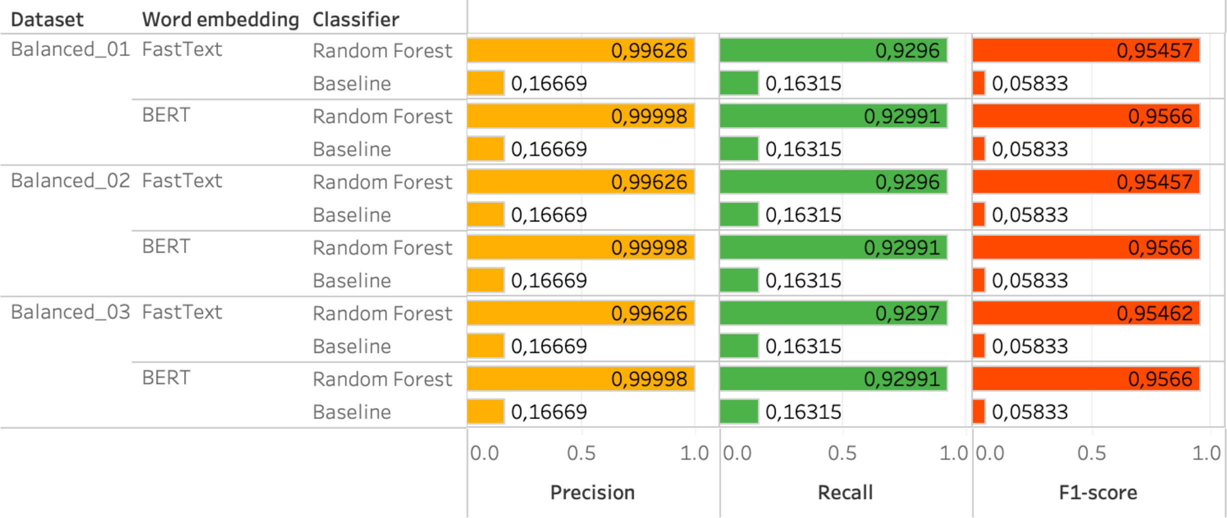


Figure 17: Class performance - Random Forest BERT and FastText word embeddings - balanced trainset

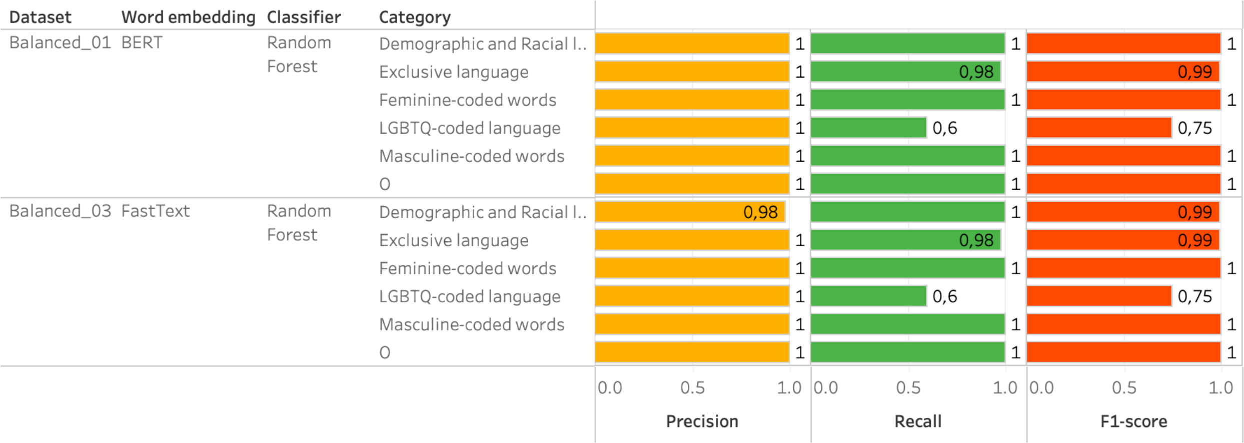


Figure 18: Semantic Text Classification models performance cross-validated - balanced trainset

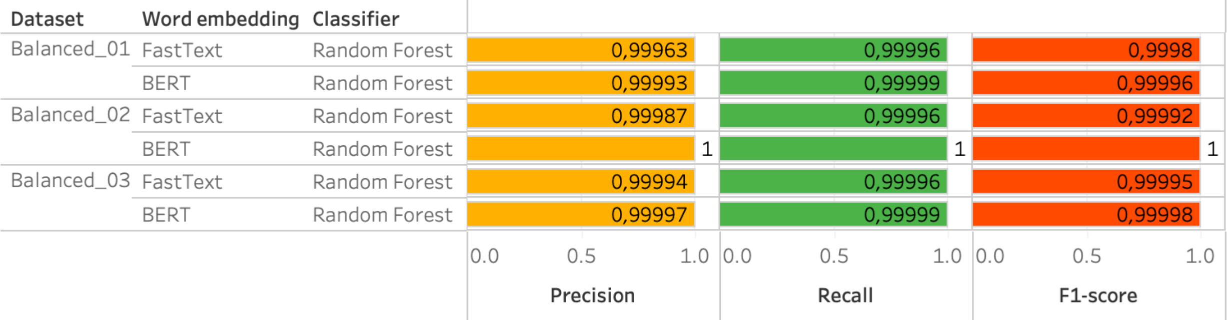


Figure 19: Confusion matrix - Random Forest BERT word embeddings - balanced trainset

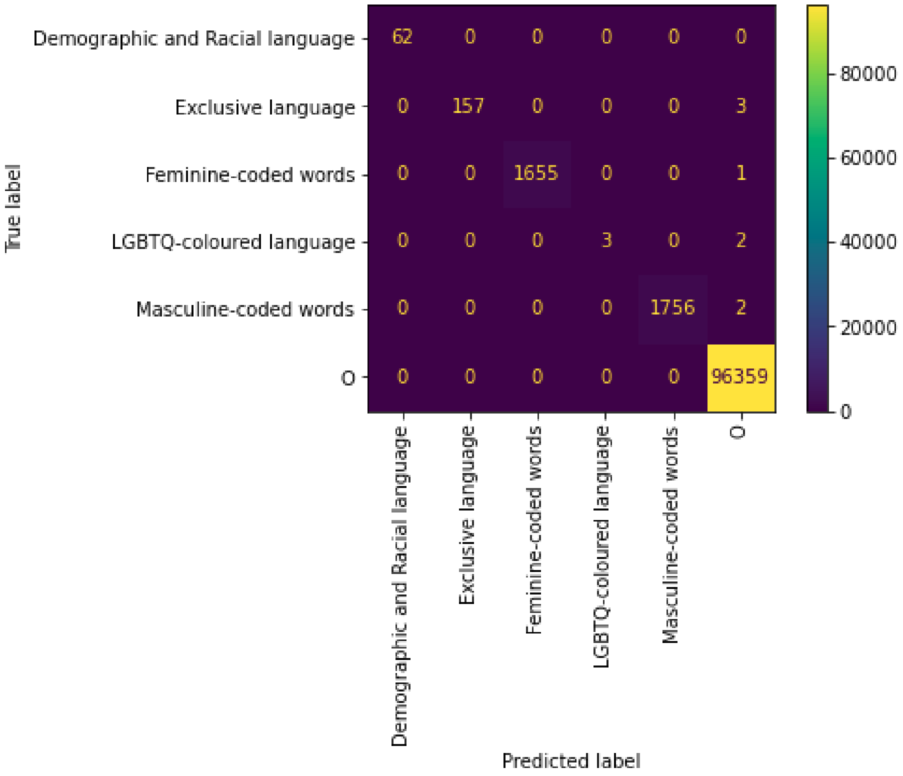
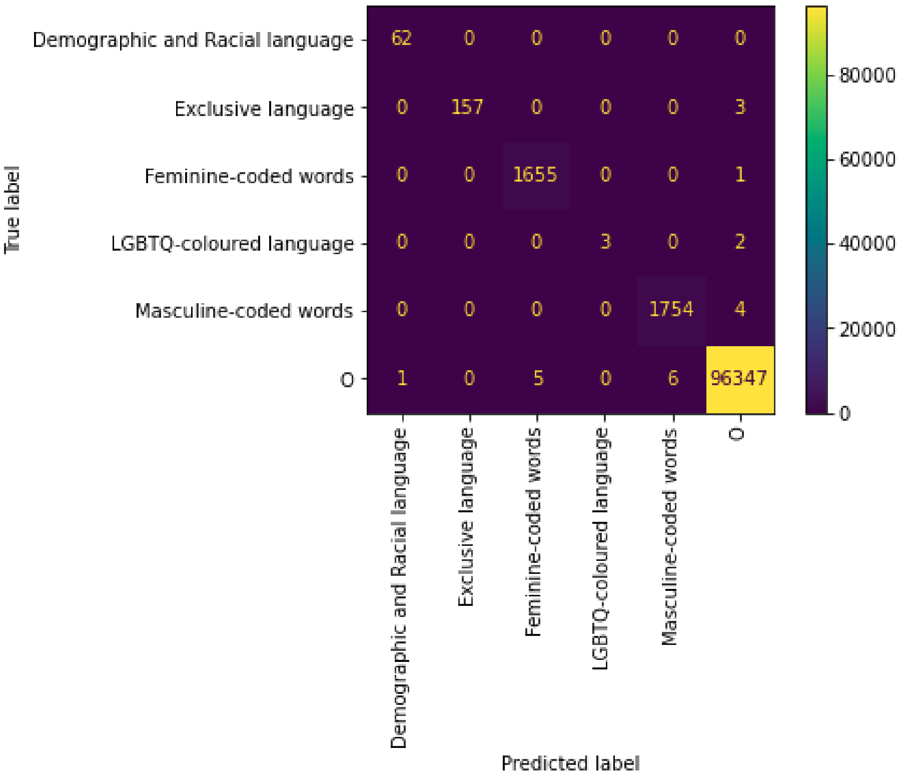


Figure 20: Confusion matrix - Random Forest FastText word embeddings - balanced trainset



True label \ Predicted label	Demographic and Racial language	Exclusive language	Feminine-coded words	LGBTQ-coloured language	Masculine-coded words	O
Demographic and Racial language	62	0	0	0	0	0
Exclusive language	0	157	0	0	0	3
Feminine-coded words	0	0	1655	0	0	1
LGBTQ-coloured language	0	0	0	3	0	2
Masculine-coded words	0	0	0	0	1754	4
O	1	0	5	0	6	96347

4.2.2 Model explanation

This section attempts to discuss how predictions were made by the best performing models. Because classification models are complex to interpret, especially when they use a classifier like Random Forest, several visualisations are presented. The best performing models were made transparent through the SHapley Additive exPlanations (SHAP) algorithm [27] [28].

Figure 21 the feature importance of the Random Forest classifier using BERT word embeddings is presented. In addition, figure 22 presents the feature importance of the cross-validated best performing model, using a Random Forest classifier with FastText word embeddings. A sample of the twenty most important features are visualized in both figures, in descending order.

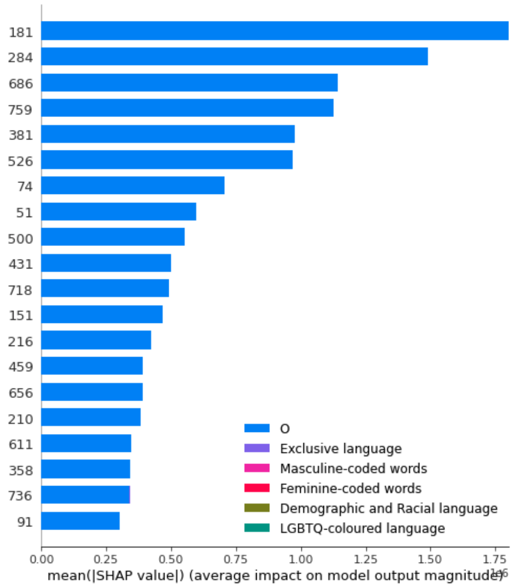


Figure 21: RF BERT - Feature importance

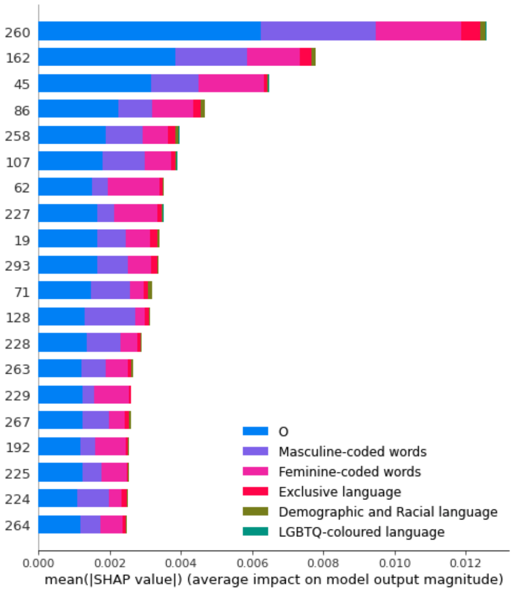


Figure 22: RF FastText - Feature importance

Each bar represents a feature on the Y-axis, the colors for each class in a bar denote the class interaction for the specific feature. The first interesting conclusion we can draw from figure 21 is that the predictions are mainly influenced by the category ”O”. In addition, we can also state that word embedding vector number 181 is the most influential with a percentile influence on the predictions made of 1.75.

The features presented in Figure 22 indicate that the interactions on which the predictions are based by the Random Forest classifier using FastText word embeddings, in contrast to the feature importances presented in figure 21, arise from a variety of classes, besides category "O". FastText word embedding vector number 260 is considered as the most important feature. In contrast to the importances shown in figure 21, the highest influence measured is only 0.0125 percentile which is very slight compared to the most influential features of the RF classifier using BERT word embeddings.

Since linguistic features were also used for both classification models of which we are examining their feature importance's, it is an interesting conclusion to draw that they are not represented in this summary and thus not belong to the twenty most important features in both models.

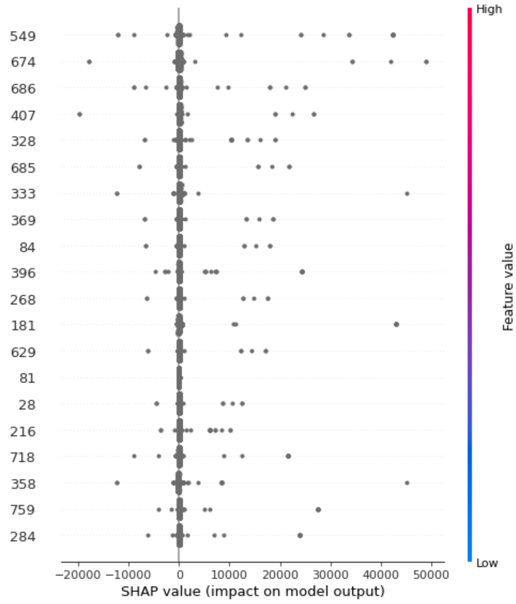


Figure 23: RF BERT - Beeswarm plot

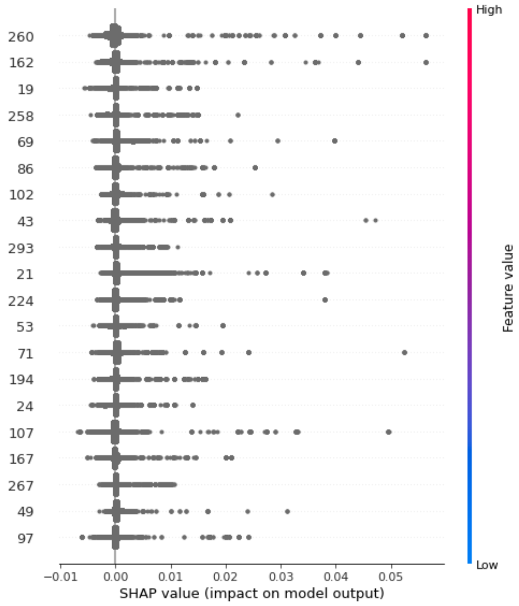


Figure 24: RF FastText - Beeswarm plot

Figure 23 and 24 present beeswarm plots that both illustrate how predictions are based for the ”Masculine-coded language” class, with Figure 23 focusing on the model RF using BERT word embeddings, and Figure 24 describing the model RF using FastText.

The X-axis of these plots demonstrate the impact of the feature on the predictions. Each point shown in the plot represents a single prediction made. The lower the value on the X-axis (negative value), the more influence on a negative outcome, resulting in a token that is not classified as Masculine-coded language. In contrast to the above, the higher the value on the X-axis (positive value) the more influence on a positive outcome, resulting in a token that is classified as Masculine-coded language. The Y-axis represents a sample of the twenty most important features specific for the class Masculine-coded language.

From the results of the Random Forest model using BERT word embeddings, presented in figure 16, we can withdraw that word embedding vector number 549 is the most important for this class. Apart from the outliers we can see in the plot, we see that the majority of classifications made by the model lead to a low influence on model output.

Figure 24, shows a larger spread of importance, however it is important to mention that the X-axis covers a much smaller range compared to the plot presented in figure 23. Again we see that the majority of tokens to be classified are located as a cluster around the turning point of zero. However, what the plot shows is that word embedding vector number 260, which is the most important feature making this classification, shows more spread on the positive half of the plot. Therefore, we can claim that it actually has positive influence on classifying a token belonging to the class Masculine-coded language.

4.3 Conclusion

For the NER models, we can can conclude that using a lemmatizer, no word vectors during preprocessing of the dataset leads to the best performing model in this situation. However, the results also indicate that utilizing the large word vectors model during for model training, will lead to the best performing model. The model scored a precision score of 99,84, a recall of 99,78 and a finally a F1 score of 99,81.

In addition, based on the presented results regarding the semantic text classification models, we conclude that the Random Forest classifier, using BERT word embeddings, yields the best performing semantic text classification model on test data, with an accuracy of 0,9992, a precision of 0,99999, a recall of 0,92992 and finally a F1-score of 0,95660. The cross-validated Random Forest classifier using FastText word embeddings shows even better performance, having a precision of 0,9976, a recall of 0,9900 and finally a F1 score of 0,9936.

Finally, comparing the performance of the NER models and the semantic text classification models using the evaluation data, we can conclude that the NER models perform better, considering the F1 score which is a good representation of the overall performance.

4.4 Summary

This chapter described the results achieved for the different NER and semantic text classification models. By using the accuracy, precision, recall and F1 scores it was possible to make a thorough evaluation of the performance of each individual model. In addition, a 10-fold cross validation was used to measure the performance of the semantic text classification models. Finally, an attempt was made to explain the models' predictions of the two best performing machine learning models by using the SHAP algorithm.

5 Discussion

This chapter discusses the results as they have been described in Chapter 4 of this thesis. First of all, the research questions will be answered. Then there is a justification regarding the validity of these results. In the following sections the technical limitations, research limitations, managerial implications, societal implications and academic implications are discussed. In these sections a further analysis is presented of the limitations of the thesis and the consequences that can be experienced per subject.

5.1 Research questions

The first research question, “In what way is bias and discrimination currently present during the recruitment process and how does this manifest itself in the context of job descriptions?”, involves the way bias and discrimination currently occurs during the recruitment process.

A study conducted by [19] shows that bias can occur both implicitly and explicitly, which is also the case during the recruitment process. The researchers also conclude that the most common types of bias and discrimination that occur during the recruitment process are: Gender, ethnic and age bias. In addition, studies conducted by [3] and [7] have shown that age and ethnic background also have a significant role during the recruitment process.

Measures to prevent bias and discrimination during the recruitment process have also been described to investigate whether similar applications have already been developed. A study by [3] presents an approach including the development of a prototype tool that can be used to identify and measure gender bias in IT job postings. Compared to the tool developed in this research, the tool described by [3] can process job advertisements directly from the internet, in the current research this is not yet possible and the development has taken place based on the provided job advertisements by the EMSCAD dataset. Also, in the current research it is not possible to obtain alternative suggested language when biased or discriminatory language has been identified by the system, this is yet possible with the prototype as described in the research of [3].

Given these findings and their likeliness of applicability within the EMSCAD dataset, it was decided to focus on five categories within this study, namely: Masculine-coded language, Feminine-coded language, Exclusive language, Demographic and Racial-coded language, and finally LGBTQ-colored language.

From these results presented in chapter 4 of this thesis, it was also possible to answer the research question "Which state-of-the-art natural language processing technologies can best be employed to achieve the optimal model to automatically identify and classify possible bias indicators in job descriptions?".

Based on the results, we can state that the application of NER Model 4, as shown in figure 4, achieves the best performance compared to the other NER models tested with a precision of 99,84, a recall of 99,78 and a finally a F1 score of 99,81.

Regarding the semantic text classification models, we can conclude that the Random Forest classifier, using FastText word embeddings, using 10-fold cross-validated, performs best. The model achieved with a precision of 0,9976, a recall of 0,9900 and finally a F1 score of 0,9936. These results were presented in figure 9.

Comparing these two aforementioned results, we can conclude that the used configuration for the NER model 4 using lemmatized words, no word vectors during pre-processing of the dataset and a large word vector containing 500k vectors during model training, leads to the best performing model.

Attempts to improve the performance by resampling the dataset and thereby creating more balance between the categories indicate that no significant improvements are evident. In fact, compared to the intitial dataset, resampling the dataset led to a decrease in performance.

Using dataset Balanced_01 to train the Random Forest classifier using BERT word embeddings led to the best performing model. The model achieved a precision score of 0,99998, a recall of 0,92991 and a F1 score of 0,9566.

5.2 Justification of findings

Given that the results were based on the real-life EMSCAD dataset, no complications were expected in this matter. The dataset consisted of a subset of legitimate and fraudulent job advertisements, only the legitimate advertisements were used.

The applicability of the results obtained through this research can thus be used to develop similar applications in the field of recruitment, in particular during the attraction phase to analyse job advertisements.

Results can also be applied to other domains, although it would be necessary to consider in detail how representative the results would be in these focused domains. Of course, it is possible to implement a similar approach whereby processes such as investigating which types of bias and discrimination occur in the domain in question, data annotation based on these findings, and model training and evaluation must be carried out again.

5.3 Technical limitations

Given the fact that the various machine learning models were trained using extensive feature sets, model training and evaluation was conducted using a sample of 3000 job advertisements. Despite this being a representative sample, it is still possible that tokens that are underrepresented in the overall dataset have not been incorporated into model training and evaluation. This led to a situation in which these tokens are even more underrepresented, which complicates classification.

5.4 Research limitations

On the subject of bias and discrimination, plenty of literature can be retrieved, however, the connection to the recruitment process is often lacking. In addition, it appears that the development of a comparable tools as described and developed and evaluated in this study, to the best of our knowledge, has only been described once in previous studies. This leads to a situation in which it was challenging to compare and in addition learn from results, limitations and implications that other researchers have faced in previous studies.

In addition, anonymization on the dataset used took place within this study. This led to the fact that data that is affected by the process of anonymization, such as email addresses, were not included in the research results. It is possible that the machine learning models developed were impacted by this. It is therefore important that this study will be tested again without using anonymization before any implementation will take place.

5.5 Managerial implications

Almost every company in the world has the need to recruit personnel to be able to execute and deliver its services to clients and customers. Due to this need, it is almost always necessary to post a job advertisement to which potential future employees can apply before a subsequent job interview between employer and applicant will take place.

It is precisely in this phase of the recruitment process that the results obtained from this study can make a valuable contribution. Many companies are often unaware that conscious or unconscious bias and discrimination occurs during their processes. A possible cause of this could be that too few insights are generated in this area.

The tool that was designed, developed and tested in this research provides the insight on an operational level. As a result, recruiters in companies where a similar application is implemented are aware of the occurrence of biased- or discriminatory language in job descriptions. Based on this, a recruiter can adjust the written description so that it is more inclusive to the overall society.

Diversity in recruiting people may increase a company's productivity and performance because different perspectives will be involved in solving problems, improving services, and so on.

5.6 Societal implications

Conducting this research will make the adoption of such an application, designed and developed in this research, easier and more accessible. Not only because of the technical approach and results, but also because of the overall execution of the research. For example, the conducted systematic literature review and the obstacles described in this research that were uncovered during the implementation of the research.

Additionally, many social groups or individuals will experience positive influence when such a tool is implemented. This will be due to the fact that it is then possible for recruiters to check job descriptions for the presence of biased or discriminatory language. When they do so, it is likely that they will adapt the text to be more inclusive and specific social groups within society will be less likely to feel uninvolved during the attraction phase of the recruitment process.

5.7 Academic implications

Based on this conducted research, other researchers can take various benefits from the findings. First of all, many different technologies and methods were applied in this research.

For the semantic text classification models, a very broad application has taken place. This is due to the fact that different types of word embeddings have been used, each in combination with 7 different machine learning classifiers.

Researchers such as [3] can take advantage of these results by applying findings in this thesis to their applications and potentially improve performance.

5.8 Summary

This chapter analyzed the results and concluded sub-conclusions based on them. An attempt has also been made to account for the class imbalance in the dataset. Results show that there is no difference between the results obtained from the initial dataset and the three different balanced datasets.

It was found that NER Model 4 has the best performance with an accuracy of 99,84, a recall of 99,78, and an F1 score of 99,81. This model does not use word vectors during the pre-processing of the training and evaluation data. In addition, the model does use a large word vector model during the training of the NER model and a token lemmatizer was used.

For the semantic text classification models, we can state that the combination of using a Random Forest classifier and BERT word embedding leads to the best performing model, measured from test data. The measured performance scores are: an accuracy of 0.99999, a recall of 0.92992 and an F1 score of 0.95660.

In addition, a 10-fold cross validation also took place which showed that the combination of the Random Forest classifier and FastText word embedding led to the best performing model. This model was evaluated using the 10-fold cross validation based on data that was in the train set. The model was tested with an accuracy score of 0.9976, a recall of 0.9900 and an F1 score of 0.9936.

The following chapter describes the conclusion of this thesis and furthermore discusses recommendations and future work.

6 Conclusion and Future work

This thesis presented a machine learning approach to identify bias and discriminatory language in job advertisements by utilizing state-of-the-art natural language processing techniques. Based on the development of different named entity recognition, and semantic text classification models, we investigated which approach achieves the best performing model to identify five different types of biased and discriminatory language.

6.1 Named entity recognition

By testing various configurations it was possible to discover which configuration resulted in the best performing model. The results indicate that the best performing model, NER Model 4, was tested with a precision of 99,84, a recall of 99,78 and a finally a F1 score of 99,81. The model is trained with lemmatized tokens, no word vectors during pre-processing but does use large word vectors while training the model.

6.2 Semantic text classification

Various popular state-of-the-art machine learning classifiers using linguistic and semantic textual features have been utilized to identify five different types of biased and discriminatory language. Based on the results obtained, we can conclude that the 10-fold cross-validated random forest classifier with FastText word embeddings achieved the best performance, testing a precision of 0,9976, a recall of 0,9900 and finally a F1 score of 0,9936 using training data. In addition, we conclude that the Random Forest classifier with BERT word embeddings, yields the best performing semantic text classification model on test data, with an accuracy of 0,9992, a precision of 0,99999, a recall of 0,92992 and finally a F1-score of 0.95660.

6.3 Recommendations and Future work

Given the limitations that applied in this study, it is first of all interesting to conduct the same study using the entire EMSCAD dataset. This might lead to different results referring to the classes that were underrepresented. Furthermore, it is of interest to include other machine learning classifiers in addition to the ones selected in this research.

Secondly, the models trained in this thesis can be made available to recruiters by publishing them on the internet using a graphical user interface in which users can insert a written job description after which the biased- and discriminatory language will be marked. This way it is easily accessible for everyone and adoption will be encouraged. This graphical representation can later be improved by offering suggestions for alternative language based on the detected biased- or discriminatory language in a inserted job description. These alternatives can be provided as was done in [3], for example using a keyword repository that contains gender-unmarked words and phrases for each biased- or discriminatory term, that can be used to replace gender biased words without changing the meaning of a sentence.

Third, in this thesis, word embeddings were determined per target token without considering the context of the sentence. It is of value to check how the results are affected when context is considered, this application can be achieved by using the word embeddings Flair, BERT and ELMo.

Fourth, for the implementation of the models trained in this thesis, it is of positive impact if the number of categories will be increased. This allows for even more control over the writing style applied by recruiters and can potentially contribute to even more inclusiveness. Recruiters can be made aware of terms that are biased is discriminated. However, as demonstrated in a study by [13], the fact exists that when processes are automated, there is less chance of mistakes in a process. This brings us to the relevance of a vision like this.

The results found in this dissertation can be applied both within organizations and in further research. For example, the performance measures testing the different configurations, can contribute to future research by providing insights into which configuration is likely to work best in a specific situation. Although it is important to note that the results obtained are representative of situations where the English language is used within the recruitment process. For example, studies such as [3] can use these results to improve the models by expanding the classes and using alternative techniques.

Another perspective in contrast to the previous one mentioned is that it is in fact interesting for future studies to look into other languages and how they influence the performance in such applications. This will provide verification whether recognizing bias and discrimination by using the same techniques and method applied in this study, lead to the same or different results.

In addition, the various named entity recognition models were all developed using SpaCy's NLP library. It would be interesting to investigate whether using a different technology would lead to the same or different results.

In addition, it will be interesting to see how the NER models perform when more categories are added. Finally, the method used in this thesis can serve as an inspiration for other organizations in which they might improve the models proposed in this thesis or their own models which are already operational within their organization, allowing recruiters to react more effectively to favorable language styles.

7 Reference list

[1] Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. Journal of Applied Psychology, 100(1), 128–161. https://doi.org/10.1037/a0036734

[2] Amin, S., Jayakar, N., Kiruthika, M., & Gurjar, A. (2020). Best fit resume predictor. International Journal of Engineering and Technology. Published. Retrieved from https://www.researchgate.net/publication/341537308_Best_Fit_Resume_Predictor

[3] Böhm, S., Linnyk, O., Kohl, J., Weber, T., Teetz, I., Bandurka, K., & Kersting, M. (2020). Analysing Gender Bias in IT Job Postings. Proceedings of the 2020 on Computers and People Research Conference. Published. https://doi.org/10.1145/3378539.3393862

[4] McKenzie Raub, Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices, 71 Ark. L. Rev. 529 (2018)

[5] Derous, E., & Decoster, J. (2017). Implicit Age Cues in Resumes: Subtle Effects on Hiring Discrimination. Frontiers in Psychology, 8. Published. https://doi.org/10.3389/fpsyg.2017.01321

[6] Derous, E., Pepermans, R., & Ryan, A. M. (2016). Ethnic discrimination during résumé screening: Interactive effects of applicants' ethnic salience with job context. Human Relations, 70(7), 860–882. https://doi.org/10.1177/0018726716676537

[7] Derous, E., & Ryan, A. M. (2018). When your resume is (not) turning you down: Modelling ethnic bias in resume screening. Human Resource Management Journal, 29(2), 113–130. https://doi.org/10.1111/1748-8583.12217

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. Published. https://doi.org/10.18653/v1/n19-1423

[9] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Future Internet, 9(1), 6. https://doi.org/10.3390/fi9010006

[10] Erkmen, C. P., Kane, L., & Cooke, D. T. (2021). Bias Mitigation in Cardiothoracic Recruitment. The Annals of Thoracic Surgery, 111(1), 12–15. https://doi.org/10.1016/j.athoracsur.2020.07.005

[11] Guo, S., Alamudun, F., & Hammond, T. (2016). RésuMatcher: A personalized résumé-job matching system. Expert Systems with Applications, 60, 169–182. https://doi.org/10.1016/j.eswa.2016.04.013

[12] Marc Bendick, Jr Marc Bendick, Jr and Ana P. Nunes: Developing the Research Basis for Controlling Bias in Hiring; Journal of Social Issues, Vol. 68, No. 2, 2012, pp. 23 —262

[13] Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

[14] Maheshwary, S., & Misra, H. (2018). Matching Resumes to Jobs via Deep Siamese Network. Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. Published. https://doi.org/10.1145/3184558.3186942

[15] McShane, S. L. (1990). Two tests of direct gender bias in job evaluation ratings. Journal of Occupational Psychology, 63(2), 129–140. https://doi.org/10.1111/j.2044-8325.1990.tb00515.x

[16] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Efficient Estimation of Word Representations in Vector Space. Published. Retrieved from https://arxiv.org/abs/1301.3781

[17] Muralidhar, S., Nguyen, L., & Gatica-Perez, D. (2018). Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes. Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Published. https://doi.org/10.18653/v1/w18-6247

[18] Oates, M. E. (2018). Optimizing Recruitment Into Radiology: Some Simple Approaches to Controlling Bias. Journal of the American College of Radiology, 15(4), 684–686. https://doi.org/10.1016/j.jacr.2017.12.002

[19] ÖSterlund, P. (2020). Preventing discrimination in recruiting through unconscious biases. Preventing Discrimination in Recruiting through Unconscious Biases. Published. Retrieved from https://www.theseus.fi/bitstream/handle/10024/346242/Osterlund_Pia.pdf?sequence=2

[20] Sanyal, S., Hazra, S., Ghosh, N., & Adhikary, S. (2017). Resume parser with natural language processing. Resume Parser with Natural Language Processing. Published. https://doi.org/10.13140/RG.2.2.11709.05607

[21] Tang, S., Zhang, X., Cryan, J., Metzger, M. J., Zheng, H., & Zhao, B. Y. (2017). Gender Bias in the Job Market. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–19. https://doi.org/10.1145/3134734

[22] Gravett WH "The Myth of Objectivity: Implicit Racial Bias and the Law (Part 1)" PER / PELJ 2017(20) - DOI http://dx.doi.org/10.17159/1727-3781/2017/v20i0a1312

[23] Bertrand M and Mullainathan S (2003) Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. National Bureau of Economic Research Cambridge MA 2003.

[24] Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. Journal of Personality and Social Psychology, 101(1), 109–128. https://doi.org/10.1037/a0022530

[25] ROBERT WALTERS. (2021). DIVERSITY AND INCLUSION IN RECRUITMENT. Retrieved from https://www.robertwalters.co.uk/content/dam/robert-walters/country/united-kingdom/files/whitepapers/Diversity-In-Recruitment-Whitepaper-web.pdf

[26] Akhil Alfons Kodiyan: An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool

[27] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[28] Lundberg, S. M., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30. Published. Retrieved from https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[29] Disability Language Style Guide. (n.d.). Retrieved March 13, 2021, from https://ncdj.org/style-guide/

[30] glaad. (2016, October). GLAAD MEDIA REFERENCE GUIDE (10). Retrieved from https://www.glaad.org/sites/default/files/GLAAD-Media-Reference-Guide-Tenth-Edition.pdf

[31] Inclusive Language Guide. (2020, September 15). Retrieved March 13, 2021, from https://nasaa-arts.org/nasaa_research/inclusive-language-guide/

[32] Inclusive Language Guide — SUNY Geneseo. (2021). Retrieved March 13, 2021, from

https://www.geneseo.edu/comm_mark/inclusive-language-guide

[33] Kanigel, R. (2021). The Diversity Style Guide. Retrieved March 13, 2021, from https://www.diversitystyleguide.com/

[34] RMIT University. (2021). Guide to Inclusive Language. Retrieved from https://www.rmit.edu.au/content/dam/rmit/documents/Students/Support_and_Facilities/dgss/guide-to-inclusive-language.pdf

[35] WGBH. (2019). Inclusive Language Guidelines. Retrieved from https://wgbh.brightspotcdn.com/ff/53/bef446844efebdc3c212a4df8083/wgbh-inclusive-language-guidelines.pdf