UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science Master thesis

Natural Language Processing in Resume Data: The Interplay Between Gender and Occupation on Resume Writing Style

First examiner:

Ayoub Bagheri

Second examiner:

Huyen Nguyen

Candidates:

Malka Aktepe

Sara Marti Marcet

Maike Lea Vaitea Weiper

July 4, 2023

Abstract

Resumes constitute an important part of forming an impression of candidate employees during the hiring process. They are shaped by the interplay between societal, personal, and occupational values. One important personal value or norm that shapes the writing style of a resume, is the gender of the candidate employee (Guadagno & Cialdini, 2007). Previous research shows that women tend to communicate in a more communal way, while men tend to communicate in a more agentic way. In our current society, the feminine gender norms (e.g. care) seem to be less in line with occupational values than the masculine gender norms (e.g. competitiveness; Eagly and Karau, 2002). Nevertheless, to our knowledge, previous research has only investigated gender differences in resumes in male-dominated occupations, like IT careers (Parasurama et al., 2022). Therefore, in this research, we have investigated the interrelationship between gender and occupation in careers that are male-dominated, gender-balanced, and female-dominated.

In our first study, we have identified the most important textual features that differentiate resumes written by men from those written by women in more than 1700 resumes (Yang et al., 2022). Our results indicate that women use more communal language in their resumes with words such as "assist" or "care" being the most predictive ones, while the features that predicted male resumes were clearly agentic or pointed out that they are the ones who apply or have experience in higher positions or technical fields. Examples of those words are "manager" or "engineering".

In a second study, we tested multiple machine learning algorithms to predict gender from the resume texts in different occupations. We found that all models can predict gender from the resume text. The traditional and word-embedding models alongside DistilBERT perform well in balanced occupations, but fall short when the data was more female- or male-dominated. RoBERTa and Longformer showed steady performance across

all occupations demonstrating the capabilities of newer transformer models.

In our third study, we investigated to what extent men and women conform to their respective gender norms and whether this differs across occupations. We found that women communicate significantly less gender-congruently in male-dominated occupations compared to gender-balanced and female-dominated occupations. Similarly, men communicate significantly less gender-congruently in female-dominated occupations compared to gender-balanced and male-dominated occupations. Thus, even though people might experience social and economic penalties if they communicate in a gender-incongruent way, they still use a different communication style depending on the occupational context to which they apply.

To sum up, in this research project, we successfully trained multiple machine learning algorithms to predict gender from textual features in resumes. We found that their performances and predictive scores differ across occupations. In our discussion, we discuss the implications of these results with regard to societal norms and values surrounding gender and occupation and how these influence the hiring process. We argue that our results highlight the need for gender- and context-aware tools to help employers in selecting appropriate candidates for hiring in a fair manner.

Contents

1. Introduction

Written together

Resumes and cover letters are an integral part of the recruitment procedure, regardless of the latest developments in future job performance assessment technologies (Risavy et al., 2017). They provide a unique first opportunity to present oneself to potential new employers in a favorable manner, which has been underlined by research on impression management in the recruitment context (Bolino et al., 2016). Impression management (or self-presentation) is the process by which individuals attempt to control or influence the perceptions of others about them (Baumeister, 1982; Gardner & Martinko, 1988). During recruitment, applicants not only adapt their appearance in job interviews to match the hiring context but they also do that in their resumes i.e. adjusting their communication style to create a positive and compelling self-image to potential employers (Bolino et al., 2016). Given that impression management in recruitment is common practice and thereby expected to some extent, candidates need to strike a balance between authenticity and strategic self-presentation to maximize their chances of success.

Gender differences can have a profound impact on how individuals navigate the delicate balance between self-presentation and societal expectations. Extensive research suggests that during the recruitment process, men and women often adopt distinct strategies in portraying themselves (Bolino et al., 2016; Parasurama et al., 2022). While men tend to emphasize their achievements, competencies, and assertiveness, aiming to project confidence and leadership qualities, women are typically found to downplay their accomplishments and display more communal traits, such as being nurturing and cooperative (Guadagno & Cialdini, 2007; Parasurama et al., 2022; Prentice & Carranza, 2002; Smith et al., 2013).

These gender differences are reflective of general communication patterns and societal expectations of the different genders. For instance, the existence of gender roles, which are consensual beliefs about the attributes of men and women, as defined by Eagly, 1987. Such roles are also normative (Eagly, 1987; Eagly et al., 2000), i.e. people are generally expected to participate and act in a manner consistent with their culturally defined role. Gender roles also describe some aspects of the behavior that are believed to be desirable for men and women (Broverman et al., 1972; Williams & Best, 1990).

Two concepts that are highly associated with gender-desirable behaviors, and therefore with gender roles, are agency and communion (Bakan, 1966; Hsu et al., 2021). [1] Traditionally, agency-related traits have been more strongly associated with masculinity or male gender roles (Bakan, 1966; Spence et al., 1975; Wood & Eagly, 2015). People high in agency may be more likely to engage in assertive communication styles, focus on achieving individual goals, and prioritize their own needs and desires in social interactions. In contrast, communion-related traits have been associated with femininity or female gender roles. People high in communion are more likely to engage in affiliative communication styles, emphasize building relationships, and prioritize the needs and emotions of others in social interactions. In a recent meta-analysis, Hsu et al., 2021 showed that even though the magnitude of gender differences in agency and communion decreased over time, the gender difference remained larger than that in many other behavioral or cognitive aspects that are associated with gender.

The traits associated with the different genders are, in fact, relevant in organizational contexts. The more 'masculine' qualities of assertiveness and leadership are often not only expected but even seen as necessary for obtaining higher-ranking positions in organizations (Bolino et al., 2016). The misalignment between expectations of individuals in managerial roles and societal gender norms adds a layer of difficulty for women to climb the ca-

[1] Agency refers to behaviors associated with assertiveness, independence, and the pursuit of individual goals and interests. Communion refers to behaviors associated with empathy, nurturing, and fostering interpersonal connections.

reer ladders (Rudman & Phelan, 2008). Furthermore, in high-ranked managerial positions, the proportion of men is much higher than women e.g. Just a 4,8% of women are CEOs in Europe (Steffens et al., 2019). The consequences of their self-presentation strategies can force people to present themselves in a more (or less) gender-normative way, both to succeed in recruitment and to avoid the backlash of not conforming to gender norms (Moss-Racusin & Rudman, 2010). More difficulties for women appear to be true in male-dominated fields like IT, where the social norms of the environment also often clash with traits and behaviors typically associated with women (Guadagno & Cialdini, 2007). Moreover, female candidates compete directly with male candidates, who generally tend to present themselves as more agentic than female applicants (Parasurama et al., 2022). This could force women to present themselves differently in this context than they would present themselves in another less male-dominated occupation. Conversely, in female-dominated fields e.g. healthcare centers, there are mixed findings on whether or not men face gender norm related challenges (Heilman & Eagly, 2008; Parasurama et al., 2022). To the best of our knowledge, little research has yet investigated differences in self-presentation across different genders and occupations simultaneously.

Therefore, this research revolves around the central research question: To what extent do people from different genders present themselves differently in recruitment contexts across different occupational groups? We will systematically answer this question in a three-fold manner. First, Sara will investigate how and to what extent people from different genders generally differ in their self-presentation. Specifically, she will train different machine learning models to examine which features in resume texts are the most predictive of gender. Second, Malka will explore how different machine learning models can accurately identify and predict gender given the resume texts across occupations. The goal of her work is to assess the reliability of these models in determining and predicting gender, thereby uncovering potential biases in different fields of work. (Yang et al., 2022). Lastly, Maike will investigate to what extent people present themselves in a gender-conforming way across the different occupational groups. Us-

ing predictive machine learning algorithms, she will investigate how and whether self-presentation patterns in resumes conform to the styles of same-gender applicants and compare this across different occupations.

The results of this research shed more light on the biases that both human recruiters, as well as those based on artificial intelligence (AI), have to face during the hiring process. We found that there are gender differences in self-presentation and that algorithms can detect implicit information when gender is removed from the models. The textual features the algorithms base their predictions on are gender roles and agency / communal writing styles. Furthermore, traditional and more advanced machine learning models are also able to accurately predict gender across occupational groups. Especially newer models like RoBERTa and Longformer showed remarkable performance even in a heavily skewed and small dataset. Lastly, we also found that people write their resumes in a less gender-congruent style when applying for a job in an occupational sector that is dominated by another gender. Specifically, women wrote less gender-congruently when applying for male-dominated jobs and men wrote less gender-congruently when applying for female-dominated jobs.

This is especially relevant due to the fact that occupational gender segregation still plays an immense role in the organizational context, starting at the time of recruitment (He & Kang, 2021). For instance, the company Amazon implemented an AI to substitute people in the recruitment process yet it was found that this technology was consistently choosing male candidates over female candidates with similar backgrounds (Dastin, 2018). The algorithm that they implemented removed the variable "gender" from the data they were using, but there were still some textual features that were enough for the algorithm to guess the gender of the candidate. Hereby in our study, we would like to provide new insights regarding the topic and if in fact, we can detect gender differences in self-presentation in resumes.

The next sections present the data used to answer our research questions and introduce the models that we trained on this data to predict gender from the text. We then discuss relevant work and go into detail about our

methods and results for each of the three sub-questions that were outlined above. Finally, we combine the results of our analyses and discuss them in the broader context of self-presentation during recruitment.

1.1 Data

To investigate our research questions, we made use of existing data collected by Yang et al., 2022. Participants of their study were recruited via Prolific in the US and asked to write their CVs as if they were applying for a promotion in their current occupational field. The data that we received from the original authors was anonymized. Access to the data can be requested from the authors of the original paper.

It is important to note that our access was limited to pre-processed data and not the raw full-text CVs. In contrast to the descriptions of the original authors (see page 2 of Yang et al., 2022), it remains unclear what type of pre-processing was done as stopwords were still present and the text had not been lemmatized yet, but simultaneously there were no full coherent sentences for us to extract.

1.1.1 Data Exploration

The data (N = 1789) contains 8 variables: gender, occupation, career objective, professional experience, education, qualifications and training (optional field for the subjects to fill in), skills/attributes, and other information (also an optional field). All of the fields except for gender, occupation, and education have missing data, especially the optional fields: "Other Information" has 1431 missing values. However, the main fields (occupation and education) are complete without any missing values. For this reason, we decided that we will not make use of any data imputation techniques.

Even though the data is roughly balanced (N_{women} = 903; N_{men} = 886) regarding gender, figure 1.1 shows that certain occupations are over-represented by one of the genders. However, the gender representations in our data are roughly in line with the benchmark data from the US Labor

Statistics 2021 as cited in Yang et al., 2022 (Appendix B).

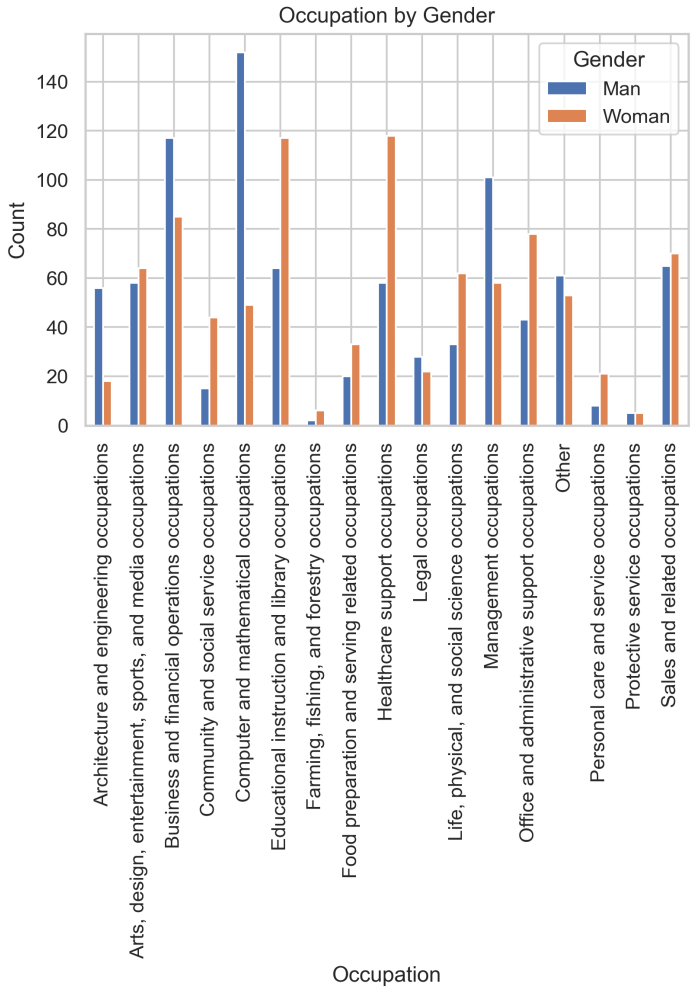


Figure 1.1: Gender distribution per occupation in the data

1.1.2 Data Cleaning

In addition to the preliminary cleaning performed by the original authors, we performed some additional cleaning of the data. We considered this an important procedure because we noticed many rows that contained non-informative and poor-quality data. Some examples will be given hereinafter. The reason to choose the steps described below is mainly due to common practice in the NLP field (such as deleting words from different languages, stopwords, lemmatization or deleting punctuation), but also we made decisions more specific to the data we had. For instance, instead of ignoring the variables with missing values, we included the information

when it was available to a general variable called "text", just to have more words to feed into the models.

Specifically, we performed the following six steps for cleaning: First, we filled in all the missing data with an empty value, to make sure all the rows contain information in all the cells. This means that we used the available information of those variables that contained missing values, as has been mentioned before, just to have more information for the models (since the amount of data itself is not large). Second, we concatenated the different (obligatory) parts of the CVs so that they became one text for the analyses. Third, we removed all non-English words by checking for each word, whether they were part of the 'words' dictionary of the nltk corpus (Bird et al., 2009). Due to restrictions during data collection (like a minimum number of characters), some participants provided very low-quality data, especially towards the end of their texts (e.g. 'bananananananananananana [...]' or 'Lorem ipsum dolor [...]'). Even though eliminating non-English words was not free of wrong classifications, the data quality did improve a lot. To confirm this, we did a qualitative manual evaluation, going through the main features again and checking if the "non-words" (words like "bananananananana" disappeared) were deleted, which was the case. Fourth in data cleaning, we removed the anonymized parts-of-speech tags (e.g. '[ORG]', '[PERSON]'). Fifth, we removed digits and punctuation. Lastly, all words were lowered and lemmatized according to the lemmatization procedure of Spacy (Honnibal et al., 2020). This version of the text was used as the input to the models.

1.2 Models

To answer the research questions, several models from the Natural Language Processing (NLP) field have been trained and tested to predict gender from our textual data. The selected models belong to several categories: statistical-based, dictionary-based, vector representations, and transformer-based.

1.2.1 Baseline Models

The baseline models are the most basic models which will be the element to compare with the more sophisticated NLP models and to see if more complex models also perform better.

The Bag of Words (BoW) is a way of representing text and extracting features that only take into consideration the simultaneous occurrence of words in documents without paying attention to the order of these words. It is the simplest method to represent textual features statistically, but it still can provide information about the content of the text. To create the BoW, the method "Count Vectorizer" from Scikit-learn in Python was used, which takes into account the frequency of the words in the text, creating this "bag" of features. The resulting frequencies are the input to feed the classification models to predict the gender of the participant.

1.2.2 Models Based on TF-IDF

The term frequency-inverse document frequency, better known as TF-IDF, is a statistical measure that tries to solve one of the main drawbacks of the BoW. In a BoW representation, all the words are considered just as relevant, since we only take into account their frequency in a corpus of documents. TF-IDF, however, does not only contemplate the term frequency within one document (TF). It also takes into account the inverse document frequency (IDF). This measure indicates how unique or rare a term is across all documents in the corpus. It is calculated as the logarithm of the total number of documents divided by the number of documents containing the term. Terms that occur in many documents have lower IDF values, while terms that occur in fewer documents have higher IDF values. The IDF component helps to give more weight to terms that are rare and potentially more informative and less weight to terms that are very frequent across all documents (like 'the', 'I', ...).

Just as it has been done with the baseline models, Logistic Regression (LR) and Support Vector Machine (SVM) were used for classification. To increase the performance of SVM, we used hyperparameter tuning for the

parameters C (extent of avoidance of classification errors), kernel (linear, rbf, and polynomial), gamma, and degree. The latter two hyperparameters are relevant for rbf and polynomial kernels respectively and determine the degree of curvature in the support vectors (scikit-learn developers, 2023). The best-performing model was a second-degree polynomial SVM with $C = 10$.

1.2.3 Dictionary-Based Models

The Language Inquiry and Word Count (LIWC) is a dictionary-based method developed by Tausczik and Pennebaker, 2010 that has been used extensively in previous research (Boyd et al., 2022; Gaucher et al., 2011; Pennebaker et al., 2015; Pietraszkiewicz et al., 2019; Ponizovskiy et al., 2020). For this research, we have used the most recent version of the original LIWC dictionary (Boyd et al., 2022) along with the dictionaries to detect agency and communion (Pietraszkiewicz et al., 2019), masculine and feminine communication (Gaucher et al., 2011), and personal values (Ponizovskiy et al., 2020). The LIWC method calculates for each keyword, the extent to which a participant used words that are part of the respective dictionary. These scores were used as the input to LR and SVM models. To improve the performance of the models in classifying each text as written by a man or a woman, in the final analyses, we included the features of all dictionaries mentioned above, except for the outdated version from 2015. Due to computational limitations, we were not able to perform a grid search on the hyperparameters.

1.2.4 Models Based on Vector-Representations

In this research study, we focused on comparing the performance of Word2Vec and GloVe, two popular word embedding models, for predicting gender based on resume text (Mikolov et al., 2013; Pennington et al., 2014). These word embedding models differ significantly from traditional bag-of-words models and TF-IDF, as they aim to capture the semantic relationships between words in a text corpus, going beyond simple word frequency-based

representations.

The Word2Vec model was implemented using the skip-gram approach. Skip-gram accomplishes this by predicting context words from the given target word, thereby accounting for the possible contextual usage of each term in the corpus. The vector length for the Word2Vec model was specified to be 100 dimensions, a value that offers a reasonable balance between computational efficiency and the level of semantic detail captured in the vectors.

On the other hand, the GloVe (Global Vectors) model, unlike Word2Vec, uses global statistical information by constructing a word co-occurrence matrix from the given corpus. It then applies matrix factorization to provide the word embeddings (Pennington et al., 2014). GloVe offers an advantage in capturing semantic meaning based on a global context. In this research, we utilized the 100-dimensional pre-trained vectors derived from Wikipedia. When it comes to classifying resume data, these 100-dimensional pre-trained vectors from Wikipedia are particularly useful. Given the broad spectrum and adaptability of these vectors, they can efficiently interpret and categorize the wide variety of terms and expressions found in resumes.

In our research, we used both Word2Vec and GloVe to analyze resume texts. For each resume, we took the average of all the word vectors generated by the models. This process allowed us to obtain a single vector representation per resume, taking into account the semantic information encoded within the word vectors and the contextual relationships between words. These averaged vectors were then used as inputs for the LR and SVM classifiers.

1.2.5 Transformer Models

Generally, transformer-based models require raw data as input. However, our study was conducted with pre-processed data, which required an additional round of cleaning due to some low-quality responses from participants (as detailed above). We refer to this post-processing data as 'cleaned

data', while the original pre-processed data we received is referred to as 'uncleaned data'. The cleaned data resembled the information typically found in resumes more closely than the uncleaned data. Therefore, we decided to train each of our transformer models using both the uncleaned and cleaned data for comparison. Since the performance metrics did not differ much from each other (see table 1.1), we decided to use the models trained on the cleaned data for further analyses.

We focused on comparing the performance of three transformer models in predicting gender based on resume text. Our analysis encompasses DistilBERT, RoBERTa, and Longformer, evaluating their distinctive characteristics, strengths, and potential limitations in relation to the task. These models are all based on the transformer architecture, a type of model that uses attention mechanisms to better understand the context of words in a sentence (Vaswani et al., 2017). Specifically, the transformer architecture allows these models to focus on different parts of the input data depending on their relevance to the task at hand, thereby enabling them to handle complex dependencies in the data.

DistilBERT, a compact and streamlined version of BERT, demonstrates proficient handling of resume text. Despite its smaller size, DistilBERT offers efficient inference, making it ideal for processing large volumes of resumes (Sanh et al., 2019). However, compared to RoBERTa and Longformer, DistilBERT may face challenges in capturing long-range dependencies and subtle contextual nuances within resumes. Despite these limitations, DistilBERT provides valuable insights into the overall resume context, although it may not extract nuanced elements as effectively.

RoBERTa, an enhanced variant of BERT, exhibits superior performance when applied to complete resumes (Liu et al., 2019). Leveraging additional training techniques, RoBERTa excels at understanding contextual relationships. It effectively extracts detailed information, identifies essential skills, and understands the complete story of the resume. With the ability to handle both short and long-range dependencies, RoBERTa is well-suited for comprehensive analysis of resume content. Its performance improvements

make it a desirable choice for various resume-related tasks, including skill extraction, summarization, and classification.

Longformer, on the other hand, is specifically designed to handle long-range dependencies and the extensive context in complete resumes (Beltagy et al., 2020). By incorporating a sliding window attention mechanism, Longformer captures information efficiently across the entire resume. This ensures a comprehensive understanding of the resume text, minimizing the risk of overlooking important details. With its adept handling of large context windows, Longformer excels in an in-depth analysis of resume content, including identifying relationships between sections, recognizing important experiences, and understanding the overall structure of the resume.

In summary, each model offers unique advantages. DistilBERT provides rapid processing for large resume volumes, RoBERTa delivers enhanced performance and a nuanced understanding of resume content, and Longformer excels in managing long-range dependencies and extensive context. The choice of a specific model depends on the specific requirements of the resume processing task, balancing considerations of speed, performance, and depth of analysis.

1.2.6 Evaluation

For the bag-of-words and word-embedding models, we used a train/test split of 80/20. Upon training these models, we applied Logistic Regression (LR) and Support Vector Machines (SVM) classifiers. These classifiers served as the basis for predicting gender based on their respective inputs.

As for the transformer models, we used a train/validation/test split of 80/10/10. After training and validating the data, we proceeded to evaluate the performance of each transformer model based on the hold-out test set to assess their potential in predicting gender accurately from resume text.

During the evaluation phase of all models, we used a variety of metrics to assess how accurately each transformer model could predict gender from resume text. Alongside accuracy, which shows how many predictions were correct, we used precision, recall, F1-score (a combination of precision and

recall into a single metric that works well with imbalanced data), and Area Under the Curve (AUC) as key performance indicators.

We also created visual representations of the model's performance through the Receiver Operating Characteristic (ROC) and Precision-Recall curves. The ROC curve shows us the balance between correct and incorrect predictions at different settings, while the Precision-Recall curve shows the relationship between precision and recall as we change these settings. These diagrams give us a deeper understanding of how the models perform under different conditions.

By using these metrics and visual aids, we aimed to thoroughly understand the strengths and limitations of each model in predicting gender from resume text. Our evaluation approach blended both numerical measurements and visual tools to provide a detailed assessment of each model's effectiveness.

1.2.7 Code Availability

The code used for this study is accessible on GitHub at https://github.com/httn22/ADS_project_2023

1.3 Classification Results

As can be seen in Table 1.1, the accuracy and F1 scores are best for the TF-IDF models as well as the Longformer model. The baseline models and models based on LIWC scores perform worst. Other performance metrics and plots are presented in Appendix A.

Model	Classifier method	Accuracy	F1
Baseline (Bow)	LR	0.65	0.65
	SVM	0.64	0.63
TF-IDF	LR	0.70	0.72
	SVM	0.72	0.73
Word2Vec	LR	0.70	0.70
	SVM	0.69	0.70
Glove	LR	0.71	0.73
	SVM	0.65	0.69
LIWC	LR	0.63	0.65
	SVM	0.63	0.64
	Data preparation		
DistilBERT	Uncleaned text	0.66	0.65
	Cleaned text	0.65	0.57
Longformer	Uncleaned text	0.73	0.74
	Cleaned text	0.68	0.69
RoBERTa	Uncleaned text	0.68	0.61
	Cleaned text	0.71	0.68

Table 1.1: Performance results for each model

2. Gender differences in self-presentations: how algorithms detect it?

Written by Sara

2.1 Related Work

Communication styles have been found to differ between genders in several fields. In the scientific sphere, DeJesus et al., 2021 found that generic language (giving statements that are timeless and give the idea that is universally true) is much more successful to achieve a higher number of citations. This is relevant because the authors also detected that men use generic language more extensively than women, and therefore it creates a gap in scientific publications for women in contrast to men. Aligned with the research conducted by DeJesus et al., 2021 on this subject, Kolev et al., 2020 discovered that the way communication patterns used by women have a direct impact on why they are less represented in STEM careers (science, technology, engineering, and math): women use more specific and narrow language expressions in comparison to men, resulting in a reduced used generic language. This contributes to the persistence of the gender gap in STEM fields.

In political sciences, research suggests that female legislators utilize arguments that involve personal experiences and concrete language expressions whereas male politicians use adversarial argumentation (persuasive communication) and abstract communication (Hargrave & Langengen, 2021). Joshi et al., 2021 also found that men in general use abstract communication more often, resulting in women having more difficulties to reach positions of power and status. Leaving behind politics, (Nguyen, 2021) also pointed out that the persuasion techniques employed by men and women

in a debate speech are different, resulting in female debaters using a more personal focus, just like Hargrave and Langengen, 2021 found out for female legislators.

The previously mentioned findings are regarding spoken communication. When it comes to writing speeches, Horbach et al., 2022 concluded that there are no significant differences between men and women. However, if we focus on the relevant topic for our research, Parasurama et al., 2022 found that there is gendered information in resumes and therefore algorithms can learn to distinguish resumes from men or women. For this reason, and considering that there is a gap in the literature regarding gender differences in self-presentation in resumes, this paper will introduce new evidence regarding the topic. To be more specific, the research question that will be tried to be answered in this paper is the following: "Which natural language processing models are most effective in identifying gender-specific differences in resumes?", and also "Which features in resume text are the most predictive of gender?".

2.2 Data

The data that has been used to investigate these research questions and how we treated it is described in section 1.1.

2.3 Method

To accomplish the objective, we pre-processed the data as it is described previously in section 1.1. This type of pre-processed data in the shape of clean tokens can only be used for the baseline models, TF-IDF, word embeddings, and LIWC. For the transformer-based models, the most optimal solution would be to use the whole text since these models learn from the context around the words. However, the received data was already partially tokenized, and thereby the outcome of those models will not reach the full potential compared to the utilization of raw text. Despite this drawback, we still pre-processed the date one step further. The details of how

this was done are explained in the "Data" section. Once the data was ready to be utilised, we applied the aforementioned NLP techniques. The choice of such a wide range of models is to see if more complex models can learn further differences compared to more simple models such as TF-IDF. This is something that (Parasurama et al., 2022) suggested in their paper as a way to have a range of comparisons. The next step was to train the classifiers. To get better performance, we used hyperparameter tuning techniques such as Grid Search from Scikit-learn. Once we had the results of the models, we printed out the metrics, such as accuracy, F1 score or AUC with the ROC curves to see the performance. We considered as a threshold that a score over 0.5 in accuracy means that the model is already learning gender differences. It is important to note that we can accept accuracy as an informative metric because the data is balanced between the two prediction labels (men/women).

Since it is of interest for this specific research question to know what are the differences in self-presentation between genders, the features that better predict the classification have also been extracted together with the coefficients of the logistic regression and the SVM models. This approach enables us to identify the words that significantly contribute to the differentiation in self-presentation between men and women if indeed there are differences.

2.4 Results

This section presents the key findings and outcomes obtained from the analysis. To better show the main results of each of the models, table 2.1 presents in an organized way each model with its corresponding performance metrics of accuracy and F1. Note that the best models are TF-IDF with SVM fine tuning (accuracy=0.72, F1=0.73) and the Longformer (accuracy=0.73, F1=0.74). To check further performance metrics such as the confusion matrixes per model or their ROC curve with the corresponding AUC measurement, check Appendix A. In general, the AUC for the models range between 0.66 and 0.82. The lowest score belongs to the dictionary-based models (LIWC) and the highest to the RoBERTa transformer-based model.

Table 2.2 presents the main features for three of the techniques used: TF-IDF, Word2Vec and Longformer. We have only chosen these three techniques for two reasons. Firstly, TF-IDF with SVM classifier and Longformer are the models that could better detect the differences between genders as it is shown in table 2.1. Secondly, it was of interest to see what it was found with the other methodologies, but considering that the baseline model and the LIWC do not perform very well, we have only chosen word embeddings. Glove is better than Word2Vec, but since it we have used a pretrained dictionary, we could not create the functions to extract the main features due to time constraints. Therefore, Word2Vec was the second better option.

Model	Classifier method	Accuracy	F1
Baseline (Bow)	LR	0.65	0.63
	SVM	0.64	0.63
TF-IDF	LR	0.70	0.72
	SVM	0.72	0.73
Word2Vec	LR	0.70	0.70
	SVM	0.70	0.71
Glove	LR	0.71	0.73
	SVM	0.65	0.69
LIWC	LR	0.63	0.65
	SVM	0.63	0.64
	Data preparation		
DistilBERT	Uncleaned text	0.66	0.65
	Cleaned text	0.65	0.57
Longformer	Uncleaned text	0.73	0.74
	Cleaned text	0.68	0.69
RoBERTa	Uncleaned text	0.68	0.61
	Cleaned text	0.71	0.68

Table 2.1: Performance results for each model

TF-IDF		Word2Vec		Longformer	
Female	Male	Female	Male	Female	Male
child	engineer	experience	patient	homemaker	accomplished
content	technology	assistant	service	passion	student
medium	engineering	operation	I	growth	look
social	tool	graduate	group	excellent	use
document	operation	learn	present	communication	time
answer	produtct	complete	include	skill	gather
research	network	event	role	experience	information
event	machine	problem	responsible	customer	skill
file	work	development	financial	service	require
club	business	quality	support	sale	serve
psychology	user	make	manager	work	people
customer	computer	communication	marketing	make	orient
skill	technical	well	level	lasting	career
organize	part	perform	company	impact	exceptionally
guest	construction	provide	education	within	organize
assist	several	proficient	ensure	company	diligent
honor	build	base	training	community	work
care	team	order	datum	acquire	experience
age	hardware	system	science	job	keep
brand	improvement	work	activity	allow	library

Table 2.2: Most relevant features that predicted gender in each NLP model

2.5 Discussion

The main goal of this project was to detect gender differences in self-presentation in resumes. As Parasurama et al., 2022 or Yang et al., 2022 have previously done, we considered a good strategy to use several NLP techniques, since we can extract a lot of information from the provided text in CVs.

The models that have been trained provided us with results that are above the random level of prediction, since the accuracy and AUC metrics score over 0.5 for all the different models (see table 2.1 and appendix A), even for the baseline (BoW). This means that even the most straightforward algorithm that only takes word frequencies, with classifier models that have not been tuned, can still detect gender differences. This is very relevant since we are still facing gender segregation in the organizational context, which can lead to gender discrimination (He & Kang, 2021) especially when we are using AI algorithms to recruit people. As it has been seen in the Amazon case (Dastin, 2018), the current algorithms still detect gender differences even with anonymized data and deleting gender from the main variables, just as we have seen in our research. This is problematic if the algorithms are the main tool to recruit people, since they are biased towards hiring more men over women, even if they are both as qualified to obtain the job. This should be a point to reflect on and address further research since we must create and use unbiased algorithms. Furthermore, this is not only relevant for the organizational context, since gender differences have been observed in communication and self-presentation in other fields such as politics (Hargrave & Langengen, 2021; Joshi et al., 2021), science (DeJesus et al., 2021; Kolev et al., 2020) or debate speeches (Nguyen, 2021).

If we focus now more in detail on the specific models that have been used in this research, one may notice that the performance metrics are not excellent (all the accuracy levels are clearly under 0.80, and there is only the AUC for RoBERTa that reaches 0.82), even if they are significant and point out that men and women do not self-present themselves the same way in resumes. This limitation of the models might have an explanation. In the

NLP field, the amount of data that is normally used to train the models is extensive, talking about hundreds of thousands of entries with raw text that is processed in further steps. In our case, instead, we only had a small sample (N = 1789) with text that was already pre-processed previously by Yang et al., 2022. This is meaningful for three reasons: with small samples, algorithms cannot learn as well as if they had a large amount of data. They might detect differences, as has been the case, but it is not as powerful. The second reason is the fact that we have used pre-trained models for the transformer-based ones. This means that there is a first step of adapting the model that may introduce previous knowledge, which can be already biased. The last reason is the fact that transformer-based models need raw un-tokenized data to perform well since it is very important for these algorithms to have proper context and learn of it. This may also explain why a simple statistical model (TF-IDF) performs just as well as a transformed-based model such as Longformer (check table 2.1) when we would normally expect better performance for the latter, just as it was found by Parasurama et al., 2022.

The second research question that we wanted to investigate was regarding the most predictive features of gender in resumes. In table 2.2, features for three of the main models are shown. In general, the words that predict gender are different for the three models. Also, we can find words that are predicting both genders, for instance, "work" in the Longformer model, which is not very informative. The most intriguing discoveries, however, emerge when examining the complete picture. For example, for the TF-IDF, among the most relevant features that predict female gender, we can find the words "care" and "assist", while for male we notice "engineer", "technical" or "improvement". Similarly, in the Word2Vec results, we can detect the word "assistant" as the second most predictive word for women, whereas for men "manager" or "science" are important. In the Longformer, the most predictive feature for women is "homemaker". All these examples make us reflect further on how much gender roles influence the way we write our resumes. This is aligned with the research made by Eagly, 1987 and Eagly et al., 2000, since it seems to be that gender roles are very well-established in

society and consequently in people's mindsets, taking us to write and communicate accordingly to the gender we define ourselves in. Moreover, the top features do not only point out the underlying gender roles but also the examples are cases of words that belong to what is desirable in behavior of a man and a woman (Broverman et al., 1972; Williams & Best, 1990). Bakan, 1966 and Hsu et al., 2021 identified that agency is more strongly associated with men and communion with women. In the same line, we could find an association of this in our research since the words "assist" or "care" that predict the label "female" belong to the behavior of nurturing and interpersonal connections, giving a sense of prioritizing the needs of others. This is a sign of communal behavior, and it is one more evidence of what has been found previously regarding how women behave in the recruitment process (Guadagno & Cialdini, 2007; Parasurama et al., 2022; Prentice & Carranza, 2002; Smith et al., 2013). On the other hand, for men, we found words that are strongly correlated to success and higher positions in the organizational environment. This is associated with agentic behavior, and also points out the inequality that stands until today between men and women in the workplace, with men still having higher positions than women (He & Kang, 2021; Steffens et al., 2019).

3. Navigating Gender Bias in Resumes: An Investigation of Word and Contextual Embedding Models

Written by Malka

3.1 Related Work

3.1.1 Gender Differences in Writing Styles

Numerous studies have been done on the differences in writing styles of men and women (Argamon et al., 2003). The study by DeJesus et al., 2021 found gender differences in writing styles within academic publications: women in lead author roles used less generic language than men. Works that utilized generic language, primarily written by men, were notably cited more frequently. However, this examination was limited to the context of peer-reviewed psychology papers and focused on scientific article language, which differs from resume writing.

Adding to DeJesus's findings, an additional study examining life sciences articles illustrated that male authors showcased their results in a more positive light compared to their female counterparts, with males using more positive descriptors such as "novel," "promising," "robust," and "excellent." (Lerchenmueller et al., 2019).

3.1.2 Gendered Writing Styles in Professional Contexts

Research into gendered writing styles primarily focuses on identifying variations in various linguistic aspects that may be indicative of the author's gender. These characteristics span across a broad range of linguistic dimen-

sions, including, but not limited to, the degree of self-promotion, readability, and specificity of written content.

In particular, a large body of research has found substantial differences in the degree of self-promotion between genders (Scharff, 2015). These investigations extend to other linguistic characteristics as well, demonstrating gender-based variations in aspects such as readability (DeJesus et al., 2021) and specificity of content (Joshi et al., 2021; Kolev et al., 2020).

In contrast to these findings, some research provides evidence of only minor and statistically insignificant gender differences in writing styles. Franco et al., 2021 is one such study. Moreover, Horbach et al., 2022 contribute to this by investigating gender differences in funding applications. Their analysis of 1560 applications in natural and technical sciences from a Danish funder considered a variety of factors including peer-review bias, language use, and the potential influence of gendered writing on funding rates. Using methodologies from previous studies, they examined the use of positive words, readability, specificity, and sentiment in the applications. Despite earlier findings pointing to gendered differences in writing styles, their study revealed only minor differences between the genders. They concluded that the writing styles likely do not explain uneven funding patterns, particularly in academic settings.

Such divergent findings underline the complex nature of gendered language use in professional contexts and further emphasize the need for more in-depth investigation.

3.1.3 Machine Learning in Gendered Information Identification

Despite extensive research on writing style differences, there is a lack of studies in the context of resumes which is a fundamental part of professional self-presentation. This suggests a need for more research in this area. Understanding how men and women present themselves differently across various job fields has seen notable progress, yet questions remain unanswered.

De-Arteaga et al., 2019 explored the growing significance of machine learning in online recruitment and automated hiring and the related outcomes of bias. They constructed a dataset comprising hundreds of thousands of English-language online biographies, from which they inferred binary gender based on third-person narratives and gendered pronouns. This dataset was used to predict occupations via multi-class classification using different semantic models: bag-of-words, word embeddings, and deep recurrent neural networks. They assessed both scenarios with and without explicit gender identifiers to evaluate fairness and legal compliance. Their findings suggested a correlation between true positive rates and existing gender imbalances in occupations, which could further amplify these inequalities.

However, the potential of machine learning models to detect gender-specific information in resumes still remains relatively unexplored. Stepping into this research gap, Parasurama et al., 2022 and Yang et al., 2022 made some of the first investigations into this topic.

Going deeper into this subject, Parasurama et al., 2022 studied if men and women with similar job-relevant characteristics write their resumes differently and how this difference affects hiring outcomes. They built upon earlier research by Streib et al., 2019 and He and Kang, 2021 who explored the presence of gendered information in resumes, although with varying results. Streib et al., 2019 did this by using hand-coded features, but find no evidence. The reason for this could be that there is no clear definition of what researchers thought to be 'gender characteristics' and that the researchers themselves specify is. He and Kang, 2021 found in their study qualitative evidence through participant interviews but found no evidence for gender differences when they used a quantitative dictionary-based method.

Building on this approach, Parasurama et al., 2022 used a sample of 248k matches resumes and trained an advanced deep learning model and a bag-of-words model to quantify the amount of gendered information in IT sector resumes. They found that even after anonymizing these resumes, signifi-

cant gendered information remained in the text. An interesting finding was that an advanced deep learning model, known as the Longformer model, could predict gender with at least 80 percent accuracy using only the resume text. However, the limitation of this study is its exclusive focus on the IT sector, a field heavily skewed toward men.

Bridging to our current study, the work of Yang et al., 2022 forms a central reference, as our study utilizes their data. Yang et al., 2022 developed a database comprising 1.8K genuine, English-language resumes from the United States, spanning 16 different occupations. Their findings revealed that women tend to use more verbs that give an impression of lower power. Furthermore, even after balancing data and eliminating pronouns and named entities, classifiers were still able to discern gender signals. This was the case for both transformer-based and linear classifiers. This consistent identification of gender signals suggests interesting consequences for both future research and real-world applications.

3.1.4 Algorithmic Bias in Recruitment

With the increasing use of advanced resume screening algorithms in businesses, it becomes ever more important to raise awareness about these tools' potential capability to detect gender-specific information within resumes, even when they have been anonymized (Chen et al., 2018; Peng et al., 2019; Raghavan et al., 2020). Adding advanced NLP models to these tools could unintentionally make these biases worse, as shown by Shah et al., 2019. For instance, research has shown that job ads often contain gendered language (Böhm et al., 2020). This means that if a resume screening tool uses document embeddings or text representations to match resumes with job descriptions, male resumes end up getting matched with ads that use male-oriented language more often (Devlin et al., 2018).

Progress in predictive technology, coupled with increased data availability and rising hiring costs, suggests a more central role for algorithms in recruitment (Perkowski, 2023). However, blindly trusting these algorithms could have significant implications, as illustrated by Amazon's 2018 hir-

ing algorithm which implicitly favored men due to historical hiring data (Dastin, 2018). This highlights the risk of unintentionally reinforcing existing inequalities. Moreover, Perkowski's 2023 study showed that despite potential performance improvements, algorithms were not widely adopted if they reduced female hires, indicating the persistent struggle to develop fair, effective hiring algorithms.

Moving forward, as the development of advanced models continues, we need to understand their performance in comparison to older, yet effective models. Therefore, this research aims to make these comparisons to better understand the capacity of machine learning techniques in identifying differences in self-presentation across gender and job fields. Furthermore, this study will shed light on the performance of these models with smaller data sets, a research gap not explored yet in the current literature.

3.1.5 Occupational differences

It is important to extend the focus to different occupational sectors to gain a better understanding of how these models function. Therefore, this research will examine three different occupational categories, using the gender proportion data from the World Economic Forum for the USA as a reference. These include male-dominated sectors (IT and engineering), female-dominated sectors (healthcare and education), and sectors with a balanced gender distribution (business and finance). The full list of occupations per category can be found in Appendix B. A similar effort was made by Zide et al., 2014, who examined differences in LinkedIn profiles, revealing variations in self-presentation by industry and gender.

The study by Zide et al., 2014 dived into the differences within HR, sales/marketing, and industrial/organizational psychology industries. It revealed not only industry-specific variations but also certain gender differences. Moreover, men tended to be more willing than women to disclose personal information on their profiles. These findings suggest the presence of both gender and industry variations in profile presentation. Nonetheless, the generalizability of these findings is limited due to the small sample

size, drawn from New York City. Additionally, LinkedIn's specific context as a platform for networking and job searches differs from the traditional resume.

3.1.6 Summary

Overall, this literature review has shown that gender differences in writing styles are evident and can be identified from academic publications to professional settings. The use of machine learning in identifying these differences, specifically in professional self-presentation such as resumes, is a growing field but remains under-explored. The existing studies have mostly focused on specific sectors or used particular machine learning models.

This study aims to address these gaps by examining the effectiveness of different machine learning techniques in identifying gender differences in self-presentation across various occupational groups. This research will also provide a comparative analysis of the performance of these models with smaller data sets, a research gap that is not yet explored in the existing literature. By doing so, we can better understand the details of gendered self-presentation in professional contexts and the implications for hiring practices.

This leads to this study's research question: Which word- and contextual-embedding model(s) are best in the detection and prediction of gender differences in self-presentation strategies across occupational groups?

3.2 Data

In Section 1.1, the methodologies and processes involved in data collection and preparation for this study were outlined. Based on the 2021 U.S. Labor Statistics, as cited in Yang et al., 2022 (see Appendix B), the professions were further categorized into three groups: male-dominated (with less than 30% women), balanced gender representation (with 30% to 70% women), and female-dominated (with over 70% women).

For the purpose of this research, participants who identified their oc-

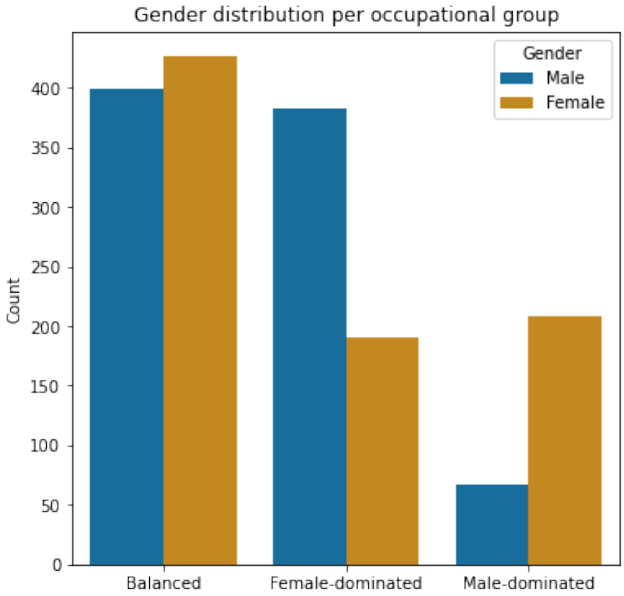


Figure 3.1: Gender Count per Occupational Group

cupation as 'Other' according to Yang et al., 2022 were excluded, resulting in the removal of a total of 114 participants from the data. A visualization of the gender distribution across each occupational category is provided in Figure 3.1.

3.3 Method

In this study, the aim was to evaluate and compare the performance of various models in the task of identifying gender-related information in resumes across different occupational groups. The models selected for evaluation included different methods:

- Traditional: Term Frequency-Inverse Document Frequency (TF-IDF)
- Word-embedding: Word2Vec and GloVe
- Contextual embedding: DistilBERT, RoBERTa, and Longformer

Detailed descriptions and justifications for the choice of these models can be found in Section 1.2.

We followed the steps outlined in Section 1.2.3 for the evaluation process. However, we had to address the class imbalance in two of the three

datasets by stratifying the train/validation/test split. This ensured that the models were trained on data that accurately represented the gender distribution of specific occupational categories.

For the baseline model, TF-IDF, we used default parameters. The classifiers logistic regression and SVM were also included in the evaluation, and their default parameters were utilized. The word-embedding models, Word2Vec and GloVe, represented each word with a 100-dimensional vector. We chose the same dimensionality for Word2Vec to make a fair comparison with GloVe, which used pre-trained Wiki vectors.

With the transformer models (DistilBERT, RoBERTa, and Longformer), we focused on tuning the number of epochs and batch size due to computational limitations. We tested epochs ranging from 1 to 8 and batch sizes of 4, 8, and 16.

To determine the best-performing combination for each model, we used cross-validation. Detailed numbers for these settings can be found in the tables below. It is important to note that due to computational constraints, this study did not extensively explore parameter tuning beyond epochs and batch size. Further research could delve deeper into these parameters.

Table 3.1: DistilBERT parameters.

Dataset	Epochs	Batch Size
Balanced	4	4
Female-dominated	2	8
Male-dominated	4	8

Table 3.2: RoBERTa parameters.

Dataset	Epochs	Batch Size
Balanced	3	16
Female-dominated	3	4
Male-dominated	3	4

Table 3.3: Longformer parameters.

Dataset	Epochs	Batch Size
Balanced	8	4
Female-dominated	8	4
Male-dominated	8	4

3.4 Results

3.4.1 Overview of the results

The objective of this study was to determine which word- and contextual-embedding model(s) are best for detecting and predicting gender differences in self-presentation strategies across different occupational groups. Six different models were assessed: TF-IDF, Word2Vec, GloVe, DistilBERT, RoBERTa, and Longformer. With TF-IDF, Word2Vec, and GloVe additionally paired with the classifiers Logistic Regression (LR) or Support Vector Machine (SVM).

Each model was evaluated using several performance metrics, including ROC curves, precision-recall curves, AUCs, F1-scores, and accuracies. For context, ROC curves are particularly useful for balanced datasets, whereas precision-recall curves often work better for imbalanced datasets, which is the case with two of our datasets.

A detailed performance comparison of the six models revealed a variety of results. To simplify comparison, the ROC and precision-recall curves are only plotted for the best-performing combination of TF-IDF, Word2Vec, and GloVe with either LR or SVM. This approach allows an in-depth analysis of top performance while reducing potential visual clutter from the less successful combinations.

Initial results indicate a range of performance across the models. TF-IDF provided a very strong baseline for the balanced group with notable accuracy, though other models showed higher performance in the other two groups. Word2Vec and GloVe, while demonstrating similar performance, were found to be outperformed by the baseline and some transformer-based

models in certain occupations. Notably, half the models struggled to perform consistently within the male-dominated group.

The transformer-based models, DistilBERT, RoBERTa, and Longformer, presented interesting results. DistilBERT had the lowest performance when considering AUC and AP for all groups compared to the other models. However, RoBERTa and Longformer showed a stronger performance overall, with a well-rounded performance across all occupational groups.

The following sections will present a more detailed analysis of each model's performance across the different metrics and occupational groups.

3.4.2 Results for balanced occupations

For a detailed overview of the results mentioned in this section, please refer to Table 3.4 and Table 3.7. Among all tested models, the baseline TF-IDF, paired with both logistic regression and SVM, emerged as the superior performer for the balanced group. Given the balanced nature of this dataset, it is particularly informative to compare metrics such as accuracies and AUCs. The TF-IDF model boasted an accuracy of 76%, coupled with F1-scores of 0.75 (LR) and 0.78 (SVM), making it the top-performing model. This is especially notable considering that it outperformed RoBERTa, which, in contrast, delivered the lowest accuracy of 54%.

When considering the AUC, the TF-IDF model, paired with logistic regression, achieved the highest score of 0.79, equaled only by RoBERTa. However, RoBERTa's comparatively lower accuracy suggests that the standard decision threshold of 0.5 may not be optimal for this specific model.

The AUCs for all models were within a relatively narrow range from 0.71 to 0.79, despite significant variation in accuracy (54-76%). This difference may indicate the potential for performance enhancement through decision threshold optimization for each model separately.

Table 3.4: Model performances on the balanced data.

Model	Gender	Precision	Recall	F1-Score	Accuracy
TF-IDF + LR	Male	0.74	0.82	0.78	0.76
	Female	0.79	0.70	0.74	
TF-IDF + SVM	Male	0.75	0.81	0.78	0.76
	Female	0.78	0.71	0.75	
Word2Vec + LR	Male	0.68	0.71	0.69	0.68
	Female	0.68	0.65	0.66	
Word2Vec + SVM	Male	0.69	0.65	0.67	0.67
	Female	0.65	0.69	0.67	
GloVe + LR	Male	0.65	0.75	0.70	0.66
	Female	0.68	0.56	0.62	
GloVe + SVM	Male	0.69	0.78	0.73	0.70
	Female	0.72	0.62	0.67	
DistilBERT	Male	0.67	0.62	0.64	0.65
	Female	0.63	0.68	0.65	
RoBERTa	Male	0.53	1.00	0.69	0.54
	Female	1.00	0.02	0.10	
Longformer	Male	0.73	0.63	0.68	0.69
	Female	0.65	0.75	0.70	

3.4.3 Results for female-dominated occupations

In this dataset, the distribution of males to females was approximately 35/65, creating a reasonably imbalanced representation. Given this skew, our analysis will focus more on the F1-score and average precision as key metrics for comparing the performance of different models. For a detailed overview of the results mentioned in this section, please refer to Table 3.5 and Table 3.7.

Notably, Word2Vec paired with both Logistic Regression (LR) and Support Vector Machine (SVM), and GloVe combined with SVM, exhibited an F1-score of 0.00 for the male category (0). However, GloVe paired with LR showed significantly better performance, recording an F1-score of 0.32 for predicting males and 0.81 for predicting females.

Remarkably, TF-IDF combined with SVM demonstrated exceptional performance in this female-dominated dataset, achieving the highest recall rate for males at 0.81. RoBERTa and Longformer were both in second place with equal recall rates of 0.42.

When assessing average precision, scores were near each other from 0.71 to 0.81. Both RoBERTa and Longformer performed the best, with an average precision of 0.81, closely followed by TF-IDF (+ SVM) at 0.79.

Table 3.5: Model performances on the female-dominated data.

Model	Gender	Precision	Recall	F1-Score	Accuracy
TF-IDF + LR	Male	0.50	0.05	0.10	0.67
	Female	0.68	0.97	0.80	
TF-IDF + SVM	Male	0.75	0.81	0.78	0.76
	Female	0.78	0.71	0.75	
Word2Vec + LR	Male	0.00	0.00	0.00	0.66
	Female	0.67	0.99	0.80	
Word2Vec + SVM	Male	0.00	0.00	0.00	0.67
	Female	0.67	1.00	0.80	
GloVe + LR	Male	0.67	0.21	0.32	0.70
	Female	0.71	0.95	0.81	
GloVe + SVM	Male	0.00	0.00	0.00	0.67
	Female	0.67	1.00	0.80	
DistilBERT	Male	0.00	0.00	0.00	0.67
	Female	0.67	1.00	0.80	
RoBERTa	Male	0.38	0.42	0.40	0.59
	Female	0.70	0.67	0.68	
Longformer	Male	0.80	0.42	0.55	0.78
	Female	0.77	0.95	0.85	

3.4.4 Results for male-dominated occupations

This dataset presented an approximately 80/20 male-to-female ratio, indicating a significant skew towards males. Similar to the analysis performed for the female-dominated dataset, our focus will primarily be on precision, recall, F1-scores, and average precision in this male-dominant context.

The Longformer model stood out as the only model to achieve a recall and F1-score above zero for predicting females (category 1), with respective scores of 0.29 and 0.40.

An examination of the average precision reveals mixed performances. TF-IDF, Word2Vec, and DistilBERT lagged behind, each scoring an average precision of 0.38, a score even lower than what might be expected from a random model. GloVe, when paired with Logistic Regression, displayed

slightly superior performance, achieving an average precision of 0.58.

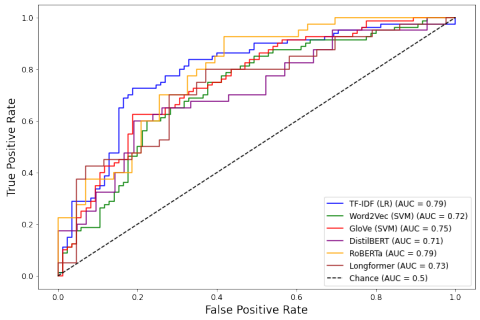
Interestingly, RoBERTa emerged as the top performer with an average precision of 0.76, closely followed by Longformer at 0.71. The precision-recall curve also highlighted RoBERTa's exceptional performance compared to the other models (see Figure 3.2(f)). For a detailed overview of the results mentioned in this section, please refer to Table 3.5 and Table 3.7.

Table 3.6: Model performances on the male-dominated data.

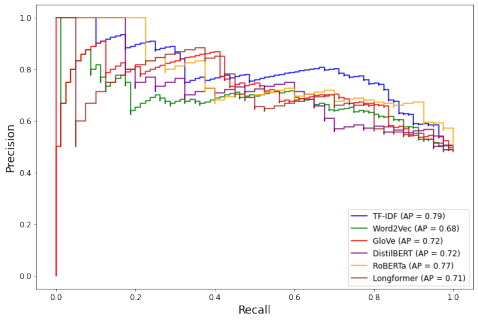
Model	Gender	Precision	Recall	F1-Score	Accuracy
TF-IDF + LR	Male	0.76	1.00	0.87	0.76
	Female	0.00	0.00	0.00	
TF-IDF + SVM	Male	0.76	1.00	0.87	0.76
	Female	0.00	0.00	0.00	
Word2Vec + LR	Male	0.76	1.00	0.87	0.76
	Female	0.00	0.00	0.00	
Word2Vec + SVM	Male	0.76	1.00	0.87	0.76
	Female	0.00	0.00	0.00	
GloVe + LR	Male	0.76	1.00	0.87	0.76
	Female	0.00	0.00	0.00	
GloVe + SVM	Male	0.76	1.00	0.87	0.76
	Female	0.00	0.00	0.00	
DistilBERT	Male	0.78	1.00	0.88	0.78
	Female	0.00	0.00	0.00	
RoBERTa	Male	0.75	1.00	0.86	0.75
	Female	0.00	0.00	0.00	
Longformer	Male	0.80	0.95	0.87	0.79
	Female	0.67	0.29	0.40	

Table 3.7: AUCs and APs for all groups.

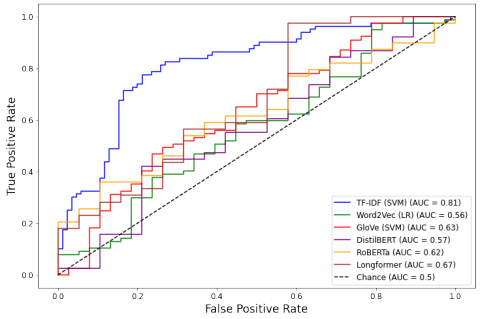
	Balanced		Female-dominated		Male-dominated	
Model	AUC	AP	AUC	AP	AUC	AP
TF-IDF + SVM	0.79	0.79	0.81	0.79	0.59	0.38
Word2Vec + SVM	0.72	0.68	0.56	0.73	0.57	0.38
GloVe + LR	0.75	0.72	0.63	0.76	0.72	0.58
DistilBERT	0.71	0.72	0.57	0.71	0.56	0.38
RoBERTa	0.79	0.77	0.62	0.81	0.89	0.76
Longformer	0.73	0.71	0.67	0.81	0.71	0.63



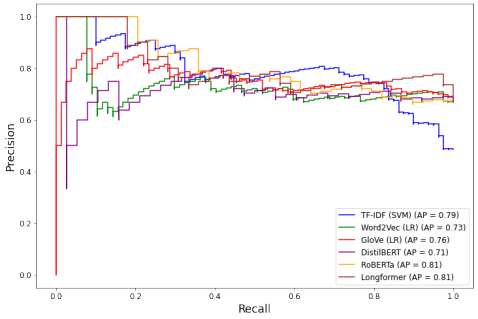
(a) Balanced



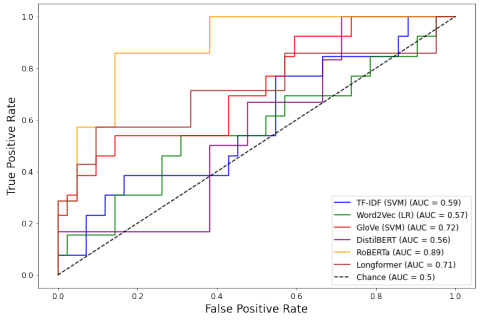
(b) Balanced



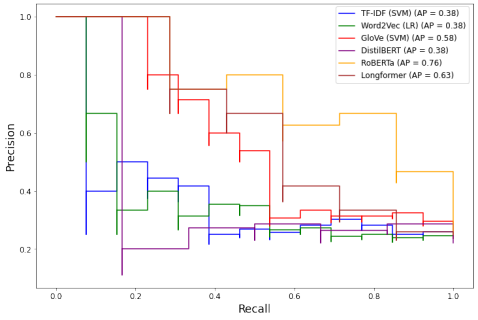
(c) Female-dominated



(d) Female-dominated



(e) Male-dominated



(f) Male-dominated

Figure 3.2: Receiver operating characteristic (ROC) (left) and precision and recall (PR) (right) curves for all groups

3.5 Discussion

In our exploration of word- and contextual-embedding models for the detection and prediction of gender differences in self-presentation strategies across different occupational groups, the baseline model, TF-IDF, in combination with either logistic regression or SVM, demonstrated impressive performance, particularly in the balanced group. However, within datasets exhibiting gender imbalances, other models showed significant strengths, offering intriguing insights for future research in this field.

Our study employed a range of traditional, word-embedding, and contextual-embedding models, including TF-IDF, Word2Vec, GloVe, DistilBERT, RoBERTa, and Longformer. These models were chosen based on their prior performance in similar text classification tasks and evaluated against the task of identifying gender-related information in resumes across different occupational groups.

In the female-dominated dataset, GloVe combined with logistic regression emerged as a promising performer in predicting both genders. In contrast, the Longformer model demonstrated its robustness within the male-dominated dataset, becoming the only model to achieve above-zero recall and F1-score for predicting females when using the default decision threshold of 0.5.

Interestingly, despite the superior performance of TF-IDF in the balanced group, it fell short in the context of gender imbalances. This outcome highlights the importance of selecting appropriate models based on the specific characteristics of the dataset. This point was made clear by the diverse performance across models in the male- and female-dominated datasets.

One significant observation across our study was the difference between models' AUCs and accuracies, suggesting potential enhancements through decision threshold optimization. Further research could focus on fine-tuning these thresholds for each model, and in particular, for models like RoBERTa, which showed a high AUC but lower accuracy. These results are in line with Yang's finding which mentioned that RoBERTa outper-

formed the linear classifiers in the full data condition, but showed unstable performance in their balanced condition presumably due to the smaller dataset leading to overfitting (Yang et al., 2022).

Given the constraints of computational resources, our study focused on tuning a limited range of parameters—epochs and batch size. The exploration of other hyperparameters might lead to further improvements in model performance and, therefore, presents an exciting avenue for future research.

Ultimately, our study provides a comparative analysis of various models for detecting gender differences in self-presentation strategies across occupational groups. While the results offer valuable insights, they also highlight the complexity of the task and the need for more targeted, context-aware methods in future research.

3.5.1 Broader Implications and Concerns

In addition, our study provides substantial evidence backing concerns regarding resume screening algorithms. These systems are capable of identifying gender-related cues from resume content, thereby sustaining any biases present in the training data. This is in alignment with the findings by Yang et al. (2022) and Parasurama et al. (2022), who also uncovered the presence of gender-related information in anonymized resumes. Even traditional models such as TF-IDF can differentiate between genders.

This matter grows increasingly pressing as HR software companies are introducing these types of algorithms into the commercial market. Often, these firms provide limited information on how their algorithms address potential biases within the training data (Raghavan et al., 2020). Our research highlights the importance of strict examination and regulatory measures on these tools to prevent the unintentional continuation of gender bias within hiring practices.

3.5.2 Limitations and Future Research

3.5.2.1 Limitations Related to Data Distribution

While the stratification of our data into balanced, male-dominated, and female-dominated groups was guided by World Economic Forum data, the actual distribution within our dataset did not perfectly mirror these categories. For instance, the "Management" occupation had a representation of only 36% females, a divergence from the anticipated 52% based on the forum's figures. This discrepancy may have impacted our models' ability to accurately detect and predict gender-based differences in self-presentation strategies. Future studies should strive to gather data that more closely aligns with established gender distributions within occupations.

3.5.2.2 Methodological Constraints

The anonymization and pre-cleaning of the data performed by Yang may have influenced the performance of our models, particularly the transformer models that rely on contextual understanding. The lack of complete, interpretable sentences within the dataset might have affected their ability to grasp context fully. Furthermore, participants were instructed to upgrade their current CV as if they were applying for a promotion, a scenario that may not fully reflect real-world resume crafting for actual job applications. Future research might consider utilizing raw, unedited resumes and instructing participants to draft CVs for genuine job application scenarios.

3.5.2.3 Limitations in Participant Demographics

The participant demographic and experience level data were not available in our dataset. Aspects such as age and years of experience can significantly influence self-presentation strategies and language use in resumes. Without this information, our analysis may not have captured certain demographic or experience-based nuances in self-presentation. Future studies should attempt to incorporate such demographic data to provide a more nuanced analysis.

3.5.2.4 Geographical Limitations

Our study was based on data collected in the United States, and occupational group gender distributions were also drawn from U.S. data. As such, the results of our study may not be generalizable to other geographic regions with different occupational gender distributions or cultural norms surrounding resume drafting. Subsequent research should consider including data from a variety of regions to establish a more globally representative understanding of gender differences in self-presentation strategies.

Additionally, the study operated within the constraints of binary gender, categorizing participants as either male or female. This binary approach overlooks the diversity and fluidity of gender identities, limiting the scope and inclusivity of our study. Future research should consider more inclusive categorizations of gender to better capture the diversity of self-presentation strategies across a broad spectrum of gender identities.

This research, despite its limitations, contributes to our understanding of gender differences in self-presentation strategies across different occupational groups. The findings underscore the importance of considering gender imbalances within occupational groups and raise intriguing questions for future research to address.

4. Occupational Differences in Gender Congruent Communication

Written by Maike

4.1 Related Work

Until now, we have investigated how resumes written by men and women differ from each other regarding their textual features as well as in different occupations. Our results are in line with previous research on gender norms in impression management (Guadagno & Cialdini, 2007). While men used words that are more agentic in nature, women communicated in a more communal way in their resumes (see section 2.4 and Hsu et al., 2021). This way, both genders communicated in ways that are congruent with the respective gender stereotypes and norms that persist across society. Even though we have expected this pattern in communication to some extent, it does limit women in promoting their work-related capabilities to potential new employers.

The difficulty that feminine gender norms are less in line with values that are believed to be important in the occupational context compared to masculine gender norms, was formalized in the gender-role congruity theory (Eagly & Karau, 2002). According to this theory, occupational values are inherently agentic in nature and encompass leadership qualities that are mainly associated with masculinity (Schein, 1973, 1975). Consequently, women face more difficulties in the occupational sector than men due to two different types of prejudice related to gender norms and occupational values. First, descriptive prejudice describes the preconception that women are less capable of performing occupational tasks, especially those associated with leadership. Second, prescriptive prejudice describes an under-

lying social norm in which women should not perform occupational tasks that do not align with qualities associated with their gender (Bakan, 1966; Hsu et al., 2021). Put differently, women face the difficult dilemma of either conforming to occupational values or conforming to feminine gender norms.

On the one hand, communicating in ways that emphasize occupational values and agentic characteristics that promote work-related success is of great benefit to women in the application phase in order to compete with other candidates applying for the same job. On the other hand, women who communicate in a way that does not conform to feminine gender norms can experience severe social and economic penalties, known as backlash effects (Rudman & Phelan, 2008). For example, women who present themselves as assertive and agentic, are often perceived as less likable than their male counterparts or women who communicate in a more gender congruent way (Bolino & Turnley, 2003; Heilman et al., 2004). This results not only in social costs during the interaction with others (Rudman & Phelan, 2008), but also in decreased chances of being hired (Rudman, 1998; Rudman & Glick, 2001), salary negotiations, getting a promotion and leadership evaluations (for a review, see Rudman and Phelan, 2008).

As a result, some women feel a certain pressure to conform to the feminine gender norms even though they often do not appeal to occupational values. The backlash avoidance model describes processes in which women often feel inhibited to promote their work-related capabilities in order to avoid the backlash that comes from agentic and self-promoting communication (Moss-Racusin & Rudman, 2010). Other studies found that women who communicate in a more communal way experience less backlash and are rated as more likable than their peers who communicate solely in an agentic way (Rudman & Glick, 2001; Rudman & Phelan, 2008). Thus, communicating in a less agentic, and more gender-congruent way seems to have certain advantages for women during the application process.

To sum up, women face a difficult dilemma with their communication in the occupational context. On the one hand, they can aim to conform to oc-

cupational values and communicate in an agentic way in order to promote their capabilities. On the other hand, women often face social and economic penalties when communicating in ways that do not conform to such gender norms. As a consequence, women continually feel a certain pressure to conform to feminine gender norms in order to avoid such backlash effects. Nevertheless, to my knowledge, there are few studies that have yet investigated the way that women deal with this important dilemma. One study has investigated a similar research question (Parasurama et al., 2022). However, the researchers only investigated resumes in the IT-related sector, a very male-dominated context (Yang et al., 2022). Therefore, in this study, I will investigate the extent to which applicants communicate in a gender-congruent way in different occupational contexts.

From the previously discussed literature, I derived three hypotheses: First, due to the general nature of the occupational context in which resumes are written, I expect that women are forced to communicate less gender-congruently than men, regardless of the occupation that they apply for. Second, due to the skewed gender distribution in male-dominated occupations, I presume that social norms are skewed towards masculine gender norms in these occupations as well (Cialdini & Trost, 1998). Therefore, I expect that women who apply for jobs in male-dominated occupations communicate even less gender-congruently than women who apply for jobs in gender-balanced or female-dominated occupations.

In contrast, masculine gender norms are generally more in line with occupational values than feminine gender norms (Eagly & Karau, 2002). An exception to this rule, I presume, is the social norm in female-dominated occupations. Not only do more women work in these occupations, which in turn shapes the social norms of this context (Cialdini & Trost, 1998). But also the type of work corresponds more to feminine gender norms. For example, the communal aspect of feminine gender norms can be found back in social service occupations as well as in diverse occupations that are related to (health-)care. Therefore, my last hypothesis is that men communicate less gender congruently in female-dominated occupations than in gender-balanced or male-dominated occupations.

4.2 Data

The data collection and preparation procedures that I used for this study were described in section 1.1. Additionally, I divided the occupations into male-dominated occupations (< 30% women), gender-balanced occupations (>= 30% and <= 70% women), and female-dominated occupations (> 70% women) according to US Labor Statistics in 2021 as cited in Yang et al., 2022 (Appendix B). Participants who reported working in an occupation that was categorized as 'Other' by Yang et al., 2022, were excluded from this study (N = 114).

4.3 Method

In order to investigate the extent to which people wrote in a gender-congruent way, I first trained different machine learning models on the whole data set to predict the gender from written text. The models are described in section 1.2. As expected, the training performance of these models was much better than the test performance of the models that were trained on only 80% of the data (and 10% validation data). The baseline models as well as the SVM on TF-IDF data had an accuracy and F1 score of 100%. The analyses based on word vectors, LIWC values, DistilBERT, and RoBERTa performed slightly worse with accuracies of roughly 70 to 80%. The best-performing transformer model was Longformer with more than 90% accuracy. Details about the performance metrics as well as ROC plots and confusion matrices can be found in Appendix C.1.

Second, I used these models to calculate the probability of being a woman for each participant. These predictive scores range between 0 (indicating a 0% chance that this participant is a woman) and 1 (indicating a 100% chance that this participant is a woman). Third, the score was reversed for men and then used as a measure of gender congruity.

To test my hypotheses, I performed a 2x3 ANOVA with gender congruity as the dependent variable and the main and interaction effects of gender and occupation as the predictors. The first hypothesis (*H1: Women communi-*

cate less gender congruently than men) was tested by contrasting the congruity scores of female resumes against the congruity scores of male resumes across all occupations. The second hypothesis (*H2: Women communicate less gender congruently in male-dominated occupations than in gender-balanced or female-dominated occupations*) was tested by contrasting the congruity scores of female resumes in male-dominated occupations against the congruity scores of female resumes in gender-balanced and female-dominated occupations. The third hypothesis (*H3: Men communicate less gender congruently in female-dominated occupations than in gender-balanced or male-dominated occupations*) was tested by contrasting the congruity scores of male resumes in female-dominated occupations against the congruity scores of male resumes in gender-balanced and male-dominated occupations.

The analysis was conducted on all gender-congruity scores separately, as well as on a weighted average of all these scores. To calculate this score, I weighted the congruity scores by the training accuracy scores of the models (see Appendix C.1) in order to account for the accuracy of the predictive scores. The reason for weighting the congruity scores by the training accuracy lies in the definition of the congruity scores in this study. The congruity score is directly related to the classification result and thereby also to the classification performance. As a result, models with high accuracy scores will also result in high congruity scores (close to 100%) and a smaller variance, while models with low accuracy scores (closer to 50%) will also result in lower congruity scores (closer to 50%) with a larger variance. Similarly, the differences in congruity scores, which form the basis for supporting or rejecting my hypotheses, should also be smaller in models with high accuracy than in models with lower accuracy. Therefore, weighting the congruity scores by the accuracy of the respective models (thus giving a higher weight to models with higher accuracy), allows me to check the robustness of the individual results.

4.4 Results

4.4.1 Hypothesis Testing

Data pre-processing	Classifier	H1	H2	H3
Baseline (BoW)	LR	-0.01***	-0.05***	-0.02***
	SVM	0.00	-0.03*	-0.03***
TF-IDF	LR	-0.02***	-0.16***	-0.14***
	SVM	0.00***	-0.00	-0.00
Word2Vec	LR	-0.02***	-0.23***	-0.20***
	SVM	-0.01	-0.10***	-0.11***
GloVe	LR	-0.02***	-0.20***	-0.19***
	SVM	-0.02***	-0.22***	-0.23***
LIWC	LR	-0.01	-0.14***	-0.15***
	SVM	-0.00	-0.10***	-0.11***
	DistilBERT	-0.22***	-0.33***	-0.21***
	Longformer	0.01	-0.09	-0.01
	RoBERTa	-0.03***	-0.00	0.01
Weighted average		-0.02***	-0.12***	-0.10***

The coefficients shown here describe the comparison of gender congruity in resumes in the following three scenarios:
H1: Female resumes in comparison with male resumes.
H2: Female resumes written for male-dominated occupations in comparison with female resumes written for gender-balanced and female-dominated occupations.
H3: Male resumes written for female-dominated occupations in comparison with male resumes written for gender-balanced and male-dominated occupations.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 4.1: Hypothesis Testing: Regression Analyses on the Effect of Gender and Occupation on Congruity Scores

As can be seen in Table 4.1, the first hypothesis is supported by a significant main effect of gender in eight out of thirteen analyses as well as by the coefficient based on the weighted average congruity score. All these coefficients are significant at $p < .001$. While one analysis indicates a significant effect in the opposite direction, the coefficient of this analysis is practically equal to zero. The negative coefficients of this effect in most analyses indicate that female resumes are classified as significantly less gender congruent

than male resumes across the occupational groups. However, this effect is small in most cases, ranging between -0.00 and -0.03 (-0.22 in DistilBERT). Thus, female resumes were rated as 0% to 3% less gender congruent in comparison with male resumes (22% in the case of DistilBERT).

The second hypothesis is supported by a significant interaction contrast in ten out of thirteen analyses as well as by the coefficient based on the weighted average congruity score. All these coefficients except for the one based on SVM with BoW pre-processing are significant at $p < .001$. This effect is greater than the main effect of gender only. Specifically, the negative coefficients indicate that female resumes written in male-dominated occupations were significantly classified as 3% to 33% less gender congruent compared to female resumes written in gender-balanced and female-dominated occupations. In fact, as can be seen in Table C.2 and Figure 4.1, female gender congruity is lowest in male-dominated occupations and highest in female-dominated occupations, with the latter congruity being even higher than male gender congruity in that occupational group.

Finally, the third hypothesis is also supported by a significant interaction contrast in ten out of thirteen analyses as well as by the coefficient based on the weighted average congruity score with all ten coefficients being significant at $p < .001$. The negative coefficients of this effect indicate that male resumes written in female-dominated occupations were significantly classified as 1% to 23% less gender congruent than male resumes written in gender-balanced or male-dominated occupations. As the exact opposite of the effects observed for female gender congruity scores, male congruity scores are highest in male-dominated occupations and lowest in female-dominated occupations (see Table C.2 and Figure 4.1).

4.4.2 Model Comparisons

The size of the effects described above varies across the models that served as the base for the congruity scores. Specifically, the differences in congruity scores between genders and occupational groups were generally more pronounced in logistic regression models compared to support vector

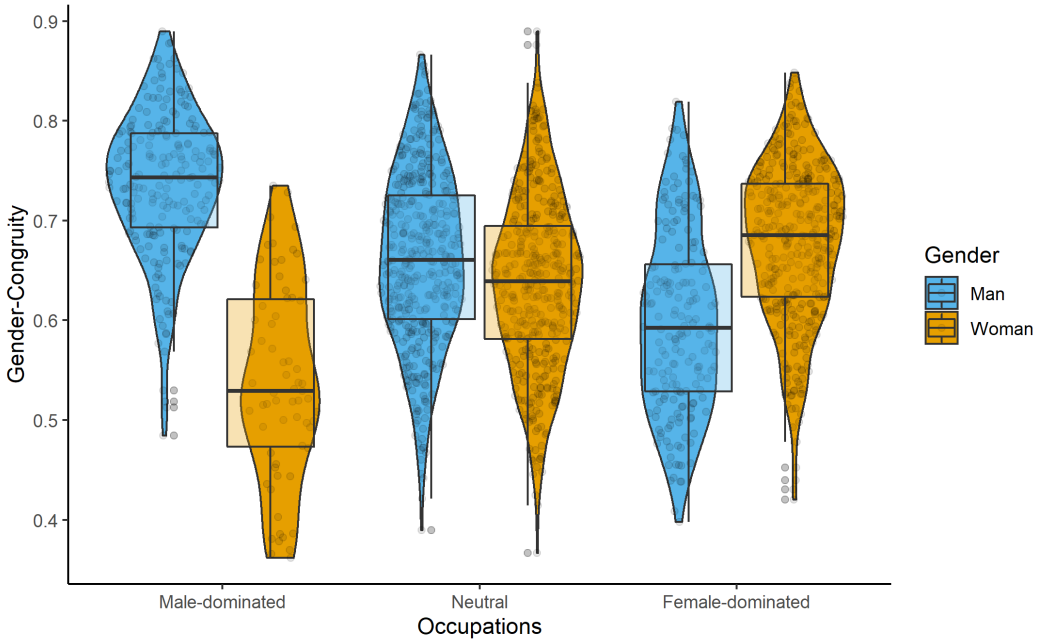


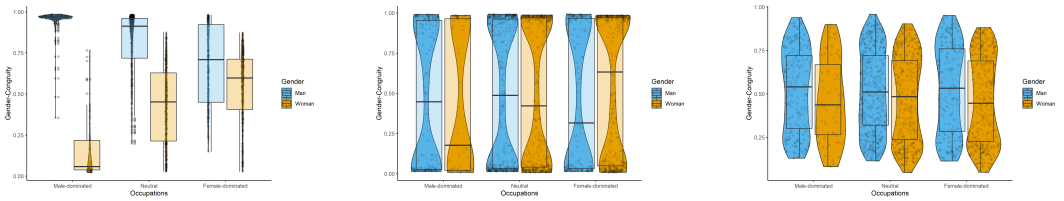
Figure 4.1: Weighted Gender Congruity Scores Sorted by Occupations and Gender.

machines, with the slight exception of the analysis of the congruity scores based on GloVe vector representations.

Regarding the transformer-based models, the relevant effects were strongest when comparing congruity scores that were based on predictions by DistilBERT. However, the effects remained non-significant when congruity scores were based on Longformer and RoBERTa predictions. As can be seen in Figure 4.2, the non-significance of the effects goes hand in hand with greater standard deviations of the congruity scores. Specifically, the gender predictions by the Longformer and RoBERTa models were much closer to either 0 (indicating 100% certainty that this person is a man) or 1 (indicating 100% certainty that this person is a woman), than predictions from any other models (see e.g. Figure C.14a).

4.4.3 Effects of Occupation and Gender on Gender Congruity

In addition to the specific contrasts that I elaborated on above, I also tested the main and interaction effects of gender and occupation in general. The re-



(a) DistilBERT **(b)** Longformer **(c)** RoBERTa

Figure 4.2: Congruity scores based on BERT predictions

sults can be found in Appendix C.2. In the following paragraphs, I will elaborate on the main effects of occupation only, since the main effect of gender and the interaction effects mimic the effects that were tested and described in section 4.4.1. The main effect of male-dominated occupations was significant in only two out of thirteen analyses and non-significant in the analysis of the weighted congruity score. This indicates that the magnitude of gender congruity in resumes written for male-dominated occupations did not differ from the gender congruity in resumes written for gender-balanced or female-dominated occupations (controlling for gender).

The main effect of female-dominated occupations was significant in six out of thirteen analyses as well as in the analysis of the weighted congruity score. However, all coefficients were quite small with none of the coefficients exceeding an absolute value of 0.02. The negative coefficients of all (significant) analyses indicate that resumes written for female-dominated occupations were marginally less gender-congruent than the resumes written for gender-balanced and male-dominated occupations (controlling for gender).

4.5 Discussion

In this study, I investigated to what extent people communicate in gender-congruent ways in their resumes in different occupations. The outcomes of multiple analyses support my first hypothesis that women communicate in a less gender-congruent way than men do. This result aligns with predictions from role congruity theory (Eagly & Karau, 2002), which states that feminine gender norms are less in line with occupational values and charac-

teristics associated with work-related success. Even though women might face severe backlash when not conforming to their gender norms (Heilman et al., 2004; Rudman, 1998; Rudman & Glick, 2001; Rudman & Phelan, 2008), they apparently resist this concern and communicate in more masculine patterns regardless. Taking into account the results of the first study of this series (see section 2.4), this outcome is not surprising. The more 'masculine' style of communicating in resumes entails impression formation techniques that are very advantageous during the application procedure (Bolino et al., 2016; Bolino & Turnley, 2003; Guadagno & Cialdini, 2007). Not only is it more common for men to promote their own capabilities in their resumes (Moss-Racusin & Rudman, 2010; Rudman, 1998), but they also use more agentic terms like 'manager' or 'improvement' instead of more communal terms like 'assist' or 'care' (see section 2.4). Thus, in the current occupational sector, it does seem more advantageous for women to communicate in a less gender-congruent way and thereby conforming to a more 'masculine' style of communication while applying for a new job or a promotion.

Nevertheless, the effect described above was comparably small (at least when controlling for occupation). The differences in gender congruity amounted to only 2% on average, even though they were significant at the $p < .001$ level. This can be explained by the fact that p-values only indicate whether the results are based on enough data to generalize to the broader population. With more than 1600 participants, it is therefore not surprising that the gender differences were significant, even though the magnitude of the effect was relatively small.

However, the magnitude of the interaction effects testing hypotheses 2 and 3 was greater and thereby hints towards effects that are more important in reality as well. Supporting my second and third hypotheses, the gender difference in congruity was significantly more pronounced in male-dominated and female-dominated occupations compared to gender-balanced occupations. Specifically, women communicated significantly less gender-congruent in male-dominated occupations compared to gender-balanced and female-dominated occupations. Similarly, men communicated significantly less gender-congruent in female-dominated occupations

compared to gender-balanced or male-dominated occupations. Both effects were comparably large with average values of 12% for the difference in female resumes and 10% in male resumes. When controlling for gender, however, the differences between occupational groups in general was negligible.

These results show that men and women communicate differently in their resumes, thereby conforming to the social norms of their occupational context (dominated by one gender). Not only did women communicate in a more masculine way, thereby conforming to general occupational values (Eagly & Karau, 2002), but also men communicated in a more feminine way when applying in female-dominated occupational fields. Even though the methods of this study do not allow for conclusions on causality, nor the mechanisms behind these differences in communication, the results again confirm predictions by the role congruity theory (Eagly & Karau, 2002). Specifically, the occupational groups of the analyses in this study were constructed by taking into account the gender distribution of the occupations according to US Labor statistics (see Appendix B). They thereby refer to a descriptive norm that holds in this context. The role congruity theory (Eagly & Karau, 2002) states that descriptive and prescriptive prejudices interact in order to maintain a social norm related to gender stereotypes (Hsu et al., 2021; Martin & Slepian, 2021). Future research should investigate the role of prescriptive norms next to descriptive roles, e.g. by taking into account the way that applicants and employers regard the importance of gender-related characteristics in applications (also see the review by Bolino et al., 2016).

4.5.1 Methodological Considerations

4.5.1.1 Model Comparisons

The effects discussed above varied across the different analyses with effects based on logistic regression models being larger than effects based on SVM in almost all cases. This can be explained by the way that gender congruity was calculated. As can be seen in Appendix C.1, SVM models performed better on most training performance measures, which resulted in congruity

scores with lower variation compared to models with lower performances (see Appendix C.1, Figures C.9 - C.13). As a consequence, the effects based on SVM models were smaller and less consistently significant compared to effects based on logistic regression models.

A similar pattern occurred in the three transformer models: While gender and occupational differences in congruity scores based on DistilBERT predictions were very large, differences in congruity scores based on Longformer and RoBERTa predictions were almost negligible. In Figure 4.2, it can be seen that the variance of congruity scores based on the latter two models was much larger than that of DistilBERT. In fact, predictions by Longformer and RoBERTa were much closer to either 0 or 1, regardless of gender, compared to predictions by any other models (see Appendix C.1), indicating that Longformer and RoBERTa were more certain about their classification, even though their performance was not considerably different from the other models. This can be explained by the fact that Longformer and RoBERTa are both designed to capture longer-range dependencies in the input sequence (Beltagy et al., 2020; Liu et al., 2019). They have larger architectures and can handle more context compared to models like DistilBERT or non-transformer-based models like TF-IDF. This ability to capture more context often leads to more accurate predictions and higher confidence levels (Beltagy et al., 2020; Liu et al., 2019), which, in turn, leads to higher variation in congruity scores and less consistent gender and occupational differences.

4.5.1.2 Operationalization of Gender Congruity

The most central concept for the research question of this part of the project was gender congruity. Following a procedure developed by Parasurama et al., 2022, I calculated the congruity scores based on the predictions of classification models. These predictions, in turn, were based on models trained on the data itself (and in the case of transformer models also on pre-trained data from the internet). Thus, the congruity scores constituted a more detailed, non-binary reflection of the model performances for each gender (see classification matrices in Appendix C.1). This way, instead of comparing the

texts to how people of different genders communicate in general, the congruity scores compared the texts only to how people in this study communicated.

On the one hand, this limits our understanding of gender-specific communication because it is not possible with this study to investigate the extent to which people communicate in a gender (in-)congruent way in the recruitment context compared to other contexts. The recruitment context is already a very male-dominated field (Bolino et al., 2016; Parasurama et al., 2022). Furthermore, it is a very formal context in comparison with e.g. text messages between friends (Newman et al., 2008). Therefore, I expect the gender difference to be even larger when comparing congruity scores that are based on data from different fields.

One way how I was able to preliminary test this idea was an analysis of the scores from the LIWC analysis; especially on the words related to masculinity | femininity and the words related to agency | communion. However, as described in Appendix C.3, the differences between the genders on these scores were almost negligible with on average less than 2% of words related to femininity and masculinity each, and only about 6.5% of words related to agency and communion each (although note that the gender-difference in the usage of communal words was significant at $p < .001$; see Hsu et al., 2021). Thus, more research is needed to investigate the actual extent to which people communicate in gender-congruent ways in the recruitment contexts compared to other contexts.

On the other hand, the construal of congruity scores allowed us to discover that even within the recruitment context, there are gender-specific differences in communication. Even though the models were trained on data from the occupational sector only, these gender differences were significant and large enough (10-12% on average) to show clear patterns also with regard to the occupational groups (see figure 4.1).

4.5.2 Limitations and Future Research

Even though the analyses in this research led to important insights into gendered communication during the application process, the data that we used in this research had many limitations. First, due to legal and ethical implications, we were not able to obtain resumes that were used in real-world application processes, such as those from linkedin.com or indeed.com (for an example of research based on real CVs, see Blommaert et al., 2014). Instead, the data of this study originated from Prolific, an online tool to collect responses for survey research. Even though participants were asked to copy and paste textual information from their resumes in the respective fields, the data was partially of very low quality (see Section 1.1) which, in turn, also resulted in relatively poor-performing classification models.

Second, the data was collected in the United States, a country that is quite divided in political views, but nevertheless scores well regarding economic participation and opportunities for women (World Economic Forum, 2022). Interestingly, Hsu et al., 2021 found that certain gender differences in communication were larger in both English-speaking countries as well as in countries with higher gender parity. Therefore, it would be interesting to investigate gender congruity in resumes in different cultural contexts.

Third, this study only focused on two self-reported genders: men and women. Studies, which also take into account other genders, or differentiate between cis- and transgender people would be able to obtain more insights into the development of gender differences in communication.

Fourth, participants in this study were people who already worked in their respective occupations. This way, they were already exposed to the gendered social norms in their occupational fields and they may have adapted to these social norms, not only to avoid backlash but also to fit in the context in general (Cialdini & Trost, 1998). For example, women who worked in a very competitive setting, which values agentic communication, might communicate in a more agentic way not only to avoid backlash (Moss-Racusin & Rudman, 2010) but also subconsciously to mimic other people in their environment and generally conform to the social norms

in their everyday life (Cialdini & Trost, 1998). Future research could investigate communication by people who are not already part of such an environment, such as students, or people who have worked in a different occupation before.

Lastly, the focus of this research was on the differences in occupations in general. However, our data did not allow us to distinguish between different hierarchical positions within a company. This could be a confounding factor in that women often hold lower-ranking positions in a company (World Economic Forum, 2022) and leadership is associated with more masculine qualities (Eagly & Karau, 2002). Future research should closely examine the interrelationship between occupation and job position on gender differences in occupations.

4.5.3 Societal Implications

The current research sheds light on the different communication styles used by men and women in their resumes for different occupations. I found that women communicate in a less gender-congruent way than men in general and that this difference is greater in male-dominated occupations, while it is the other way around in female-dominated occupations. Previous research found that gender-incongruent communication can cause social and economic penalties, both in women (Heilman et al., 2004; Rudman & Glick, 2001; Rudman & Phelan, 2008) and in men (Rudman, 1998; Rudman & Glick, 2001; Rudman & Phelan, 2008). However, it remains up to future research to investigate the way that occupational and gender norms interact with each other and influence the backlash on hiring outcomes on the basis of communicational differences in resumes.

Furthermore, even though we found that people tend to adapt their communication styles to the organizational context that they apply for, the classification models that we used were still able to differentiate between male and female resumes, and thus gender congruity turned out to be quite high in most models as well (see Appendix C.2). Thus, my research shows that differences between the genders still widely exist and are further fuelled

by fear of backlash for example (Moss-Racusin & Rudman, 2010; Rudman, 1998). Consequently, if people keep communicating in gender-congruent ways in order to avoid backlash, they continue to reproduce gender segregation and workplace discrimination against the minority genders (Guadagno & Cialdini, 2007; Heilman & Eagly, 2008; Rudman & Phelan, 2008). Therefore, future research needs to continue investigating ways to implement strategies that reduce workplace discrimination. These strategies can allow women, men, and people of other genders alike to communicate in ways that enhance their chances of workplace success, liberated from gender norms and associated prejudice.

5. Conclusion

Written together

In this research, we investigated how men and women communicate in their resumes in different occupations.

Model wise, we showed that traditional and newer machine learning algorithms can distinguish between text written by men and women across occupational groups with high levels of accuracy. Noteworthy are the performances of the RoBERTa and Longformer models which showed good results even with a small imbalanced dataset. We also emphasized the importance of optimizing decision thresholds to enhance these models' performance. Additionally, we highlighted the role of selecting suitable performance metrics, which depend on whether the data at hand is balanced or imbalanced. This choice of metrics is important, because it helps us interpret and compare results accurately across different datasets and models.

Transitioning into a more detailed analysis of gender-specific communication patterns, we observed some interesting trends. We found that overall, women communicate in a less gender-congruent way. Importantly, this effect is amplified in male-dominated occupations, while it is the other way around in female-dominated occupations. This demonstrates the broad existence of gender norms across occupations, which highlights the importance of accounting for communication styles between genders when evaluating resumes of potential future employees.

A limitation of this work is that we cannot investigate hiring decisions given the different textual features that we found. Nevertheless, we were able to show that women used more communal and less agentic terms (see Section 2.4) even though masculine and self-promoting communication styles are often seen as the key to work-related success (Eagly & Karau, 2002). Therefore, as long as these gender norms continue to exist in the

minds of both employees and employers, the occupational gender segregation seen across the world will be maintained (World Economic Forum, 2022).

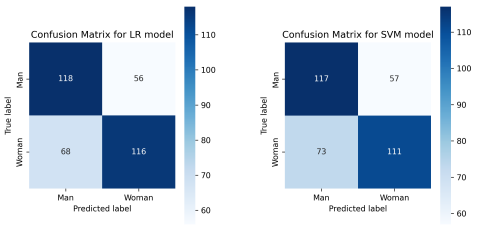
There are multiple ways to reduce the impact of these gender differences in the hiring process. First, our research has focused on the candidate employees and how they communicate in their resumes. Using psychological interventions, such as information campaigns, the gender gap in self-promotion might be reduced and women could feel more secure in promoting their abilities in their resumes (Kessel et al., 2021).

Second, it is important that employers know about these gender differences and about the potential effects that they have on hiring outcomes (Rudman & Glick, 2001; Rudman & Phelan, 2008). Only with knowledge about these gender differences, employers can select a fair instrument for evaluating candidates such as application forms (Risavy et al., 2017; Risavy et al., 2022), or an appropriate, transparent, and gender-aware artificial intelligence (Gonen & Goldberg, 2019).

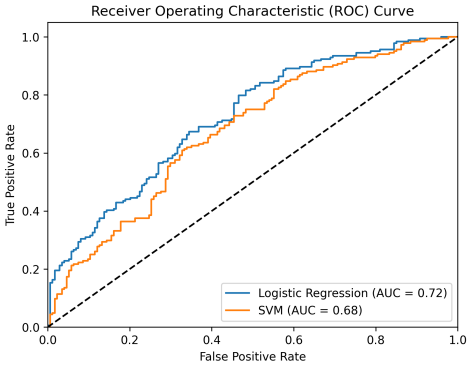
Lastly, a societal change in values and norms surrounding gender, as well as occupations can only come from society itself. The established norms in Western societies often place a high value on competitiveness and individualism, patterns that subtly perpetuate gender inequalities. Such standards, predominantly shaped by a historically male-dominated workforce, highlight the need for change in order to foster gender equality. Despite current efforts, including Equity, Diversity, and Inclusion (EDI) initiatives, gender quota laws, and parental leave policies, these standards still often subtly favor one gender over the other, particularly in the recruitment process and job applications. Thus, a more sustainable change toward gender equality requires us to acknowledge and actively address these persistent, gendered patterns in our society. As reported by the World Economic Forum, at the current pace, it could take an estimated 136 more years to achieve gender parity (Armstrong, 2021), indicating the urgency of reevaluating our societal norms.

Appendices

A. Model Performances

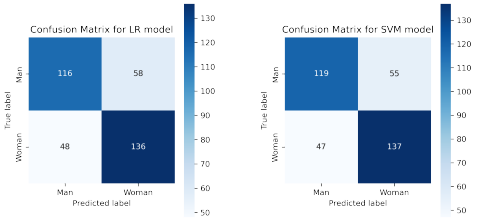


(a) Confusion Matrix

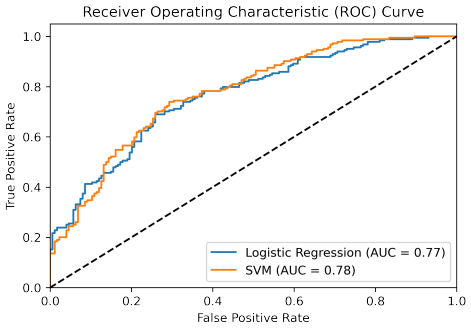


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.1: Performance of Baseline (BoW) Models

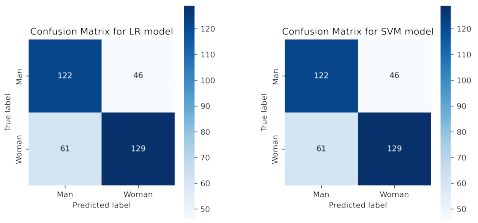


(a) Confusion Matrix

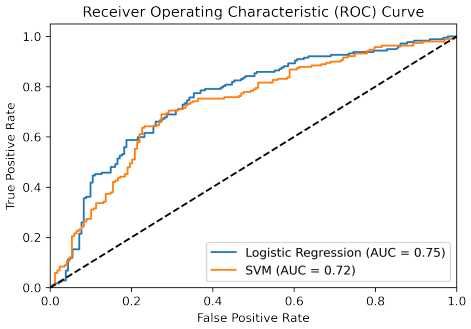


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.2: Performance of TF-IDF Models

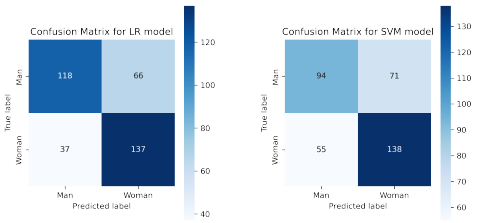


(a) Confusion Matrix

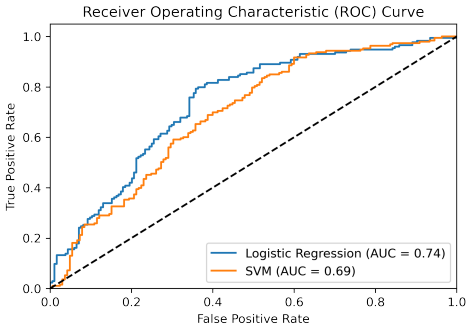


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.3: Performance of Word2Vec Models

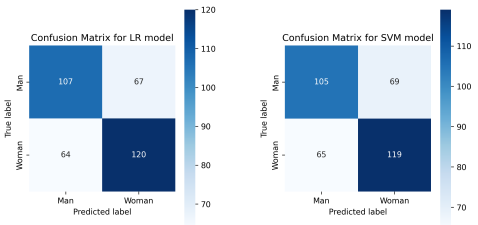


(a) Confusion Matrix

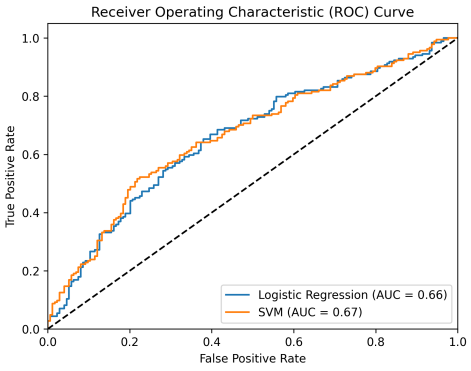


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.4: Performance of GloVe Models

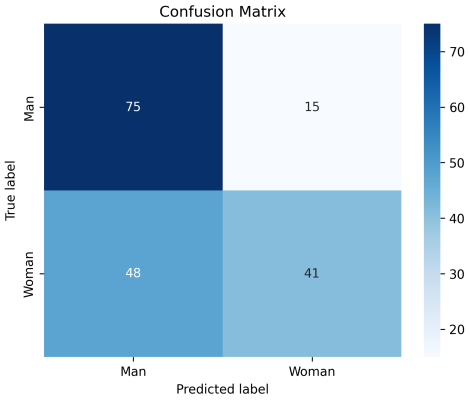


(a) Confusion Matrix

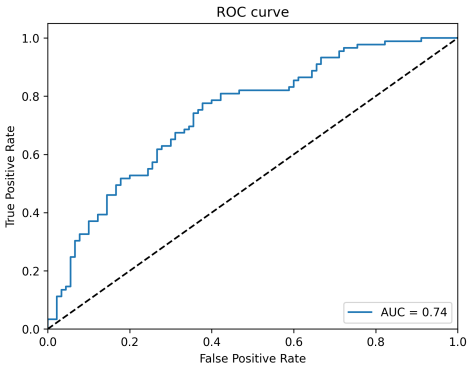


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.5: Performance of LIWC Models

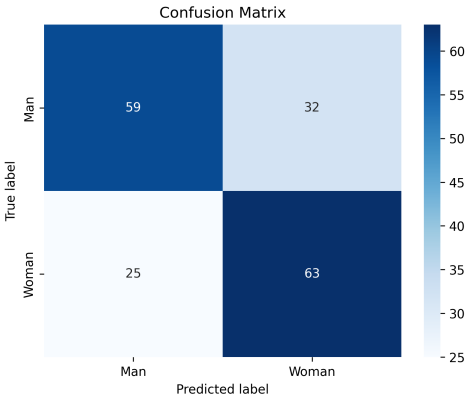


(a) Confusion Matrix

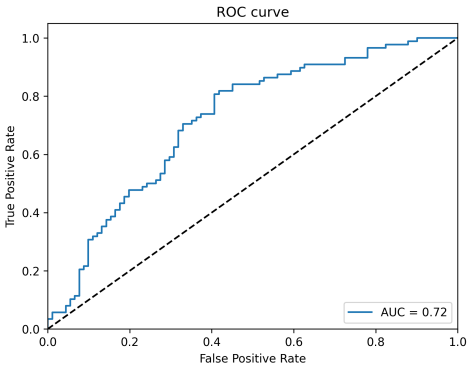


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.6: Performance of DistilBERT Model

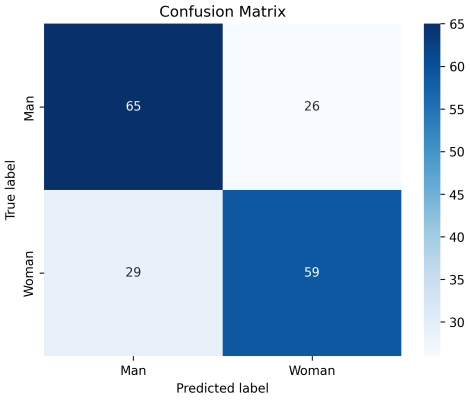


(a) Confusion Matrix

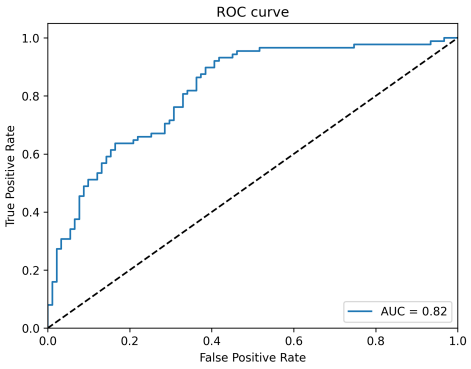


(b) Receiver Operating Characteristic (ROC) Curve

Figure A.7: Performance of Longformer Model



(a) Confusion Matrix



(b) Receiver Operating Characteristic (ROC) Curve

Figure A.8: Performance of RoBERTa Model

B. Benchmark Occupations

Occupational Group	Occupation	Men	Women	% Women	Total	% F USLS
Male-Dominated	Architecture and engineering	56	18	24%	74	24%
	Computer and mathematical	152	49	24%	201	26%
Gender-Balanced	Legal	28	22	44%	50	42%
	Arts, design, entertainment, sports, and media	58	64	52%	122	50%
	Protective service	5	5	50%	10	50%
	Management	101	58	36%	159	52%
	Business and financial operations	117	85	42%	202	55%
	Food preparation and serving related	20	33	62%	53	59%
	Life, physical, and social science	33	62	65%	95	61%
	Sales and related	65	70	52%	135	62%
Female-Dominated	Office and administrative support	43	78	64%	121	72%
	Community and social service	15	44	75%	59	72%
	Personal care and service	8	21	72%	29	72%
	Educational instruction and library	64	117	65%	181	74%
	Farming, fishing, and forestry	2	6	75%	8	75%
	Healthcare support	58	118	67%	176	85%
	Other	61	53	46%	114	
Total		886	903	50%	1789	53%

Occupations with number of CVs and proportion of female participants (%F) in our CV data set.
% F USLS are the official percentages of female employees per occupation taken from the US Labor Statistics (2021).
The table is adopted from Yang et al., 2022.

Table B.1: Occupational Groups

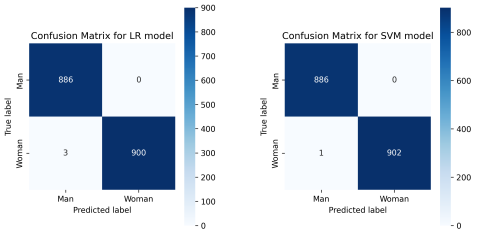
C. Occupational Differences in Gender Congruent Communication

Written by Maike

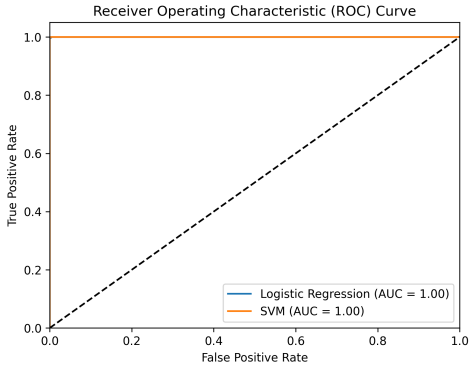
C.1 Performance Metrics of Models Trained on all Data

Model	Classifier method	Accuracy	F1
Baseline (Bow)	LR	1.00	1.00
	SVM	1.00	1.00
TF-IDF	LR	0.87	0.87
	SVM	1.00	1.00
Word2Vec	LR	0.69	0.70
	SVM	0.78	0.79
Glove	LR	0.70	0.71
	SVM	0.83	0.83
LIWC	LR	0.67	0.68
	SVM	0.68	0.68
	Data preparation		
DistilBERT	Cleaned text	0.77	0.76
Longformer	Cleaned text	0.93	0.94
RoBERTa	Cleaned text	0.78	0.77

Table C.1: Performance Results for each Model Trained on all Data

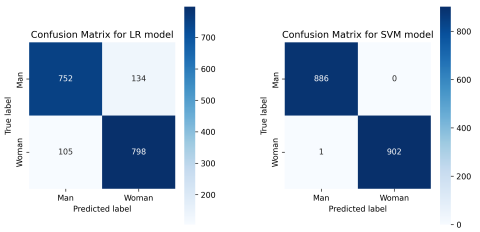


(a) Confusion Matrix

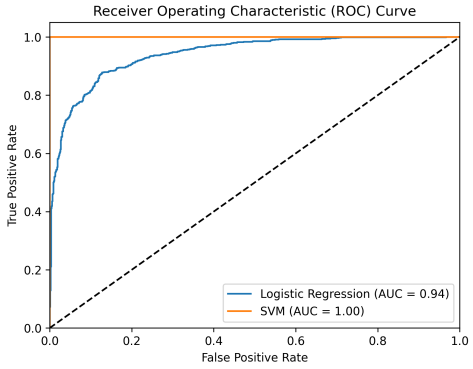


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.1: Performance of Baseline (BoW) Models Trained on all Data

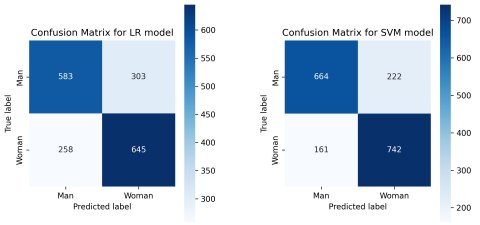


(a) Confusion Matrix

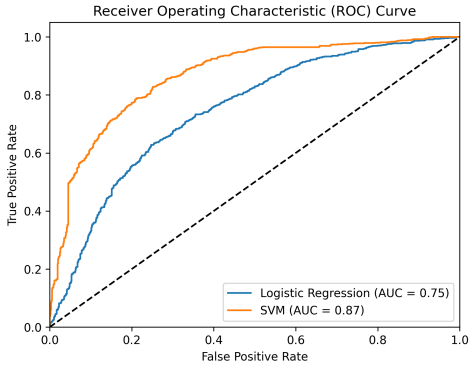


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.2: Performance of TF-IDF Models Trained on all Data

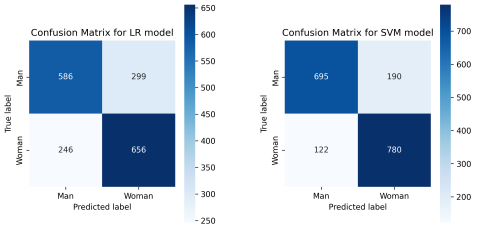


(a) Confusion Matrix

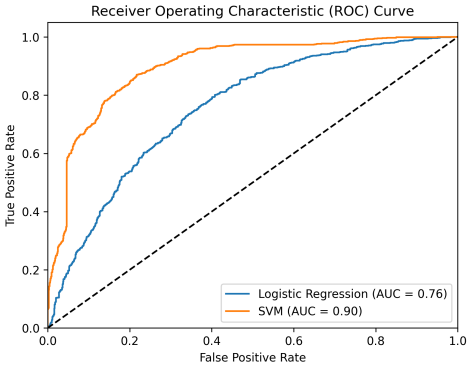


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.3: Performance of Word2Vec Models Trained on all Data

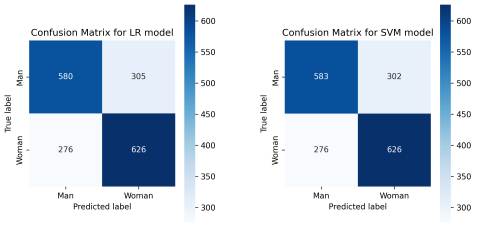


(a) Confusion Matrix

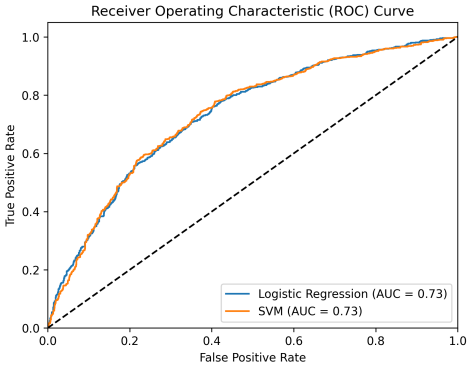


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.4: Performance of GloVe Models Trained on all Data

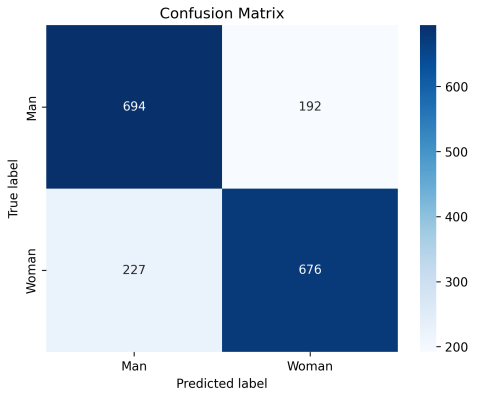


(a) Confusion Matrix

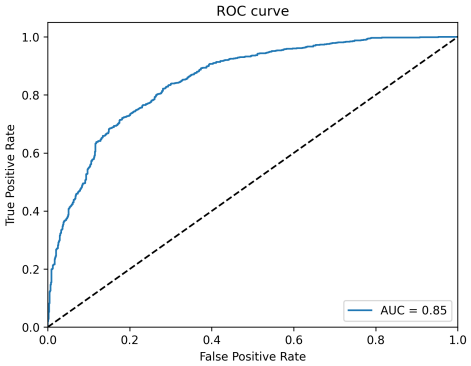


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.5: Performance of LIWC Models Trained on all Data

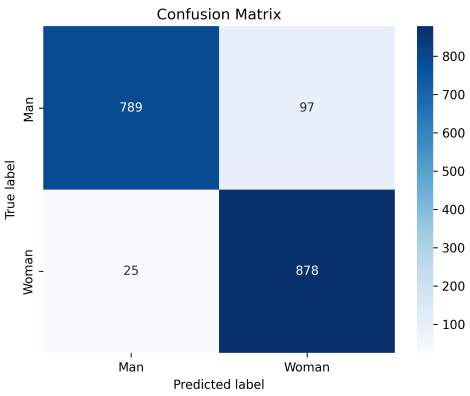


(a) Confusion Matrix

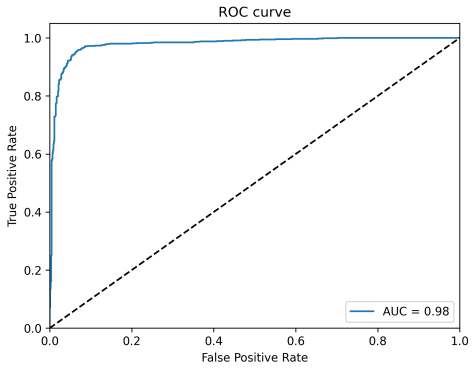


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.6: Performance of DistilBERT Model Trained on all Data

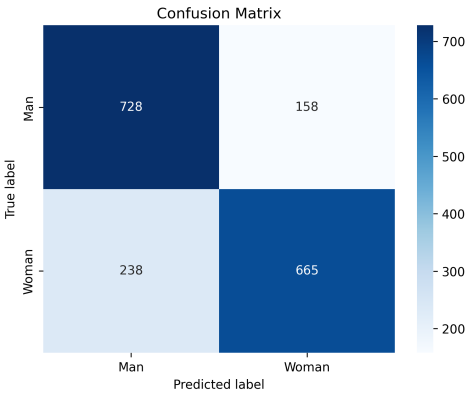


(a) Confusion Matrix

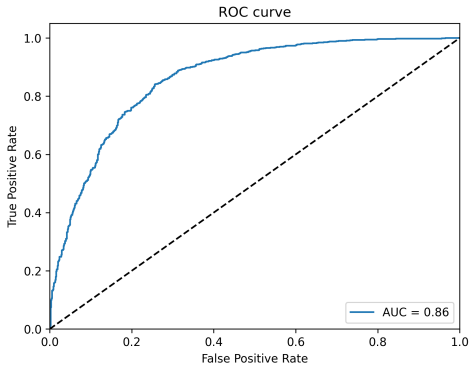


(b) Receiver Operating Characteristic (ROC) Curve

Figure C.7: Performance of Longformer Model Trained on all Data



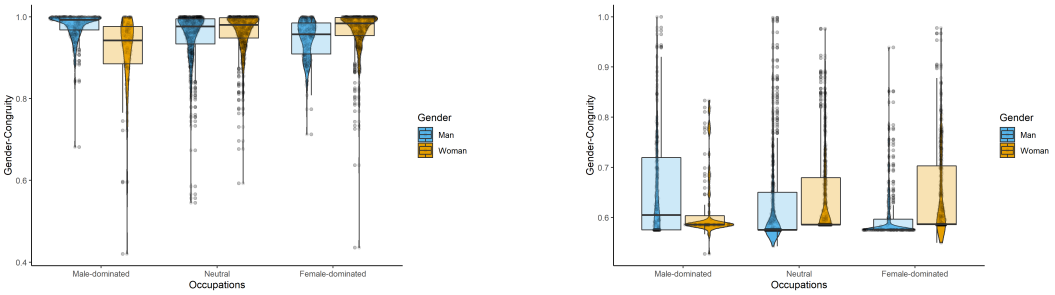
(a) Confusion Matrix



(b) Receiver Operating Characteristic (ROC) Curve

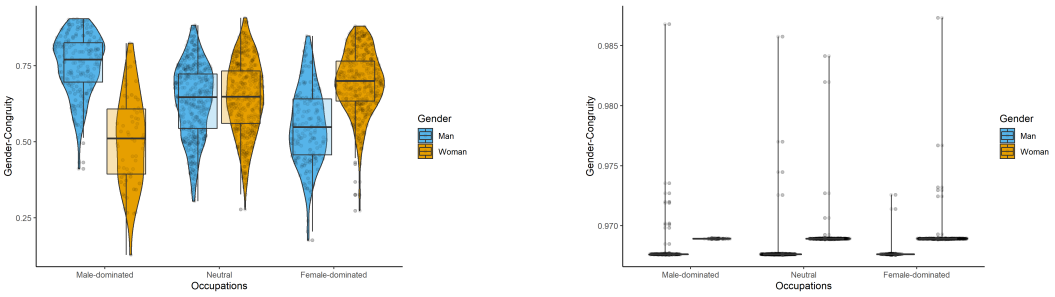
Figure C.8: Performance of RoBERTa Model Trained on all Data

C.2 Congruity Analyses



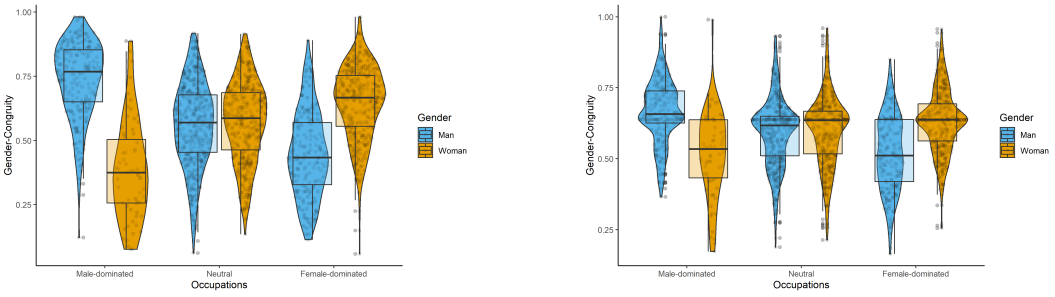
(a) Logistic Regression Classification **(b)** SVM Classification

Figure C.9: Congruity Scores Based on BoW Predictions



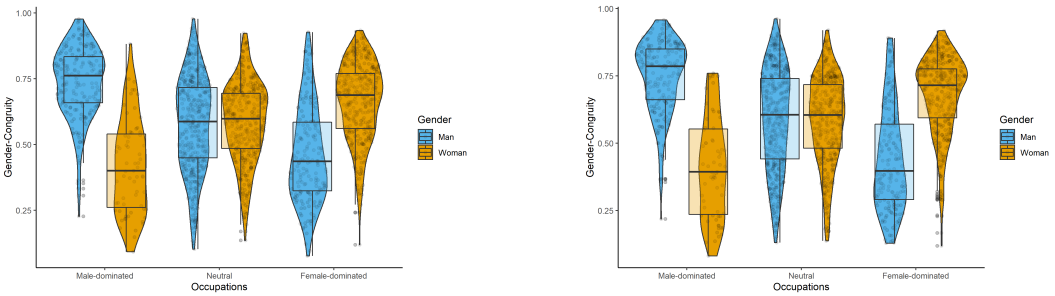
(a) Logistic Regression Classification **(b)** SVM Classification

Figure C.10: Congruity Scores Based on TF-IDF Predictions



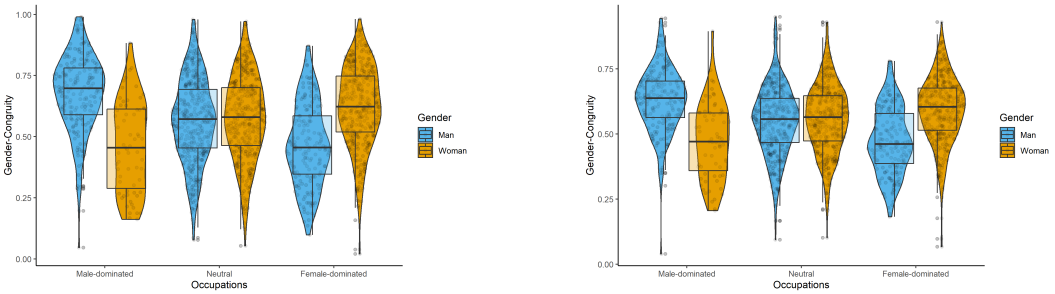
(a) Logistic Regression Classification **(b)** SVM Classification

Figure C.11: Congruity Scores Based on Word2Vec Predictions



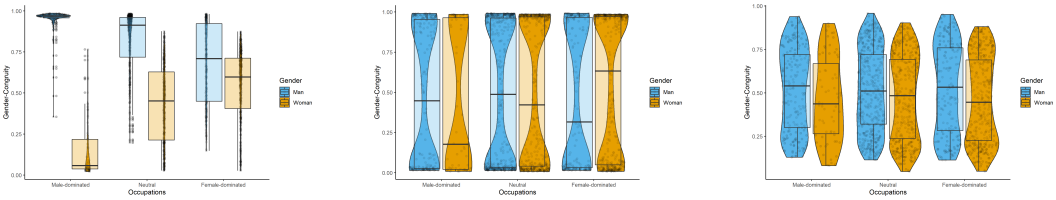
(a) Logistic Regression Classification **(b)** SVM Classification

Figure C.12: Congruity Scores Based on GloVe Predictions



(a) Logistic Regression Classification **(b)** SVM Classification

Figure C.13: Congruity Scores Based on LIWC Predictions



(a) DistilBERT **(b)** Longformer **(c)** RoBERTa

Figure C.14: Congruity Scores Based on BERT Predictions

Data preprocessing	Classifier	Intercept	MDO	FDO	Gender (f)	MDO : gender	FDO : gender
Baseline (BoW)	LR	0.95***	-0.01*	-0.00	-0.01***	-0.02***	0.00
	SVM	0.63***	0.00	-0.01	-0.00	-0.02***	0.01**
TF-IDF	LR	0.63***	-0.01	-0.01*	-0.02***	-0.09***	0.04***
	SVM	0.97***	0.00	-0.00	0.00***	-0.00	0.00
Word2Vec	LR	0.56***	-0.00	-0.01*	-0.02***	-0.12***	0.06***
	SVM	0.59***	-0.00	-0.01**	-0.01	-0.05***	0.04***
GloVe	LR	0.57***	-0.00	-0.01	-0.02***	-0.11***	0.06***
	SVM	0.57***	-0.00	-0.02**	-0.02***	-0.11***	0.08***
LIWC	LR	0.56***	-0.00	-0.02**	-0.01	-0.08***	0.05***
	SVM	0.55***	-0.00	-0.01**	-0.00	-0.05***	0.04***
	DistilBERT	0.60***	-0.05***	-0.01	-0.22***	-0.13***	0.09***
	Longformer	0.49***	-0.03	0.01	0.01	-0.01	0.02
	RoBERTa	0.49***	-0.00	-0.00	-0.03***	-0.01	0.01
Weighted average		0.64***	-0.01	-0.01*	-0.02***	-0.06***	0.03***

Coefficients of the occupational groups refer to the comparison between the respective group and the other two groups simultaneously (the gender-balanced occupations and the occupations dominated by the other gender), while coefficients of gender refer to the congruity of female resumes in comparison with male resumes. All coefficients are sum-to-zero coded. The last two columns display the interaction coefficients of occupational group and gender.
MDO: Male-dominated occupations. FDO: Female-dominated occupations.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table C.2: Results of Regression Analyses of Congruence Scores

C.3 LIWC Regression Analyses

Multiple linear regression analyses were conducted to investigate gender and occupational differences in masculinity, femininity, agency, and communion, according to the LIWC values. Results are presented in tables C.3 and C.4.

Variable	$M(SD)_{men}$	$M(SD)_{women}$
Masculinity	1.85 (1.73)	1.79 (2.01)
Femininity	1.30 (1.64)	1.50 (1.78)
Agency	6.26 (3.88)	6.52 (3.83)
Communion	6.63 (4.38)	8.15 (4.05)

Table C.3: Means of LIWC Features for Men and Women

Variable	Intercept	MDO	FDO	Gender (f)	MDO : gender	FDO : gender
Masculinity	1.91***	0.25*	-0.13	0.09	0.27**	0.07
Femininity	1.41***	0.20*	0.21**	0.06	-0.16	0.11
Agency	6.29***	-0.29	0.17	0.06	-0.11	-0.13
Communion	7.26***	-0.63**	1.16***	0.36**	-0.35	-0.07

Coefficients of the occupational groups refer to the comparison between the respective group and the other two groups simultaneously (the gender-balanced occupations and the occupations dominated by the other gender), while coefficients of gender refer to the congruity of female resumes in comparison with male resumes. All coefficients are sum-to-zero coded. The last two columns display the interaction coefficients of occupational group and gender.
MDO: Male-dominated occupations. FDO: Female-dominated occupations.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table C.4: Regression Analysis Results LIWC

Bibliography

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts [Published on January 23, 2006]. *Text & Talk*, *23*(3), 321–346. https://doi.org/10.1515/text.2003.014

Armstrong, M. (2021). It will take another 136 years to close the global gender gap. *World Economic Forum. Accessed November*, *12*, 2021.

Bakan, D. (1966). *The duality of human existence : Isolation and communion in western man*. Beacon Press. https://cir.nii.ac.jp/crid/1130282270046949376

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, *91*(1), 3–26. https://doi.org/10.1037/0033-2909.91.1.3

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Blommaert, L., Coenders, M., & Van Tubergen, F. (2014). Discrimination of arabic-named applicants in the netherlands: An internet-based field experiment examining different phases in online recruitment procedures. *Social forces*, *92*(3), 957–982.

Böhm, S., Linnyk, O., Kohl, J., Weber, T., Teetz, I., Bandurka, K., & Kersting, M. (2020). Analysing gender bias in it job postings: A pre-study based on samples from the german job market. *Proceedings of the 2020 on computers and people research conference*, 72–80.

Bolino, M. C., Long, D., & Turnley, W. H. (2016). Impression management in organizations: Critical questions, answers, and areas for future research. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 377–406.

Bolino, M. C., & Turnley, W. H. (2003). More than one way to make an impression: Exploring profiles of impression management. *Journal of Management*, *29*(2), 141–160.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 1–47.

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal 1. *Journal of Social issues*, *28*(2), 59–78.

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women [Visited on 01/06/2023].

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.

DeJesus, J. M., Umscheid, V. A., & Gelman, S. A. (2021). When gender matters in scientific communication: The role of generic language. *Sex Roles*, *85*(9-10), 577–586. https://doi.org/10.1007/s11199-021-01240-7

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Hillsdale, NJ: Erlbaum.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological review*, *109*(3), 573.

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender*, *12*(174), 123–174.

Franco, M. C., Rice, D. B., Schuch, H. S., Dellagostin, O. A., Cenci, M. S., & Moher, D. (2021). The impact of gender on scientific writing: An observational study of grant proposals. *Journal of Clinical Epidemiology*, *136*, 37–43.

Gardner, W. L., & Martinko, M. J. (1988). Impression management in organizations. *Journal of management*, *14*(2), 321–338.

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.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Guadagno, R. E., & Cialdini, R. B. (2007). Gender differences in impression management in organizations: A qualitative review. *Sex Roles*, *56*, 483–494.

Hargrave, L., & Langengen, T. (2021). The gendered debate: Do men and women communicate differently in the house of commons? *Politics amp; Gender*, *17*(4), 580–606. https://doi.org/10.1017/S1743923X20000100

He, J. C., & Kang, S. K. (2021). Covering in cover letters: Gender and self-presentation in job applications. *Academy of Management Journal*, *64*(4), 1097–1126. https://doi.org/10.5465/amj.2018.1280

Heilman, M. E., & Eagly, A. H. (2008). Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology*, *1*(4), 393–398. https://doi.org/10.1111/j.1754-9434.2008.00072.x

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male gender-typed tasks. *Journal of applied psychology*, *89*(3), 416.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Horbach, S. P., Schneider, J. W., & Sainte-Marie, M. (2022). Ungendered writing: Writing styles are unlikely to account for gender differences in funding rates in the natural and technical sciences. *Journal of Informetrics*, *16*(4), 101332. https://doi.org/https://doi.org/10.1016/j.joi.2022.101332

Hsu, N., Badura, K. L., Newman, D. A., & Speach, M. E. P. (2021). Gender, "masculinity," and "femininity": A meta-analytic review of gender differences in agency and communion [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *147*(10), 987–1011. https://doi.org/10.1037/bul0000343

Joshi, P. D., Wakslak, C. J., Huang, L., & Appel, G. (2021). Gender differences in communicative abstraction and their organizational implications. *Rutgers Business Review*, *6*(2).

Kessel, D., Mollerstrom, J., & Van Veldhuizen, R. (2021). Can simple advice eliminate the gender gap in willingness to compete? *European Economic Review*, *138*, 103777.

Kolev, J., Fuentes-Medel, Y., & Murray, F. (2020). Gender differences in scientific communication and their impact on grant funding decisions. *AEA Papers and Proceedings*, *110*, 245–49. https://doi.org/10.1257/pandp.20201043

Lerchenmueller, M. J., Sorenson, O., & Jena, A. B. (2019). Gender differences in how scientists present the importance of their research: Observational study. *bmj*, *367*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martin, A. E., & Slepian, M. L. (2021). The primacy of gender: Gendered cognition underlies the big two dimensions of social cognition. *Perspectives on Psychological Science*, *16*(6), 1143–1158.

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

Moss-Racusin, C. A., & Rudman, L. A. (2010). Disruptions in women's self-promotion: The backlash avoidance model. *Psychology of women quarterly*, *34*(2), 186–202.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, *45*(3), 211–236.

Nguyen, H. (2021). The (great) persuasion divide? gender disparities in debate speeches and evaluations, 1–109.

Parasurama, P., Sedoc, J., & Ghose, A. (2022). Gendered information in resumes and hiring bias: A predictive modeling approach. *Available at SSRN 4074976*.

Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., & Kamar, E. (2019). What you see is what you get? the impact of representation criteria on human bias in hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 125–134.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (tech. rep.).

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

Perkowski, P. (2023). Gender representation and the adoption of hiring algorithms: Evidence from mba students and executives. *Available at SSRN 4367113*.

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.

Ponizovskiy, V., Ardag, M., Grigoryan, L., Boyd, R., Dobewall, H., & Holtz, P. (2020). Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality*, *34*(5), 885–902.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly*, *26*(4), 269–281.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.

Risavy, S. D., et al. (2017). The resume research literature: Where have we been and where should we go next. *Journal of Educational and Developmental Psychology*, *7*(1), 169–187. https://doi.org/10.5539/jedp.v7n1p169

Risavy, S. D., Robie, C., Fisher, P. A., & Rasheed, S. (2022). Resumes vs. application forms: Why the stubborn reliance on resumes? *Frontiers in Psychology*, 4622.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of personality and social psychology*, *74*(3), 629.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of social issues*, *57*(4), 743–762.

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in organizational behavior*, *28*, 61–79.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Scharff, C. (2015). Blowing your own trumpet: Exploring the gendered dynamics of self-promotion in the classical music profession. *The Sociological Review*, *63*(1_suppl), 97–112. https://doi.org/10.1111/1467-954X.12243

Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *Journal of applied psychology*, *57*(2), 95.

Schein, V. E. (1975). Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of applied psychology*, *60*(3), 340.

scikit-learn developers. (2023). RBF SVM parameters. Retrieved June 19, 2023, from https://scikit-learn/stable/auto_examples/svm/plot_rbf_parameters.html

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Smith, A. N., Watkins, M. B., Burke, M. J., Christian, M. S., Smith, C. E., Hall, A., & Simms, S. (2013). Gendered influence: A gender role perspective on the use and effectiveness of influence tactics. *Journal of Management*, *39*(5), 1156–1183.

Spence, J. T., Helmreich, R., & Stapp, J. (1975). Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of personality and social psychology*, *32*(1), 29.

Steffens, M. C., Viladot, M. A., & Scheifele, C. (2019). Male majority, female majority, or gender diversity in organizations: How do proportions affect gender stereotyping and women leaders' well-being? *Frontiers in psychology*, *10*, 1037.

Streib, J., Rochmes, J., Arriaga, F., Tavares, C., & Weed, E. (2019). Presenting their gendered selves? how women and men describe who they are, what they have done, and why they want the job in their written applications. *Sex Roles*, *81*, 610–626.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study, rev*. Sage Publications, Inc.

Wood, W., & Eagly, A. H. (2015). Two traditions of research on gender identity. *Sex Roles*, *73*, 461–473.

World Economic Forum. (2022). *Global gender gap report* (tech. rep. July 2022). Retrieved June 4, 2023, from https://www.weforum.org/reports/global-gender-gap-report-2022/in-full/

Yang, J., Njoto, S., Cheong, M., Ruppanner, L., & Frermann, L. (2022). Professional presentation and projected power: A case study of implicit gender information in english cvs. *arXiv preprint arXiv:2211.09942*.

Zide, J., Elman, B., & Shahani-Denning, C. (2014). Linkedin and recruitment: How profiles differ across occupations. *Employee relations*, *36*(5), 583–604.