Bachelor Thesis

Examining and Mitigating Gender Bias in German Word Embeddings

---

Author

David Prokosch

Supervisor

Dr. Matthias Aßenmacher
Department of Statistics

Department of Statistics

Ludwig-Maximilians-Universität München

Munich, 10th of February 2023

Abstract

Word embeddings are a common method of natural language processing that assigns a mathematical representation to words. Recent research shows that word embeddings adopt and reinforce gender biases. This problem has been widely addressed for the English language, but less in other languages. In this thesis methods are investigated to examine and mitigate bias in word embeddings for the German language. For this purpose, the specific characteristics of the German language such as the three grammatical genders or the generic masculine are considered. The word embedding association test and the proposed word pair embedding association tests were used to show that bias can be detected in word embeddings. Additionally, it is discussed how concepts such as (grammatical) gender can be identified and interpreted in word embeddings using dimensionality reduction techniques. Furthermore, different methods that had previously been used to debias word embeddings in languages with grammatical gender are applied to the German language and evaluated. It is shown that some methods can reduce gender bias in the word embeddings for the German language. Nevertheless, some issues remain and are discussed.

Contents

1. Introduction

For humans, learning is not about memorizing examples, but generalizing on them. For example, a child does not learn what a cat looks like by remembering examples of cats, but by being able to distinguish the general concept of "cat" from other types of animals. The aim of machine learning algorithms is to learn the same way.

More specifically, Natural Language Processing (NLP) models are designed to learn this way, too. They are created in order to reproduce how humans write, speak, and treat natural language. For example, taking Wikipedia articles as input, they generalize these masses of text to develop an imitation of language sense. Natural language processing models are applied in language translation, hate speech detection, spam detection, and various other cases.

As machine learning is used more and more for decision-making, it is starting to have a huge impact on society. Especially, the importance of NLP models has increased in recent years with exceeding growth of available computing power. This could be viewed as a positive trend for overall fairness: If an algorithm is non-discriminatory, then it might be fairer than a human with their own bias or their arbitrary instinct. On the other hand, observational data is used for training models in nearly all cases of NLP. Barocas et al. (2019) explain that observational data likely reflects historical prejudice, social inequalities, or cultural stereotypes. Note that the Wikipedia text corpus, as often used for NLP, is an observation of how humans apply language. Historically, scientists have often been white males, and thus Wikipedia articles about scientists could lead to the generalized interpretation of a scientist as primarily white and male. Another example would be the belittlement of young girls in comparison to young boys in the German language. der deutschen Sprache (2023) stated that *Mädchen* (girl) was the diminutive form of the old German word *Magd*, which was the pair word to *Junge* (boy). In the last 300 years the term maid has disappeared from the German language and the diminutive *Mädchen* is now the opposite word to *Junge*. The fact that girls are perceived as cuter

and that they are taken less seriously in the German language is an example of historical discrimination. In the context of machine learning, generalizing data, in the NLP case on text corpora, means including these problematic historical dynamics in the predictor. As a result, the use of machine learning can reproduce discrimination.

To make a machine understand languages, parts of text must first be assigned a mathematical representation. This thesis will primarily focus on word embeddings that learn mathematical representations for words from huge text corpora. As previously emphasized, the models could be discriminatory. In their paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", Bolukbasi et al. (2016) addressed this issue. They recognize the relevance of gender bias in NLP models and emphasize "how blatantly sexist the embeddings are". To address the issue, they propose several methods that build on the closeness of words and word concepts to each other. They construct subspaces of the embeddings respectively directions that represent gender. Granted that, they define bias in word embeddings based on the projection of a word on a gender subspace. The results show that for example, *Computer Programmer* is associated with the male gender concept, while *Homemaker* is associated with the female gender concept. The work was very revealing and subsequently, more researchers started to investigate the existence and importance of biases in word embeddings.

Since the paper was published, a lot of research on the topic has been done, although research has been focused mainly on the English language. This research may be insufficient for many other languages. For example, languages like Spanish, Dutch, and German have a grammatical gender, which provides a justified explanation for the closeness of a word to a gender concept. Therefore, the definition of closeness as an indicator of gender bias proposed by Bolukbasi et al. (2016) is not entirely applicable to such languages. For example, *Programierer* (male programmer) should be closer to the concept of male than *Programmiererin* (female programmer).

Trying to overcome this challenge, Zhou et al. (2019) propose new definitions and methods. They focus on Spanish and French embeddings, where both adjectives and nouns are either masculine or feminine. In addition, they show that bilingual embeddings also contain gender bias. Based on this, they propose an approach for debiasing bilingual word embeddings where Spanish or French embeddings are aligned together with English

word embeddings.

The aim of this thesis is to examine gender bias in German word embeddings. Then, several mitigating methods will be used to reduce the bias in monolingual German word embeddings and in bilingual German and English word embeddings.

2. Problems and Definitions

2.1. Fairness in Machine Learning

For Barocas et al. (2019), assessing fairness in machine learning means shifting the goal from faithfully reflecting the data to questioning the data and developing systems in a way that corresponds to a certain idea of fair behavior. To understand the need for fairness in machine learning in general and the mechanisms of unfair algorithms, the following example will be considered:

In the USA, the justice system and the police are increasingly using algorithms to assess offenders who recidivate. One of these algorithms is called COMPAS, a commercial tool made by Northpointe, Inc. which assigns scores from 1 for "low risk of recidivism" to 10 for "high risk of recidivism" to defendants. It is one of the most widely used assessment tools in the United States. The company does not publish its score calculations and the methodology of the algorithm. However, Angwin et al. (2016b) investigate that the COMPAS score is derived from 137 questions, which are answered by the defendants or are derived by their Crime Index. The race of the defendant is not explicitly asked.

Keep in mind that simply removing race from the training data cannot solve discrimination in most cases. There are almost certainly variables in the data that correlate with race. Unfortunately, these variables often cannot be omitted because they can contain relevant information about the defendant's recidivism risk. For example, one question on the test is "How often did you get in fights while at school?". This question clearly refers to the potential danger that a person poses to society. On the other hand, people in poorer, more socially disadvantaged neighborhoods are more likely to be involved in a fight through no fault of their own. More blacks than whites live in such neighborhoods, which could even make this question a potential proxy for racial profiling.

The ProPublica journalists Angwin et al. (2016a) investigate which defendants were

reincarcerated within two years. Their studies find that the COMPAS score was able to predict the true defendant's recidivism 61% of the time. Furthermore, black defendants were classified wrongly as future criminals twice as often as white defendants, and white defendants were classified as low-risk more often than black defendants:

	White	Black
Classified as high risk, but did not reoffend	23.5%	44.9%
Classified as low risk, but did reoffend	47.7%	28.0%

Table 2.1.. False positives and false negatives by race. Angwin et al. (2016b)

The analysis of Angwin et al. (2016a) also shows that black defendants, even after controlling for age, gender, future recidivism, and prior offenses, are 45% more likely to receive a higher risk score than white defendants.

This example accentuates the need to assess fairness in machine learning. Although the used COMPAS algorithm is likely quite simple, the example shows that the underlying social dynamics of recidivism can nevertheless be very complex. Additionally, due to its public nondisclosure, the algorithm lacks the necessary interpretability to determine the cause of its unfairness, which is a hard requirement for readjusting the predictor without the loss of useful information.

2.2. Fairness in Word Embeddings

Word embeddings face similar problems as algorithms that are not publicly available since they are hard to interpret. One way to interpret how embeddings work is to perform arithmetic operations on them. For example, the vector representation for $\overrightarrow{king}$ could be approximately represented by $\overrightarrow{queen} + (\overrightarrow{man} - \overrightarrow{woman})$). Two abstract concepts are involved in this example: One concept is a shift in royalty, the switch from $\overrightarrow{man}$ to $\overrightarrow{king}$ and $\overrightarrow{woman}$ to $\overrightarrow{queen}$. The other is the concept of maleness, which is represented by the subtraction of $\overrightarrow{man}$ and $\overrightarrow{woman}$. Adding the maleness concept to $\overrightarrow{queen}$ changes its gender. Bolukbasi et al. (2016) examine which word pairs are approximately the same if the term $\overrightarrow{man} - \overrightarrow{woman}$ is added to one side. One result of their research is that $\overrightarrow{computerprogrammer}$ is approximately $\overrightarrow{homemaker} + \overrightarrow{man} - \overrightarrow{woman}$.

Similarity and dimension reduction are approaches that help to understand word embeddings. One way to examine gender bias is to analyse the similarity of words to one gender concept in relation to another one. Another way, proposed by Bolukbasi et al. (2016), is to construct a gender subspace with dimension reduction methods, using word vectors from word pairs that are defined by gender ($\overrightarrow{she} : \overrightarrow{he}; \overrightarrow{woman} : \overrightarrow{man}$). Accordingly, gender bias is defined by the projection of a word vector on this constructed gender subspace and gender bias could be mitigated by shifting the word vector on the gender subspace.

Using gender definition words to construct a gender subspace does not work that well for languages with grammatical genders, like French, Spanish, or German. Such languages with grammatical gender are called gendered languages. Word vectors in such languages naturally lead toward their grammatical gender. For example, *Kindergärtnerin* (female kindergarten teacher) is naturally more associated with the female gender concept, since it only refers to females. Nevertheless, the concept of teaching little kids is likely more associated with the female concept than with the male concept. Consequently, the common definition cannot be applied reasonably.

Note that the grammatical information in a word vector is necessary for word embeddings to work meaningfully. Trying to completely remove the gender meaning of a word could lead to grammatical mistakes.

It is also necessary to distinguish between words that only have a grammatical gender and words that have a grammatical gender and a respective justified semantic gender. The first type of noun is called inanimate noun. One example is *Leiter* (ladder, feminine). Vectors of such types of nouns should not have a close relation to the semantic gender concepts. The second type of noun is called animate noun. Vectors of animate nouns like *Lehrerin* or *Lehrer* (female/male teacher) should have a closeness to gender concepts since they explicitly represent teaching people.

Zhou et al. (2019) propose the idea to construct a grammatical gender subspace and use it to orthogonalize the gender subspace constructed with gender definition words. This procedure should create a subspace that represents the purely semantic gender, which means it does not represent that words need to appear in a specific context

through linguistic rules. Theoretically, shifting words on this subspace should not cause the grammatical rules to be violated. They try to reduce the bias in French and Spanish word embeddings, where adjectives and nouns are either masculine or feminine. Accordingly, they define their grammatical and semantic subspace in one dimension. In the German language, nouns, articles, and some pronouns are either masculine, feminine, or neuter. In this thesis, a grammatical and semantic subspace is identified based on the approach of Zhou et al. (2019). To debias the embedding, words and word pairs are shifted on the semantic subspace and projected back into the embedding.

3. Materials and Methods

3.1. Word Embeddings

Word embeddings are the main object of study of this thesis. Lendave (2021) consider them to be approaches to provide vector representations of words that are trained to obtain a word via context. Unlike word counts or frequency-based methods, words are represented by vectors with many fewer dimensions. The words are represented by a vector in the vector space where similar vectors should be semantically connected. Similar meanings are clustered within the vector space.

3.1.1. Word2Vec Skipgram model

The Skipgram model with negative sampling has been introduced by Mikolov et al. (2013). Their main idea is to predict a word $w \in 1, ..., W$ given its context. The objective of the Skipgram model is formalized by maximizing the likelihood

$$\sum_{t=1}^{T} \sum_{c \in C_t} log(p(w_c|w_t)), \tag{3.1}$$

where $w_1, ..., w_T$ are words of a large training corpus and C_t is the set of indices of words surrounding the word w_t. Predicting context words can be viewed as a set of independent binary classification tasks. From this perspective, the presence or absence of context words is predicted. For the word w_t, the context words are considered positive examples and random words are used as negative examples. Accordingly, the objective is defined as

$$\sum_{t=1}^{T} \left[\sum_{c \in C_t} \mathcal{L}(s(w_t, w_c)) + \sum_{n \in N_{t,c}} \mathcal{L}(-s(w_t, n)) \right], \tag{3.2}$$

where $N_{t,c}$ is the set of negative samples from the vocabulary and

$$\mathcal{L} : x \rightarrow log(1 + e^{-s(w_t, w_c)}) \tag{3.3}$$

is the logistic loss function. It follows that for each word w_t, the loss of score between the word and its context is maximized, while the loss of the word and a random word is simultaneously minimized.

An intuitive way to parameterize the scoring function is to use word vectors. Therefore, two vectors $u_w, v_w \in \mathbb{R}^d$ are defined for each word of the vocabulary. The score is computed as the scalar product $s(w_t, w_c) = u_{w_t}^T v_{w_c}$.

3.1.2. FastText Model

The Word2Vec Skipgram approach works suboptimally for rare words or words that are not in the corpus at all. Furthermore, it does not learn from the composition of the letters in a word, which contains valuable information as well. For example, *friend* and *friendship* are very similar. To solve these issues, Bojanowski et al. (2016) propose FastText models that are derived from Word2Vec Skipgram models with negative sampling. These FastText models represent each word as a bag of character n-grams. Character n-grams are defined as a sequence of characters of length n. The n-gram of a word produced by this operator consists of all the sequences of characters of that word of length n. Through the addition of $<$ at the beginning and $>$ at the ending of a word, models can distinguish between prefixes, suffixes, and other n-grams. They are also able to distinguish the sequence *her* in *where* from the sequence *<her>* representing the word *her*.

The n-grams are denoted in a dictionary of size G. Let $G_w \subset 1, ..., G$ be the set of n-grams that appears in a word w. Then a vector representation z_g is defined for each n-gram. A word is represented by the sum of the vector representations of its n-grams. Accordingly, the scoring function is

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c. \tag{3.4}$$

FastText embeddings solve the disadvantages of Skipgram that are stated above. They share the representation across words and allow the model to learn rare words reliably, too.

For experiments in section 4, English and German FastText embeddings [1] are used. They use vectors in dimension 300 that were obtained using the FastText model described by Bojanowski et al. (2016) with default parameters.

3.1.3. GloVe

In another approach, Pennington et al. (2014) use a matrix to capture the context of words. Let X be the matrix for co-occurrence counts of word pairs. Its entry X_{ij} counts the occurrence of word j in the context of the word i. Let $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ be the probability of word j appearing in the context of the word i. We start with a general model

$$F(w_i, w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}} \tag{3.5}$$

where $\tilde{w_k} \in \mathbb{R}^d$ are context words. With more assumptions, the function F gets specialized to

$$F((w_i - w_j)^T \tilde{w_k}) = \frac{F(w_i^T \tilde{w_k})}{F(w_i^T \tilde{w_k})} \tag{3.6}$$

with

$$F(w_i^T \tilde{w_k}) = P_{ik} = \frac{X_{ik}}{X_i}. \tag{3.7}$$

To solve equation 3.7, $F = exp$ is set:

$$w_i^T \tilde{w_k} = log(P_{ik}) = log(X_{ik}) - log(X_i). \tag{3.8}$$

$log(X_i)$ can be transformed to an intercept b_i, since the term is independent of k. To restore the symmetry, an additional bias b_k for w_k is added. The following simplified model emerges:

$$w_i^T \tilde{w_k} + b_i + \tilde{b_k} = log(X_{ik}) \tag{3.9}$$

For the experiments, a German GloVe embedding [2] derived from the Wikipedia corpus is used.

3.1.4. Bilingual Word Embeddings

Before 2018, modern techniques for learning cross-lingual word embeddings primarily depended on dictionaries in both languages or parallel corpora. Conneau et al. (2017)

[1] Downloaded from https://fasttext.cc/docs/en/pretrained-vectors.html
[2] Downloaded from https://www.deepset.ai/german-word-embeddings

proposed a way of aligning two languages in an unsupervised way, only relying on unaligned datasets of each language. The method provided by Conneau et al. (2017) and Lample et al. (2017) [3] is used to align the German FastText embedding to the English one. Similarly, it is used to align the same German embedding to the debiased (3.4.2) English embedding.

3.2. Quantification of Gender Bias

The approach to quantify gender bias is to compare the similarities of words to semantically masculine and semantically feminine sets of words. The similarity between the word vectors is measured by applying the cosine similarity *cos*.

$$cos(\overrightarrow{v}, \overrightarrow{w}) = cos(\theta) = \frac{\langle \overrightarrow{v}, \overrightarrow{w} \rangle}{||\overrightarrow{v}||\ ||\overrightarrow{w}||} \tag{3.10}$$

3.2.1. Word Embedding Association Test

To quantify the gender bias in word embeddings, the Word Embedding Association Test (WEAT) proposed by Caliskan et al. (2017) is used. For this test, two sets of attribute words A and B are collected. The sets are intended to represent social concepts. For individual words, the difference in the association to the concepts is of interest. Then, two sets of equally sized target words X and Y are used, too. For each word in the target sets, each similarity to each word in the attribute sets is computed. $s(\overrightarrow{w}, A, B)$ measures the difference of the similarity of the word vector $\overrightarrow{w}$ and attribute sets A and B:

$$s(\overrightarrow{w}, A, B) = \frac{1}{|A|} \sum_{\overrightarrow{a} \in A} cos(\overrightarrow{w}, \overrightarrow{a}) - \frac{1}{|B|} \sum_{\overrightarrow{b} \in B} cos(\overrightarrow{w}, \overrightarrow{b}) \tag{3.11}$$

Finally, the test statistic is defined by measuring the difference of the sums over $s(\overrightarrow{w}, A, B)$

$$s(X, Y, A, B) = \sum_{\overrightarrow{x} \in X} s(\overrightarrow{x}, A, B) - \sum_{\overrightarrow{y} \in Y} s(\overrightarrow{y}, A, B). \tag{3.12}$$

Accordingly, the null hypothesis is that no difference exists between the similarity of the target groups X and Y to either group of attribute words A and B. With this test

[3] https://github.com/facebookresearch/MUSE

statistic, the one-sided p-value of the permutation test is defined by

$$P_i(s(X_i, Y_i, A, B) > s(X, Y, A, B)), \tag{3.13}$$

where $\{(X_i, Y_i)\}_i$ denote all the partitions of $X \cup Y$ into two sets of equal size. Kauermann et al. (2021) explain that the permutation test is a method where the p-value is calculated by sampling without replacement using a Monte-Carlo procedure. Caliskan et al. (2017) state, that the permutation test measures the (un)likelihood of the null hypothesis.

For A and B, gender-definition word pairs for females and males are chosen, because the focus of attention is on gender bias. Such a pair would be *Mann:Frau* (man:woman). As target groups, words that belong to the field of families and careers are used. These words can be used in English as well as in German.

3.2.2. Word Embedding Association Test for Individual Words and Word Pairs

Zhou et al. (2019) extend this approach to measure gender bias for individual words and word pairs as well. For words like *Staat* (state, masculine), inanimate nouns, no association with the gender concepts A and B is justified. Accordingly, only the strength of the absolute association between the word $\overrightarrow{w}$ and concepts is measured:

$$b_w = |s(\overrightarrow{w}, A, B)| \tag{3.14}$$

The greater the relationship between $\overrightarrow{w}$ and gender notions is, the larger the b_w.

For animate noun pairs with two gender forms, like *Jäger* ([male] hunter) and *Jägerin* ([female] hunter), it is tested if the word forms $\overrightarrow{w_f}$ and $\overrightarrow{w_m}$ are symmetric with respect to the gender definition terms. For a fair pair, it is assumed that the difference in the association between the gender concepts is equal for the male and the female word form. This means, that the gender bias for a word pair is defined by

$$b_{wp} = ||s(\overrightarrow{w_m}, A, B)| - |s(\overrightarrow{w_f}, A, B)||. \tag{3.15}$$

Now, a test statistic should be defined that tests for a set of male occupation words M and a set of female occupation words F if the respective entries have symmetric associations with the word concepts. Granted that the Word Pair Association Test WPEAT is defined by

$$s(M, F, A, B) = |\sum_{\overrightarrow{w_m} \in M} |s(\overrightarrow{w_m}, A, B)| - \sum_{\overrightarrow{w_f} \in F} |s(\overrightarrow{w_f}, A, B)||. \tag{3.16}$$

The larger the value, the stronger the gender bias. For example, if $\overrightarrow{\text{Jäger}}$ had a stronger absolute difference in its association to the gender concepts than $\overrightarrow{\text{Jägerin}}$, the word pair would be considered as biased. Equation 3.16 differs from the approach of Zhou et al. (2019). This makes no difference in the calculation under the assumption that all male words are more associated with M and all female words are more associated with F. However, their approach leads to another interpretation of the p-values. This is due to the imbalance of the gender groups in the partition sets. The p-value of the WPEAT is defined as in equation 3.13, but now bootstrapping instead of permutation is used. For the WPEAT, opposite sets for occupation words are used. For example, is *Jäger* in M and *Jägerin* in F.

3.3. Constructing Gender Subspaces

One key point of this thesis is the utilization of dimensionality reduction. Typically, dimension reduction is used to reduce the number of features in the dataset to obtain a smaller dataset that still contains as much of the original information as possible, depending on the number of dimensions. In this thesis, we use dimensionality reduction on a specific, partly preprocessed subset of our data to reduce the dimension to subspaces of the embeddings, namely the grammatical gender subspace and the semantic gender subspace.

3.3.1. Grammatical Gender Subspace

To construct the grammatical gender subspace, Linear Discriminant Analysis (LDA), proposed by Fisher (1936), is used. Zhou et al. (2019) try to identify a grammatical subspace for French and Spanish, predict the gender for the nouns of two classes masculine and feminine, and reduce the embedding to one dimension. For the German embedding, the approach is to predict the three classes of masculine, feminine, and neuter for

nouns. Therefore, the dimensionality of the word embedding is reduced from 300 to a two-dimensional plane. For German, it is not sensible to reduce the embedding to a single dimension since neuter does not only lay between masculine and feminine but also contains information on e.g. belittlements, such as *das Kätzchen* (the kitten).

Through LDA, the dimension of vectors x_i gets reduced to y_i while preserving as much of the class-discriminatory information as possible. Gutierrez-Osuna (2005) emphasize that a measure of separation needs to be defined to find a good projection matrix W. For each class, we define the scatter S_i, an equivalent of the variance

$$S_i = \sum_{x \in \omega_i} (x - \tilde{\mu_i})(x - \tilde{\mu_i})^T. \tag{3.17}$$

The within-class scatter matrix is defined as the sum of the scatters $S_W = \sum_i^3 {S_i}^2$. On the contrary, the between-scatter matrix is defined by the sum over the squared differences of class means and the general mean: $S_B = \sum_{i=1}^3 N_i(\mu_i - \mu)(\mu_i - \mu)^T$.

A projection where elements of the same class are close to each other and elements of different classes are far away from each other is pursued. The main goal is to maximize the ratio of the between-class scatter to the within-class scatter $max(\frac{S_B}{S_W})$. Since this is a three-class problem, it should be sought for two projection vectors that should be arranged by columns into a projection matrix $W = [w_1|w_2]$ so that

$$y_i = w_i^T x \Rightarrow y = W^T x. \tag{3.18}$$

$\tilde{S}_W$ and $\tilde{S}_B$ are the corresponding scatter matrices for the projected samples. It can be shown that $\tilde{S}_W = W^T S_W W$ and $\tilde{S}_B = W^T S_B W$ hold, respectively. The projection matrix W^* is sought that maximizes the equation

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{3.19}$$

To avoid overfitting and to evaluate the model, 5-fold cross-validation is used. With this resampling method, the data is split into a training set containing 80% of the total data and a test set containing the rest for five times. Using these five variants of train-test split one model is fit on each respective training dataset. Afterward, the models are evaluated on the corresponding test data. As a test metric, accuracy is chosen, which is

the percentage of the truly predicted test data. Finally, the mean of the obtained five different accuracies gets evaluated. The mean measures how well the models work on average. This way, all available data is used without the risk of overfitting.

For the identification of the grammatical subspace, a subset of word vectors for frequently used nouns with their respective grammatical gender is used. A score for words based on their frequency has been established by the Institut für Deutsche Sprache: Programmbereich Korpuslinguistik (2007). A python module of Weichbrodt (2022) is used as well. It provides a list of 100,000 German nouns and their grammatical properties compiled from WiktionaryDE [4], to obtain the respective gender of the words. In addition, Weichbrodt added a module to look up the data and parse compound words. Consequently, the module is applied to get the grammatical gender of the most frequent words provided by the Institut für Deutsche Sprache: Programmbereich Korpuslinguistik (2007). Finally, to conduct the LDA for each gender, about 2600 of the most common words are used with their respective gender as the target class.

3.3.2. Semantic Gender Subspace

Principle Component Analysis

To get the semantic gender subspace of the respective embeddings, Principal Component Analysis (PCA) is used. PCA is an unsupervised method that finds the directions of the maximum variance in the data and projects it into a subspace of the data. The orthogonal axis spanning the subspace (**principal components**) can be interpreted as directions of the maximal variance.

Furthermore, spectral decomposition can be used to identify the principal components. Keep in mind that every covariance matrix is symmetric and thus diagonalizable. It follows that the covariance matrix can be decomposed into $A\Lambda A^T$, where A is the matrix of the orthonormal eigenvectors and Λ the diagonal matrix of the ordered eigenvalues. The ordered eigenvalues can be viewed as the ordered principal components. The first component d_1 explains the largest variance, while all other components explain the largest variance possible under the restriction that they are orthogonal to the components before.

[4] urlhttps://de.wiktionary.org/wiki/Deutsch

2×19 vectors of word pairs that are opposite and differ in gender meaning are selected as features (e.g. $\overrightarrow{he}$ and $\overrightarrow{she}$ or $\overrightarrow{man}$ and $\overrightarrow{woman}$). The difference in the word pairs can be interpreted as the difference in their gender. Therefore the word pairs can be reduced to their difference by centering them around the origin.

$$\begin{aligned} \overrightarrow{w_{m\_cen}} &= \overrightarrow{w_m} - \frac{\overrightarrow{w_m} + \overrightarrow{w_f}}{2} \\ \overrightarrow{w_{f\_cen}} &= \overrightarrow{w_f} - \frac{\overrightarrow{w_m} + \overrightarrow{w_f}}{2} \end{aligned} \tag{3.20}$$

The shifted word pair is now symmetric with respect to the origin. Afterward, PCA is conducted on these shifted word pairs to reduce the information of the vector space.

As the vectors are already centered around the origin now, no further data preprocessing is required. Features are usually scaled before applying PCA in order to have every feature provide the same contribution to the total variance. In this case, scaling is not sensible since less importance should be ascribed to features, where the pairs are not very different. Such features probably contain information about other concepts of the words, but only the semantic gender concept will be elaborated. It can be safely assumed that these features do not include much information about the semantic gender.

To conduct the PCA, a transformation matrix W is constructed to reduce the dimensionality of the 300-dimensional embedding to d dimensions:

$$x_i^{1,300} W^{300,d} = d_i^{d,1}. \tag{3.21}$$

The reduced feature space is called d_{pca}. The dimension d of d_{pca} is chosen by analyzing how well the components separate the gender of the features.

Adjustments towards the Semantic Gender Subspace

Note, that it can be assumed that d_{pca} and d_g are similar, since nouns that have a gender definition, e.g. *Mann* (man, masculine), almost always have the same grammatical gender. It is a hindrance to the semantic gender subspace that it also explains parts of the grammatical meaning of word vectors. This is because if words are shifted in

one direction of the semantic gender subspace, the grammatical meaning of the words should not get changed. Accordingly, we obtain the pure semantic gender subspace d_s by transforming each axis d_{pca_i} of the semantic gender subspace to be orthogonal to the grammatical gender subspace:

$$\begin{aligned} w_{g_1} &= v_{g_1} - \frac{\langle v_{g_1}, v_{g_2} \rangle}{\langle v_{g_2}, v_{g_2} \rangle} \cdot v_{g_2} \\ d_{s_i} &= d_{pca_i} - \frac{\langle v_{g_2}, d_{pca_i} \rangle}{\langle v_{g_2}, v_{g_2} \rangle} \cdot w_{g_1} - \frac{\langle w_{g_1}, d_{pca_i} \rangle}{\langle w_{g_1}, w_{g_1} \rangle} \cdot w_{g_1} \end{aligned} \tag{3.22}$$

In equation 3.22 v_{g_1} and v_{g_2} are the columns of the matrix from 3.18. Note, that the columns of the matrix are not orthogonal to each other therefore v_{g_1} needs to be transformed to w_{g_1} first so that $\langle v_{g_2}, w_{g_1} \rangle = 0$. The equations in 3.22 are inspired by the Gram–Schmidt process.

3.4. Mitigation Objectives

For the mitigation of bias in word embeddings, several methods proposed by Zhou et al. (2019) are used. All the methods aim to reduce the bias in the embeddings by shifting the word vectors of specific nouns along the semantic gender direction $\overrightarrow{d_s}$. The approaches differ depending on the setting. In the monolingual setting, only the German embedding, which is gendered, can be used. However, in the bilingual setting, the embedding of the German language gets aligned with the English embedding, which does not have a grammatical gender. This is helpful because the aligned English embeddings can be used to mitigate the bias of the word embedding. Respectively for both settings, objectives are distinguished by the types of nouns. For inanimate nouns, it is assumed that they should not carry any semantic gender information.

To motivate this assumption, suppose that we are using an NLP algorithm to filter job applications for a job related to cars. Men or women could be discriminated against in such job applications if $\overrightarrow{car}$ is closer to the male gender concept. Therefore, $\overrightarrow{car}$ should not carry any semantic gender information. For this reason, the inner product

$$\langle \overrightarrow{w}, \overrightarrow{d}_s \rangle \tag{3.23}$$

between the inanimate word vector $\overrightarrow{w}$ and the semantic gender direction $\overrightarrow{d_s}$ should be minimized.

For nouns with both gender forms, a more complex approach is needed. As previously concluded, their gender form should lay close to the respective gender subset since they represent a specific gender. Therefore, the aim is not to remove the total gender meaning of the word pair, but the inequality of its meaning. Specifically, the difference between the distance of the word vector $\overrightarrow{w_f}$ of the feminine form to an anchor point $\overrightarrow{w_a}$ and the distance of the word vector $\overrightarrow{w_m}$ of the masculine form to the same anchor point is observed. Such an anchor point could be the origin point of the gender subspace. To sum up, the difference in the distance between male and female word vectors to an anchor point gets minimized with respect to an anchor point:

$$D = ||\langle \overrightarrow{w_m}, \overrightarrow{d_s} \rangle - \langle \overrightarrow{w_a}, \overrightarrow{d_s} \rangle| - |\langle \overrightarrow{w_f}, \overrightarrow{d_s} \rangle - \langle \overrightarrow{w_a}, \overrightarrow{d_s} \rangle|| \tag{3.24}$$

It is sensible to assume both forms lay on opposite sites of the anchor point since the anchor point should be designed to represent some sense of a justified center. Consequently, 3.24 can be transformed to

$$D = |\langle \overrightarrow{w_m}, \overrightarrow{d_s} \rangle + \langle \overrightarrow{w_f}, \overrightarrow{d_s} \rangle - 2 \cdot \langle \overrightarrow{w_a}, \overrightarrow{d_s} \rangle| \tag{3.25}$$

This equation adds some intuition. It measures the need to move the word pair on the semantic gender direction subspace so that the word pair is symmetric with respect to the anchor point.

3.4.1. Monolingual Setting

Shifting Along Semantic Gender Direction ShiftOri

For the inanimate nouns, equation 3.23 gets minimized. This means that the semantic gender meaning is removed from such words. For nouns with two gender forms in the monolingual settings, the origin of the gender subspace d_s is used as the anchor point w_a in equation 3.24. With this choice, $\overrightarrow{w_a} = 0$ and therefore $\langle \overrightarrow{w_a}, \overrightarrow{d_s} \rangle = 0$, too. This reduces equation 3.25 to

$$|\langle \overrightarrow{w_m}, \overrightarrow{d_s} \rangle + \langle \overrightarrow{w_f}, \overrightarrow{d_s} \rangle|. \tag{3.26}$$

Consequently, the *male* inner product should be the negative of the *female* inner product, which means that after projection w_m and w_f are symmetrical with respect to the origin.

3.4.2. Bilingual Setting

As Zhao et al. (2020) show, gender bias can be seen when gendered languages are aligned to languages without grammatical gender. Therefore, the English embedding can be used to facilitate both the bias in German embeddings and the cross-lingual bias. Based on this, four approaches are defined. For some of these approaches, a debiased English embedding is used. The approach to debias the English embedding follows.

Hard debiasing the English embedding

In contrast to Bolukbasi et al. (2016), the gender subspace d_{en} for the English embedding is identified by performing PCA on the vectors of centered word pairs as in 3.3.2. In contrast to the German language, the subspace d_{en} is already sufficient for the English language since there is no grammatical gender. Like Bolukbasi et al. (2016), only the first component is chosen.

After the identification of the gender subspace, the hard debiasing approach by Bolukbasi et al. (2016) can be used. For that, the projection of a vector w onto d_{en} is defined by

$$\overrightarrow{w_{d_{en}}} = \sum_{j=1}^{k} (\overrightarrow{w} \cdot d_{en_j}) \cdot d_{en_j}. \tag{3.27}$$

Now each gender-neutral word $\overrightarrow{w}$ of a subset of the embedding gets neutralized and re-embedded.

$$\overrightarrow{w_{new}} = \frac{\overrightarrow{w} - \overrightarrow{w_{d_{en}}}}{||\overrightarrow{w} - \overrightarrow{w_{d_{en}}}||}. \tag{3.28}$$

With this approach, we ensure that gender-neutral words are zero in the gender subspace.

Each chosen word pair E, which primarily differs in its gender, gets equalized and re-embedded. First, the center of the set E, $\mu = \sum_{w \in E} w/|E|$ is computed and ν is defined as the difference of μ and the projection of μ_B. After that, a new representation for the word pair is defined for each $\overrightarrow{w} \in E$:

$$\overrightarrow{w_{new}} = \nu - \sqrt{1 - |\nu|^2} \frac{\overrightarrow{w_B} - \mu_B}{||\overrightarrow{w_B} - \mu_B||} \tag{3.29}$$

Mitigating Before Alignment DeAlign

In the DeAlign approach, the German embedding simply gets aligned to the debiased English one. It should be noted that most words with both gender forms align with the same word in English. For example, $\overrightarrow{\text{Jäger}}$ and $\overrightarrow{\text{Jägerin}}$ in German both align to $\overrightarrow{hunter}$ in English. For this reason, after debiasing, the alignment places the two gender forms of the words in a more symmetric position in the vector space.

Shifting Along Semantic Gender Direction ShiftEN

For ShiftEN, the debiased English embedding is not utilized. Rather, the English translation of the German occupation word pairs is used as anchor point w_a in Equation 3.24. In this manner, the German word pair is shifted, until it is symmetric to the corresponding English word, which lies in the same space.

Hybrid Methods

HybridOri and HybridEN are hybrid methods designed to combine aspects of the previously mentioned methods. For both methods, the German embedding is once more aligned to the debiased English embedding. Additionally, the occupation words also get shifted along the semantic gender axis. For these approaches, a new semantic direction is fit, as presented in Section 3.3.2. Furthermore, the word vectors of the German embedding are used, which is aligned to the debiased English embedding. The two hybrid methods differ in what is chosen as the anchor point. For the HybridOri method, the zero point of the gender subspace is chosen as the origin. On the contrary, for the HybridEN method, the English word from the debiased embedding is used as the anchor position.

3.5. Performance Evaluation

To analyze the remaining bias of the embeddings, WPEAT (3.2) is used with occupation words. It is also tested if inanimate family and career words are still biased using the not-modified WEAT and the respective p-values are calculated.

In another sense, it is tested if the debiased embeddings lose their functionality. To repeat, an embedding works if words with similar semantics lay near each other. For

this reason, the task proposed by Camacho-Collados et al. (2017) is applied to investigate the word embeddings (still) catches multilingual semantics. They provide a data set that gives a human-labeled score of similarity for German and English words.

This similarity is now compared with the cosine similarity (3.10) of the same words in the respective embedding. Then, the Pearson correlation coefficient is used to measure the linear correlation of these two scores.

4. Results

In this chapter, the analysis described in section 3 was carried out. The results of the proposed methods on the German GloVe embedding and the FastText embeddings as described in section 3.1 are discussed. First, the bias in the Word embeddings is examined using the WPEAT. Then the gender subspaces for both embeddings are examined in more detail, visualized, and interpreted. Finally, the performance of the mitigation procedure is evaluated.

4.1. Word Embedding Association

For specially selected words, the difference of the association to the respective attribute words is computed according to equation 3.11. For five occupation word pairs, the visualization of the results for FastText can be observed in figure 4.1. It can be observed in figure 4.1 that feminine words are more strongly associated with the female gender definition set than with the male. For male words, the converse holds true. However, it is noticeable that masculine words have a smaller disparity in difference to the gender concept than feminine ones. More precisely, feminine words are a lot more associated with the feminine concept with respect to the masculine concept than masculine words are associated with the masculine concept with respect to the feminine context. The absolute difference in the association to the attribute sets for the male words ranges from 0.04 and 0.12, while it ranges from 0.39 to 0.50 for the female words. It is also noticeable that the pairs of *dancer* and *worker* are further together than the pairs of *doctor*, *educator*, and *hunter*.

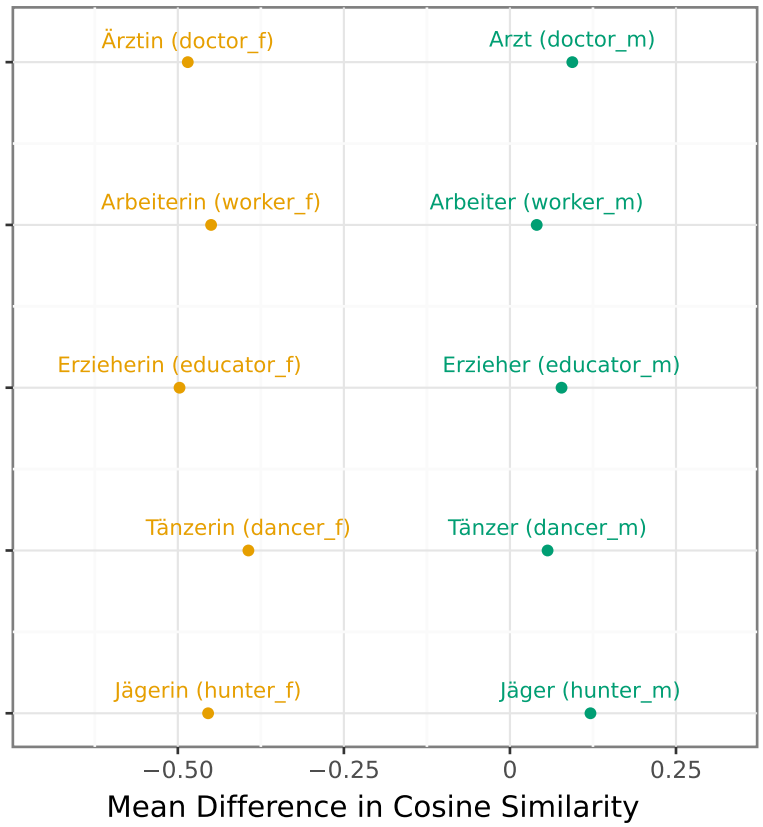

Figure 4.1.. For each word, the difference in the association to the attribute sets of the gender definition words for the FastText Embedding is visualized. For illustration purposes, the respective word pairs are placed next to each other.

hunter	dancer	educator	worker	doctor
0.33	0.34	0.42	0.41	0.39

Table 4.1.. The table shows the bias for selected German word pairs for the FastText embedding.

Table 4.1 shows the bias of word pairs defined in equation 3.15. According to the definition, it shows, that the German word pairs for *hunter*, and *dancer* are less biased than the pairs for *educator*, *worker*, and *doctor* because they are more symmetric with respect to zero.

Now, the same plot is examined for the GloVe embedding.

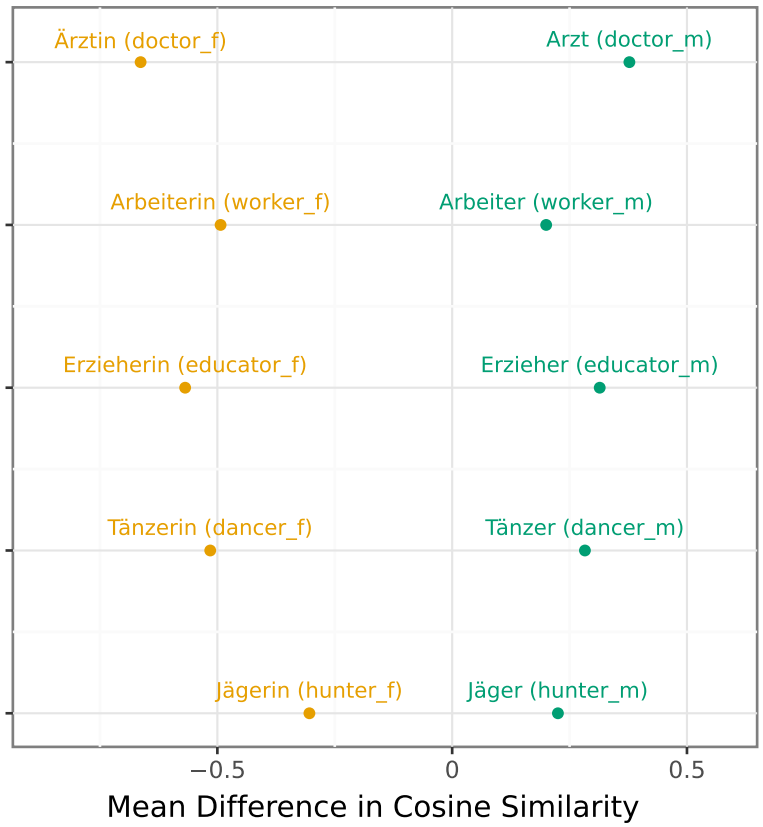

Figure 4.2.. For each word, the difference in the association to the attribute sets of the gender definition words for the GloVe Embedding is visualized. For illustration purposes, the respective word pairs are placed next to each other.

Like in the FastText embedding, it is observable that male words are closer to the origin than female words. This means that male words have a smaller disparity in the difference of the gender concepts than female words. However, this effect is less pronounced for GloVe. The absolute difference in the association to the attribute sets for the male words ranged from 0.20 and 0.38, while it ranged from 0.30 to 0.67 for the female words. It is noticeable that the range within the respective gender groups is larger compared to FastText, especially for the female words. In general, the words appear more symmetrical.

hunter	dancer	educator	worker	doctor
0.08	0.23	0.25	0.29	0.29

Table 4.2.. The table shows the bias for selected German word pairs for the GloVe embedding.

Table 4.2 shows that the bias for the GloVe embedding is far lower than the bias for the FastText embedding. Foremost the word pair to hunter has a particularly low bias of 0.08.

4.2. Grammatical Gender Subspace

As stated in 3.3.1, a grammatical gender subspace is identified with LDA.

4.2.1. FastText

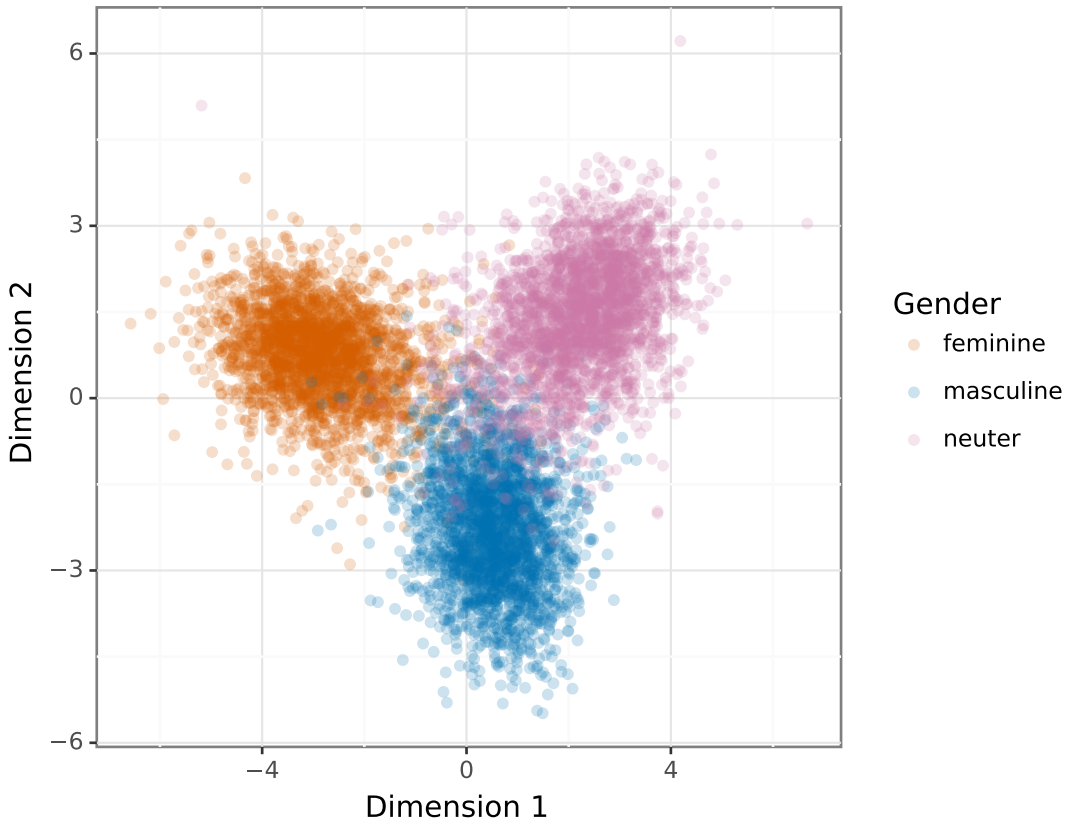

Figure 4.3.. The grammatical subspace d_g for the FastText embedding. The subspace is estimated with linear discriminant analysis. The points represent the transformed training data, which were used to fit the model.

The mean train accuracy of this model is *0.95* and the mean test accuracy is *0.94*. The grammatical gender is thus better predicted in German than in Spanish (*0.92*) and in

French (*0.83*)[1]. This is surprising because three classes instead of two are predicted and therefore one would expect it to be lower.
First of all, it is noticeable that the neuter does not lie clearly between masculine and feminine. This proves the assumption that one dimension would not be sufficient for the grammatical subspace. We also see that the three classes are clearly separated, even if there are some outliers.

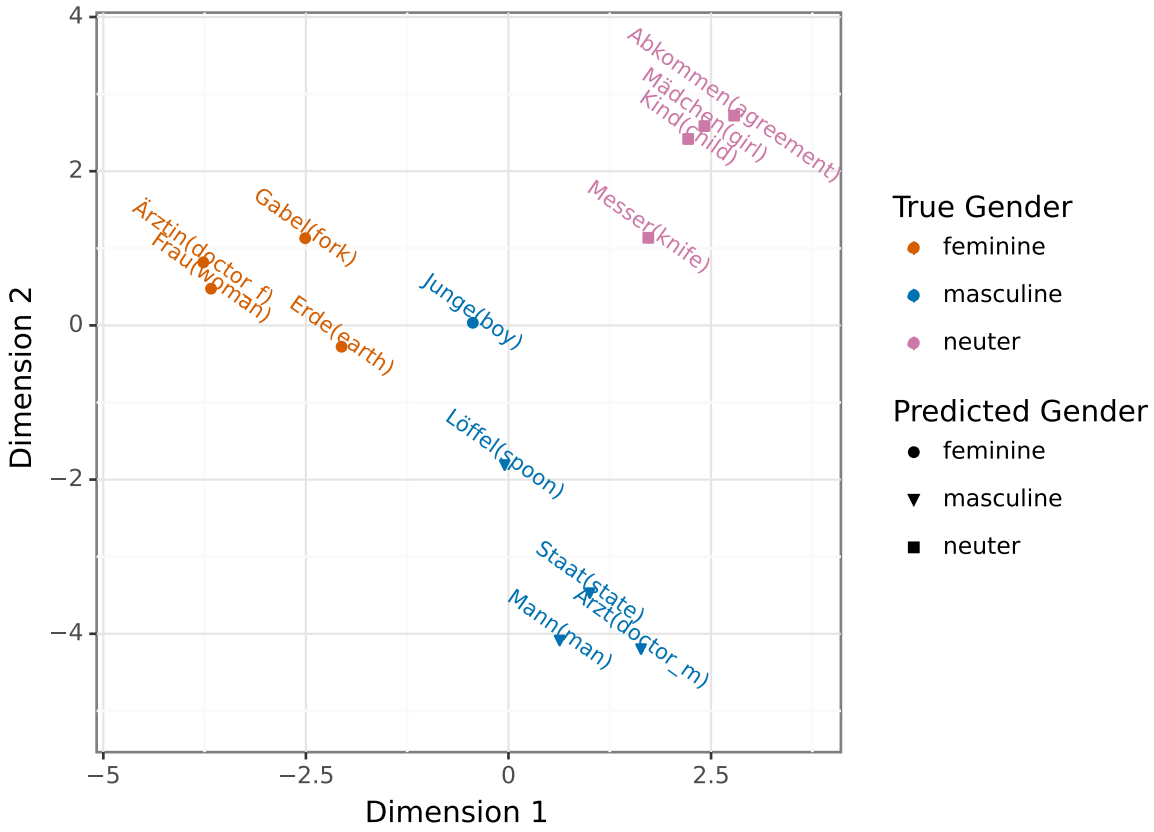

Figure 4.4.. Specific words projected in the grammatical subspace d_g for the FastText embedding.

In order to examine the subspace more closely, the projection of individual words is observed in figure 4.4. In addition, the predicted gender and the true gender are provided. An interesting result is that the gender for *Junge* (boy, masculine) was incorrectly predicted as feminine. This is especially surprising because *Junge* is not only grammatically, but also semantically masculine. On the other hand, the semantic and grammatical

[1]Zhou et al. (2019)

words *Ärztin* ([female] doctor, feminine), *Arzt* ([male] doctor, masculine), *Frau* (woman, feminine), and *Man* (man, masculine) are, as expected, rather in the extreme of their grammatical direction, especially in comparison with the inanimate nouns. Surprisingly, *Mädchen* (girl, neuter) was correctly predicted, although for *Mädchen* the grammatical gender is different from the semantic gender. Words that are normally next to each other in word embeddings, such as *Gabel* (fork), *Messer* (knife), and *Löffel* (spoon) or *Mann* (man) and *Frau* (woman) are now apart from each other because they have different grammatical gender. In summary, the subspace appears to be a good representation of the grammatical gender.

4.2.2. GloVe

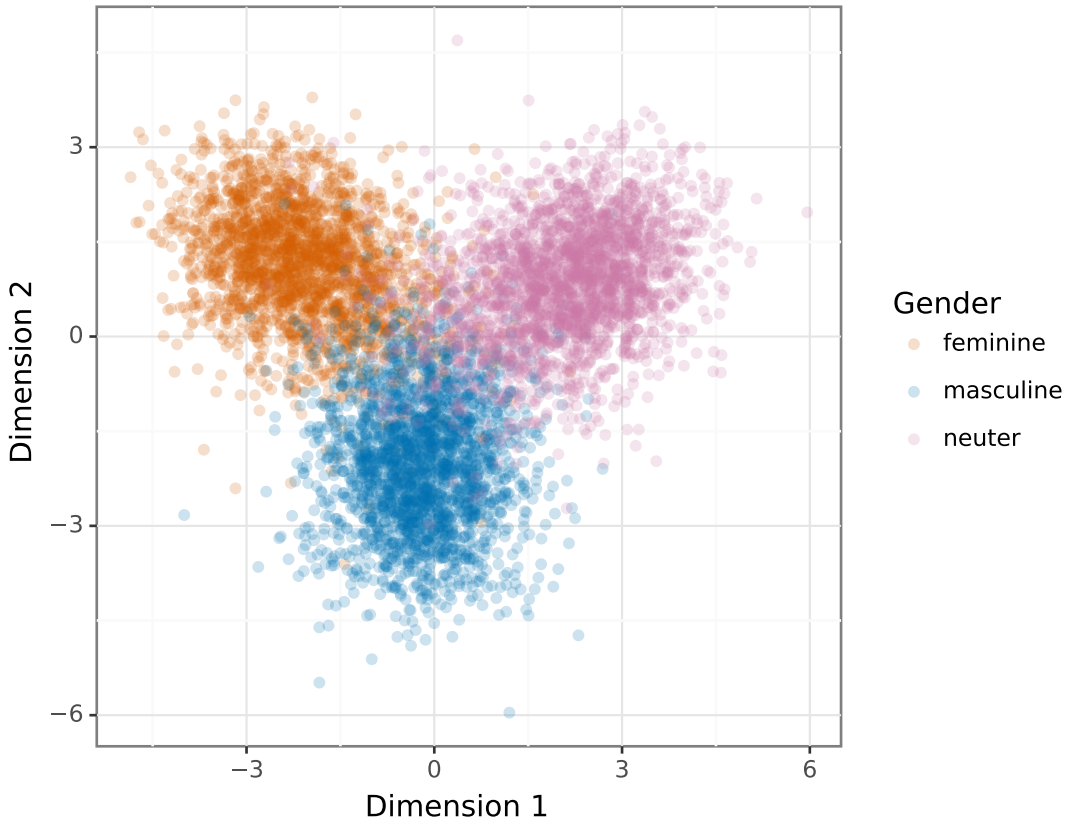

Figure 4.5.. The grammatical subspace d_g for the GloVe Embedding. The subspace is estimated with linear discriminant analysis. The points represent the transformed training data, which were used to fit the model.

The mean train accuracy of this model is 0.93 and the mean test accuracy is 0.90. This suggests that for the German language, the grammatical subspace can be identified

better using FastText than with GloVe. In contrast, to figure 4.3, a stronger scatter with fewer data near the center is observed in figure 4.5. The reason for this might be that the n-grams give meaning to the suffixes in the FastText embedding. The grammatical gender can often be identified with the suffix. For example, nouns ending with 'chen' are always neutral, like *Mädchen* (girl, feminine).

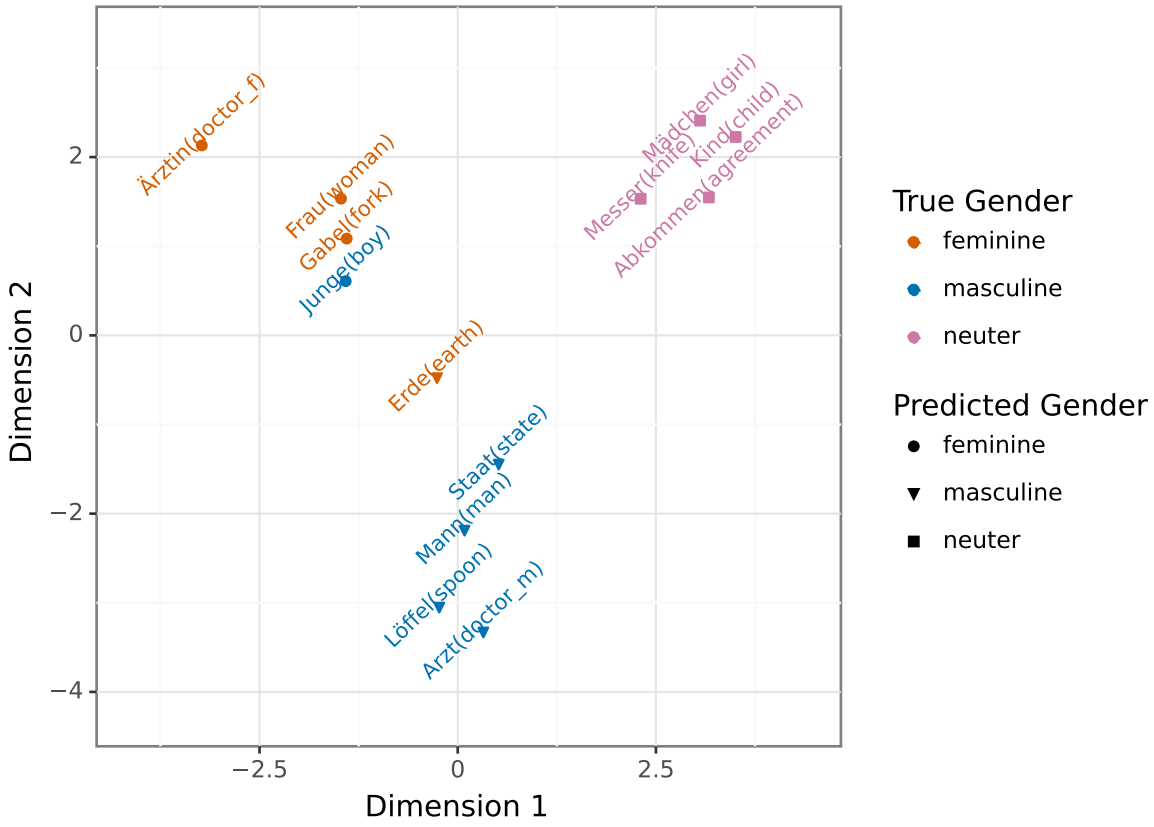

Figure 4.6.. Specific words projected in the grammatical subspace for the FastText embedding.

Figure 4.6 shows the projection on the grammatical subspace for some selected words. For these, the neuter is clearly distinguished from the other grammatical genders. In this figure, two falsely classified words can be observed. *Junge* (boy, masculine) is again classified as feminine, yet *Erde* (earth, feminine) is classified as masculine. It is recognizable that *Arzt* ([male] doctor, masculine) and *Ärztin* ([female] doctor, feminine) are leaning towards the extreme of their gender, like in FastText. However, in contrast to FastText, this is not recognizable for *Mann* (man, masculine) and *Frau* (woman, feminine). These words are rather close to inanimate nouns. In conclusion, the grammatical subspace of the GloVe embedding is a good representation of the grammatical gender,

too. However, the separation of gender concepts is less pronounced and the accuracy is lower than with the grammatical subspace for the German FastText embedding.

4.3. Semantic Gender Subspace

4.3.1. FastText

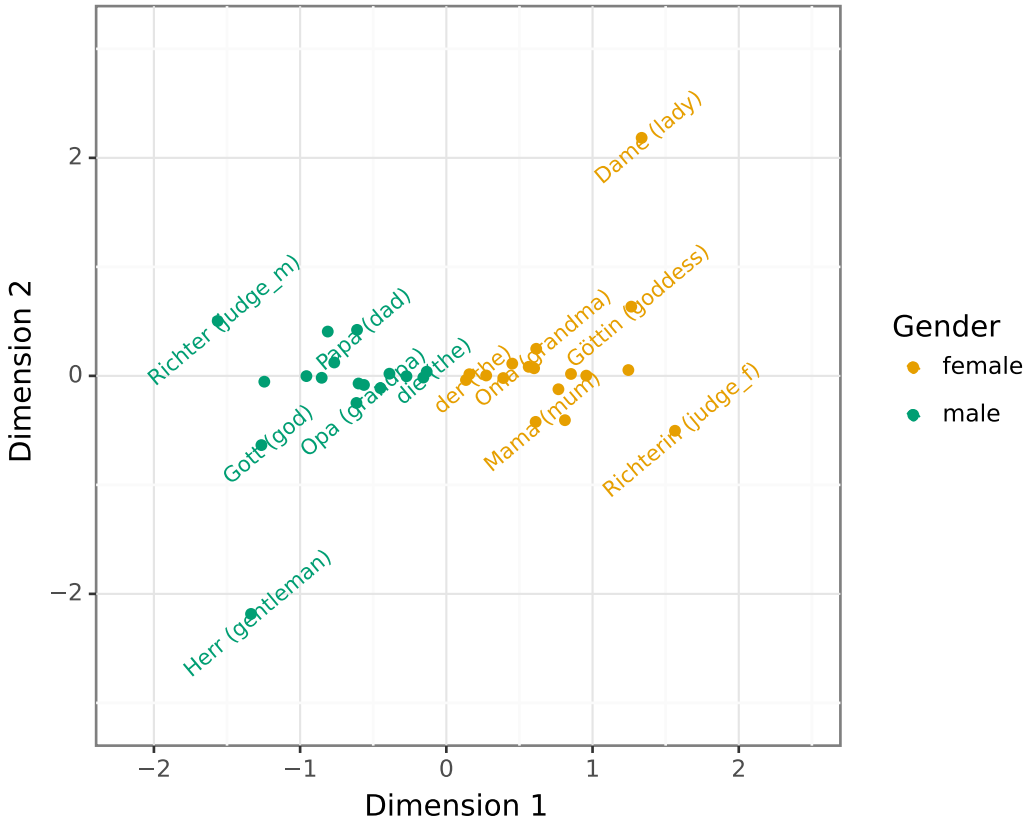

Figure 4.7.. d_{pca} for the FastText embedding. The figure shows the projection of the 19 **centered** word pairs that constructed the d_{pca}. Each dot represents a word but for clarity, not every point is labeled.

The projection of the words used in the process of fitting d_{pca} is shown in figure 4.7. It is noticeable that the word pairs are symmetrical to the origin. The reason for this is that they were centered in pre-processing and the dimensionality reduction is linear, so the symmetry is maintained. The figure shows that the first component separates the pre-processed gender pairs perfectly. All male words are projected in the negative part of the first dimension and all female words are projected in the positive part.

The second component does not seem to explain the semantic gender. For this component, most of the word pairs lay near the origin of the component. Rather, this component seems dominated by *Dame* (lady) in the positive part and *Herr* (gentleman) in the negative part.

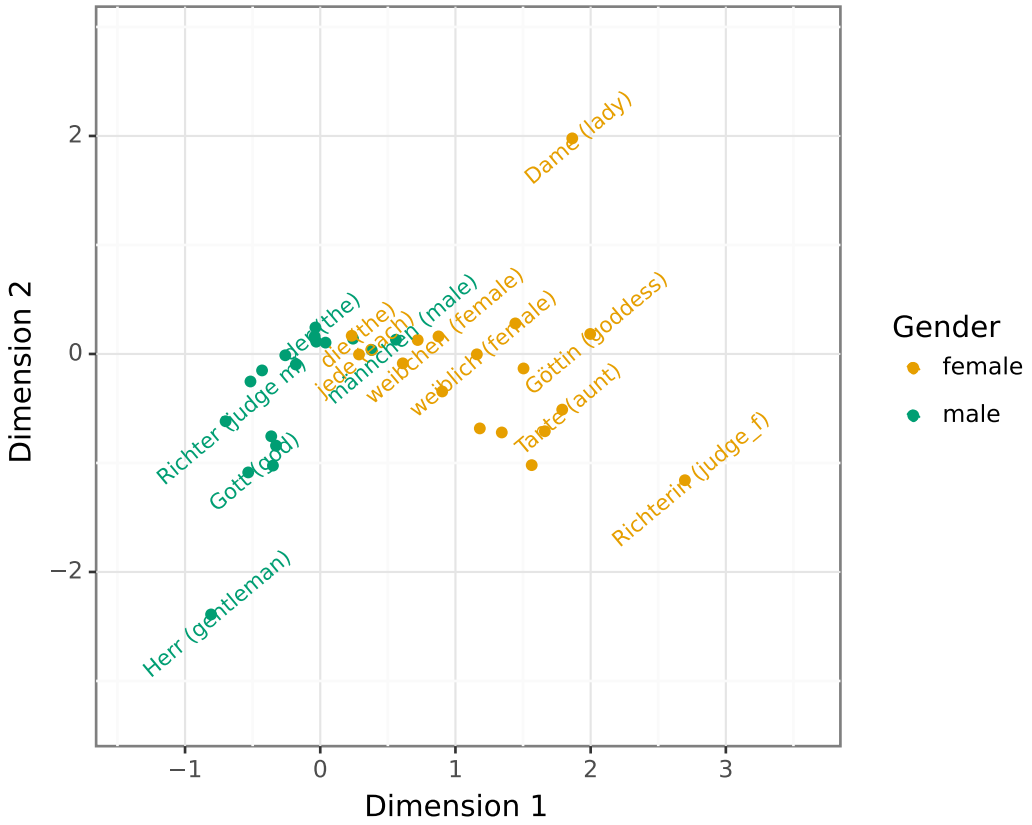

Figure 4.8.. d_{pca} for the FastText Embedding. The figure shows the projection of the 19 **not-centered** word pairs that constructed the d_{pca}. Each dot represents a word but for clarity, not every point is labeled.

In figure 4.8, it stands out that the word pairs are no longer symmetrical because they are no longer centered. It is noticeable that the separation, even in the first dimension, does not work as well as for the centered pairs. For the first component, a separation of the two genders can be observed at 0.6, but *die* (the, for feminine nouns) and *jede* (each, for feminine words) are in the male cluster.

For *die*, this can be explained intuitively, because *die* can also be used in the plural

for all genders in the nominative, the frequentist grammatical case[2] in German. In this manner, *"Der Mann schläft"* (the man sleeps) switches to *"Die Männer schlafen"* (the men sleep). In contrast, the word *jede* can only represent feminine words. *Männchen* (male) and *Weibchen* (female) are close to each other in the grammatical subspace. These words are primarily used in reference to animals. This could be an explanation for why they are not well distinguished from the subspace.

Additionally, it can be observed that the male words are much closer to the origin in contrast to the female words. Especially the words *Richterin* (judge, female) and *Göttin* (goddess) lie farther than two units away from the origin in the female direction. In contrast, no word is farther than one unit away from the origin in the male direction.

Even if the non-centered vectors are considered, it can be observed that the second component does not appear to explain the semantic gender. Therefore, only the first component is chosen to construct d_s for the FastText embedding. As described in section 3.3.2, if the semantic subspace contains information about the grammatical gender, this is a problem. The cosine similarity of the semantic gender direction and the first axis of the grammatical gender subspace is 0.28. For the second axis of the grammatical gender subspace, it is -0.28. Therefore, in some sense, the axes are connected. Figure 4.9 shows the semantic gender directions orthogonalized with the axes of the grammatical gender subspace. The first dimension is denoted as d_s and used later for mitigation of the gender bias.

[2]In German, there are four grammatical cases that modify the word or suffix, depending on where the word occurs. The cases influence nouns and noun modifiers (articles, some pronouns, adjectives, and participles). For English, only pronouns are influenced by the grammatical case. The change from *i* to *me* and from *he* to *him* shows this phenomenon: 'I teach him'; 'He teaches me'

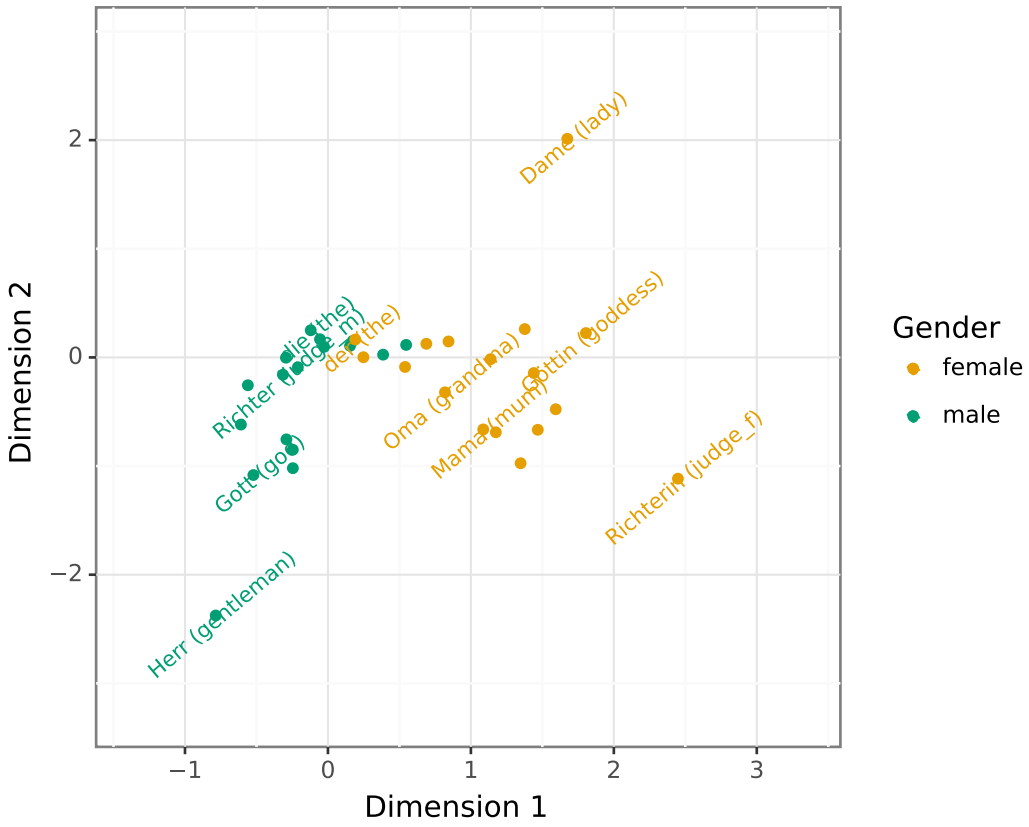

Figure 4.9.. d_s for the FastText Embedding. The figure shows the projection of the 19 word pairs that constructed the d_{pca}. Each dot represents a word but for clarity, not every point is labelled. Even if only the first dimension is used, the second is included for illustration purposes.

Figure 4.9 shows the d_s, where the projected word vectors, which were used to construct d_{pca} are shown. First of all, it is noticeable that figure 4.9 and figure 4.8 are not very different. The overall illustration remains relatively similar. Nevertheless, it can be observed that *Richter* ([male] judge) and *Richterin* ([female] judge) move towards the origin. This can be interpreted in a way that this pair of words is now less distinguished by their grammatical gender. It is also worth noting that *Richter* now has hardly any semantic meaning to it, while *Richterin* still has a strong feminine representation. In summary, words that do not necessarily refer to humans, such as *jede*/*jeder*, *die*/*der* and *Weibchen*/*Männchen* are poorly distinguished by the semantic subspace. However, especially nouns describing people in their roles or family relationships are well distinguished.

4.3.2. GloVe

For the most part, the same visual analysis can also be carried out for GloVe.

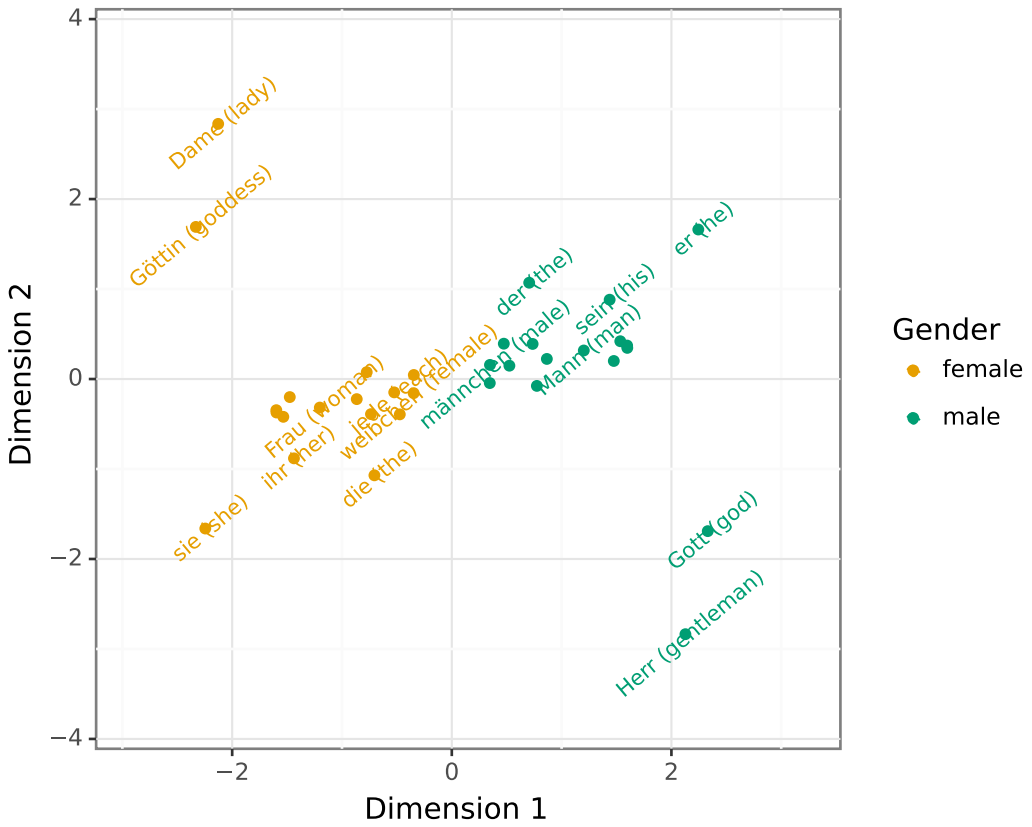

Figure 4.10.. d_{pca} for the GloVe Embedding. This figure shows the projection of the **centered** word pairs, that constructed the *pca*. Each dot represents a word but for clarity, not every word is written.

Figure 4.10 shows, that d_{pca} separates the 19 word pairs perfect in the first dimension. In this axis, all male words are negative and all female words are positive. Even in the second dimension, the word pairs are separated quite well. Especially the pronouns and articles which were not well separated in the FastText embedding seem to be better represented in the second component of the GloVe embedding. However, *Herr* (gentleman), *Gott* ([male] god), *Dame* (lady), and *Göttin* (goddess) are located in the opposite direction with respect to the other word pairs. Nevertheless, two dimensions for the semantic subspace are used for the GloVe embedding. It is also noticeable that *Richterin* ([female] judge) does not has an outgoing value for the GloVe embedding, although it was the outlier in the FastText Embedding.

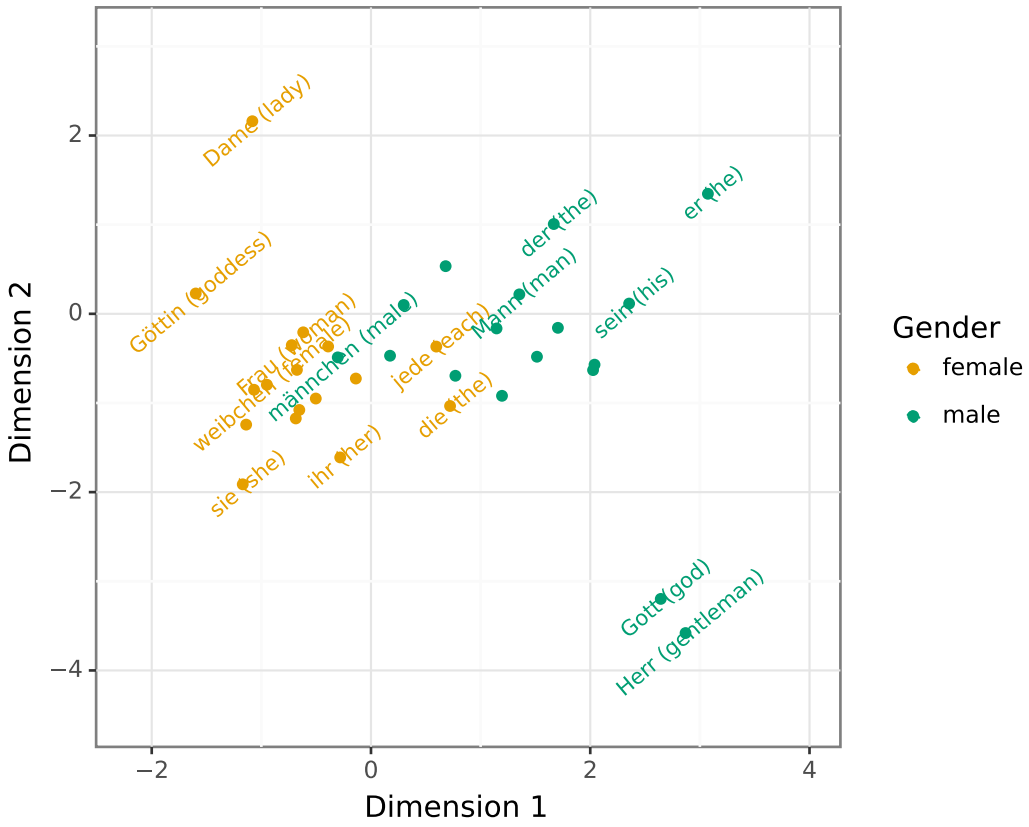

Figure 4.11.. d_s for the GloVe Embedding. Here, the projection of the **not-centered** word pairs that constructed the d_{pca}, on d_s is shown. Each dot represents a word but for clarity, not every point is labeled.

Figure 4.11 shows that the genders are not separated that well for the non-centered vectors projected on d_s. On the one hand, *Männchen* (male) lays in the cluster where most of the female words are located. On the other hand, the female words *jede* (each) and *die* (the) are far from their male equivalents but still lie between other male words. It is also noticeable that *Herr* (gentleman) and *Gott* ([male] god) are closer to each other than *Dame* (lady) and *Göttin* (goddess). An explanation for this could be potentially that in the German translation of the bible, god is often referred to as *der Herr*.

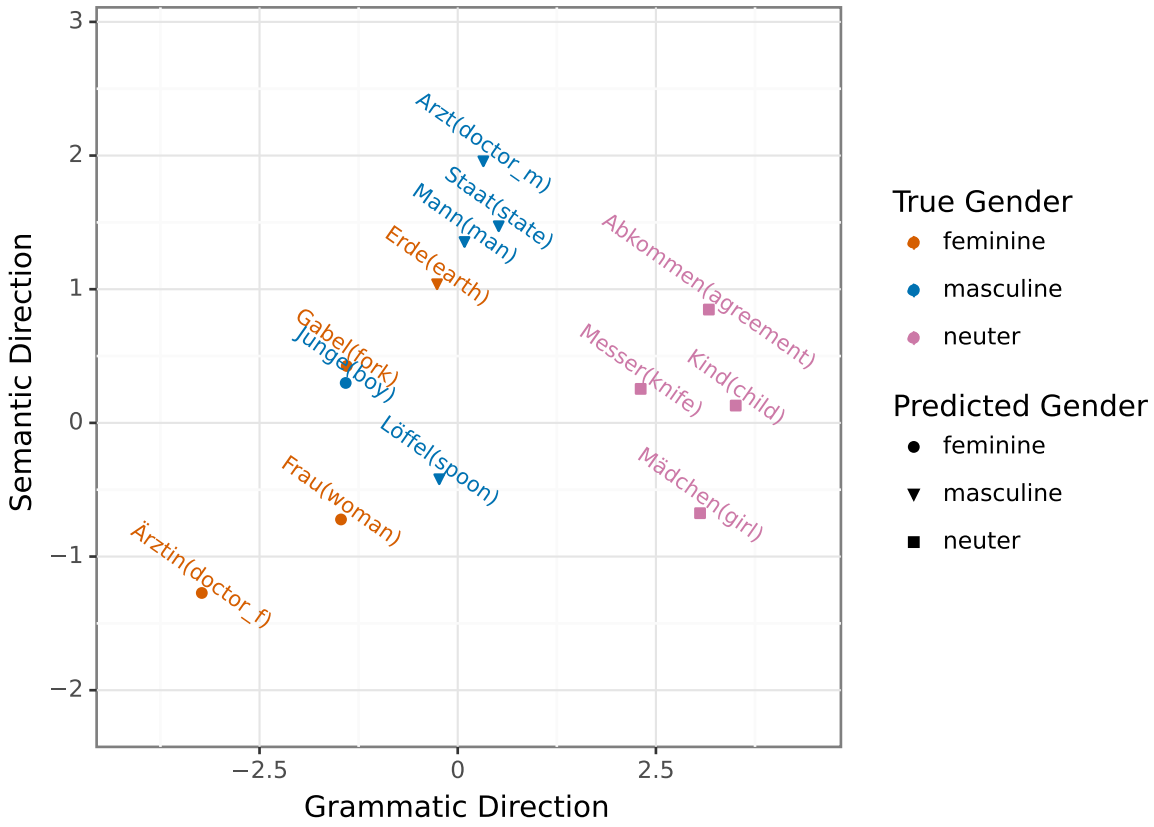

Figure 4.12.. Specific words that are projected on the first semantic and the first grammatical axis of the respective subspaces.

Figure 4.12 shows the relation of the first semantic and the first grammatical axis for some words. There is no correlation between the results because they were orthogonalized. It can be observed that *Junge* (boy) is not only incorrectly predicted for the grammatical gender, the semantic gender subspace also does not project him to the cluster of the male words. Although *Mädchen* (girl, neuter) is not grammatically feminine, it is clustered with the other semantically feminine words on the semantic direction. It is noticeable that the inanimate words *Abkommen* (agreement) and *Staat* (state), which are connected to reign, are clustered as semantically male. This could be explained by the historical fact that men have ruled most of German history. These words can be considered examples of biased inanimate words.

4.4. Debiasing the Word Embeddings

For debiasing, the GloVe and the FastText embedding for the German language are used in the monolingual setting. The used method is provided in section 3.4.1. For debiasing in the bilingual setting as in section 3.4.2, only the German FastText embedding aligned to the English FastText embedding is used. For the debiasing experiments approximately 50 occupation words are debiased and eight words per family and career words group. As described in section 3.4.2, the English Fasttext embedding is debiased first for some mitigation methods. Afterward, the German embedding is aligned with the debiased embedding to archive a bilingual subspace. For the debiasing of the English embedding, the same words are debiased as for the german embedding. In addition, the gender definition words that are used to construct the English gender subspace are also debiased. This is done because the words are also partially used to construct the German semantic gender subspace.

Embedding	WPEAT	p-val WPEAT	WEAT	p-val WEAT	WordSim
GloVe	1.47	0.0000	0.40	0.0189	NaN
ShiftOri_gv	1.37	0.0000	0.34	0.0009	NaN
FastText	3.64	0.0000	0.433	0.0624	0.7127
ShiftOri_ft	1.90	0.0000	0.197	0.3441	0.7126
DeAlign	4.06	0.0000	0.43	0.0624	0.7089
ShiftEN	1.26	0.0000	0.20	0.3441	0.7127
HybridEN	0.38	0.0511	0.18	0.3389	0.7091
HybridOri	0.28	0.1456	0.18	0.3389	0.7091

Table 4.3.. The table shows the test statistics for the Word Pair Embedding Association Test on sets of occupations words for gender pairs. It also shows the Word Embedding Association Test for sets of career and family words. In addition, the values of the word similarity task are shown. Note that the word similarity task is not done for the GloVe embedding, because only the German embeddings were observed for it.

The performance of the debiasing methods is evaluated with the methods proposed in section 3.5. Table 4.3 shows that ShiftOri for the GloVe embedding does not decrease the bias for the WPEAT and the WEAT test. For the FastText embedding, it can be observed that the DeAlign approach does not increase the p-value for the WEAT and WPEAT tests, but does decrease the word similarity score. Therefore the approach does not manage to reduce gender bias, but the functionality of the embedding got worse.

ShiftOri for the FastText embedding reduced the bias for inanimate nouns but fails to reduce the bias for animate nouns. The ShiftEN embedding is the only embedding that does not decrease the word similarity score of the original FastText embedding. It also performs well at debiasing the family and career words but fails to debias animate nouns. The hybrid methods are the only ones that reduce the bias for animate nouns to the insignificant level of 0.05. But both embeddings have a lower word similarity score than the original FastText embedding. It can be noted, however, that the two hybrid methods that build on the DeAlign embedding improve the word similarity score by debiasing with respect to the DeAlign embedding. In summary, when the performance values are compared, it can be seen that HybridOri leads to the best performance.

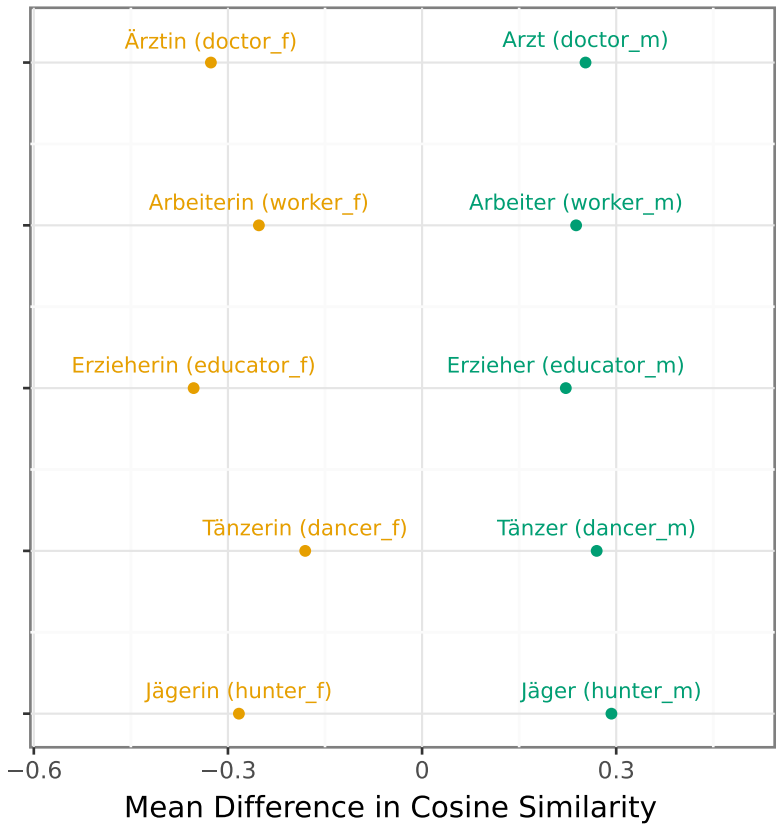

Figure 4.13.. For each word, the difference in the association to the attribute sets of the gender definition words for the HybridOri Embedding is visualized. For illustration purposes, the respective word pairs are placed next to each other.

In contrast to figure 4.1, figure 4.13 shows the difference in the association to the respective attribute words. The figure shows that the word pairs have become much more symmetrical to the origin. In particular, the german word pairs corresponding to *hunter*

and *worker* appear to be nearly perfectly symmetrical. It is also apparent that the absolute mean difference in cosine similarity to the gender terms is not that different for all the words considered.

5. Discussion and Limitations

This chapter discusses the results and methods in the context of other work. It points out the limitations of the definition and the methods and, based on this, makes suggestions for future research.

5.1. General Classification of the Results

As seen in the previous chapter, debiasing was performed only for some example words. With the manual classification of all nouns into animate word pairs and inanimate nouns, it could be investigated more precisely whether the performance of the embedding deteriorates.
Zhou et al. (2019) state that the proposed methods are capable of reducing cross-lingual gender bias. They propose a word analogy test to observe if the gender of persons is predicted right if the grammatical context is given. They also test the cross-lingual performance of the embedding in general. Future work could investigate the cross-lingual bias for German and English word embeddings.

It should be kept in mind that nouns are the main object of the investigation for the gender bias analysis of this thesis. The proposed methods are only used for debiasing nouns, which could leave gender bias in the embedding for other word types. For example, words like *programming* might be more strongly associated with male concept words than with female concept words. However, the approaches for debiasing inanimate nouns could be reused or tested for verbs, adjectives, and other word types in future work.

It is an oversimplification for inanimate nouns to assume that they have no similarity to gender concepts. For example, the association of beard and menstruation to gender concepts can be considered justified, because they are correlated to the physical characteristics of males or females. Future debiasing approaches could manually filter

out such words and not change their representation in terms of gender definition words.

To make things easier, the proposed methods only address the gender bias for males and females, not for non-binary gender identities. In the German language, non-binary people usually want to be addressed by their name or gender-neutral terms like *Person* (person) instead of man or woman or they prefer *Kind* (child) instead of daughter or son. Therefore, gender bias, as defined here, cannot be meaningfully examined for non-binary gender identities in German, since most words are used much more frequently in contexts that are not specifically referring to non-binary people. Such gender-neutral terms in the context of non-binary people and gender bias could be explored in the future.

By using LDA, the same covariance matrix for each grammatical gender class is assumed. The result can be observed in figure 4.3 and figure 4.5. All projected words within the classes have a similar structure for the respective embedding. This assumption might be violated. Unlike the masculine and feminine, the neuter does not appear in contexts where the grammatical gender suggests the semantic gender. Therefore, the linguistic structure of the neuter could be very different from that of the masculine and feminine. For future work, quadratic discriminant analysis might be used to identify the grammatical gender subspace, where different covariance matrices among the classes are estimated. Velilla (2008) showed that quadratic discriminant analysis can also be used for dimensionality reduction.

5.2. Classification of the Results in Relation to the Generic Masculine

In the German language, the generic masculine exists. In this manner, the masculine is chosen to designate groups in which there is at least one male. The grammatical masculine is also chosen if the gender is not specified explicitly. The Journalistinnenbund (2020) states that this is a linguistic convention that "increasingly no longer works".

In recent years, the effect of this linguistic property has been widely discussed socially and there are proposals for linguistic changes such as the so-called *Gendersternchen* (gender star), which is supposed to replace the generic masculine. Some use this ty-

pographic design when writing in gender-neutral language in German. It is created by adding an asterisk and the feminine plural suffix "-innen" after the stem. For instance, *Programierer* ([male] programmer or programmers) becomes *Programierer*innen* (programmers). The gender star allows for the inclusion of non-binary individuals as well as all genders.

These proposed changes are perceived as unnecessary by some parts of society. The opinion of this groups is that the generic masculine is a grammatical form that includes all genders in its semantic meaning. Other groups reference studies, for example by Stahlberg et al. (2001), which shows that people are more likely to think of men than women when the generic masculine is used.

Surprisingly, the mitigation methods for animate nouns are constructed by Zhou et al. (2019) in such a manner that they cause similar effects as the use of fair language, with respect to gender bias. The methods transform the word pairs to be symmetrical to a proper center, like the origin of the semantic gender subspace. It should be noted, however, that the generic masculine leads to more frequent use of masculine word forms and therefore more frequent male appearances in text corpora, which makes gender debiasing in German an unbalanced problem. This could mean that NLP methods work worse for female animate nouns than for male animate nouns. Future research could address this issue in more detail.

5.3. Criticism of the approach to identify gender bias

Gonen and Goldberg (2019) show that the definition of gender bias in word embeddings proposed by Bolukbasi et al. (2016) might not be sufficient. Indeed, they claim that the removal of the association with gender concepts is only superficial. Furthermore, they argue that (semantic) gender subspaces only measure gender bias rather than determining it. According to the used definitions of bias (3.2), the bias is actually reduced, but this effect may hide the bias rather than eliminate it. They explore that algorithmic discrimination is more likely to occur by the association of one implicitly gendered term with other implicitly gendered terms.

In addition, gender bias might occur when the models learn to condition on gender-biased words and generalize to other gender-biased phrases. According to the authors, the spacing between the "gender-neutralized" words in the mitigated embeddings still reflects the information about gender bias, and it is possible to extract gender bias from them. They conclude that the methods currently in use to remove bias are inadequate and should not be relied upon to create gender-neutral models. For the future, they advise that other characteristics of the bias also need to be taken into account.

The approach of Gonen and Goldberg (2019) is reminiscent of that of Kilbertus et al. (2017), which referred to machine learning in general. This study was written when using fairness criteria for machine learning models, which are designed to evaluate how fair a model is, was common. In response to this kind of fairness criteria, they outlined the reasons why and how frequently such criteria fail, reveal previously unnoticed nuances, and clarified why these nuances are essential to fairness in models. By analyzing problems through the lens of causal inference, they focus more on consciously including and questioning the data-generating process during modeling.

Questioning the data-generating process in NLP could mean questioning the cause and effect of language in general. This implies reasoning over specific contexts in society and how they are reflected in word embeddings. This point of view would respect that addressing fairness is always a balancing act that requires distinguishing between patterns that should be learned and other patterns that represent stereotypes that are strived to be avoided in prediction. In the author's opinion, the used definition of gender bias might not be sufficient for capturing the total gender bias. Nevertheless, despite all the criticism, it can be assumed that a noticeable part can be mitigated with these methods. The methods can be considered the first step of a long journey of research in the future.

6. Conclusion

The proposed subspaces allow interpretations for the entirety of the word embedding and for individual words. The visualization shows that the dimensionality reduction of word embeddings to subspaces allows interpretations of the otherwise difficult-to-interpret high-dimensional word embeddings. Based on this, gender bias could be examined and interpreted in more detail. The investigated methods use the subspaces to reduce gender bias according to a provided definition. The results show that the reduction of the bias for certain words works with the HybridEN and HybridOri methods. After debiasing, the WEAT for the debiased family and career words and the WPEAT for the debiased occupation word pairs differing in gender was computed. Both embeddings reject the null hypothesis for both of these tests at 0.05 insignificance level using a bootstrap test. Future work could focus on expanding the definition of gender bias in word embeddings and debiasing all words in the total embedding. Another relevant question for future research is how language models are affected by training on debiased word embeddings.

7. Acknowledgements

This thesis would not have been possible without Dr. Matthias Aßenmacher. His contributions, support, and time have been essential to my research and are greatly appreciated. Furthermore, I wish to express my gratitude to Sabine, Norbert, Max, Cosima, Michi, Philip, Julius, Eugen, Nina, Harald, Sophia, Lea, Jan, Salo, and Maya who have provided me with all the support and energy I required for my scientific workings. A special "thank you" goes to Chris and Jan who have provided remarkable constructive feedback.

List of Figures

List of Tables

A. Appendix

A.1. Code for the Production of the Results

The pyhton code of the implementation mentioned in this thesis is provided in the corresponding GitLab [1] repository. The `data/` folder contains .txt and .json files, which provide words that were used for various tasks. The embeddings are not part of the repository this directory, due to their large size. Rather, the file GET_DATA explains how to download and align the word embeddings. The file get_nouns.py shows the code, which creates files for the most common words for each gender. Note that this data is already saved in the `data` folder.

The `result/` folder contains plots which are included in this thesis. The thesis folder contains the bachelor thesis and the latex files to output it.

The code folder contains the python files which provide the methods and output. The `code/gender_subspaces` folder contains the method to construct the respective subspaces for the FastText, the GloVe, and the FastText embedding, which was aligned to the debiased English embedding. The `code/mitigation` folder contains the method for debiasing the English embedding and all methods for debiasing the German embeddings in separate files. The `code/results` folder contains three files: The file measures.py performs the evaluation methods on the respective embedding, the plots.py file produces the plots stored under `results/` and the file numbers_in_text.py stores the computation of the results that had been proposed in the thesis.

[1] https://gitlab.lrz.de/statistics/winter22/ba_prokosch

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016a). How we analyzed the compas recidivism algorithm. Available online at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, retrieved February 10, 2023.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016b). Machine bias. Available online at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing , retrieved February 10, 2023.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. Available online at http://www.fairmlbook.org, retrieved February 10, 2023.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (2017). SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

der deutschen Sprache, D. W. (2023). „mädchen“. Available online at https://www.dwds.de/wb/M%C3%A4dchen, retrieved February 10, 2023.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Gonen, H. and 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*.

Gutierrez-Osuna, R. (2005). Introduction to pattern recognition. Available online at https://people.engr.tamu.edu/rgutier/web_courses/cs790_w02/l6.pdf, retrieved February 10, 2023.

Institut für Deutsche Sprache: Programmbereich Korpuslinguistik (2007). Korpusbasierte wortgrundformenliste derewo, v-30000g 2007-12-31-0.1, mit benutzerdokumentation. Available online at https://www.ids-mannheim.de/digspra/kl/projekte/methoden/derewo/, retrieved February 10, 2023. ©.

Journalistinnenbund (2020). Generisches maskulinum. Available online at https://www.genderleicht.de/generisches-maskulinum/, retrieved February 10, 2023.

Kauermann, G., Küchenhoff, H., and Heumann, C. (2021). *Statistical Foundations, Reasoning and Inference: For Science and Data Science*. Springer Series in Statistics. Springer International Publishing.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lendave, V. (2021). Hands-on guide to word embeddings using glove. Available online at https://analyticsindiamag.com/hands-on-guide-to-word-embeddings-using-glove/, retrieved February 10, 2023.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Stahlberg, D., Sczesny, S., and Braun, F. (2001). Name your favorite musician: Effects of masculine generics and of their alternatives in german. *Journal of Language and Social Psychology*, 20(4):464–469.

Velilla, S. (2008). A method for dimension reduction in quadratic classification problems. *Journal of Computational and Graphical Statistics*, 17(3):572–589.

Weichbrodt, G. (2022). German nouns. Available online at https://github.com/gambolputty/german-nouns, retrieved February 10, 2023.

Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W., and Awadallah, A. H. (2020). Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., and Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. *arXiv preprint arXiv:1909.02224*.

Declaration of Authenticity

The work contained in this thesis is original and has not been previously submitted for examination which has led to the award of a degree.

To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made.

David Prokosch