Master Thesis

Erasmus School of Economics

MSc Data Science and Marketing Analytics

Beyond the gender bias

Identifying gendered language in job advertisements and predicting organisational diversity outcomes using embedding methods, POS tagging and predictive modelling

K.A. (Kars-Jan) Giesen

445739

Supervisor: dr. S.L. Malek

Second assessor: prof. dr. P.H.B.F. Franses

Company supervisor: Jochem Dogger

Date final version: 2 August, 2021

Abstract

Women remain largely underrepresented throughout US and EU companies, while gender diversity has multiple beneficial outcomes, including increased innovation, higher creativity levels, improved firm performance, and hereby competitive advantage. This research focuses on empirically validating findings of existing experimental research that suggest that a negative relationship exists between female job attraction and wording in job advertisements that includes male stereotypical characteristics. Furthermore, based on word embedding models GloVe, Word2Vec and LSA, this study proposes additional words to be added to existing masculine and feminine word lists in two languages. Additionally, by means of POS tagging an analysis is provided of these gendered words' grammatical function since phrasing in verbs, instead of nouns or adjectives, has been theorised to influence female applicant rates positively. In predictive analyses on real-world data, in this study no effect has been found of gendered words and grammatical forms on female applicant rates as opposed to the findings of existing experimental research. Given the potential impact of this relationship for firms, further empirical research is needed to (in)validate these results.

Keywords: *gender inequality, diversity, gendered wording, job advertisements, NLP*

Table of Contents

1. Introduction

Throughout all levels in both US and EU companies, women remain underrepresented (Heilman & Caleo, 2018; McKinsey, 2019). This underrepresentation is more prevalent within management positions and traditionally male-dominated industries (Gaucher, Friesen, & Kay, 2011; Heilman, 2012; Heilman & Caleo, 2018). While gender diversity is essential from an ethical perspective on the need for representative equity in society, several other organisational benefits result from a gender diverse workforce (Campbell & Mínguez-Vera, 2008; De Cabo, Gimeno, & Nieto, 2012). Among other, a gender diverse team positively impacts several workplace processes, such as creativity, innovation, problem-solving and decision-making, which in turn have positive consequences for team performance (Galinsky et al., 2015). In marketing teams in which such processes are vital, this results in innovative, creative concepts and targeted initiatives (Pless & Maak, 2004; Turban et al., 2019). Also, through an improved positive corporate reputation and image, and through improved representation of potential employees, gender diverse firms can attract and retain talent from diverse groups, making gender diversity efforts a self-reinforcing mechanism (Bear, Rahman, & Post, 2010; Bond & Haynes, 2014; Campbell & Mínguez-Vera, 2008). Additionally, through wider representation of customers, customer relationships improve, as well as financial results of marketing campaigns through incorporating enhanced emotional intelligence of diverse audiences (Herring, 2009). An example of improved corporate gender diversity along with marketing campaigns focussed on diverse audiences can be found at Proctor and Gamble (P&G). Over the past years, their marketing campaigns empowered individuals that contest bias and inequality, and emphasised the rights of mothers, while the company also substantively improved (gender) diversity levels throughout their workforce (Brownfield, 2020; Shadrach, 2021).

Furthermore, improved corporate reputation resulting from gender diversity has benefits for investor relationships through signalling competent management (Dezsö & Ross, 2012; Herring, 2009; Turban, Wu, & Zhang, 2019). More specifically, for listed companies a higher share of women through all levels appears to be associated with stronger share-price performance, Goldman Sachs reported in the Financial Times when comparing the stock price development for European listed companies (Bell, 2019). Additionally, public statements of improved gender diversity levels

led to same day increases of stock prices for listed companies (Daniels, Dannals, & Neale, 2021). When comparing US tech companies with Google, the widely considered industry leader, investor reactions were even stronger when these firms outperformed Google in terms of gender diversity levels (Daniels et al., 2021). This emphasises that gender diversity can be considered a key area to gain competitive advantage and that improving gender diversity levels at the industry pace should be considered as a bare minimum. While attaining gender diversity is an essential practice, female underrepresentation and barriers to advancement remain prevalent partly by organisations being more likely to attract individuals that are similar to existing personnel.

To a large extent this underrepresentation is maintained by gender-based stereotypes and bias that exist throughout job advertisements, resulting in negative consequences for female job attraction (Gaucher et al., 2011; Heilman, 2012). Stereotypic defining characteristics of women are communal and warm, while for men these are agentic and competitive (Cuddy, Fiske, & Click, 2008; Heilman, 2012). In job advertisements stereotypical male characteristics are expressed through masculine-themed words and phrasing, such as 'leader', 'competitive' and 'dominant' (Gaucher et al., 2011). Women have been found to be less likely to apply to job advertisements that contained masculine-coded language as it resulted in lower anticipated feelings of belongingness and job interest, while not influencing perceptions of their skills (Gaucher et al., 2011). Additionally, when male characteristics were emphasised in job advertisements, women felt more attracted to jobs if these characteristics were expressed as behaviours (verbs), indicating a mediating effect of grammatical word class (Born & Taris, 2010). Hence, using feminine language and verbs when describing gender stereotypical characteristics in job advertisements can improve gender diversity within organisations for all positions and can lead to fairer hiring practices (Born & Taris, 2010; Gaucher et al., 2011). Herein, creating gender fair job advertisements is an important stage in the self-reinforcing process of reducing gender bias and improving organisational gender diversity as it is an essential bottom-up approach needed for further diversity efforts.

Existing research has expanded upon the findings of Born and Taris (2010) and Gaucher et al. (2011) from the (social) psychology or linguistic perspective. In contrast, this study focuses on using unsupervised and supervised machine learning (ML) methods to quantify and investigate the relationship between gendered language and female application rates. In order to classify job

advertisements into feminine and masculine language on the basis of words and grammatical class, and to draw subsequent conclusions on job applicants gender diversity rates, the following research question is proposed:

How can ML techniques be used to identify gendered text in job advertisements and to predict its effects on gender diversity of applicants?

This study contributes to existing literature on gendered language in several ways. Firstly, it adds to the literature on gendered language and psychology by using word embedding methods to complement and fortify existing gendered wording dictionaries by providing a data-driven approach. Secondly, it contributes to literature on dictionaries by using the more specific context of job advertisements that is naturally different from many other sources of text. Thirdly, the focus of this study is on the relatively gender diverse industry of consultancy, while current research focuses on industries in which there is gender overrepresentation. Fourthly, it adds to quantitative literature by applying more complex methods for classification of job advertisements into feminine and masculine beyond dictionaries, and by using features derived from previous steps to predict gender diversity of job applicants. Lastly, while existing literature uses experimental methods to measure the response likelihood of women and men relative to different degrees of gendered text in job advertisements, this study uses predictive methods based on real-world data to investigate whether the hypothesized effects exist in a business context.

To examine the degree to which gendered text is present in job advertisements, this study firstly provides a theoretical framework on gendered text and appropriate methods, after which explorative and predictive methodologies for text data are set out in more detail. Hereafter, explorative analyses tools are conducted using Natural Language Processing (NLP) tools to discover the degree of gendered text in consultancy job advertisements from Indeed.com. Lastly, predictive analyses are used to predict gender diversity of applicants by features related to gendered language in a real business context.

2. Literature Review

This literature review consists of both a substantive and a methodological part. Regarding the substantive part, firstly, the effects of gender diversity are set out from both an ethical and an economic perspective after which the underlying sources of persistent gender inequality are discussed. Hereafter, HR practices are considered as these can be used as an indicator of and a potential solution for widespread gender inequality within firms. Within these practices, a focus is put on job advertisements within the HR process of recruitment, because it appears to be an important source for attracting gender diverse job applicants. Regarding the methodological part, existing descriptive methods are set out, which are used to detect gendered text and features related to gendered text. Furthermore, various predictive methods are reviewed that have been or can be applied in the context of gendered text.

2.1 Substantive literature review

2.1.1 Benefits of gender diversity

In the broadest sense, diversity refers to all types of differences in individual characteristics, such as race, gender, age, ethnicity, sexual orientation, physical ability, religion, and national origin (Herring, 2009). Consequently, diversity policies and practices refer to all actions and strategies aimed at creating a culture of inclusion for people with various individual characteristics that are to some extent different from traditional members, resulting in using talents of all potential members (Herring, 2009). There is a growing consensus that implementing diversity policies and practices results in organisational benefits from both an ethical as well as an economic perspective that are, in turn, related (Annabi & Lebovitz, 2018; Bear et al., 2010; Campbell & Mínguez-Vera, 2008; De Cabo, Gimeno, & Nieto, 2012; Heilman, 2012; Herring, 2009; Galinsky et al., 2015; Turban et al., 2019).

From an ethical perspective, creating an inclusive culture with equal opportunities for people with various individual characteristics empowers the non-dominant groups to advance on a societal level (Annabi & Lebovitz, 2018; De Cabo et al., 2012). Furthermore, it is argued that excluding non-dominant groups both indirectly and directly is immoral and by improving their status it enables individuals from such groups to exercise their human rights (Campbell &

Mínguez-Vera, 2008). From this perspective, equitable representation resulting from organisational diversity policies and practices should be regarded as a goal in itself as it is related to corporate social responsibility (CSR) (Annabi & Lebovitz, 2018; Bear et al., 2010; Campbell & Mínguez-Vera, 2008; De Cabo et al., 2012). However, its effects reach beyond ethical benefits as spill-over effects to corporate reputation and corporate diversity itself are generally regarded to result in multiple economic benefits as well (Bear et al., 2010; Campbell & Mínguez-Vera, 2008).

While ethical benefits hold for diversity regarding all types of individual characteristics, existing research focuses separately on potential economic benefits that result from diversity in general, as well as diversity regarding specific types of demographics, such as race and gender. Furthermore, existing research distinguishes between the economic effects of diversity for all levels in the company and specific levels in the company, which concerns mainly the board level. Although diversity concerns a wide variety of individual characteristics, the focus of this research is on diversity regarding gender and the effects of gendered language on workforce gender composition in general. Hence, the economic effects of or the 'business case' for gender diversity as well as diversity in general are set out hereafter and illustrated in Table 1.

In general, diversity positively impacts 1) several team and workplace processes that in turn influence larger business processes, 2) HR practices, 3) customer relations, 4) external relations with investors, and potentially 5) firm performance and value (Annabi & Lebovitz, 2018; Bear et al., 2010; Campbell & Mínguez-Vera, 2008; De Cabo et al., 2012; Dezsö & Ross, 2012; Herring, 2009; Galinsky et al., 2015; Pless & Maak, 2004; Turban et al., 2019).

Firstly, a diverse team tends to perform better compared to a homogeneous team as a variety of backgrounds, qualities, experiences and perspectives is combined (Herring, 2009; Galinsky et al., 2015). This variety of individual characteristics leads to consideration of a wider range of ideas and better alternatives, while preventing narrowmindedness (Pless & Maak, 2004). These broader incorporated perspectives in turn benefit creativity, innovation, problem-solving, decision-making and, ultimately, team performance and quality of work (Galinsky et al., 2015). However, as Turban et al. (2019) point out, a safe environment and culture of openness to different perspectives is essential for firms to benefit optimally from diverse teams.

Secondly, firm diversity has several benefits to HR practices of attracting and retaining talent from both dominant and non-dominant group members by having an attractive work

environment, by having a positive corporate reputation and image, and by representing a wide range of potential employees (Bear et al., 2010; Campbell & Mínguez-Vera, 2008). Representation of a diverse range of individuals leads to attraction and retention of diverse talent as people have strong in-group preferences (Herring, 2009). Hence, it is important to recognize that HR diversity policies and practices reinforce firm diversity and its positive effects (Turban et al., 2019).

Thirdly, employee diversity can potentially lead to positive effects for customer relationships, especially when the firm operates near the consumers of the product or service offered, and there is substantial customer-worker interaction, such as in banking, media, and retail (Campbell & Mínguez-Vera, 2008; Herring, 2009). Contrarywise, firms in industries that do not maintain direct relationships with final consumers, such as resources and engineering, are traditionally less gender diverse and benefit less from gender diversity with respect to customer relationships (Campbell & Mínguez-Vera, 2008; Herring, 2009). Additionally, the benefits of diversity regarding customer relationships seem to occur particularly within service industries as it improves serving the needs of a broader group of customers (Annabi & Lebovitz, 2018). These benefits of diverse employees are reinforced by improved corporate reputation (Bear et al., 2010; Campbell & Mínguez-Vera, 2008).

Fourthly, employee diversity results in positive effects on investor relationships in two ways. On a general level, a (gender) diverse workforce signals to investors that management is competent in employing diversity policies, which is increasingly recognized by investors as a driver of value and a 'best practice' for firm success (Dezsö & Ross, 2012; Turban et al., 2019). While this relationship has empirically been established in the US by positive effects of diversity awards and public statements on company stock prices, the institutional context influences whether it is considered a 'best practice', which is increasingly true (Daniels et al., 2021; Turban et al., 2019). On a board level, gender diversity results in improved monitoring of management through increased independence, which is beneficial to investors as it improves protection of their interest (De Cabo et al., 2012).

Lastly, while the aforementioned effects of having (gender) diverse employees are widely recognized throughout literature and are generally regarded as moderately positive or leading to competitive advantage, in existing literature the effect of (gender) diversity on firm value and performance is profoundly disputed. Some authors that use various types of data find positive

effects of board level gender diversity on firm performance and value relative to a homogenous board, some with limitations of this effect to an innovation focused firm strategy or a dynamic environment (Campbell & Mínguez-Vera, 2008; Dezsö & Ross, 2012; Herring, 2009; Nguyen, Locke, & Reddy, 2015). However, other authors find no (clear) effects of board gender diversity on firm performance and value, and associated metrics such as revenues or profits (Chapple & Humphrey, 2014; Marinova Plantenga, & Remery, 2016). This ambiguous evidence on the direct relationship can potentially be explained by the presence of many other influencing factors on firm performance and value, making it difficult to establish a direct causal relationship (Turban et al., 2019).

This ambiguity is a recurring theme in literature on the effects of gender diversity on economic gains. Also, the effects appear to be increasingly ambiguous when the relationship with (gender) diversity becomes more indirect, which is, for example, the case for firm performance, but less so for increased problem-solving abilities in teams due to a larger variety of perspectives. As Zhang (2020) points out in a cross-industry and cross-country analysis on this relationship, the magnitude of positive diversity effects seems to depend on the institutional context and, in particular, on the normative acceptance and expectance of diversity in a national context. Additionally, by creating an inclusive firm culture and by emphasising the benefits of diversity to employees, firms can optimally reap the benefits of a diverse workforce relative to a homogeneous one (Galinsky et al., 2015). However, as gender diversity comes with some benefits regardless of the institutional context, and some benefits that occur especially in contexts and countries in which gender diversity is regarded as important, such as the Netherlands, managing gender diversity and diversity in general appropriately should be a priority for businesses and HR departments specifically.

Table 1. Effects of gender diversity and diversity in general

Effects	**Authors**
Benefits for team and workplace processes	
Creativity	Campbell & Mínguez-Vera, 2008; Herring, 2009; Pless & Maak, 2004; Turban, Wu, & Zhang, 2019;
Innovation	Annabi & Lebovitz, 2018; Campbell & Mínguez-Vera, 2008; Dezsö & Ross, 2012; Galinsky et al., 2015; Herring, 2009; Pless & Maak, 2004; Turban, Wu, & Zhang, 2019
Problem-solving abilities	Campbell & Mínguez-Vera, 2008; De Cabo, Gimeno, & Nieto, 2012; Galinsky et al., 2015; Herring, 2009; Pless & Maak, 2004
Decision-making	De Cabo, Gimeno, & Nieto, 2012; Galinsky et al., 2015
Quality of team work	Annabi & Lebovitz, 2018; Herring, 2009
Benefits for HR practices and customer relations	
Enhanced representation of customers	Annabi & Lebovitz, 2018; Campbell & Mínguez-Vera, 2008; Herring, 2009; Pless & Maak, 2004
Enhanced representation of job candidates	Campbell & Mínguez-Vera, 2008; Herring, 2009
Signaling an attractive work environment	Annabi & Lebovitz, 2018; Bear, Rahman, & Post, 2010; Turban, Wu, & Zhang, 2019
Access to all possible talent	Annabi & Lebovitz, 2018; Bear, Rahman, & Post, 2010; Campbell & Mínguez-Vera, 2008; De Cabo, Gimeno, & Nieto, 2012

Positive reputation and image	Campbell & Mínguez-Vera, 2008; Bear, Rahman, & Post, 2010
Improvements of investor relations	
Signaling competent management to investors	Dezsö & Ross, 2012; Turban, Wu, & Zhang, 2019
Improving monitoring and controlling of management	De Cabo, Gimeno, & Nieto, 2012
Ambiguous effects on firm performance and value	
Positive effects on firm performance	Campbell & Mínguez-Vera, 2008; Daniels, Dannals, & Neale, 2021; Dezsö & Ross, 2012; Herring, 2009; Nguyen, Locke, & Reddy, 2015
Negative effects on firm performance	Chapple & Humphrey, 2014; Marinova, Plantenga, & Remery, 2016

2.1.2 Obstacles to gender diversity

2.1.2.1 Underlying mechanisms

Theories for persistence of gender inequality are described in sociological and social psychological literature, and many theories suggest that gender bias results from gender stereotyping as a underlying cause. Gender stereotypes are generalisations about individuals based on preconceptions about traits and characteristics of men and women (Burgess & Borgida, 1999; Heilman, 2012). These generalisations entail characteristics that are conceived to be more related to men or women and upon which both are subsequently expected to behave (Heilman, 2012). For women this defining characteristic is communality, while for men this is agency, and both characteristics are conceived to be lacking in the opposite gender (Cuddy et al., 2008; Diekman & Eagly, 2000; Gaucher et al., 2011; Heilman, 2012). On one hand, communality is associated with selflessness and concern for others, including related traits, such as kindness, understanding, warmth and respectfulness (Cuddy et al., 2008; Gaucher et al., 2011; Heilman, 2012). Agency, on the other hand, is associated with assertiveness, independency, self-confidence, and competitiveness (Cuddy et al., 2008; Gaucher et al., 2011; Heilman, 2012).

Both descriptive and prescriptive gender stereotypes result from these generalisations and have negative effects on workplace advancement of women (Burgess & Borgida, 1999; Heilman, 2012). Descriptive gender stereotypes are beliefs about characteristics and behaviours of men and women (Burgess & Borgida, 1999; Heilman, 2012). These aid people to form quicker impressions about men and women, and, hereby, serve as unintentional shortcuts (Burgess & Borgida, 1999; Heilman, 2012). Descriptive gender stereotypes are widely shared throughout different cultures and their impact depends on the context (Heilman, 2012).

Prescriptive gender stereotypes are beliefs about characteristics and behaviours of men and women to which they should conform, and consist of expectations how men and women should behave (Burgess & Borgida, 1999; Heilman, 2012). This stereotype can lead to negative bias when individuals demonstrate competence in attributes that are expected to belong to the opposite gender (Bond & Haynes, 2014; Heilman, 2012). For example, men that are competent in feminine roles are often disapproved, while women are regarded as less socially appealing when showing masculine behaviour, such as self-promotion (Bond & Haynes, 2014; Heilman, 2012).

One of the sources of workplace gender inequality is driven by descriptive gender stereotypes that cause a perceived lack of fit. This lack of fit occurs when stereotypical male characteristics associated with agentic traits are regarded as essential for high-level roles or occupations in male-dominated industries (Heilman, 2012). In such occupations, traits such as competitiveness and self-assertiveness are regarded as necessary to be successful at the job, while undervaluing the importance of female traits (Heilman, 2012).

Similarly, social dominance theory explains persistent gender inequality in the workplace and throughout society (Gaucher et al., 2011). In the light of this theory, institutional-level mechanisms reinforce existing societal inequalities and hierarchies (Gaucher et al., 2011). According to this theory, descriptive gender stereotypes by which men are associated with attributes such as leadership and competitiveness, naturally lead to hierarchy in society at large and at lower levels, such as firms (Gaucher et al., 2011). The lack of fit model could be regarded as fitting within this theory, as it is one of the reinforcing mechanisms of gender inequalities at a smaller scale.

Lastly, similarity to other individuals is one of the sources of gender inequality through preference for similar others and similarity attraction (Bond & Haynes, 2014; Born & Taris, 2010). In general, similarity attraction theory entails that individuals are attracted to environments and other individuals that are similar to them in terms of personal characteristics (Born & Taris, 2010). More specifically, firm attraction of job candidates is influenced by the degree to which the organisation and its members appear to be similar, which is influenced by the communication of the organisational identity (Born & Taris, 2010; Graves & Powel, 1995). Resulting from this, the attraction-selection-attrition (ASA) model implies that individuals dissimilar to other members in an organisation are less likely to be attracted as employees and when attracted and selected, they are more likely to leave the organisation (Bond & Haynes, 2014). For firms and firm levels that are dominated by men, the ASA model implies the existence of difficulties for females to enter and to remain.

Hence, gender inequalities persist on a societal and firm level through both descriptive and prescriptive gender stereotypes, a perceived lack of fit when male stereotypical characteristics are deemed necessary for a role, existing institutional mechanisms that reinforce social dominance of

men, and, lastly, similarity attraction that prevents women from being attracted to and to succeed in penetrating male-dominated areas.

2.1.2.2 The influence of HR practices

As the ASA model indicates, gender diversity is difficult to attain in a male-dominated organisation as various mechanisms reinforce the homogeneity of members. However, several HR practices actively prevent female advancement within organisations: recruitment, selection, evaluation and promotion (Born & Taris, 2010; Heilman, 2012). Regarding selection practices, male résumés with equal experience and qualifications tend to be judged more favourably than female résumés (Born & Taris, 2010; Heilman & Caleo, 2018). This can be substantially improved by hiding direct indicators of gender of applicants before and - if possible - during selection (Born & Taris, 2010). Furthermore, in evaluating and assessing promotion eligibility of female employees, the aforementioned perceived lack of fit between stereotypical male characteristics and female attributes tends to dominate (Bond & Haynes, 2012; Heilman & Caleo, 2018). This can be improved by training employees to detect and decrease (gender) bias in judgements (Heilman & Caleo, 2018). While such barriers can negatively influence female advancement in organisations, they can to a large extent be combatted by adequate HR interventions, such as decreasing exposure of assessors to information revealing gender, training, and education (Heilman & Caleo, 2018). However, having a sufficient gender diverse workforce can mainly be influenced by creating and attracting a gender diverse pool of applicants. Attracting gender diverse job candidates is an essential bottom-up practice, which forms the core of recruitment practices.

Recruitment includes an employer's actions with the purpose of attracting attention of potential candidates, signalling whether these candidates are eligible, and maintaining the interest of these candidates with the goal of them accepting a job offer (Breaugh, 2013). In existing literature, most attention has been given to the latter phases of recruitment, while disregarding the first phase that involves attracting eligible job candidates (Breaugh, 2013). However, it generally is in the interest of the organisation to create a rather large pool of competent job candidates as this increases the probability that the best candidate for the role is considered (Herring, 2009).

Additionally, in order to hire a job candidate that contributes to organisational heterogeneity and diversity, it is essential that sufficient diverse candidates are considered for the role (Bond & Heynes, 2014; Herring, 2009). Namely, when individuals in a pool of applicants are similarly competent and eligible for the role, but there is a disproportionate gender representation, it tends to result in hiring an individual from the over-represented gender (Johnson et al., 2016). Hence, ensuring a proportional gender diverse pool of applicants is essential to attain a gender diverse workforce.

2.1.2.3 The role of job advertisements

Job advertisements are one of the most important ways to attract job candidates and to subsequently create a gender diverse pool of applicants. In line with similarity-attraction theory and job attraction theory, its text poses a means to signal organisational culture and values which can enhance job and firm attraction to individuals when these values are in line with those of job candidates, hereby also increasing the likeliness of job application (Bond & Haynes, 2014; Born & Taris, 2010). It often contains a job title, a company description, a job description, required personal characteristics and a company contact person (Born & Taris, 2010). Additionally, companies can include a statement about equal opportunities, which appears to enhance attractiveness (Born & Taris, 2010). As stereotypic defining characteristics of men are frequently regarded as essential job characteristics for roles in higher levels or in male-dominated industries, job advertisements for such roles tend to emphasise these characteristics (Gaucher et al., 2011; Heilman, 2012).

Multiple studies have found that women are less likely to apply to job advertisements that contain masculine-coded language related through stereotypical masculine characteristics in words and phrases (Born & Taris, 2010; Gaucher et al., 2011). Gendered wording is based on stereotypic gender traits of men and women (Heilman, 2012). Throughout different cultures adjectives related to communality, such as supportive and committed, are perceived to be more associated to women, while adjectives related to agency, such as competitive and dominant, are perceived to be more associated with men (Heilman, 2012). In addition to words based on agentic and communal characteristics, to identify gendered language existing research uses a list of masculine and feminine words that concerns other masculine traits and feminine traits, which can be found in

Appendix A (Gaucher et al., 2011; Sczesny et al., 2016; Dinan et al., 2020). For women, in existing studies masculine wording resulted in both less interest in the job and lower feelings of expected belongingness to the job and organisation (Gaucher et al., 2011). However, this did not result from a perceived lack of skills needed for the job (Gaucher et al., 2011). Feminine-coded language, on the opposite, did not result in any negative effects on job attraction for men (Born & Taris, 2010; Gaucher et al., 2011).

Next to words associated with gender stereotypical characteristics, for women job attraction appears to be influenced by phrasing of personal characteristics in a job advertisement (Born & Taris, 2010). When personal characteristics are presented as traits, candidates will consider whether they fit or do not fit with the required characteristic, whereas when characteristics are presented as behaviours, candidates will assess the degree to which they are able and motivated to conduct the task (Born & Taris, 2010). A trait is generally described in the form of an adjective or a noun, such as 'decisiveness', while a behaviour is generally written as a verb, such as 'you can decide (on important issues)' (Born & Taris, 2010). While the wording itself as well as the form have no significant effects on job application tendencies of men, for women the inclination to apply is greater when feminine-coded language is used and when verbs are used when describing stereotypical masculine-coded language (Born & Taris, 2010; Gaucher et al., 2011).

Thus, in order to create job advertisements that are equally attractive to qualified men and women, and, consequently, to generate a gender diverse pool of job applicants, using gender-fair language in job advertisements is important (Breaugh, 2013; Gaucher et al., 2011; Johnson et al., 2016). However, thus far these results have mostly been supported in an experimental setting, and not in a real-business context as will be done in the current research (Born & Taris, 2010; Gaucher et al., 2011; Johnson et al., 2016).

2.2 Methodological literature review

Natural Language Processing (NLP) methods are a category from the wide range of machine learning (ML) techniques that use computers to process and analyse natural language, such as speech and text (Moreno & Redondo, 2016). A subcategory of NLP is text analytics or text mining, of which the focus is on analysing patterns and information from unstructured written text, such as e-mails, literature, or job advertisements (Moreno & Redondo, 2016). As the first goal of this research is to recognise gender bias in job advertisements, literature is revised on both dictionary-based methods, as well as methods to expand on existing dictionaries using semantic approaches, such as Latent Semantic Analysis (LSA) and Word2Vec models. Hereafter, several forms of the syntactic method of Part-Of-Speech tagging are reviewed that can be used to identify grammatical functions of words as the grammatical wording of job characteristics, i.e. using a verb or noun/adjective to describe a stereotypical male characteristic, has been found to impact job attraction for women. As the second goal of this research is to predict gender diversity rates of job applicants using textual features indicating gender bias, several methods are reviewed: (penalized) regression, Neural Nets and Random Forests. An overview of the few methodological studies that have focussed on this topic can be found in Table 2.

2.2.1 Dictionary-based methods

Dictionary-based methods can be used to gain insights in frequencies of words, and to quantify psychological constructs in documents or sentences related to these words (Pietraszkiewicz et al. 2019). In sentiment analysis, such dictionaries are widely used to classify whether a word in a document or sentence is positive or negative (Wilson, Wiebe, & Hoffmann, 2005). Similarly, to assess the degree of genderedness, existing word lists or dictionaries of masculine and feminine words have been used by several authors to assess whether a word or document is masculine or feminine, such as Gaucher et al. (2011) or by Dinan et al. (2020), which can be found in Appendix A. While Gaucher et al. (2011) use the percentage of masculine and feminine wording of the total number of words in a job advertisement, Dinan et al. (2020) classify a text masculine if it contains more masculine than feminine words, and vice versa. Additionally, if the number of masculine and gendered words are equal, including zero, it was labelled as neutral (Dinan et al., 2020). Whereas this approach is transparent, intuitive, and provides a solid basis

for gendered text analysis, it is limited to explicitly binary gendered words and fails to include words that are less explicitly gender biased (Dinan et al., 2020; Pietraszkiewicz et al. 2019). Also, these lists of gendered words contain mostly adjectives, while verbs appear to better capture gendered wordings related to the concepts of agency and communality (Pietraszkiewicz et al. 2019). Moreover, as dictionaries are based on subjectivity, they are generally presumed to be incomplete and to contain inconsistencies (Gladkova & Drozd, 2016).

2.2.2 Expanding on existing gendered-wording dictionaries

In order to expand upon the limited approach of classification using dictionaries of feminine and masculine words, several authors have made advancements in expanding this approach. By performing a multi-task classification on words regarding the person that speaks, is spoken about and spoken to and for word lists, high classification accuracy was reached by Dinan et al. (2020). However, this approach might be less suited to job advertisements, because the gender of the person that speaks, and that is spoken about and to, is less clear-cut and generally missing.

Another approach to expanding the gendered dictionary itself instead of the gender of a person that speaks, is spoken about and spoken to, was used by Pietraszkiewicz et al. (2019). In one of their studies, they scored advertisements on gendered wording using several dictionaries that capture agency and communality, and subsequently averaging those scores. In creating a dictionary for agency and communality specifically, the dictionary creation process of Pennebaker et al. (2015) was used. An important step in adding words to existing psychological dictionaries is to correlate frequently occurring words to dictionaries, in order to assess whether they should be included (Pennebaker et al., 2015; Pietraszkiewicz et al., 2019). Additionally, in assessing similarity of words in dictionaries Latent Semantic Analysis (LSA) has been employed, by which the semantic similarity of words and groups of words can be determined. LSA is a method originally theorised by Landauer, Foltz and Laham (1998) which uses the distance between vector representations of words or groups of words to attain a measure of semantic similarity. The vector representations in a multidimensional space are based on the natural textual context in which these (groups of) words are present and absent (Landauer, Foltz & Laham, 1998). While LSA performs relatively well in assessing similarity for smaller corpora, in learning semantic similarities between individual words for large databases the methods Word2Vec or GloVe tend to be more

efficient compared to LSA (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013a; Mikolov, Sutskever, Chen, & Corrado, 2013b; Pennington, Socher, & Manning, 2014). Both these methods are unsupervised deep learning algorithms for text vectorization or word embedding that construct word vector representations from term co-occurrences. A clear disadvantage of LSA is that it does not include the context of the word, while GloVe and Word2Vec do include this. Furthermore, while both these methods and LSA are based on measuring distance between vectors, LSA is a count-based model that uses a term-document matrix and subsequent dimensionality reduction, whereas Word2Vec is a prediction-based model that uses neural networks to predict words based on their local context (Altszyler, Sigman, Ribeiro, & Slezak, 2016; Naili, Chaibi, & Ghezala, 2017). GloVe is generally regarded as an extension of count-based models as it uses counts of co-occurrence a global level (Pennington et al., 2014). However, the performance of all three methods in finding word similarity and synonyms depends on the context and application, but while used by multiple (recent) studies on dictionary creation LSA appears to be less efficient for large corpora and has a disadvantage of failing to account for the context of a word (Naili et al., 2017). Hence, this paper leverages the advantages of different methods by combining results of several methods.

2.2.3 Identifying grammatical functions of words

As aforementioned, Born and Taris (2010) found that the grammatical function of male characteristics in job advertisements affect the degree to which women are attracted by those advertisements. However, their approach has not been based on algorithmic decision-making as is used in the current research. In order to detect the grammatical function or class of words in previously unseen text, syntactic parsing is often used as it identifies the grammatical arrangement of words and their relationships (Ratnaparkhi, 1996; Tayal, Raghuwanshi, & Malik, 2014). More specifically, for algorithmic syntactic parsing Part-Of-Speech (POS) tagging is widely used, which assigns a class tag to each word (Paroubek, 2007; Ratnaparkhi, 1996). By using POS tagging, relatively high accuracy rates in grammatical tagging can be attained in a relatively low amount of time, which makes it a reliable method (Ratnaparkhi, 1996). There are several approaches to POS tagging that differ mostly on assigning a tag to ambiguous words, i.e. disambiguation: rule-based, statistical and hybrid approaches (Perez-Ortiz & Forcada, 2001).

Rule-based methods apply a predefined set of linguistic rules to a word context to assign a POS tag to an ambiguous word (Perez-Ortiz & Forcada, 2001; Voutilainen, 2003). While such approaches do not provide any probabilities on tag assignment, statistical methods on the contrary use the maximisation of probabilities that a word belongs to a certain tag (Perez-Ortiz & Forcada, 2001). In statistical POS tagging, the algorithm is generally trained on a pre-tagged corpus and uses contextual features on a sentence-level to attain probabilities of a word belonging to a certain POS tag (Paroubek, 2007; Ratnaparkhi, 1996). By choosing the POS tag with the highest grammatical probability, often high accuracy rates can already be attained, but this can be improved by including probabilities based on word sequence and contextual factors (Paroubek, 2007). Lastly, hybrid approaches to statistical and rule-based POS tagging combine both methods, such as by using a small amount of rule-based data to train an algorithm to establish new rules (Brill & Pop, 1999).

2.2.4 Predictive methods for textual features of gender bias

In using gendered wordings based on communality and agency to predict organisational gender diversity, Pietraszkiewicz et al. (2019) found significant effects when using the percentage of agentic and communal language in job advertisements to predict the percentage of men in professions of those advertisements, and vice versa. However, while significant results were obtained for both directions using regressions, this approach was limited to using percentages of agentic and communal words, and did not include the grammatical function of words. Also, it did not predict the gender of applicants, but merely the current gender diversity states of professions. In order to discover which gendered words and which grammatical forms are specifically important in predicting organisational gender diversity outcomes, methods should be used that allow for inclusion of many variables.

When exploring the use of many variables in a regression, a balance is generally sought between a good fit of the model and a minimal number of explanatory variables to improve model interpretation and out-of-sample performance. Model overfitting entails that a model performs well on the sample data, but has ill-performance when predicting using out-of-sample data, which decreases the generalisability of conclusions (McNeish, 2015). As model overfitting often results from including many variables in a regression, several methods exist to select relevant variables

and to prevent overfitting (McNeish, 2015). While common methods, such as forward or backward stepwise selection, tend to improve the R^2 and address some issues of overfitting, they can lead to inaccurate standard errors and p-values as these are not suitable for adaptively choosing predictors (Lockhart, Taylor, Tibshirani, & Tibshirani, 2014; McNeish, 2015). With the goal of using many words and grammatical structures while aiming to have a limited number of explanatory variables, regularisation methods can be used to decrease complexity of the model by adding a penalty term to the model (McNeish, 2015). Various methods of regularisation or penalisation, such as Ridge or Lasso, address the disadvantages of previously discussed methods and tend to perform better in decreasing model overfitting (Hastie, Tibshirani, & Tibshirani, 2020; Lockhart et al, 2014; McNeish, 2015). Firstly, in Ridge regression a penalty term is added to the sum of squared regression coefficients that results in a shrinkage proportional to the size of the coefficient estimate and that leads to shrinkage towards zero (McNeish, 2015). As the shrinkage penalty is non-zero, all predictors are retained and consequently this does not lead to disadvantages of variable selection methods, such as unreliable standard errors (McNeish, 2015). Secondly, least absolute shrinkage and selection operator (Lasso) is a similar method to ridge regression, but it does impose a penalty term that can result in a shrinkage of coefficients to exactly zero, and, therefore, is a variable selection method (Tibshirani, 1996). The benefit of variable selection is that a lower number of variables improves model interpretation (Owen, 2007; Tibshirani, 1996). A shrinkage towards zero and a possibility that coefficients become zero results from introducing a penalty term to the absolute value of the sum of regression coefficients (McNeish, 2015; Owen, 2007; Tibshirani, 1996). However, while p-values of Lasso regressions cannot be interpreted for similar reasons to other variable selection methods, Lasso regressions generally outperform these methods in relation to overfitting and Ridge with regards to interpretation purposes (Lockhart et al, 2014; McNeish, 2015).

In addition to single (penalized) regression methods, more advanced Random Forest models and Neural Nets are often able to reach high predictive accuracy on out-of-sample observations with many features. Firstly, Random Forests are an ensemble learning method in which results of single classification or regression tasks are aggregated (Liaw & Wiener, 2002). By aggregating results of a number of decision trees and by subsequent majority voting or averaging of outcomes of these decision trees for classification and regression, respectively, robust

classification or prediction outcomes can be attained (Liaw & Wiener, 2002). Secondly, Artificial Neural Networks (ANNs) can be trained to model complex non-linear relationships between inputs and outputs for a wide variety of tasks, such as classification and prediction (Lantz, 2019, p.206). Similar to the functioning of a biological brain, in between inputs and outputs (hidden) artificial neurons or nodes process information and solve learning problems (Lantz, 2019, p.206). While ANNs are powerful machine learning algorithms, they are considered black-box methods due to their complex internal mathematical systems (Lantz, 2019, p.205). Random Forests are less generally also considered black-box methods due to the combination of many single prediction or classification tasks (Liaw & Wiener, 2002). With such black-box models a trade-off should be made between accuracy and interpretation. When minimal improvements in accuracy occur compared to non-black-box models, while interpretation may be substantively more difficult due to their black-box nature, usage of these models might not be preferred. Hence, in interpretation of the results in this research a careful consideration is made between usage of black-box and non-black-box methods.

Table 2. Comparison of existing studies

Study	Type of data	Focus on job ads	Focus on gendered words	Focus on gendered grammatical structure	Uses word embeddings			Uses predictive models		
					LSA	**GloVe**	**Word2Vec**	**Regression**	**RF**	**ANN**
Born and Taris (2010)	Experimental	Yes	Yes	Yes	No	No	No	No	No	No
Gaucher, Friesen and Kay (2011)										
Study 1 and 2	Empirical	Yes	Yes	No	No	No	No	No	No	No
Study 3, 4, 5	Experimental	Yes	Yes	No	No	No	No	No	No	No
Pietraszkiewicz et al. (2019)										
Study 1 and 2	Experimental	No	Yes	No	Yes[a]	No	No	No	No	No
Study 3	Survey	No	Yes	No	Yes	No	No	No	No	No
Study 4	Empirical	Yes	Yes	No	No	No	No	Yes[b]	No	No
Dinan et al. (2020)	Empirical	No	Yes	Yes	No	No	No	No	No	No
The current study	Empirical	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

[a] They used LSA to validate similarity assessment of a panel judgement.

[b] They conducted a regression to predict the percentage of male employees in jobs in general.

3. Methodology

In this section, a more technical explanation of the aforementioned methods will be provided: word embedding models LSA, Word2Vec and GloVe, Part-Of-Speech (POS) tagging based on Hidden Markov and Maximum Entropy Models, and the predictive models Lasso regression, ANNs, and Random Forest prediction.

3.1 Models for obtaining semantic word similarity

Three methods for obtaining semantic word similarity are set out hereafter: LSA as it has been used in existing literature in a similar context, and Word2Vec and GloVe as both tend to outperform LSA with the goal of obtaining semantic word similarity and take the word context into account. Furthermore, in general it is suggested that reliance on single embedding models for similarity calculations should be minimised as these tend to be rather unstable (Antoniak & Mimno, 2018). Regarding Word2Vec and GloVe, in existing literature neither method is seen as outperforming the other throughout all contexts.

3.1.1 LSA

Latent Semantic Analysis (LSA) is a Bag-Of-Words method that was proposed originally by Landauer and Dumais (1997) and that determines word embeddings by looking at global word co-occurrences. In this model, a term-document co-occurrence matrix is constructed from all documents and dimensionality reduction is applied in order to obtain vector representations (Altszyler et al., 2017). The term-document matrix M is constructed where:

$$M_{i,j} = n_i \cdot m_j. \quad (1)$$

In this matrix, the rows represent each unique word or term $n_1, n_2, \dots n_i$, and the columns represent each document $m_1, m_2, \dots m_j$ (Landauer, Foltz & Laham, 1998). Each cell represents the frequency of term n_i in each document m_j. Hereafter, dimensionality reduction is applied to this matrix using Singular Value Decomposition (SVD), which results in three separate matrices that result in matrix M when their product is taken (Landauer et al., 1998). This results in the following semantic space:

$$M_{(n*m)} = U_{(m*r)} \cdot S_{(r*r)} \cdot V^T_{(n*r)}. \quad (2)$$

In this space U and V^T are orthogonal matrices, in which documents and terms are represented respectively against latent (hidden) concepts (r) (Landauer et al., 1998; Naila et al., 2017). In turn, U is a diagonal matrix that includes a multiplication of all r latent concepts represented on the diagonal (Landauer et al., 1998; Naila et al., 2017). Dimensionality reduction results from reducing the number of r latent concepts by retaining only the largest singular values in these three matrices (Landauer et al., 1998). The vector of a term in the semantic space represents the word embedding, which can be used to assess word similarity by using, for example, cosine similarity (Naila et al., 2017). A vector space for LSA can be created in R using the package **lsa** (Wild, 2007).

3.1.2 Word2Vec

Word2Vec is a local word embedding model proposed by Mikolov et al. (2013a) that uses neural networks to gain vector representations that capture both semantic as well as syntactic relationships between words. Word similarity is calculated by evaluating the cosine similarity between vectors (Naili et al., 2017).

More specifically, two approaches exist within Word2Vec that use neural networks: Continuous Bag-Of-Words (CBOW) that predicts a focal word based on the context, and Skip-Gram that predicts the context based on the focal word (Mikolov et al., 2013a; Naili et al., 2017). In CBOW, the input is the context, which is a chosen window containing words before and after a target word, and the output is the target word (Mikolov et al., 2013a). It should be noted that it uses a continuous distributed representation of the context words, different than the previously described bag-of-words method of LSA (Mikolov et al., 2013a). Given a word sequence $w_1, w_2, \dots, w_V$, the aim of the CBOW model is to maximize the average log-likelihood function:

$$\frac{1}{V}\sum_{i=1}^{V} \log\, p(w_i|w_{icx})\,. \qquad (3)$$

In this equation, w_{cxt} forms the context of word w_i for the context window size c. The log-likelihood is calculated of correctly predicting word w_i given the words in the context window (Mikolov, Le, & Sutskever, 2013; Naili et al., 2017).

The Skip-Gram model uses similar logic, but is the opposite of the CBOW model regarding its inputs and outputs. The Skip-Gram model is defined as follows:

$$\frac{1}{V}\sum_{i=1}^{V}\sum_{j=i-c,\ j\neq i}^{i+c}\log\ p(w_j|w_i). \qquad (4)$$

In this equation, c represents of the size of the context window around word w_i. In turn, the log likelihood is calculated of correctly predicting word w_j, given the middle word w_i (Mikolov et al., 2013c). For little variation in language and small datasets the Skip-Gram model appears to give more accurate word presentations, while CBOW performs better for large datasets (Mikolov et al., 2013c).

In both models, the transformation from the hidden layers to the output vectors results in probability distribution that constitute the (neural) word embeddings (Mikolov et al., 2013a). Both the word embedding vectors from the CBOW and the Skip-Gram model can subsequently be compared in terms of cosine similarity to determine word similarity. In R, the Word2Vec algorithm for CBOW and Skip-Gram can be trained by using the **word2vec** package in the **text2vec** library (Selivanov & Wang, 2020).

3.1.3 Global Vectors (GloVe)

GloVe is a model based on word co-occurrences in a fixed context window, hereby combining both the local word context as in Word2Vec, as well as the global word co-occurrences as in LSA and leveraging the advantages of both approaches (Pennington et al., 2014). By relying on word-word co-occurrence counts, the method is more efficient than Word2Vec, as in the latter each context window is considered, while repetition in documents tends to occur (Pennington et al., 2014). Also, by considering the word count of a word in a local context window, it reduces the impact of frequent words and tends to perform better on word analogy tasks compared to LSA as it reflects different meanings of words (Pennington et al., 2014). The basis is constructing the word-word co-occurrence matrix (X), in which X_{ij} is frequency in which word i co-occurs with word j. Then, the co-occurrence of any word k in the context of word i is defined as:

$$X_i = \sum_k X_k\,. \qquad (5)$$

The probability P that word j co-occurs in the context of word i is:

$$P_{ij} = \frac{X_{ij}}{X_i}\,. \qquad (6)$$

When words i and j are similar, these should co-occur with similar words. Consequently, the general model of GloVe is defined as:

$$F(w_i - w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}}\ , \qquad (7)$$

in which w_i and w_j represent vectors of word i and word j, respectively, and P_{ik} and P_{jk} show the probability that any word k co-occurs with word i and word j. When the difference between the linear vectors of w_i and w_j is small, they co-occur with the same words $\widetilde{w}_k$. In order to match the odds ratio and to ensure a linear structure, the scalar product is taken as follows:

$$F\left(\left(w_i - w_j\right)^T \widetilde{w}_k\right) = \frac{P_{ik}}{P_{jk}} = \frac{F\left(w_i^T\ \widetilde{w}_k\right)}{F\left(w_j^T\ \widetilde{w}_k\right)}. \qquad (8)$$

The last ratio within equation 8 results from the assumption that the choice is arbitrary whether word i or word j is the focal and context word, or vice versa, and symmetry should exist. This assumption is represented by assuming a homomorphism or structure-preservation, which means that the relationship between word i or word j will hold when switched around. By incorporating equation 8 in equation 6, the following equation results:

$$F\left(w_i^T\ \widetilde{w}_k\right) = P_{ik} = \frac{X_{ik}}{X_i} \qquad (9)$$

Then, the exponential of function F is taken and X_i is absorbed into a bias term b_i as it is independent from k (Pennington et al., 2014). Also, a bias term $\tilde{b}_k$ is added for $\widetilde{w}_k$ as symmetry should hold as described above, and the following equation results:

$$w_i^T\ \widetilde{w}_k + b_i + \tilde{b}_k = \log\left(X_{ik}\right) \qquad (10)$$

Lastly, in order to correct for very infrequent or words that do not co-occur, and for highly frequent words in X_{ij}, a weighting function $f(X_{ij})$ is introduced. This results in the following weighted least squares problem, in which V is the vocabulary size:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T\ \widetilde{w}_j + b_i + \tilde{b}_j - \log\left(X_{ij}\right))^2. \qquad (11)$$

Hence, by means of the objection function J represented in equation 11, GloVe results in multi-dimensional representation of words, which can be used to capture similar words when words have similar embeddings in the latent space. In R, a GloVe model can be trained using the **GlobalVectors** function in the **text2vec** library (Selivanov & Wang, 2020).

3.1.4 Comparing embedding results

To enhance comparison, the following hyperparameters have been set to similar values for Word2Vec Skip-Gram, Word2Vec CBOW, GloVe, which is also a skip-gram algorithm, and LSA when possible:

(i) For all four models, the number of dimensions in the latent space has been set to 50 given the rather small size of vocabulary as discussed in section 4.2.2. Hereby, this default value of 50 dimensions (in GloVe and Word2Vec in R) is not required to be greater, which is also recognised as sufficient for most NLP tasks (Lai, Liu, He, & Zhao, 2016).

(ii) The number of training iterations has been set to 200 for the first three models after which embedding models tend to be sufficiently trained while overfitting should be prevented (Lai et al., 2016).

(iii) A window size of 5 is applied to the first three models, indicating that five words before and after each context word are taken into consideration. As stop words are removed, a larger window size is not deemed necessary as related words should then be found within five words distance to a word.

When focussing on comparing words, in the three methods LSA, Word2Vec and GloVe the output is a vector representation of a word. Consequently, words that are similar are presumed to have a similar vector or embedding. A commonly used measure to calculate similarity between two words is the cosine similarity measure, which can be employed in R by the **sim2** function in the **text2vec** library (Naili et al., 2017; Selivanov & Wang, 2020). The cosine similarity between the vector of word i ($\overrightarrow{w_i}$) and the vector of word j ($\overrightarrow{w_j}$) is in a simple form defined as follows:

$$similarity(w_i, w_j) = \cos(\theta) = \frac{\overrightarrow{w_i}\,\overrightarrow{w_j}}{|\overrightarrow{w_i}||\overrightarrow{w_j}|}. \quad (12)$$

However, this approach concerns comparison of individual terms, and the goal of this paper is to find terms similar to the masculine and feminine word lists. Hence, for each embedding method the cosine similarity of each term to each term in existing gendered word lists has been computed and averaged, to create an average cosine similarity measure to assess similarity between individual terms and existing gendered word lists. Furthermore, as for each embedding method the cosine similarity measure has been used to assess similarity, direct comparison of these

similarities for each method can be performed. More specifically, in order to compare these results, the following three methods are used.

Firstly, a comparison of the top ten words is provided for each method. Herein, the embedding methods are assessed separately, which is also used by Antoniak and Mimno (2018) to evaluate stability of embedding-based word similarities. Secondly, a mean of word embeddings over all methods is provided as is suggested by Antoniak and Mimno (2018) to enhance robustness of findings. The words with the highest mean score are supposed to have high similarity to existing masculine and feminine word lists throughout all methods. Lastly, to evaluate space consistency of the embedding spaces resulting from the four embedding methods, the cosine similarity is also used as suggested by Bloem, Fokkens, and Herbelot (2019) when using a relatively small dataset.

3.2 Part-Of-Speech (POS) Tagging

Part-Of-Speech (POS) or word class tagging is a classification technique that aims at assigning a grammatical function to a word which reflects its role in a sentence (Kupiec, 1992). While some words can be unambiguously tagged as they only have one grammatical role, the challenge of POS tagging is to tag ambiguous words that can have various grammatical classes, the process which is called disambiguation (Kupiec, 1992). The goal of statistical POS tagging is to find the sequence of n POS tags $t_1, t_2, \dots t_n$ for a given sequence of n words $w_1, w_2, \dots w_n$ with the greatest posterior probability by maximizing (Highfill, 2011):

$$P(t^n|w^n) = P(w^n|t^n)\, P(t^n), \qquad (13)$$

in which the second part is rewritten based on Bayesian models (Lee, Tsujii, & Rim, 2000). To find these posterior probabilities, Hidden Markov Models and maximum entropy are common statistical models for POS tagging that are described hereafter.

Hidden Markov Models (HMMs) are probabilistic models that use Markov chains with unobserved hidden states, which in this application are POS tags and their transition probabilities, to choose the most probable tags for a sequence of words (Jurafsky & Martin, 2020). Based on a training corpus, HMMs use probabilities that one POS tag (state) transitions to another tag (state) by assuming that this transition depends solely on the current tag (state) (Kupiec, 1992). Given this assumption, the HMM aims to maximize:

$$\prod_{i=1}^{n} P(w_i|t_i)\, P(t_i|t_{i-1}). \quad (14)$$

In this maximisation model, the first probability corresponds to the emission probability, which is the probability of observing word w_i when observing state t_i, while the second probability is the transition probability, which is the probability of transitioning to state t_i given the previous state t_i (Jurafsky & Martin, 2020). By solving the maximisation function, each word is given posterior tag probabilities (Kumar & Paul, 2016, p. 18). As the states or tags are hidden, a Baum-Welch or Viterbi algorithm can be applied to uncover the most likely underlying sequence of states and transition probabilities (Cutting, Kupiec, Pedersen, & Sibun, 1992).

Instead of estimating the posterior probability of a sequence of tags $t_1, t_2, \ldots t_n$ for a given sequence of words $w_1, w_2, \ldots w_n$ by maximising the joint probability, the Maximum Entropy (MaxEnt) model directly calculates the posterior probability and does not account for interaction between states (Ratnaparkhi, 1996; Ratnaparkhi, 1997). MaxEnt entails that entropy or uncertainty should be maximised to constraints that are known, such as probability distributions (Ratnaparkhi, 1997). Compared to HMMs, MaxEnt has the benefit of choice of other contextual features that are included after which the weight of these features is automatically determined (Ratnaparkhi, 1997). The benefits of the MaxEnt and the HMM can be leveraged by using the MaxEnt model to attain the optimal POS-tags that can be considered in the HMM (Highfill, 2011).

In R, various packages exist by which POS tagging can be performed. However, many packages offer only a pretrained POS tagging model for English, while the goal of this paper is to statistically POS tag both Dutch and English. However, the R package **UDpipe** that has been pretrained on large annotated text datasets performs relatively well for both Dutch (91% accuracy) and English (94% accuracy) while being relatively efficient in terms of computational power (Straka, Hajic & Straková, 2014). This package uses a neural algorithm to train an HMM model that relies on Viterbi decoding for attaining the probabilities of a POS tag (Straka et al., 2014).

3.3 Predictive models

The aim of the following predictive models is to predict job applicant gender diversity which is operationalised as the share of female job applicants from the total number of job applicants. As independent variables, textual features are included that result from existing gendered wording lists, expanded gendered wording lists as described in section 4.1 and grammatical features of these words as described in section 4.2.

3.3.1 Lasso regression

In this paper, many features are used to predict gender diversity of job applicants. Lasso regression is chosen over Ridge regression given the benefit of variable selection, which improves subsequent interpretation as less variables need to be interpreted in such a sparse model. Additionally, prediction accuracy can be improved by variance reduction of the outcome variable, which results from introducing bias (Tibshirani, 1995). Lasso or least absolute shrinkage and selection operator is built upon a linear regression model that can be defined as follows (excluding the constant):

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j. \qquad (15)$$

In this model, y_i is the value of the dependent variable and x_{ij} the value of the predictor variable for each observation $i_1, i_2, \dots i_n$ and variable $j_1, j_2, \dots j_p$. In turn, β_j is the coefficient estimate for each variable $j_1, j_2, \dots j_p$. In the Lasso regression model, a penalty term is introduced to the standard linear regression model, and the model aims to minimize the following regression problem:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|. \qquad (16)$$

In this model, the first term equals the linear regression term, and the second term is the penalty term in which λ or lambda is the penalty parameter that penalizes the absolute value of each β_j (Tibshirani, 1995). While $\lambda = 0$ results in a linear regression model, increasing the value of the tuning parameter λ results in greater shrinkage of the coefficient estimates and if great enough it will lead to coefficient estimates equal to zero. Hence, the value of λ is highly influential and can be chosen by the commonly used cross-validation (Roberts & Nowak, 2014). In cross-

validation different values of λ are used to predict the outcome variable for random hold-out subsets (folds) of the data and, subsequently, the λ that is chosen results in the lowest prediction error (Roberts & Nowak, 2014). While data is randomly split in a predefined number of n folds and this split is random, cross validating the tuning parameter λ generally results in a good balance between bias and variance (Roberts & Nowak, 2014). A Lasso regression with a 5-fold cross-validation of λ is implemented in R with the package **glmnet** (Friedman, Hastie & Tibshirani, 2020).

3.3.2 Random Forest prediction

Random Forest (RF) is a method that aggregates individual decision trees for classification or regression, and given the goal of prediction the focus hereafter is on RF regression. In RF analysis, a large number of decorrelated tree predictors is generated after which the predicted value of all individual trees is averaged to attain an output (Breiman, 2001; Friedman, Hastie, & Tibshirani, 2017). The outcome of the kth individual tree is defined as follows:

$$h(\boldsymbol{x}, \Theta_k). \qquad (17)$$

In this definition, h represents the prediction for the kth tree given the input vector $\boldsymbol{x}$. For the kth tree, Θ_k is a random vector created independently from other random vectors $\Theta_1, \ldots \Theta_{k-1}$ and denotes firstly the random choice of m predictors and secondly the randomly bootstrapped subset of n observations that is considered in the kth tree (Breiman, 1999). The first part of this random vector Θ_k ensures decorrelation of individual decision trees, which is performed by taking the random subset of m predictors out of all considered p predictor variables, in which $m < p$ (Breiman, 2001; Friedman et al., 2017). At each node within each tree, the best variable to split is selected while considering the randomly chosen m predictors. Reducing m will decrease variance as the correlation between each pair of individual trees is reduced (Breiman, 1999; Friedman et al., 2017). The second part of the random vector Θ_k, indicates inclusion of a randomly bootstrapped sample out of all n observations in the training dataset, which is usually two-third of all observations (Breiman, 1999; Friedman et al., 2017). One-third of the observations that is randomly left out are the out-of-bag (OOB) observations and can be used to obtain the OOB error by predicting the outcome for the OOB observations using the trained model (Breiman, 1999; Friedman et al., 2017). This error is similarly accurate to using a separate test dataset and

is commonly used to assess the RF performance (Breiman, 1999). The aggregated outcome over all K trees results from averaging each individual tree as follows:

$$\hat{f}^B(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} h(\boldsymbol{x}, \Theta_k)\,. \qquad (18)$$

As this RF prediction involves aggregation of many individual decision trees over random subsets of the data resulting in a large size, it is often considered a black-box model (Zhang & Wang, 2009). The RF regression algorithm is performed in R by using the package **randomForest** (Liaw & Wiener, 2002). The following hyperparameters were tuned using grid search: predictors m, the number of K trees, and the minimum size of terminal nodes, by which the depth of each tree is determined.

3.3.3 Artificial Neural Network (ANN)

While Random Forests are frequently seen as a black-box method due to their size, Artificial Neural Networks (ANNs) are more generally considered a black-box method due to their complex mathematical systems (Lantz, 2019, p. 205). The idea of ANNs is based upon the biological brain in which a network of interconnected neurons processes sensory inputs into outputs. For such a transformation a network of artificial neurons or nodes is used, which results in a wide range of applications, such as classification or prediction (Lantz, 2019, p. 208). When focusing on one neuron and a single-layer network, as in the biological brain the dendrites receive the inputs $(x_1, x_2, \dots x_n)$ and attach a weight $(w_1, w_2, \dots w_n)$ according to their relative importance. Afterwards, an activation function (f) is applied that transforms these weighted inputs into an output (y), but only when a specified threshold is reached (Lantz, 2019, p. 208). For one neuron or node, these factors result in the following definition:

$$y\,(x) = f\left(\sum_{i=1}^{n} w_i x_i\right). \qquad (19)$$

There are three main characteristics of ANNs that can be tuned or selected when building an ANN: the activation function (f), a network architecture and a training algorithm (Lantz, 2019, p. 208). Firstly, there are various forms of the threshold activation function of which a sigmoid activation function is the most common form $(f(x) = \frac{1}{1+e^{-x}})$, while other forms include a (saturated) linear or Gaussian activation functions (Lantz, 2019, p. 210).

Secondly, the neural network architecture includes the number of (hidden) layers and the number of hidden nodes therein, and the direction of information travel (Lantz, 2019, p. 211). While for rather simple tasks a single layer can suffice, which has the form as described above, for more complicated tasks one or multiple layer(s) of hidden nodes can be included through which input signals are processed before attaining an output. Also, the number of hidden nodes within those layers can be tuned to reach the optimal performance, but a high number can result in overfitting the model on the training data. Next to the hidden layers and nodes, the direction of information travel through the network can be chosen. While a feedforward ANN allows information to travel from the input to the output in one direction, a feedback ANN uses loops by which information can travel in both directions and more complex patterns can be learned (Lantz, 2019, p. 213-214).

Lastly, a training algorithm is required to improve the processing of information and to train the ANN to perform a similar task on unseen data. The backpropagation algorithm is the most widely used and involves both a forward phase in which input signals travel in one direction to attain and output, and a backward phase in which the weights are modified to reduce error produced in the forward phase (Lantz, 2019, p. 216). In addition, decaying the weight can prevent overfitting as large weights can result in overfitting the model on noise in the training dataset (Krogh & Hertz, 1992). While the backpropagation algorithm is generally accurate in many different tasks and makes few assumptions on the relationships in the data, its results are difficult or impossible to interpret and can be computationally intensive to train (Lantz, 2019, p. 216). Also, with increasing complexity ANNs can lead to overfitting on the training data, which can result in less accuracy when using the ANN model on the test data. ANN is implemented in R using the package **nnet** which uses a backpropagation algorithm as described above (Ripley, Venables, & Ripley, 2016). In the ANN, the number of nodes in the hidden layer and the weight decay were tuned using grid search while decreasing complexity of the model by using a sigmoid activation function with a single hidden layer.

3.3.4 Evaluation of predictive methods

In order to assess the performance of predictive regression methods and select the model with the highest performance, the trained model is used to predict the outcome variable using the

test data that was excluded in training the model. The performance of the model for both the test and training are evaluated with the commonly used Mean Absolute Error (MAE), the Root mean square error (RMSE) and for validation, the coefficient of determination (R^2). The MAE is the average absolute difference between the ith value of the dependent variable (y) and its predicted value ($\hat{y}$) by the model:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \quad (20)$$

The RMSE is defined as the root of the average squared difference between the ith value of the dependent variable (y) and its predicted value ($\hat{y}$) by the model:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}\,. \quad (21)$$

Lastly, the R^2 is defined as the sum of squared differences between the ith value of the dependent variable (y) and its predicted value ($\hat{y}$), divided by the sum of squared differences between the ith value of the dependent variable (y) and the mean value of the dependent variable ($\bar{y}$). In other words, the R^2 indicates the amount of variance in the dependent variable accounted for by the model:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \quad (22)$$

4. Data

4.1 Data collection

In this research, two sources of data are used and analysed: consultancy job advertisements from the Dutch version of Indeed.com[1] and consultancy job advertisements from an international consultancy firm along with the gender of applicants. As Indeed.com is one of the largest job search sites in the world and in the Netherlands, it is expected that consultancy job advertisements are representative for all consultancy job advertisements in the Dutch job market. The job advertisements of Indeed.com have been extracted using web scraping, which is a technique by which unstructured data is extracted from websites and stored as structured data (Sirisuriya, 2015). In web scraping, it is important to work ethically by, among other, not engaging in scraping privacy sensitive user data without permission, which this study refrains from.

Scraping of job advertisements has been conducted using the library **Selenium** within Python, which is a tool for automating website browsing using Application Programming Interfaces (APIs) (Lawson, 2015). An advantage of **Selenium** is that it is less detectable to be non-human by controlling the browser of the user (Sirisuriya, 2015). In turn, this contributes to common challenges in scraping, such as captchas that detect bots from humans, and getting (temporarily) banned because of sending too many requests to the website. The latter also relates to unethical scraping as it can lead to overcharging a website. By including a limit on server requests and a waiting time, this was prevented. Elements that have been scraped are job title, company, location, and job description. Also, data has been scraped several times during three months in order to gain as much data as possible. Afterwards, duplicate job advertisements with the second dataset concerning job advertisements of a consultancy firm were checked and removed, because these were partly included on Indeed.com.

To gather the second dataset, which contains consultancy company data, Selenium was also used to scrape data from the internal HR website of the consultancy firm. Here, the job title, job description, first name and gender of applicants is scraped as the goal is to use gendered wording to predict the share of female applicants.

[1] https://nl.indeed.com/

4.2 Data pre-processing

4.2.1 General data pre-processing and selection

While the original Indeed.com dataset based on a search for 'consultant' contained over 4000 job advertisements, many job advertisements have been discarded as these were duplicated or did actually not concern a consultancy related position, but appeared in the search nonetheless. Regarding the consultancy HR dataset, which contains data from an international consultancy firm, only job advertisements were kept when including over ten applicants in order to increase reliability of analyses. Also, job advertisements were removed when the gender was unknown for over 40% of the applicants and hereafter described gender determination tools could not be applied as data had been anonymised according to the EU General Data Protection Regulation (GDPR). These data removal steps have resulted in the number of Dutch and English job advertisements that can be found in Table 3.

Table 3. Frequency table

	Dutch	**English**	**Total**
Indeed.com dataset	1304	432	1736
Consulting HR dataset	164	15	179

Next to removal of data, data was transformed for the second dataset. When the gender of applicants was absent while the name was available, the gender was predicted using the **genderizeR** package in R that bases its prediction on many data sources, including census data and social network profiles (Wais, 2016). Furthermore, data was transformed from individual applicant data to gender share per job advertisement.

For the predictive analyses, the English observations in the consultancy HR dataset were not included due a low number of job advertisements ($n = 15$). Furthermore, the Dutch observations of the consulting HR dataset have been split in an 80% training dataset and a 20% test dataset to assess the predictive performance of the model created on the training dataset on previously unseen data. This is used to validate that the predictive models do not overfit on the training data.

4.2.2 Pre-processing of job advertisement texts

For each of the three phases of analysis, different text pre-processing steps have been conducted depending on the respective goal of the phase. For the goal of gaining insights into similar words to the existing word lists of masculine and feminine language in both Dutch and English, the words appearing in the Indeed.com dataset have been stemmed. This is in accordance with the existing word lists about which Gaucher et al. (2011) have emphasised that a word is masculine or feminine regardless of its grammatical form. Hence, the goal is to discover stemmed words from the gendered dictionary that are similar to words in the dataset.

Furthermore, stop words have been removed as these do not contain relevant information about the job advertisement text and their frequency can disproportionately influence mostly LSA, but also GloVe to a lesser extent. While for English most NLP packages in R contain a list of stop words with relatively many acknowledged stop words, for the Dutch language several packages have been combined to increase the list of stop words as it is limited in most R packages. Furthermore, punctuation, digits, and URL's have been removed, and all text has been decapitalized. Also, based on the scraped company name, company names have been removed from the job advertisement as the focus is not on comparing individual companies.

Lastly, the job advertisements have been tokenized while keeping the terms only when occurring at least five times, to remove any noise in the data, such as misspellings, but to include somewhat more uncommon words than with the default value of ten times. This has resulted in a vocabulary of 4,415 words in Dutch and 1,746 words in English. For GloVe this tokenized set has been converted to a term co-occurrence matrix (TCM) in which every value represents the frequency of co-occurrence. For LSA this set has been transformed into a document-term matrix (DTM) in which every row represents a job advertisement and each column is a term. Every value represents the frequency of occurrence within an advertisement.

For the second phase of POS tagging, pre-processing has been minimised as POS tagging relies to a large extent on the grammatical form of the word and whether it is capitalized. Consequently, words have not been stemmed and removal of stop words has not been done as this could lead to unnatural textual structures for which the model has not been trained. Furthermore, as named entities are recognized by, for example, their capitalization and textual location, it was unnecessary to remove company names.

Regarding the last phase of predictive analyses, for creating independent variables, each term has been POS tagged for which pre-processing has been minimized. For each tagged word that appears in the masculine or feminine word list for Dutch and English, respectively, a variable has been created including their POS tag, as well as the amount of nouns, adjectives and verbs used to describe these masculine and feminine words in each job advertisement.

4.2.3 Pre-processing of dictionaries

In order to assess the prevalence of both English and Dutch gendered wording dictionaries, a stemmed version of existing dictionaries has been used to detect which words truly occur in vacancies in the Dutch consulting job market. For English, the dictionary of Gaucher et al. (2011) has been used and stemmed (Appendix A). For Dutch, no academic literature exists on gendered word lists, but there are non-academic examples of word lists[2]. While these contain translations of English gendered wording, some words are included that do not appear in the English dictionary created by Gaucher et al. (2011) and do not reflect characteristics related to agency or communality, such as 'technical' or 'commercial'. Hence, by means of **translateR** package in R the English gendered dictionaries have been translated to Dutch. Words were left out when translation implied different meaning or when no consensus over translations was found, such as every word starting with 'together' (samen). For example, the Dutch translation of this word, 'samen', is also used to indicate 'samenstelling' (composition) and 'samenvatting' (summary). Hence, only 'samenwerken' (collaborate) was used for this word. For the cases with unclarity in which translations were insufficient, word embeddings could serve to fill gaps. After composing the list, the remaining Dutch words have been stemmed. The words from both the Dutch and English gendered wording dictionaries that appear in Dutch job advertisements for consulting can be found in Appendix B.

[2] https://www.leiderschapontwikkelen.nl/wp-content/uploads/eBoek-Inclusief-Werven-en-Selecteren-1.pdf

5. Results

5.1 Word embeddings

For both Dutch and English wordings separately, word embeddings have been created using GloVe, Word2Vec Skip-Gram (SG), Word2Vec Continuous-Bag-of-Words (CBOW), and LSA. Afterwards for each of these embedding spaces, the average cosine similarity has been computed between each term vector in the latent space and each term vector from the existing dictionary in the same space in order to assess semantic similarity. For all methods, the most similar ten words to the vectors of the gendered word lists can be found in Appendix C, which are based on the cosine similarity measure. Additionally, the top ten words based on the mean cosine similarity of all four methods has been added.

These quantitative embedding methods have not resulted in high similarities of 'similar' words throughout methods, which could be due to relative instability of single embedding methods or due to ambiguity in the meaning of words (Antoniak & Mimno, 2018; Gladkova & Drozd, 2016). Furthermore, some 'similar' words in methods cannot be intuitively related to existing dictionaries of gendered words. As interpretability of word similarities resulting from word embeddings is generally considered as driving their value, hereafter the cosine similarity of embeddings is used as a data-driven tool to aid and improve subjective dictionary addition, hereby balancing between computational and theoretical linguistic theory (Gladkova & Drozd, 2016).

Hence, in order to be included in gendered word lists, as mentioned in section 3.1.4 a term has to be included in the top ten of cosine similarities for one or – preferably – more methods. Additionally, a criterion for addition is to be associated with (a part of) the definition of communality and agency, respectively, in which communality includes selflessness and concern for others, kindness, understanding, warmth and respectfulness, and agency includes assertiveness, individuality, independency, self-confidence, and competitiveness (Cuddy et al., 2008; Gaucher et al., 2011; Heilman, 2012). As this process involves subjectivity in decision-making, words are only conservatively included in case of an obvious relationship to communality or agency.

Hereafter, the results for both English and Dutch consulting job advertisements are set out separately in order to discuss inclusion of new words into existing masculine and feminine dictionaries for both languages and can be found in Table 4 as well as in Appendix C extensively.

Table 4. Proposed words to be added to existing gendered dictionaries

	English	**English**	**Dutch**	**Dutch**
	Masculine	Feminine	Masculine	Feminine
1.	Steer	Team(-/work/mate)	Bepalend	Samenspraak
2.	Persuas(ive/ion/uade)	Colleague	Sturend	Samenleving
3.	Driver/drive	Like (*verb*/*adj*)	Stevig	Samengewerkt
4.	Led	Encourage	Aanpakk(er/en)	Gezamen(-/lijk)
5.			Slagvaardig	Coachend
6.			Topper	Hecht
7.			Visie	Respect

5.1.1 English word embeddings

For all methods, a mean of the similarity measure for each term has been created, indicating the extent to which the methods have produced similar results. For all English words, the cosine similarity between the four methods' embedding spaces has been found to be between 0.66 and 0.67, indicating a stable but consequent difference in cosine similarities between methods.

Most masculine terms are not specifically related to agency, but can also be related to communality. For example, the term 'approach' that appears multiple times in the top ten is used in some job advertisements to describe a 'hands-on approach', while other advertisements describe the need for an 'open and friendly approach'. However, four words have been identified which are evidently related to agency and agentic words, but have not been recognised by Gaucher et al. (2011) in their stemmed form, or are closely related to other words that have been included in existing word lists. These words are: 1) 'steer', which refers to 'dominant' and 'lead', 2) 'persuas(ive/ion/uade)', which refers to 'strength' and 'power', 3) 'driver' and 'drive', which is similar to 'leader' and 'determinant', and 4) 'led' which is the past tense of 'lead', but only words in the form of 'lead' have been included in accordance with Gaucher et al. (2011).

In opposite to masculine terms, the feminine terms that have high cosine similarities with existing female dictionaries are more directly related to words in these dictionaries. Some words can be found to appear in equal opportunity statements, such as 'multicultural', 'ethical', and

'diverse'. However, as these are rather characteristics of open work environments instead of female stereotypical characteristics, they are not included. However, several words qualify for inclusion in female dictionaries: 1) 'team(-/work/mate)', which refers to 'together', 'interdependency', and 'concern for others', 2) 'colleague', which refers to similar concepts and is used often in job advertisements to refer to 'cooperating with colleagues' similar to the previous word, 3) 'like' as a verb and 'likeable' as an adjective, which are similar to 'warm', 'pleasant' and 'kind', and 4) 'encourage', which is similar to 'support'. For further analysis, a stemmed version of the above-mentioned words (Table 4) has been added to the existing gendered word lists.

5.1.2 Dutch word embeddings

For Dutch word embeddings, the same procedure has been applied as for the English word embeddings. For all Dutch words, the cosine similarity between the four methods has been found to be between 0.63 and 0.70, indicating a less stable difference in cosine similarities between methods compared to English. The results of the top 10 similar words can be found in Appendix C as well as in Table 4 for the words that are proposed to be included. For masculine terms, cosine similarity to English translations of masculine dictionaries has resulted in some interesting words that do not appear directly as the first translation of English words, but are closely related in meaning. These are as follows: 1) 'bepalend', which is closely related to 'lead' and to 'decide', 2) 'stevig' that captures the meaning of English words 'force' and 'decisiveness', 3) 'sturend', which is closely related to 'lead', 4) 'aanpakk(er/en)', which is closely related to 'assertiveness', but is a more accurate translation of the concept in the Dutch language, which literally means 'to tackle (something)', 5) 'slagvaardig', which is a less common translation of 'decisive', 6) 'topper', which is a more commonly used Dutch version of 'superior', literally meaning 'the best', and lastly 7) 'visie', which is related to 'objective', and literally means 'vision'. The above discussed terms indicate that direct translations of English masculine words do not necessarily reflect similar concepts in the Dutch language, but word embeddings offer a means to reveal similar concepts.

Regarding feminine terms, it is notable that words that are relevant translations of 'together', starting with 'samen' occur several times, hereby complementing the limited translation into Dutch as described in section 4.2.3. These words are: 1) (in) 'samenspraak' which occurs mostly in the context of (in) 'consultation', 2) 'samenleving', which literally means 'society', but

it can refer to 'community' which is included by Gaucher et al. (2011); 3) 'samengewerkt', which is a passive form of 'cooperating', and 4) 'gezamen(lijk)', which literally means 'together', but it is a less common translation. Additionally, four other relevant words have been identified that are closely related to communality. These are 5) 'coachend', which is closely related to 'supporting' and literally means 'coaching', 6) 'hecht', which is generally an indicator of 'closeness' of a community and relates to 'warm' and 'communal', and 7) 'respect', which also is an English word related to 'submissive' and 'consideration'. All in all, through assessment of cosine similarity, the different embedding models have supported expansion of existing Dutch gendered dictionaries over direct translations. Next to the existing dictionaries, these words have been added for further analyses. A comparison of percentages of gendered wording in job advertisements including standard errors can be found in Appendix D for both the proposed Dutch and English word lists.

5.2 POS tagging

Each of the words including the proposed words in section 5.1 have been POS tagged. The focus of the POS tag results is on the frequency of verbs versus adjectives and nouns when using gendered words as Born and Taris (2010) suggest that using the former leads to higher inclination of females to apply. Verbs, namely, refer to a behaviour of which a potential candidate will assess the degree to which they fit and the motivation needed for this, as in the case of a noun or adjective a binary assessment is made of whether a potential candidate would fit or not, and consequently gives less room for consideration. The results of relative frequencies of verbs versus adjectives and nouns for both English and Dutch job advertisements can be found in Table 5. In comparing English and Dutch job advertisements for consulting jobs in the Netherlands, for both languages, the rate of noun and adjective usage versus verb usage of masculine words is similar with over 70% being a noun or adjective, while feminine words tend to be expressed as verbs to a greater extent.

Furthermore, in Appendix E it can be seen that some words are used many times throughout job advertisements, while especially the masculine words in noun or adjective form could have negative consequences for attracting women. In the top ten most used Dutch masculine words, most words are described in noun or adjective form, with only three words being verbs. In the top ten for most used English words this is similar with only two words being verbs. These

results are similar for feminine words in both languages. This indicates that there is room for improvement with respect to usage of verbs when using masculine gendered wording in consulting job advertisements in the Netherlands. For firms operating on this market, it is recommended that verbs are used to a larger extent when expressing masculine wordings in job advertisements as this will benefit female consideration of these advertisements.

Table 5. Percentages of nouns, adjectives and verbs used for masculine and feminine wording

	English	**English**	**Dutch**	**Dutch**
	Masculine	Feminine	Masculine	Feminine
Noun	51.2%	54.0%	49.6%	35.5%
Adjective	20.2%	12.7%	21.3%	36.4%
Verb	20.0%	27.3%	20.6%	24.3%
Other	8.6%	6.0%	8.6%	3.8%

As can also be found in Appendix E, the most used masculine words for Dutch and English are comparable with dominant words being mostly noun and adjective forms of 'analysis' (English) or 'analyse' (Dutch), 'lead' or 'leid', 'challenge' or 'uitdagend', 'driven' or 'gedreven', 'active' or 'actief'. The most used feminine words are also comparable with dominant words being 'support' or 'ondersteun', 'responsible' or 'verantwoordelijk', and 'collaborate' or 'samenwerken'. However, for feminine words there are some differences between English and Dutch. For example, 'team' that has not been included in Dutch word lists, is the most used feminine term in English advertisements. Vice versa, 'betrokken' (concerned) that has not been included in English word lists is used many times in Dutch job advertisements. Regarding words that are present in both word lists, 'enthousiast' (enthusiastic) and 'afhankelijk' (dependent) in Dutch are only very limitedly present in English advertisements, but often used throughout Dutch advertisements.

5.3 Predictive analyses

For the predictive analyses, the following independent variables have been compiled to predict the share of female applicants to total applicants: the frequency in each advertisement of all gendered words, their grammatical function in the sentence, the total number of masculine and

feminine words, the total number of words, and the total number of masculine and feminine verbs, adjectives and nouns. On the dependent variable, a random stratified split of the data into 80% training set and 20% test set was performed. The prediction evaluation criteria for a cross-validated Lasso regression model, a tuned Random Forest (RF) model and a tuned Artificial Neural Net (ANN) model can be found in Table 6.

Firstly, for the Lasso General Linear regression model (GLM), given the rather small sample size of the training set ($n = 107$), a 5-fold cross-validation of the tuning parameter λ was performed for the lowest RMSE value, instead of the default 10-fold cross-validation. This has resulted in an optimal value of $\lambda = 0.05$. In the final Lasso model, only three variables were found to have any predictive value for the share of female applicants to total applicants: the adjective form of 'leid' (lead), 'enthousiast' (enthusiastic), and 'prettig' (pleasant). While predictions of the Lasso model on the training data have resulted in a positive R^2, the model overfitted on the training data as regarding the R^2 it had no improvements over taking the mean value of the share of female applicants in predicting the share of female applicants in the test dataset.

Secondly, the hyperparameters of the Random Forest (RF) model have been tuned for the lowest Out-Of-Bag (OOB) RMSE using a grid search, resulting in an optimal minimum node size of 83, consideration of 55 predictors per split and a RF size of 100 trees. With these hyperparameter values a final RF model was created in which important variables are mostly the adjective of 'enthousiast' (enthusiastic), the noun of 'plezier' (fun), and the total number of words. When using this RF model to predict using both the training and test dataset, for both datasets no improvement regarding the R^2 was found over taking the mean value of the dependent variable, the share of female applicants. Hence, as with the Lasso regression, the RF model indicates no predictive value resulting from gendered wording variables.

Lastly, for the ANN the number of nodes in the hidden layer and the weight decay were tuned using grid search while decreasing complexity of the model by using a sigmoid activation function with a single hidden layer. By aiming for the lowest value of the RMSE on the training dataset in the grid search, an optimal weight decay of 10^{-4} and 3 nodes in the hidden layer were found with severe improvements over taking the mean value of the share of applicants. This model can be found in Table 6. However, when performing the same grid search on the training dataset while using these hyperparameter values to predict on the test dataset within the grid search, no

hyperparameter value positively improved over taking the mean value of the share of applicants. Hence, each ANN model with any of the grid searched hyperparameter values would have overfitted on the training dataset. As with the previous models, the ANN model indicates that no predictive value of the job advertisement results from gendered wording variables.

Hence, the use of black-box methods did not result in better models compared to the Lasso regression. More notably, all results suggest that no empirical relation can be observed between the share of female applicants and gendered wording through words and their grammatical form. The implications and potential causes of this are discussed in chapter 6.

Table 6. Results of predicting the share of female applicants

	Train			**Test**		
	RMSE	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
Lasso GLM	0.16	0.13	0.07	0.17	0.12	0
Random Forest	0.17	0.13	0	0.17	0.13	0
ANN	0.001	0.0004	0.99	0.33	0.25	0
using mean value	0.20	0.15	0	0.20	0.15	0

6. Discussion

Both word embeddings and POS tagging offer novel approaches to detect and analyse gendered language in text and specifically in job advertisements. These approaches have been based on existing experimental research in the field of social psychology in which negative effects have been found for female job attraction resulting from masculine wording through stereotypical male characteristics, and expression of masculine characteristics through nouns and adjectives (Born & Taris, 2010; Gaucher et al., 2011). While experimentally these findings have been confirmed several times, empirically no relation has been found between gendered wording and the share of female applicants in the real-world setting of this paper. Potential reasons for absence of such a relationship can be split in two sorts of arguments: data and context factors of this research, and limited external validity of existing experimental research.

With regards to factors related to data and context of the present research, results might have been influenced by several causes. Firstly, with regards to textual data a potential cause could have been the inclusion of an equal opportunity statement in all job advertisements, which has been confirmed to enhance job advertisement attractiveness (Born & Taris, 2010). Secondly, a potential reason for absence of the relationship can be traced back to a self-selection bias for applicant inclusion in the data. As it is merely known who has applied to the job, it is impossible to control for people that have read the job advertisement, but chose to not apply. This intensifies reliance on the assumption that no effect of feminine wording exists for men and makes it impossible to compare gender ratios for applicants and non-applicants as was done in existing experimental research (Born & Taris, 2010; Gaucher et al., 2011). If available, a possible control variable could be gender ratios throughout the consulting industry in the Netherlands, but these are only limitedly available and regarded as relatively equal. Also, when these would have been fully available, it is impossible to correct for people that chose another industry after reading such a job advertisement. All in all, while the experimental research design allows for controlling of many external variables, inability to control for all external factors is inherent to the empirical research design.

Furthermore, with respect to the context of the data, inclusion of HR data of multiple companies with a greater variety of job advertisements could have potentially led to different

predictive results. Also, inclusion of a wider variety of industries could have had different implications as existing research has concluded that masculine and feminine language is present to a larger extent in either male or female dominated industries, respectively.

While potential reasons for absence of a relationship between gendered language and the share of female applicants can be found in the present research, it is important to also consider limited validity of existing experimental research. Three reasons for limited external validity can be identified in two studies by Born and Taris (2010) and Gaucher et al. (2011) mostly with regards to the experimental parts upon which the negative influence of masculine wordings, and nouns and adjectives, respectively, on female job advertisement attraction has been based. Firstly, both studies involved participants to indicate whether they would have the inclination or tendency to apply. It is not implausible that the inclination to apply for a job in an experimental setting without any real-worlds consequences is different from filing an actual application when looking for actual employment which has real-world consequences. Secondly, the population of both studies involved students, which have been argued in the abovementioned studies to be seeking for employment in the near future. However, it is possible that students make different considerations than actual job seekers, which may potentially be dependent on the level of urgency in the latter group. Additionally, in contrast to the student sample in experimental research, the present research involves job seekers in all ages for all positions. Lastly, there is a possibility that the effects of language in a job advertisement are overestimated, potentially due to applicants solely focussing on certain textual areas of a job advertisement when assessing their fit to the job.

While various reasons could have influenced the absence of a relationship between gendered language and the share of female applications in the current research, the attractiveness of and level of inclusion in job advertisements remains important, may it be through the relationship that has been hypothesised by Born and Taris (2010) and Gaucher et al. (2011), or may it be to establish a competitive advantage in being attractive as a company for a diverse range of job seekers. This can be attained by phrasing masculine wordings to a larger extent as verbs instead of nouns and adjectives, which is now only done for 20% of both Dutch and English masculine wordings in consultancy job advertisements in the Netherlands. Namely, phrasing such a characteristic as a verb will incline potential job applicants to consider the extent to which they are able and have the motivation to fulfil the described behaviour, whereas nouns and adjectives

result in a binary consideration of fit. For example, it is recommended that commonly found phrases 'you are analytical', 'you are a leader' and 'you are competitive', are changed to phrases such as 'you can analyse [...]', 'you are able to lead', and 'you can compete'. However, it is recommended that such masculine wording is only used when deemed to be essential for fulfilling the role. Furthermore, as existing research has indicated, essentiality of feminine qualities to a job is often neglected. Hence, emphasising that job requirements include commonly found feminine words 'collaborating', 'understanding', 'support', 'sharing', 'connecting', 'responsibilities' and 'trust', would be beneficial to increase recognition of feminine qualities throughout a company, to decrease the perceived 'lack of fit' between female stereotypical characteristics and typically found masculine job characteristics, and to subsequentially establish a culture in which the importance of gender diversity is recognised. Namely, firms with a culture in which the importance of diversity is emphasised, will optimally reap the benefits of this diversity.

7. Conclusion

In conclusion, while women remain largely underrepresented throughout western companies, attaining a gender diverse workforce should be a central objective due to the vast number of benefits associated with gender diversity. In order to attain corporate gender diversity, female attraction for job openings is essential. Existing experimental research has reported a negative relationship between female job attraction and wording in job advertisements that includes male stereotypical characteristics related to agency, competitiveness, and dominance, as well as phrasing these words as nouns and adjectives, instead of verbs. In the present research word embedding methods LSA, GloVe, and Word2Vec have been employed in order to expand the list of stereotypical masculine and feminine wording in both Dutch and English. Also, by means of an HMM-based POS tagger, an analysis of job advertisements for consultancy jobs in the Netherlands has been made, by which has been concluded that masculine wording in the grammatical form of nouns and adjectives is overrepresented throughout job advertisements. Hence, it is recommended that masculine wordings are increasingly expressed as verbs and that feminine wordings are used increasingly in job advertisements to decrease the perceived 'lack of fit' between job characteristics and female characteristics. Hereby, firms can reap competitive advantage of a more inclusive culture in which feminine qualities are recognised to be essential.

Next to these findings, in this research predictive analyses have been used to assess whether the experimentally established relationship between masculine wordings and female job application rates and job attraction holds in a real-world context. However, throughout several analyses this relationship has not been confirmed in the empirical context of this study.

7.1 Limitations and further research

As existence of the relationship between wordings in job advertisements and gender diversity would have noteworthy consequences for firms, more research is needed to investigate the relationship between masculine wordings and female job attraction and application in a real-world context. As the present study concerns data on one company in the consulting industry for a relatively limited number of job advertisements, it is recommended that future studies examine the relationship for a wider variety of companies and industries, and especially focus on comparing male- and female-dominated with non-gender dominated industries, such as consulting. Also, for

future studies on this topic it is recommended that more data is included on personal characteristics of reviewers and applicants as effects for different ages or work experience are unknown, and as existing research focuses on students with little to no work experience. Regarding the first stages of this research that concern the addition of words based on word embeddings, it is recommended that theoretical linguists are involved to assess the degree to which proposed words fit within existing dictionaries.

With the goal to attain a gender diverse workforce, various other applications of NLP methods could be explored for other stages of the HR process. With regards to the selection stage, word embeddings can be used to find cover letters or résumés that are similar to current successful female (and male) employees. Also, predictive models can be trained on textual features for successful and unsuccessful candidates to predict the hiring chances of new applicants. With such predictive analyses, it should be carefully considered that specific groups of employees are not overrepresented to avoid bias, and such models should be used to assist hiring decisions. However, identifying potential successful female candidates can contribute to improving gender diversity for all levels as chances of promotion are greater.

Another potential HR related application of NLP methods can be found in creating an organisational culture of inclusion, by which the benefits of (gender) diversity can be optimally reaped. For example, in analysing employee surveys and feedback on the culture of inclusion and appreciation for gender diversity, word embeddings and sentiment analysis can be used to measure the development and state of the desired culture and atmosphere. Alternatively, word embeddings can be used to categorize different opinions to address and identify key areas of interest.

Lastly, regarding equitable evaluation and promotion of all employees, evaluation of male versus female employees can be compared, and potential consistent and problematic differences in evaluation can be identified through NLP methods. For example, sentiment, gendered wording through specific words, and usage of grammatical function can be analysed to aid in identifying key issues to address and potential barriers to equitable advancement.

8. References

8.1 Substantive articles

Annabi, H., & Lebovitz, S. (2018). Improving the retention of women in the IT workforce: An investigation of gender diversity interventions in the USA. *Information Systems Journal, 28*(6), 1049-1081.

Bear, S., Rahman, N., & Post, C. (2010). The impact of board diversity and gender composition on corporate social responsibility and firm reputation. *Journal of business ethics*, *97*(2), 207-221.

Bell, S. (2020, September 9). The stock market boost from having more women in management. *Financial Times*. https://www.ft.com/content/12aefc8c-9d5b-4a02-bb88-4eef62d79443

Bond, M. A., & Haynes, M. C. (2014). Workplace diversity: A social–ecological framework and policy implications. *Social Issues and Policy Review*, *8*(1), 167-201.

Born, M. P., & Taris, T. W. (2010). The impact of the wording of employment advertisements on students' inclination to apply for a job. *The Journal of social psychology*, *150*(5), 485-502.

Breaugh, J. A. (2013). Employee recruitment. *Annual review of psychology*, *64*, 389-416.

Brownfield, A. (2020, September 23). P&G shares diversity data of its workforce for first time as it efforts a 40% multicultural workforce. *Business Journal*. https://www.bizjournals.com/cincinnati/news/2020/09/23/p-g-releases-diversity-of-its-own-workforce-for.html

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, public policy, and law*, *5*(3), 665.

Campbell, K., & Mínguez-Vera, A. (2008). Gender diversity in the boardroom and firm financial performance. *Journal of Business Ethics*, *83*(3), 435-451.

Chapple, L., & Humphrey, J. E. (2014). Does board gender diversity have a financial impact? Evidence using stock portfolio performance. *Journal of business ethics*, *122*(4), 709-723.

Daniels, D. P., Dannals, J., & Neale, M. A. (2021). *Do Investors Value Diversity?*. Stanford Business School Working Paper No. 3943.

De Cabo, R. M., Gimeno, R., & Nieto, M. J. (2012). Gender diversity on European banks' boards of directors. *Journal of Business Ethics*, *109*(2), 145-162.

Dezsö, C. L., & Ross, D. G. (2012). Does female representation in top management improve firm performance? A panel data investigation. *Strategic management journal*, *33*(9), 1072-1089.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and social psychology bulletin*, *26*(10), 1171-1188.

Galinsky, A. D., Todd, A. R., Homan, A. C., Phillips, K. W., Apfelbaum, E. P., Sasaki, S. J., ... & Maddux, W. W. (2015). Maximizing the gains and minimizing the pains of diversity: A policy perspective. *Perspectives on Psychological Science*, *10*(6), 742-748.

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.

Graves, L. M., & Powell, G. N. (1995). The effect of sex similarity on recruiters' evaluations of actual applicants: A test of the... *Personnel Psychology*, *48*(1), 85-98.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in organizational Behavior*, *32*, 113-135.

Heilman, M. E., & Caleo, S. (2018). Combatting gender discrimination: A lack of fit framework. *Group Processes & Intergroup Relations*, *21*(5), 725-744.

Herring, C. (2009). Does diversity pay?: Race, gender, and the business case for diversity. *American sociological review*, *74*(2), 208-224.

Johnson, S. K., Hekman, D. R., & Chan, E. T. (2016). If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, *26*(04).

Kossek, E. E., Su, R., & Wu, L. (2017). "Opting out" or "pushed out"? Integrating perspectives on women's career equality for gender inclusion and interventions. *Journal of Management*, *43*(1), 228-254.

Lawson, R. (2015). *Web scraping with Python*. Packt Publishing Ltd.

Marinova, J., Plantenga, J., & Remery, C. (2016). Gender diversity and firm performance: Evidence from Dutch and Danish boardrooms. *The International Journal of Human Resource Management, 27*(15), 1777-1790.

Nguyen, T., Locke, S., & Reddy, K. (2015). Does boardroom gender diversity matter? Evidence from a transitional economy. *International Review of Economics & Finance*, *37*, 184-202.

Pless, N., & Maak, T. (2004). Building an inclusive diversity culture: Principles, processes and practice. *Journal of Business Ethics*, *54*(2), 129-147.

Rubini, M., & Menegatti, M. (2014). Hindering women's careers in academia: Gender linguistic bias in personnel selection. *Journal of Language and Social Psychology*, *33*(6), 632-650.

Sczesny, S., Formanowicz, M., & Moser, F. (2016). Can gender-fair language reduce gender stereotyping and discrimination?. *Frontiers in psychology*, *7*, 25.

Shadrach, D. (2021, May 7). P&G pays tribute to parents in Tokyo 2020 Olympic Games campaign. *Marketing Interactive*. https://www.marketing-interactive.com/p-g-pays-tribute-to-parents-in-tokyo-2020-olympic-games-campaign

Turban, S., Wu, D., & Zhang, L. (2019). When gender diversity makes firms more productive. *Harvard Business Review*, *11*.

Zhang, L. (2020). An institutional approach to gender diversity and firm performance. *Organization Science*, *31*(2), 439-457.

8.2 Methodology articles

Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107-119.

Bloem, J., Fokkens, A., & Herbelot, A. (2019, September). Evaluating the consistency of word embeddings from small data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 132-141).

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., & Viviani, M. (2018). WoLMIS: a labor market intelligence system for classifying web job vacancies. *Journal of Intelligent Information Systems, 51*(3), 477-502.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brill, E., & Pop, M. (1999). Unsupervised learning of disambiguation rules for part-of-speech tagging. In *Natural language processing using very large corpora* (pp. 27-42). Springer, Dordrecht.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992, March). A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing* (pp. 133-140).

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Friedman, J., Hastie, T., & Tibshirani, R. (2017). *The elements of statistical learning* (Vol. 2, No. 10). New York: Springer series in statistics.

Friedman, J. H., Hastie, T. J., & Tibshirani, R. J. (2020). glmnet: lasso and elastic-net regularized generalized linear models, 2010b. *http://CRAN. R-project. org/package= glmnet. R package version*, 4.1-2.

Gladkova, A., & Drozd, A. (2016, August). Intrinsic evaluations of word embeddings: What can we do better?. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 36-42).

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, *35*(4), 579-592.

Highfill, B. (2011). Part of Speech Tagging with Discriminatively Re-ranked Hidden Markov Models.

Hoyle, A., Wallach, H., Augenstein, I., & Cotterell, R. (2019). Unsupervised discovery of gendered language through latent-variable modeling. *arXiv preprint arXiv:1906.04760*.

Jurafsky, D., & Martin, J. H. (2020). Speech and language processing. *Chapter 8: Sequence Labeling for Parts of Speech and Named Entities (Draft of December 30, 2020). Retrieved May 26*, 2021.

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems* (pp. 950-957).

Kumar, A., & Paul, A. (2016). *Mastering text mining with R*. Packt Publishing Ltd.

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, *31*(6), 5-14.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing Ltd.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Lee, S. Z., Tsujii, J. I., & Rim, H. C. (2000, October). Part-of-speech tagging based on hidden Markov model assuming joint independence. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 263-269).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Lovaglio, P. G., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 11*(2), 78-91.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471-484.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, *3*(6), 57-64.

Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, *112*, 340-349.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, *443*(7), 59-72.

Paroubek, P. (2007). Evaluating Part-of-Speech Tagging and Parsing Patrick Paroubek. In *Evaluation of Text and Speech Systems* (pp. 99-124). Springer, Dordrecht.

Perez-Ortiz, J. A., & Forcada, M. L. (2001, July). Part-of-speech tagging with recurrent neural networks. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)* (Vol. 3, pp. 1588-1592). IEEE.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European journal of social psychology*, *49*(5), 871-887.

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327-356.

Ripley, B., Venables, W., & Ripley, M. B. (2016). Package 'nnet'. *R package version*, *7*(3-12), 700.

Roberts, S., & Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, *70*, 198-211.

Selivanov, D., & Wang, Q. (2020). text2vec: Modern text mining framework for r. *Computer software manual] (R package version 0.6). Retrieved from https://CRAN. R-project. org/package= text2vec*.

Sirisuriya, D. S. (2015). A comparative study on web scraping.

Tayal, M. A., Raghuwanshi, M. M., & Malik, L. (2014, January). Syntax Parsing: Implementation Using Grammar-Rules for English Language. In *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies* (pp. 376-381). IEEE.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, *42*(2), 413.

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, 219-232.

Wild, F. (2007, March). An LSA package for R. In *Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07)* (pp. 11-12).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 347-354).

Zhang, H., & Wang, M. (2009). Search for the smallest random forest. *Statistics and its Interface*, *2*(3), 381.

9. Appendices

Appendix A: Existing list of gendered wording

Below, the widely accepted English word lists for feminine and masculine words from Gaucher et al. (2011) can be found. The '-' indicates that all versions of a word (stemmed versions) are considered to be gendered.

Table A1. Existing gendered word list (Gaucher et al., 2011)

Masculine		**Feminine**
activ-	greedy-	agree-
adventurous-	head-strong-	affectionate-
aggress-	headstrong-	child-
ambitio-	hierarch-	cheer-
analy-	hostil-	collab-
assert-	impulsive-	commit-
athlet-	independen-	communal-
autonom-	individual-	compassion-
battle-	intellect-	connect-
boast-	lead-	considerate-
challeng-	logic-	cooperat-
champion-	objective-	co-operat-
compet-	opinion-	depend-
confident-	outspoken-	emotiona-
courag-	persist-	empath-
decid-	principle-	feel-
decision-	reckless-	flatterable-
decisive-	self-confiden-	gentle-
defend-	self-relian-	honest-
determin-	self-sufficien-	interpersonal-
domina-	stubborn-	interdependen-
dominant-	superior-	interpersona-
driven-	unreasonab-	inter-personal-
fearless-		inter-dependen-
fight-		inter-persona-
force-		kind-

Appendix B: Gendered words found in job advertisements

The '-' results from stemming the words and indicates inclusion of all words that start with the stem.

Table B1. Gendered words found in consulting job advertisements

English Masculine	English Feminine	Dutch Masculine	Dutch Feminine
activ-	agreement	actief	afhank-
adventur-	collabor-	actiev-	attentie-
analys-	commit-	analys-	betrok-
analysi-	committe-	analyseert	betrokken
analyst-	connect-	analyser-	betrouw-
analyt-	consider-	analyses-	eerlijk-
analyz-	depend-	analyst-	empathisch-
autonom-	enthusiasm	analytisch-	enthousiast-
autonomi-	enthusiast-	assertief	gemeenschapp-
challeng-	feel-	asstertiev-	gevoel-
compet-	honesti-	assertiviteit	gezell-
competit-	inclus-	autonom-	interperson-
competitor-	independ-	autonomie	klantvriend-
confid-	interperson-	avontur-	loyal-
courag-	manag-	bepal-	loyalty-
decid-	nurtur-	besluitvaard-	ondersteun-
decis-	pleasant-	competenties	ondersteund
determin-	polit-	competitie-	ondersteunt
driven-	respond-	competitief	plezier-
encourag-	respons-	deskund-	prettig-
fight-	share-	doelgericht-	samenwerk-
forc-	sharp-	doeltreff-	samenwerkingsgericht-
individu-	support-	gedrev-	samenwerkingsoploss-
intellectu-	trust-	individu-	samenwerkingsverband
lead-	understand-	individueel	samenwerkt
leader-		individuel-	steun-
leadership-		kracht	stil-
opinion-		krachtig-	toegewijd-
persist-		lef-	verantwoord-
principl-		leid-	verantwoordelijkheidsgevoel
proactiv-		leidend	verbind-
superior-		leider-	verbinder
		leiderschap-	verbindt
		leiderschapsontwikkel-	vertrouw-
		leiding-	vertrouwd-

	leidinggev-	verzorg-
	leidt-	verzorgd-
	logisch-	verzorgt
	mening-	vriendelijk-
	moedig-	warm-
	onafhank-	warmt-
	toonaangev-	
	uitdag-	
	zelfredzam-	
	zelfstur-	
	zelfverzekerd-	

Appendix C: Similar words from word embeddings

In bold, one can find the words that are discussed in section 5.1 and that are consequently added to the existing gendered dictionaries.

Table C1. English words similar to existing masculine wording dictionaries

	GloVe	**Word2Vec: CBOW**	**Word2Vec: SkipGram**	**LSA**	**Mean top 10**
1.	translat	nonsens	**steer**	**drive**	take
2.	enabl	pictur	displai	approach	approach
3.	take	**steer**	ambigu	take	curios
4.	comprehens	add	theme	insight	pictur
5.	insight	broader	implic	mid	chapter
6.	innov	craft	geographi	focus	comfort
7.	approach	differenti	holder	discoveri	enabl
8.	impact	valuabl	pictur	econom	**persuas**
9.	creativ	**driver**	**persuas**	leav	geographi
10.	colleagu	**led**	pitch	consist	**drive**

Table C2. English words similar to existing feminine wording dictionaries

	GloVe	**Word2Vec: CBOW**	**Word2Vec: SkipGram**	**LSA**	**Mean top 10**
1.	**colleagu**	stand	like	**encourag**	**teamwork**
2.	motiv	authent	stand	world	**encourag**
3.	opportun	pride	**teammat**	equal	cultiv
4.	differ	cultiv	fundament	bring	express
5.	environ	dream	theme	divers	huge
6.	abroad	**like**	talk	marit	multicultur
7.	proactiv	exchang	ethic	gender	divers
8.	outstand	accomplish	pride	express	energet
9.	**team**	energ	nonsens	workplac	authent
10.	authoris	fantast	divers	race	valu

Table C3. Dutch words similar to existing masculine wording dictionaries

	GloVe	Word2Vec: CBOW	Word2Vec: SkipGram	LSA	Mean top 10
1.	**visie**	databyt	synergie	structurel	**visie**
2.	**bepaalt**	gefocust	gunfactor	positie	coachend
3.	**aanpak**	klantomgev	inventief	economisch	inventief
4.	**stevig**	inventief	**sturend**	rak	betekenis
5.	positiev	synergie	gecreerd	waarder	argument
6.	oplossingsgericht	realiteit	expertisegebied	strategisch	stijl
7.	argument	**topper**	**slagvaard**	stimuleert	synergie
8.	trackrecord	betekenis	bedrijfsinformatie	procent	alignment
9.	win	geïnspireerd	meedenkt	agro	trackrecord
10.	werkelijk	**eigenwijs**	tikkeltj	cultuurverander	**sturend**

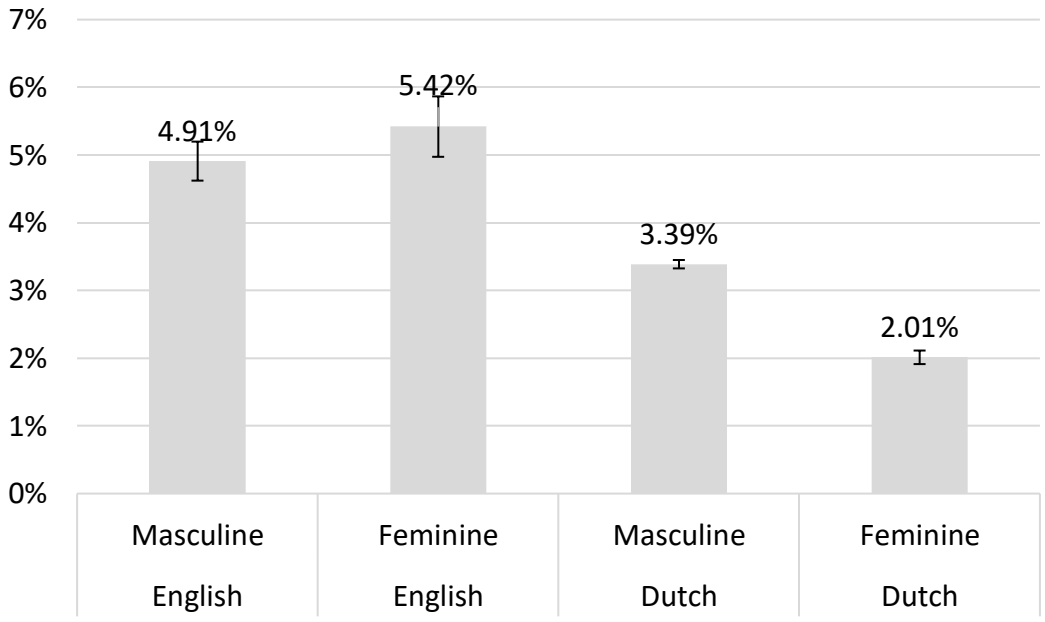
Table C4. Dutch words similar to existing feminine wording dictionaries

	GloVe	Word2Vec: CBOW	Word2Vec: SkipGram	LSA	Mean top 10
1.	schakel	synergie	beweegt	**coachend**	**coachend**
2.	managementniveau	hoogstaand	gunfactor	**gezamen**	welvaart
3.	**hecht**	**respect**	vakkund	directeur	gesprekspartner
4.	werkt	decentral	hoogstaand	primair	managementniveau
5.	fantastisch	databyt	slagvaard	rad	slagkracht
6.	jong	werkdruk	verkoper	belang	stijl
7.	aanpak	klantomgev	bediend	niveau	**samengewerkt**
8.	stabiliteit	oor	synergie	spil	doortast
9.	tegelijkertijd	slagvaard	dichtbij	**samensprak**	razendsnel
10.	continuteit	expertisegebied	bovenan	**samenlev**	**samensprak**

Appendix D: Descriptive statistics

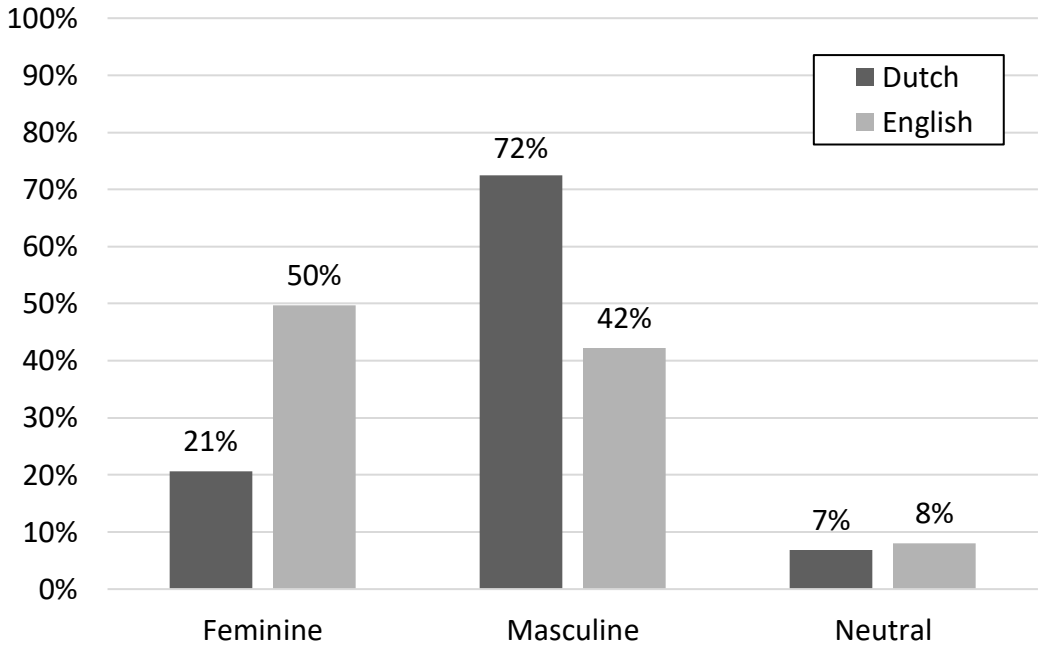
According to the method of Gaucher et al. (2011) the prevalence of gendered language as percentage of total words can be found below. In Figure 1, it can be seen that gendered language is more prevalent in English job advertisements than in Dutch job advertisements in the Netherlands.

Figure D1. Prevalence of gendered words in consulting advertisements (percentage of total meaningful words). Standard errors are represented in bars.



According to a descriptive method of Dinan et al. (2020) each job advertisement was classified as masculine, feminine or neutral. It was classified as masculine if it contained more masculine than feminine words, and vice versa, and when the number of masculine and gendered words was equal, including zero, it was labelled as neutral (Dinan et al., 2020). In Figure 2, it can be seen that masculine-coded job advertisements are most prevalent in Dutch job advertisements, while in English job advertisements 50% of observations are coded as feminine. It is good to note that this arbitrary cut-off point does not indicate the actual number of masculine or feminine words, but merely the ratio of this word usage.

Figure D2. Gendered classification of consulting job advertisements (percentage of job advertisements classified as feminine/masculine/neutral)



Appendix E: POS tags

Relative frequencies in percentages of different grammatical functions of gendered wording can be found below. N.B. 'adj.' is short for adjective.

Table E1. General tag division for gendered words

	English	**English**	**Dutch**	**Dutch**
	Masculine	Feminine	Masculine	Feminine
Noun	51.2%	54.0%	49.6%	35.5%
Adjective	20.2%	12.7%	21.3%	36.4%
Verb	20.0%	27.3%	20.6%	24.3%
Other	8.6%	6.0%	8.6%	3.8%

Table E2. Dutch masculine wordings

Tag division per gendered word for words occurring over 25 times (*n = 1304)*

Form	**Word**	**Term Count**	**Document Count**
noun	uitdag	588	475
adj	analytisch	495	434
verb	uitdag	448	392
verb	gedrev	383	309
adj	actief	344	281
noun	analys	316	258
noun	leid	239	184
noun	aanpak	214	177
verb	leid	209	194
noun	visie	201	143
verb	analys	195	178
noun	kracht	142	129
adj	stevig	134	126
noun	competenties	119	103
verb	bepal	113	98
adj	individu	98	87
adj	bepaald	94	92
adj	onafhank	81	75
adj	analys	71	68
noun	lef	71	70
verb	toonaangev	65	65
noun	analytic	63	46

noun	mening	56	53
verb	bepaald	53	50
verb	bepaalt	50	48
adj	deskund	45	43
adj	competitie	28	28

Table E3. Dutch feminine wordings

Tag division per gendered word for words occurring over 25 times (*n = 1304)*

Form	Word	Term Count	Document Count
adj	verantwoord	574	437
noun	verantwoord	475	400
verb	ondersteun	458	408
noun	samenwerk	455	371
adj	enthousiast	409	369
noun	ondersteun	342	288
verb	betrok	307	269
adj	afhank	277	247
verb	samenwerk	256	235
adj	gezell	197	185
noun	plezier	187	161
verb	verbind	163	148
adj	hecht	134	128
verb	verantwoord	131	128
verb	verzorg	128	118
adj	prettig	126	122
adj	gezamen	121	115
noun	vertrouw	121	107
noun	gevoel	107	100
noun	verbind	106	97
noun	enthousiast	105	97
noun	betrok	82	81
adj	betrouw	67	61
verb	hecht	63	61
noun	gezell	61	58
adj	warm	54	52
adj	eerlijk	48	40
noun	samenlev	46	42
adj	betrok	44	44
noun	warm	42	26
verb	vertrouw	36	36
noun	betrouw	30	30
adj	ondersteun	25	25

Table E4. English masculine wordings

Tag division per gendered word for words occurring over 25 times (*n = 432)*

Form	Word	Term Count	Document Count
verb	lead	350	249
noun	lead	334	252
noun	challeng	280	209
noun	analys	273	196
verb	drive	233	182
noun	activ	221	154
adj	analyt	180	146
noun	analyt	174	74
adj	challeng	150	124
adj	compet	126	111
noun	decis	122	103
noun	compet	101	90
noun	drive	84	80
noun	individu	69	62
verb	challeng	69	67
adj	activ	59	49
adj	individu	53	46
verb	analyz	53	49
verb	determin	47	42
noun	principl	37	32
adj	lead	36	30
noun	confid	27	25

Table E5. English feminine wordings

Tag division per gendered word for words occurring over 25 times (*n = 432)*

Form	Word	Term Count	Document Count
noun	team	1683	635
verb	support	478	326
noun	respons	341	291
noun	support	311	161
adj	respons	214	157
noun	understand	169	121
verb	like	169	128
verb	understand	130	106
noun	collabor	112	95
verb	share	111	101
verb	collabor	88	77
verb	trust	83	69
verb	consider	81	70
verb	feel	81	75
adj	enthusiast	73	60
noun	share	73	70
verb	connect	71	67
adj	commit	70	56
adj	inclus	68	47
noun	connect	58	55
verb	encourag	52	50
verb	depend	48	48
noun	commit	47	43
verb	respond	43	39
adj	interperson	42	42
noun	trust	37	35
noun	consider	35	34
adj	collabor	34	31
verb	commit	30	26
noun	inclus	29	25
noun	agreement	28	26
verb	nurtur	27	26