Analysing and Mitigating Origin Bias in German Word Embeddings

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Bachelor Thesis

23rd January 2024

Aileen Kim Reichelt

reichelt@cl.uni-heidelberg.de

Institut für Computerlinguistik

Ruprecht-Karls-Universität Heidelberg

Supervisor	Prof. Dr. Katja Markert
Reviewers	Prof. Dr. Katja Markert
	Prof. Dr. Andreas Witt

Abstract

Word embeddings have been shown to contain biases against certain social groups such as women or Black people, which can lead to discriminatory outputs in downstream tasks. In recent years, research on the measurement and reduction of bias in word embeddings has been a growing field, but has mainly been focused on English embeddings and gender bias. This thesis examines origin bias in German word embeddings.

I differentiated origin bias between bias against people of Turkish, Polish, and Italian origin for a more fine-grained analysis. To quantify the bias contained in German embeddings, I adapted the Word Embedding Association Test (Caliskan et al., 2017), which measures the association of first names with pleasant versus unpleasant terms using a statistical test. For this test, I created a new data set of typically German, Turkish, Polish, and Italian first names based on various statistics.

To mitigate origin bias in German embeddings, I adapted two debiasing methods: *Hard Debiasing* (Bolukbasi et al., 2016) and *DD-GloVe* (An et al., 2022). Hard Debiasing is a post-processing approach which reduces bias in embeddings by identifying which components of an embedding are biased based on defining word pairs, and then removing these components. DD-GloVe modifies the loss functions of GloVe to encourage embeddings to become bias-reduced at train time based on bias-free dictionary definitions.

My results showed that German word embeddings do contain origin bias, but also that this is not the case for all embeddings and biases analysed. GloVe embeddings contained considerably higher biases than fastText embeddings, which were only biased against people of Turkish origin. Debiasing with DD-GloVe did not result in a decrease in bias, whereas Hard Debiasing was able to entirely remove statistically significant bias in some cases and slightly mitigate it in others.

I concluded that analysing bias in more fine-grained categories yields new insights into the precise nature of bias in word embeddings, and that adapting debiasing methods to different bias attributes or languages is challenging due to the seed words which need to be defined for these methods. I encourage future work to focus on methods for non-gender bias, to ensure reproducibility of results, to take the issue of seed word frequency into consideration and to carefully examine how bias can be defined.

Zusammenfassung

Die Forschung der vergangen Jahre hat gezeigt, dass Worteinbettungen Vorurteile gegenüber bestimmten sozialen Gruppen wie Frauen oder Schwarzen Menschen enthalten, was zu diskriminierendem Verhalten von Algorithmen führen kann. Es existieren mittlerweile bereits einige Arbeiten zur Analyse und Reduktion von Bias in Worteinbettungen, jedoch wurde sich bisher hauptsächlich auf englische Worteinbettungen und geschlechtsspezifischen Bias konzentriert. In der vorliegenden Arbeit untersuche ich nun Herkunftsbias in deutschen Word Embeddings.

Für eine differenziertere Analyse unterschied ich zwischen Bias gegenüber Menschen türkischer, polnischer und italienischer Herkunft. Um den Bias in deutschen Worteinbettungen zu quantifizieren, adaptierte ich den Word Embedding Association Test (Caliskan et al., 2017), welcher die Assoziation von Vornamen mit positiven versus negativen Begriffen mithilfe eines statistischen Tests misst. Für diesen Zweck erstellte ich einen neuen Datensatz mit typisch deutschen, türkischen, polnischen und italienischen Vornamen auf der Grundlage diverser Statistiken.

Um den Herkunftsbias in deutschen Worteinbettungen zu verringern, adaptierte ich zwei Methoden zur Biasreduktion: *Hard Debiasing* (Bolukbasi et al., 2016) sowie *DD-GloVe* (An et al., 2022). Hard Debiasing ist ein Postprocessing-Ansatz, welcher Bias in Embeddings reduziert, indem er mithilfe von Definitionswortpaaren ermittelt, welche Worteinbettungskomponenten Bias enthalten und diese Komponenten dann entfernt. DD-GloVe hingegen verändert die Verlustfunktionen von GloVe so, dass die Worteinbettungen während der Trainingszeit auf der Grundlage von Wörterbuchdefinitionen in ihrem Bias reduziert werden.

Meine Resultate zeigten, dass deutsche Worteinbettungen tatsächlich Herkunftsbias enthalten, allerdings nicht für alle betrachteten Worteinbettungen und Nationalitäten. *GloVe*-Worteinbettungen enthielten einen wesentlich höheren Bias als *fastText*-Worteinbettungen, bei denen ein Bias nur für türkische Namen messbar war. Die DD-GloVe-Methode führte nicht zu einer Reduktion des Bias, wohingegen die Hard-Debiasing-Methode statistisch signifikanten Bias in manchen Fällen vollständig beseitigen und in anderen Fällen leicht abschwächen konnte.

Ich kam zu dem Schluss, dass eine feingliedrigere Bias-Analyse neue Einblicke in die genauen Eigenarten von Bias in Worteinbettungen liefern kann und dass die Anpassung von Debiasing-Methoden an andere Bias-Attribute oder andere Sprachen aufgrund der für diese Methoden zu

definierenden sog. Seed Words eine Herausforderung darstellt. Für zukünftige Arbeiten empfehle ich, sich nicht nur auf Methoden für geschlechtsspezifischen Bias zu beschränken und sich auf die Reproduzierbarkeit der Ergebnisse zu konzentrieren. Des Weiteren empfehle ich, Probleme im Zusammenhang mit der Häufigkeit von sog. Seed Words verstärkt zu berücksichtigen sowie sorgfältig zu prüfen, wie Bias definiert werden sollte.

Acknowledgements

I would like to express my gratitude to my siblings Melanie Reichelt and Thorsten Frank for proofreading this thesis; to Jakob Moser for his insightful feedback; and especially to Valentin Höpfl for his comprehensive review. Thank you as well to Evgeni Ulanov, Jana Engelmann, Alina Brand, and Laura Hepp for your support.

Contents

1 Introduction

How Bias Affects Us In a 2003 sociological field study on racial discrimination in the labour market, Bertrand and Mullainathan (2004) sent close to 5,000 fictitious resumes to over 1,300 different job listings. For each resume sent, a virtually identical one was sent to the same job, except that this duplicate resume differed in the applicant's race and minor additional details. Resumes with White-sounding names received 50 % more callbacks than their Black-sounding counterparts. Whether or not the employers claimed to be "equal employers" did not make a significant difference. The basic idea of this study is shown in Figure 1.

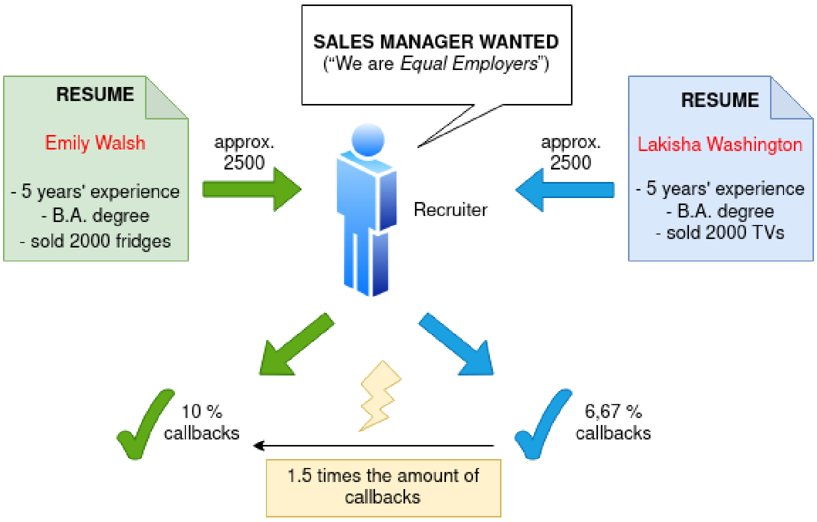

Figure 1: Illustration summarizing field study by Bertrand and Mullainathan (2004) on discrimination on the labour market. 5000 virtually identical resumes except in race were sent to the same job listings, but callback rates were significantly worse for Black-sounding names.

This study constitutes a striking example of prejudices still present in modern-day society and how these prejudices can be unconscious to the people propagating them. Humans, intentionally or not, can be harmfully biased towards members of a group with certain protected attributes (such as race, gender or religion) over other groups.

Since humans are biased in such ways, it is perhaps unsurprising that technologies created by humans exhibit the same biases. In 2016, Reuters garnered public attention by exposing Amazon's planned recruitment algorithm to be biased against women: The algorithm was rejecting resumes

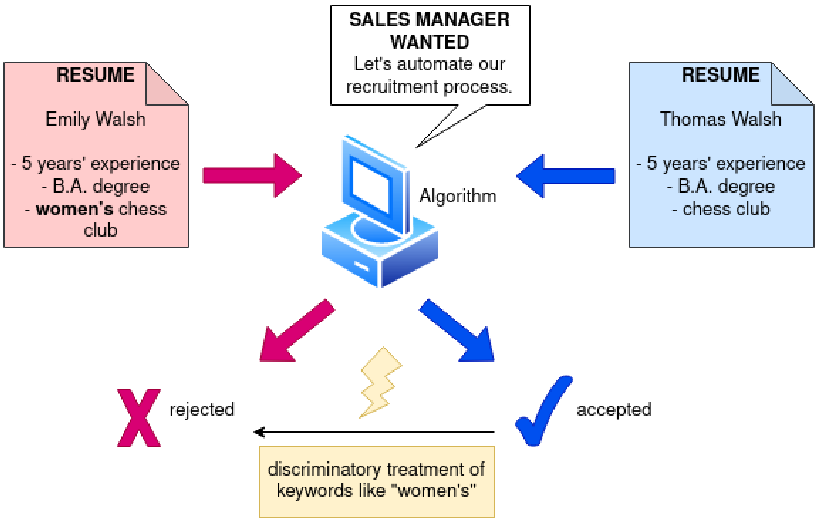

Figure 2: Illustration of Amazon's former recruitment algorithm as explained by Reuters (Dastin, 2018). Treatment of women's (pink) vs. men's (blue) applications.

which contained language hinting towards the applicant being female (Dastin, 2018). Figure 2 shows the problematic behaviour of this algorithm.

This study illustrates how biased algorithms can have unwelcome real-world consequences, but is just one of many troubling discoveries made in recent years. Other cases more relevant to the subject matter of this thesis include how tweets in African American English are more likely to be labelled as offensive (Sap et al., 2019), how named entity recognition systems are able to identify male names as persons but label female names as objects or locations (Mehrabi et al., 2020), or how automatic speech recognition systems make more mistakes for African American speakers.

Motivation It is safe to assume that these algorithms were not created with the purpose of outputting discriminatory results. Rather, the reason for this behaviour is usually that they were trained using historic and imperfect data which encodes these biases. In the Amazon example, their algorithm used existing company data on hiring decisions, which has historically favoured men. Wagner et al. (2015), among others, have shown that even Wikipedia, a common data source which one might believe to be relatively neutral, contains hidden biases.

Standard computational techniques in machine learning not only reproduce stereotypes present in their training data. This would be problematic enough, since it is misaligned with the goal of equitable judgement defined below. However, they also amplify biases. This amplification can occur in various ways. Language models can make harmful patterns found in their training data more obvious in their outputs since they are typically specifically trained to recognise and emphasise data patterns (Ethayarajh et al., 2019). Where a human might have overlooked biased data, an algorithm might find a strong data pattern and base its decisions on this pattern.

Additionally, algorithms are perceived to be neutral and objective decision-making methods and as such their output might not be questioned as much as a human's judgement, which harbours the danger of humans not recognizing bias where it is present in an algorithm's output (Ajunwa, 2019). If such algorithms are then employed in real-world scenarios, they can systematically discriminate against certain groups by assigning negative associations to those groups, under-representing them in terms of output frequency, or producing less reliable results for users of that group.

Briefly speaking, algorithms have the potential to reinforce harmful stereotypes. It has been examined that this disparate impact of algorithms is not in accordance with American discrimination law (Barocas and Selbst, 2016), and it is reasonable to assume that this observation also holds for other national and international law. It is therefore desirable to create algorithms which produce less biased results. In order to be able to study this subject, I now first define what exactly is meant by "bias" within the context of this thesis.

Definitions

Definition 1: Brookings Institution (2023)

An algorithm is biased if it is not predicting its target accurately and equitably.

Definition 2: Friedman and Nissenbaum (1996)

"[C]omputer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" are called biased.

In everyday language, bias is commonly interpreted as a personal tendency or perspective of judgement that is sometimes unfounded (Merriam-Webster, 2023). In the field of Machine Learning, bias is an overloaded term. It can mean statistical bias; it can also refer to the bias term in neural networks. The bias this thesis is about, however, is algorithmic bias "of moral import" (Friedman and Nissenbaum, 1996). In particular, I understand an algorithm, method, system or model to be biased if it behaves as specified by the working definitions presented above.

Definitions of bias in word embedding research are often reductive or even remain unstated. I further expand on this in §2.3. In this thesis, I stipulate a broad bias definition since social bias is a complex issue. There exist some additional terms which are relevant for discussing bias in this context, and I now briefly introduce these terms.

Humans group people according to characteristics like race, ethnicity, gender, social class, sexual orientation, nationality, religion, or disability (Lai et al., 2013). Such characteristics which "may not be used as the basis for [algorithmic] decisions" (Fletcher et al., 2020), as agreed upon by law or societal standards, are often called *protected attributes*.

Stereotypes are people's "beliefs about the characteristics, attributes, and behaviors of members of certain groups. [...] [T]hey are also theories about how and why certain attributes go together." (Hilton and von Hippel, 1996).

Prejudices can be defined as "feeling[s], favorable or unfavorable, toward a person or thing, prior to, or not based on, actual experience" (Allport et al., 1954). These prejudices can be *implicit*, meaning that the prejudiced person is not consciously aware of their biased view (Lai et al., 2013).

Scope and Contribution While biased algorithms are a broad research area, this thesis is specifically concerned with natural language models, a specific type of algorithms. Natural language technologies have enjoyed great popularity and research successes in the last decades, with applications such as chat bots, machine translation, voice assistants and more enjoying wide-spread and mainstream use(Jin et al., 2021). Addressing bias in Natural Language Processing (NLP) therefore has the potential for making substantial real-world improvements.

A core part of almost all current language computer systems are word embeddings (Almeida and Xexéo, 2019), i.e., vector space representations of words or other language units. Word embeddings have many useful properties, such as the ability to compute word analogies. For example, Mikolov et al. (2013) famously show that "man" is to "woman" as "king" is to "queen".

However, in the same vein, researchers have also found that embeddings contain problematic content: the same kind of word analogy task as above also computes that "man" is to "woman" as "computer programmer" is to "homemaker" (Bolukbasi et al., 2016). This is an indication that word embeddings might contain bias, i.e. that they do not fulfill their purpose of representing words equitably (see definition 1) and instead treat different groups in a discriminatory way (see definition 2).

In recent years, much research has been conducted on the topic of bias in word embeddings. So far, this research has been predominantly focused on English language embeddings and gender bias. English data is readily available and gender bias seemingly less complex than other types of bias.[1] However, bias differs across different cultures and languages (Kurpicz-Briki, 2020) and the bias present in embeddings as well as bias mitigation methods might not necessarily be directly transferable from English to other languages. In order to study this issue further, this thesis examines German word embeddings.

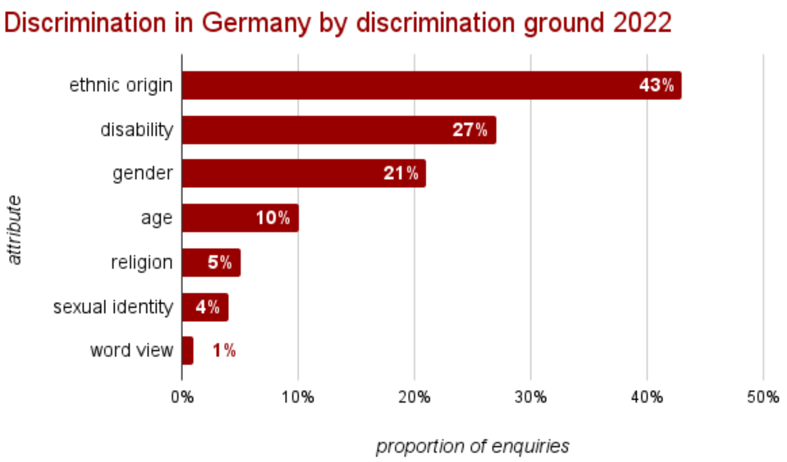

Figure 3: Distribution of 6627 counselling requests to the Federal Anti-Discrimination Agency. Translation of original diagram.

Similarly, there are also more protected attributes than just gender. A recent statistic (Figure 3) by the Statista Research Department shows that by far the most common attribute for which people in Germany experience discrimination is ethnic origin. This is the attribute I intend to focus on in this thesis.

It should be noted that in my research I refer to *origin* instead of *race*. This is due to the fact that in German, the term "Rasse" (race) is not generally accepted as an appropriate description and therefore the concept of race is not normally used.[2]

[1] It should be noted that gender bias is typically simplified to be binary, female and male. Other possible manifestations of gender are usually not addressed.

[2] See this statement of the Heidelberg Ethnology Institute on the concept and term "Rasse": https://www.

Research Questions I aim to answer the following three main research questions.

Research Question 1

How can origin bias be measured in static German word embeddings?

In order to analyse and mitigate bias, I first need to consider how bias can be measured, which is a non-trivial problem. I consider the strengths and weaknesses of various bias metrics and adapt one method to German. My hypothesis is that despite cultural differences, existing bias metrics can be adapted to German language embeddings by substituting language data. Furthermore, I make the fundamental assumption that bias contained in word embeddings can be captured using mathematical methods.

Research Question 2

Are static German embeddings origin-biased?

For English embeddings, the existence of racial bias has been detected in prior work. After defining robust bias metrics, I intend to examine whether this is also the case for German embeddings. My hypothesis is that German embeddings contain origin bias similar to the racial bias found in English embeddings. Additionally, it is likely that the exact nature of bias varies for subgroups of different ethnic origin, meaning that there might not be one generalisable "origin bias" against all people of non-German origin. Instead, there could be different biases against people from different specific countries of origin.

Research Question 3

If German embeddings are origin-biased, how can such bias be reduced?

By now, a multitude of approaches to so-called debiasing algorithms exist (Pfisterer, 2022). I examine multiple questions related to these methods. Do existing methods work in reducing origin bias in German embeddings and why or why not? Are the algorithms able to remove bias according to the initial definitions (see definitions 1 and 2) set? How do different methods compare to each other? To answer these questions, I implement and evaluate two debiasing methods. I hypothesize that any examined bias can be reduced by adapting existing debiasing algorithms to German origin bias.

eth.uni-heidelberg.de/md/eth/institut/statement_zu_konzept_und_begriff_rasse_fin.pdf

Thesis structure The remainder of this thesis is structured as follows: Chapter 2 reviews existing literature on bias and provides additional background. In chapter 3, I explain the methods I use to answer my research questions, give reasons for my methodological choices, and introduce metrics, models and data sets. The results of my experiments, including tabular data and diagrams, are presented in chapter 4 and discussed in chapter 5. Finally, chapter 6 concludes my thesis.

2 Related Research

In this chapter, I first give a brief outline of different static word embeddings, their significance, and how to evaluate them semantically. I then review prior work on different bias measurement and bias mitigation methods, which form the main theoretical foundation of my own work.

2.1 Word Embeddings

Word embeddings can be understood to be "dense, distributed, fixed-length word vectors", built using word co-occurrence statistics as per the distributional hypothesis (Almeida and Xexéo, 2019). The question of how best to turn language into vectors is one that has kept computational linguists busy for the last decades and out of which many different models have emerged.

The Role of Static Embeddings We can differentiate between two categories of embeddings. In the earlier years of word embedding research (arguably from Bengio et al. (2003) onwards), each distinct token is represented by its own fixed vector (Almeida and Xexéo, 2019). These are called static embeddings. In recent years, instead of static embeddings, pre-trained contextualised embeddings have seen wide-spread adaption across many fields in NLP (Bommasani et al., 2020), propelled forward by the publication of the first transformer model in Vaswani et al. (2017). Contextualised models create embeddings for a word based on the context in which it appears and in many areas outperform static embeddings to the point where one might wonder if static embeddings still have a place in today's research.

Two issues of contextualised embeddings are that they are challenging to interpret (Zini and Awad, 2022; Bommasani et al., 2020) and hard to train (Li, 2020). Static embeddings, in opposition, have the advantage of being more interpretable by humans and also being more economic in their training time and resources.

For this reason, Bommasani et al. (2020) convert contextualized embeddings to static embeddings for the purpose of interpretation. This enables interpretation methods for static embeddings —

such as bias measurement — to also be applied to contextualised embeddings. In another work by Dufter et al. (2021), the authors show that static embeddings perform at least as well as or better than contextualised embeddings on factual knowledge tasks due to their large vocabulary. Besides, since static embeddings were the standard for many years, it stands to reason that they are still in use in many applications today. Considering these factors, new research on static embeddings can still be a valuable addition to research.

In the following, I present the three most prominent static word embedding algorithms.

Word2Vec Mikolov et al. (2013) presented the pioneering Word2Vec (W2V) model, which is one of the first to successfully learn word embeddings with a neural network architecture. The authors devise two model architectures: *Continuous Bag of Words* (CBOW) and *Skip-gram*. Put briefly, the CBOW training objective is to maximize the model's likelihood of predicting target words from their context words, and Skip-gram works vice versa, i.e. by predicting context words from a target word (Kurpicz-Briki, 2020).

fastText FastText is an extension of Word2Vec introduced by Bojanowski et al. (2017) and expands upon Word2Vec by training embeddings on subword tokens. This change results in fastText embeddings excelling at capturing morphological variations of words, which is useful for morphologically rich languages (such as German) and words which would otherwise be outside the model's vocabulary. This characteristic and the fact that pre-trained fastText vectors have been publicly released in 157 languages[1] makes fastText a popular embedding choice.

GloVe Another influential work is the introduction of GloVe by Pennington et al. (2014). While Word2Vec and fastText optimise embeddings based on the local information surrounding a word, GloVe focuses on the global statistical information the corpus provides. It works by constructing a co-occurrence matrix of vocabulary tokens and optimises embeddings with the goal of capturing inter-word relationships based on this co-occurrence information. The resulting vectors have found extensive application in a myriad of NLP tasks due to their ability to express meaningful semantic relationships.

[1] See fastText website: https://fasttext.cc/docs/en/crawl-vectors.html.

2.2 Semantic Evaluation of Embeddings

Before I examine word embeddings for origin bias, the first question is whether they fulfill their original purpose of representing language at all. Finding a metric to evaluate this will later also be necessary for assessing whether the quality of the vectors decreased during the attempt of reducing their bias content.

How best to determine an embedding's quality remains an open research question. Embeddings are supposed to capture lexical semantics, and there exist a variety of methods to test their semantic properties, ranging from applying the embeddings in downstream tasks to cognitive science experiments (Bakarov, 2018). Downstream tasks here mean practical natural language applications in which embeddings are used, such as text classification.

As this thesis concentrates on bias analysis, I choose to focus on word pair similarity, which is a simple, commonly-used tactic for semantic evaluation. With this metric, a data set of word pairs is created and annotated for similarity by humans. The word embeddings which are to be evaluated are used to also calculate similarities for each word pair. A correlation test then determines if the model judges semantic similarity in the same way humans do.

The first human-annotated data set like this was created by Rubenstein and Goodenough (1965) as a psychological test, but contains only 65 word pairs. Finkelstein et al. (2001) created a larger set of 353 words by instructing annotators to rate each pair on a scale of one to ten. This data set is known as WordSim-353 or WS353 and constitutes one of the most popular data sets for semantic evaluation of English language embeddings. It is noteworthy that this data set contains two perhaps unexpected items, the first being a duplicate entry of the word pair `money-cash` with two different similarity scores, and the second being the word pair `tiger-tiger` with the obvious maximum score of ten.

Agirre et al. (2009) state that "different [word embedding] techniques are more appropriate to calculate either similarity or relatedness", and consequently split the WS353 data set into two (not mutually exclusive) subsets of 252 relatedness and 204 similarity word pairs. Furthermore, they advocate for Spearman Rank Correlation (Spearman, 1904) to be used as the correlation test since it is not dependent on the two data sets being linearly correlated.

Similar data sets for other languages are scarce but exist. Leviant and Reichart (2015) translated WS353 into German, Italian and Russian. To ensure consistency in their annotation, they devised new annotation guidelines and re-annotated the English as well as the new multilingual data sets. They kept the duplicate money-cash word pair, but assigned the same score to both. However, they excluded three word pairs which they deemed untranslatable, resulting in a data set size of 350 word pairs. Gurevych (2006) created another German data set independent from WS353 using a corpus-based approach, in which the author extracted word pairs of varying parts of speech from GermaNet. Unlike the other data sets mentioned, GUR350 therefore includes more than just nouns. This data set contains 350 word pairs and is sometimes referred to as GUR350.

2.3 Bias Measurement Approaches

After giving an outline of one possible method of semantic evaluation, I now present an overview of recent work on bias evaluation, which is one of the core parts of my thesis.

Bolukbasi et al. (2016)

The groundwork for bias assessment and mitigation was laid in a seminal work by Bolukbasi et al. (2016). They uncovered the fact that English word embeddings contain gender bias by critically examining word analogies and the relation of Word2Vec and GloVe embeddings to gendered terms.

Their methodology is based on the hypothesis that the concept of gender is linearly separable in the embedding space. This means that they assume there is a specific, removable, linear component in word vectors which makes, for instance, $\overrightarrow{king}$ different from $\overrightarrow{queen}$.[2] Bolukbasi et al. (2016) refrain from explicitly testing this hypothesis, but it has since been mathematically proven by Vargas and Cotterell (2020).

To go into more detail, Bolukbasi et al. (2016) collected *definitional pairs* of words such as <he, she> or <man, woman> and then used those pairs to define a *gender-direction* in the embedding space. They did so by performing a principal component analysis of the pairs' aggregated difference vectors and then taking the first principal component, which they argue contains most gender information. This component, which is a vector of the same dimensionality as the model's embeddings, is the gender direction. Calculating the bias-direction is an important concept,

[2] Throughout this thesis, a word's vector will be distinguished from the word itself by arrow notation.

both in their work and as an assumption in subsequent research. It can be viewed as a vector that captures the concept of gender in the embedding space. Bolukbasi et al. (2016) used it to define the bias of words by calculating a given word vector's projection onto the gender-direction. Assuming all vectors are normalised, they interpreted a higher absolute resulting value as higher bias. In other words, if a word vector is similar to the gender-direction, it is likely related to the concept of gender.

Apart from these calculations, the authors also carried out a word analogy test in which they let their model predict answers to tasks such as `she:he::nurse:x`[3], to which the model's most likely estimates were gender-stereotypical answers like, in this example, `surgeon`.

Critique of Bolukbasi et al. (2016) metrics

The bias evaluation step is often effectively equated with defining bias, meaning that in many works, the bias metric determines what is understood to be "bias". This is due to the fact that researchers will restrict their analysis of bias to only their, perhaps singular, quantifiable measure they employ and neglect to consider the broader implications of bias. This practice limits the extent to which bias can be understood because authors extrapolate general statements about bias from narrow bias definitions. However, it is questionable whether any single metric is able to cover all aspects of bias in word embeddings.

Gonen and Goldberg (2019) addressed this issue in their "Lipstick on a Pig" paper and went so far as to label the metrics presented in Bolukbasi et al. (2016) and later works (Zhao et al., 2018a) as important, but ultimately being "party tricks". In their words, "gender-direction is a great indicator of bias, [but] it is only an indicator and not the complete manifestation of this bias". The authors assessed that restriction to this bias definition has led Bolukbasi et al. (2016) and other researchers to report great success in their debiasing methods since they were able to calculate high bias values before debiasing and low values afterwards, but that the "debiased" embeddings may still contain bias in other, previously undetected ways.

In particular, they explored what Bolukbasi et al. (2016) call "indirect bias" in more depth. Indirect bias essentially describes the bias measured in word embeddings with second order similarities instead of first order similarities. Instead of measuring how related a word vector is to a certain bias attribute, Gonen and Goldberg (2019) analysed how associated the neighbours of two words

[3] Read: "She is to he as nurse is to what?"

are to each other. For example, it might be the case that $\overrightarrow{nurse}$ is not be directly associated with $\overrightarrow{she}$ or $\overrightarrow{he}$, but is associated with $\overrightarrow{receptionist}$ and $\overrightarrow{housekeeper}$. This could indicate that $\overrightarrow{nurse}$ is biased, if the association between these three professions can only be explained by the fact that they are all stereotypically female occupations.

Gonen and Goldberg (2019) carried out various experiments relating to these indirectly biased associations. They concluded that they are pervasive in word embeddings, explainable solely through undesired gender stereotypes, and not captured by other bias metrics. Consequently, Gonen and Goldberg (2019) proposed a new bias metric in which they calculate how many of an embedding's k nearest neighbours are "socially biased", i.e. have a strong stereotypical association with the bias attribute.

Some subsequent works (e.g. An et al., 2022; Manzini et al., 2019; Aekula et al., 2021) have implemented this new metric in the form of a clustering task. In this implementation, the most biased words are determined before debiasing according to gender-direction bias and after debiasing, a clustering algorithm is tasked with clustering these previously highly biased embeddings into two clusters. Then, the accuracy of the clustering algorithm is measured in regard to whether the two clusters correspond to the two bias characteristics, e.g. male and female. If the accuracy is high, this indicates that the embeddings still contain information that allows them to easily be separated by gender. This metric is sometimes called the *neighbourhood metric*.

Further criticism of Bolukbasi et al.'s bias evaluation relates to their use of word analogies. Nissim et al. (2020) agree with Gonen and Goldberg (2019) on calling word analogy metrics "party tricks" and warn against using this method. They point out that in many word analogy tasks, the analogy equation is formulated in such a way that it is impossible for the word embedding model to return one of the analogy input words. What this means is that for a query such as "*X* is to *woman* as *doctor* is to *man*" the answer is — per definition — not allowed to be *doctor*, since *doctor* is also an input term. Nissim et al. (2020) demonstrated that many famous word analogies which seemingly prove the existence of bias are rendered spurious by taking constraints such as this into consideration. Furthermore, the authors criticised the subjectivity which is introduced into bias measurement by contriving the queries used in word pair analogy tasks.

Caliskan et al. (2017)

Aside from calculating direct bias with projections onto the bias-direction as described above, by far the most common bias evaluation metric for static embeddings is the Word Embedding Association Test (WEAT) introduced by Caliskan et al. (2017). It measures the strength of association between word embeddings of different predefined categories.

The WEAT is based on the Implicit Association Test (IAT), which is a psychological study by Greenwald et al. (1998) designed to assess implicit social biases. Participants were asked to sort words into different categories and their reaction time for this task was measured. For example, in one experiment participants were given African American names and pleasant words like `joy`, `love`, or `peace`. In the next step, those African American names were substituted with European American names. The participants' reaction times for assigning pleasant terms to African American names were slower than when presented with European American names, which indicates a subconscious racial bias.

Caliskan et al. (2017) adapted this test for the use case of embeddings evaluation by substituting reaction time with cosine similarity as the measure of association. The WEAT quantifies bias by testing whether words belonging to one attribute group, e.g. "Black words", are associated with words belonging to two target groups, e.g. "pleasant words" and "unpleasant words", differently than words belonging to another attribute group, e.g. "White words". It does so with a statistical test which systematically compares the different groups. The procedure is more fully explained in §3.2.2. The original IAT was carried out for a variety of different attributes. Caliskan et al. (2017) adapt all original IAT experiments and thereby provide a metric and reference values for gender and race bias.

Modifications of Caliskan et al. (2017)

The WEAT has been translated to other languages before. Most pertinent to my research questions are Kurpicz-Briki's translated WEAT seed words for four of the original WEAT experiments. The author published translations to German, specifically Swiss German, and French. Using these translated seed sets, she carried out experiments measuring gender and origin bias in German fastText embeddings and was able to reproduce prior bias discoveries for this setting. In terms of translation methodology, the author kept relatively close to the original seeds, translating terms directly wherever possible. For the attribute sets consisting of names, Kurpicz-Briki (2020) used official federal lists of the most common names in Switzerland and manually selected names of "originally Swiss origin" versus "of different origin".

There exist a host of other WEAT translations such as Lauscher et al. (2020) for Arabic ("AraWEAT"), Biasion et al. (2020) for Italian, and Qin et al. (2023) or Jiao (2021) for Chinese. I will not discuss them in more detail in this thesis, since my focus is on German embeddings.

Manzini et al. (2019) studied multi-class bias, i.e. bias attributes where there are more than two manifestations like religion or race.[4] To this end, they proposed a variation to the WEAT which they named Mean Average Cosine Similarity (MAC). Ultimately, this metric simply performs the WEAT for all attribute sets (e.g. Islam, Judaism, Christianity), then normalises and averages the resulting scores.

Critique of Caliskan et al. (2017)

While the WEAT is a widely used metric and generally agreed upon to be a methodologically well-executed work, researchers have also found faults with it.

Van Loon et al. (2022) found that for WEAT experiments which use names as attributes, results are highly dependent on the relative frequency of the names in the embedding training corpus. The researchers used geo-tagged data from X (formerly Twitter) to train word embeddings for different geographical areas. For each area, they then measured the correlation between WEAT results and a collection of sociological metrics for "anti-Black sentiment" in that area, controlling for various factors. The study revealed a strong and significant correlation between the WEAT and "anti-Black sentiment" metrics without control variables or with standard controls, but this correlation became weak and non-significant when controlling for relative Black name frequency. This suggests that the ability of the WEAT to predict anti-Black bias depends on access to name frequency information.

The authors hypothesise that this is likely due to most word embedding models clustering tokens together in the vector space not just based on semantic information, but also based on factors like frequency (Mu et al. (2017), Gong et al. (2018)). Since humans use positive words more

[4] Again, let it be noted that gender, too, exists outside a binary spectrum. However, to the best of my knowledge, no work exists in which gender bias is not simplified to be a two-dimensional problem. Similarly, race is also often treated as a black/white dichotomy.

frequently than negative words (van Loon et al., 2022), this may lead to positive and frequent words, and negative and infrequent words respectively, to be grouped together in the embedding space. For example, if a Black name is rare in the data, the embedding model might estimate its vector to be closer to negative than positive words partly because negative words are also more rarely used in the corpus. That Black names are indeed more rare in corpora has been examined in numerous works such as Wagner et al. (2015) who showed that Wikipedia is biased. To address such frequency issues, the authors recommend measuring and controlling for relative seed word frequency when using the WEAT.

Ethayarajh et al. (2019) also challenged the effectiveness of the WEAT, pointing out theoretical flaws leading to a systematic overestimation of bias. They, too, criticised its dependency on frequency in terms of skewed results when attribute words in the two sets have unequal frequencies. Additionally, they remark that results can easily be manipulated by contriving the seed words used.

The authors introduced a new metric called RIPA (Relational Inner Product Association) which is very similar to the gender-direction method used by Bolukbasi et al. (2016), except that Ethayarajh et al. (2019) introduced it more formally and explored the mathematical background of this metric more closely. It is perhaps surprising that despite Bolukbasi et al. (2016) themselves stating that their bias definition could theoretically be extended to account for weighted frequency, Ethayarajh et al. (2019) did not introduce such frequency weights despite noting the issues that arise from frequency.

Interestingly, the authors furthermore compared the bias measured with RIPA on Word2Vec embeddings to what the bias would be "under perfect reconstruction", i.e. using the actual word distributions seen in the training corpus.[5] Their empirical testing revealed that words which are gender-neutral in the corpus (such as "potato") remain un-gendered in the trained embeddings, whereas for gendered words the model tends to amplify their genderedness. They attribute this amplification to the model's definitional ability of grouping words in similar contexts together. This observation points to a potential trade-off between emphasizing semantic associations, which is usually desired for language models, and inadvertently reinforcing harmful associations.

[5] See Ethayarajh et al. (2019) for the mathematical background of this, as a workup would extend beyond the constraints of this thesis.

Antoniak and Mimno (2021)

Finally, I want to direct attention to the issue of "bad seeds" studied by Antoniak and Mimno (2021). All existing bias metrics rely on lists of seed words, usually to define bias lexically. However, as Antoniak and Mimno (2021) pointed out, these lists are problematic: "The rationale for choosing specific seeds is often unclear, [...] the impact of the seeds is not well-understood, and many previous seed sets have serious limitations".

The authors criticised multiple points: First, the seeds themselves can inadvertently contain biases if, for example, ugliness is defined with terms such as fat, chubby, disfigured and wrinkled. Second, seed sets are often created for a specific purpose and then unthinkingly reused for different domains, different embeddings or different bias definitions. Third, researchers usually use just one set of seed words without testing its stability by comparing the results to different possible seed sets.

In their experiments, Antoniak and Mimno (2021) showed that for the same embeddings and measurement method, different seeds can lead to widely varying results. As a method of determining the coherence of a seed set, they measured how far apart the two groups of seeds belonging to a set (e.g. female vs. male words) were in the embedding space. They found that coherence varies greatly.

The authors identified different factors which can lead to seed instability. Among these factors are reductive bias definitions, the frequency and part-of-speech category of individual seed words, the size of seed sets, and similarity of seed groups to each other. They recommended future researchers to trace the origins of seed sets, examine seed features, document all seeds publicly, and generally warned of the dangers of following research precedents without critical examination.

Further Literature More approaches to bias measurement exist. For example, Caliskan et al. (2022) looked beyond the WEAT method and provide an in-depth bias analysis with heuristics like frequency and POS tags. Readers interested in a holistic overview may refer to literature reviews by Sun et al. (2019), Blodgett et al. (2020) or Papakyriakopoulos et al. (2020). Furthermore, Jin et al. (2021) provided a broad analysis of the social impact of NLP research.

2.4 Debiasing Methods

Upon the discovery of bias in word embeddings, much research on mitigating this bias has sprung forth. I will here focus on categorising approaches and presenting examples of each category, especially in so far as they are seminal works or fundamental to my own debiasing experiments.

Debiasing with Post-Processing

Some of the first researchers to address the issue of "debiasing" embeddings were Bolukbasi et al. (2016) in the same paper as introduced above. Building upon their hypothesis that gender is a linearly separable subspace of static embeddings, they proposed to simply remove this subspace from the trained embeddings.

Bias Subspace Removal The authors presented two slightly different debiasing algorithms. "Neutralize and Equalize" aligns all gender-neutral words, which is the vast majority of the vocabulary, to be orthogonal to the gender subspace ("Neutralize") and also to be equidistant to the two words of a word pair in a predefined list of word pairs ("Equalize"). For example, $potato$ should have zero association with the $she - he$ direction, and also be equidistant from $king$ and $queen$. This algorithm is referred to under different names, but usually called *Hard Debiasing* (Gonen and Goldberg (2019), Wang et al. (2020)). The second algorithm proposed, "Soften" contains a parameter which regulates the degree of equalisation, so that the embeddings do not need to be perfectly equidistant in the "Equalize" step.

Bolukbasi et al.'s method is an example of a *post-processing approach* to debiasing — they manipulated embeddings after they were already trained. Interestingly, Ethayarajh et al. (2019) argued that for embeddings which are constructed directly or indirectly using matrix factorisation of a co-occurrence matrix, removing a linear component from the finished embeddings can be equivalent to manipulating the training data before training, since the bias subspace could theoretically also be factored out from the original co-occurrence matrix.

Wang et al. (2020) put forward an advancement of the Hard Debiasing algorithm which takes into consideration the issue of word frequency discussed in §2.3. The authors argued that word frequency can distort the gender-direction and therefore first performed a step in which they remove frequency information from embeddings before performing Hard Debiasing. They showed that their debiasing also performed well according to the neighbourhood metric (Gonen and Goldberg, 2019), indicating that their so-called *Double-Hard Debiasing* algorithm outperforms other post-hoc methods.

Manzini et al. (2019) not only measured multi-class bias, but also adapted Hard Debiasing for multi-class scenarios such as race. They did this by calculating the bias subspace that Hard Debiasing uses with additional seed words for each additional attribute category. For example, instead of defining the bias subspace with word pairs such as <`woman, man`>, they defined it with word sets such as <`church, mosque, synagogue`>. The additional terms were simply concatenated in the bias calculations. For instance, instead of forming a 300×2 matrix for 300-dimensional embeddings and two bias characteristics, a 300×3 matrix was formed to accommodate three bias characteristics.

Post-hoc Dictionary Debiasing Kaneko and Bollegala (2021) suggested another post-processing method and introduced a new concept to debiasing research, leveraging dictionary definitions as external unbiased resources. Their approach is based on the hypothesis that dictionary definitions contain relatively neutral information about words. They trained an encoder-decoder model with the purpose of altering input word embeddings in three ways. First, the embeddings were encouraged to be orthogonal to a bias space (inspired by Bolukbasi et al. (2016)), second, the embeddings should become similar to an embedding which represents that word's dictionary definition, and third, all other information in the embeddings should be preserved. This way, the authors attempted to enrich prior methods with additional external information. However, An et al. (2022) judged that Kaneko and Bollegala (2021) did not achieve convincing results. This is characterised by many of their reported improvements being either marginal or not statistically significant, and their paper containing multiple unclear or contradictory results.

Debiasing with Adjusted Training Objectives

A different debiasing strategy is adjusting the training objective of an embedding model at train time. A pioneering example of this type of debiasing is the work by Zhao et al. (2018a), who introduce their algorithm *GN-GloVe* (gender-neutral GloVe).

GN-GloVe Zhao et al. (2018a) criticised the debiasing method of Bolukbasi et al. (2016). They stated that "[f]irst, [Bolukbasi et al.'s] method is essentially a pipeline approach and requires the gender-neutral words to be identified by a classifier before employing the projection. If the classifier makes a mistake, the error will be propagated and affect the performance of the model. Second, their method completely removes gender information from those words which are essential in some domains such as medicine and social science". Zhao et al.'s method intends to alleviate these issues in a number of ways.

First, instead of manipulating the embeddings in a multi-step post-processing pipeline, they executed debiasing during one cohesive step while the embeddings are being trained. Second, they did not completely remove gendered information from certain words as Bolukbasi et al. (2016) do, but instead made alterations to GloVe's training objectives so that gender information is contained to a limited number of embedding features and the remaining vector dimensions can be neutralised. One can then later exclude the gender features if one wishes to do so. The authors achieved this by defining female and male seed words and encouraging the model to differ in the "gender coordinates" (Gonen and Goldberg, 2019) for words of the female versus male set. The remainder of the features was encouraged to be orthogonal to the gender-direction, similar to Bolukbasi et al. (2016).

Dictionary Debiasing at Train Time DD-GloVe (Dictionary-Debiasing GloVe) by An et al. (2022) constitutes another method for leveraging dictionary definitions for debiasing, comparable to Kaneko and Bollegala (2021), except that DD-GloVe adjusts embeddings at train time. The authors directly adapted the training objectives of the GloVe model with the goal of reducing gender and racial bias. They proposed four new loss functions, which together serve to remove gendered or racial information that is not present in dictionary definitions from the embeddings, and generally bring a word's embedding closer to its dictionary definition. The precise loss functions are explained in §3.3.2.

An et al. (2022) criticised the approach of Kaneko and Bollegala (2021), stating that the assumption of definition embeddings, i.e., embeddings that capture definition texts, as neutral reference points "is a major flawed assumption in post-processing debiasing. Due to the biases in pretrained word vectors, the definition embeddings also contain biases".

An et al.'s own method addresses this issue by training the definition embeddings at the same time as all other embeddings. In their model, definition embeddings are simply an average of the word vectors contained in a definition. As the word vectors begin to gain semantic meaning, so do the definition embeddings. An et al. (2022) claimed that in every training iteration, the embedding definitions become more neutral because their building blocks — the word embeddings — get increasingly debiased in each iteration, and in return the definition embeddings again influence the word embeddings to become more similar to their definition.

This approach is reported by the authors to have promising results. DD-GloVe is claimed to perform on par or better than multiple other debiasing methods on metrics such as the WEAT or the neighbourhood metric.

Debiasing with Data Substitution

Apart from adjusting the embedding creation process after or during training, there also exist debiasing methods which alter the training data used to create embeddings.

Lu et al. (2020) were among the first to propose such a solution. They performed Counterfactual Data Augmentation (CDA) wherein they duplicated a training corpus and then substituted gendered words with their opposites in the duplicate corpus. For example, they substituted "king" with "queen" in the duplicate corpus. While doing so, they followed some additional substitution rules which prevent the creation of semantically incoherent or grammatically incorrect sentences. Using this method, they aimed to create a gender-balanced training corpus, which they argue would lead to unbiased embeddings. The authors measured bias with a coreference resolution task in which the model had to match occupations to genders. Using this metric, they reported success of their debiasing method. However, they did not report more widely used metrics such as the WEAT.

An advancement of CDA was put forward by Maudslay et al. (2019), who introduced a method they call Counterfactual Data Substitution (CDS). CDS improves upon CDA by employing a "Names Intervention", which is a name-pairing technique with which first names are no longer ignored in the substitution process. Additionally, instead of duplicating the corpus and substituting every gendered term, Maudslay et al. (2019) substituted potentially biased text within the same original corpus, but only with a 50 % likelihood. According to the authors, this prevents the creation of unnatural duplicate text. Furthermore, substitutions were performed at a document level instead of sentence level to improve coherence. The authors reported debiasing success measured on the WEAT, but also in terms of indirect bias measured by a clustering task.

One drawback of training data substitution methods is that they are time-consuming because they require the user to define substitution word pairs, execute the substitution over an entire training corpus, and then also re-train the embeddings.

I have presented an overview of pertinent debiasing methods and strengths and weaknesses of those methods. For a more complete survey of existing methods, interested readers may again refer to literature reviews by Papakyriakopoulos et al. (2020) and Sun et al. (2019). Comparisons between the performance of different methods can furthermore also be found in the results sections of some of the papers mentioned in this chapter, e.g. in An et al. (2022) or Wang et al. (2020).

3 Methods

The methodology in this study consists of experiments carried out with the goal of quantitatively answering my three leading research questions (see chapter 1). In this chapter, I will expand upon the metrics and models used in my research. This includes presenting the materials such as text corpora or other data sources used by these methods.

3.1 Choice of Embeddings

The bias present in word embeddings differs between different types of embeddings and the way in which bias can be analysed and mitigated also depends on the embedding model used. Therefore I first present the embeddings I selected for my experiments.

Selecting Pre-Trained Embeddings

In §2.1, I present works supporting the relevance of static embeddings. For multiple reasons, this thesis only examines static embeddings. There is more sophisticated prior bias research available for these embeddings than for contextualised embeddings, they are more explainable, easier to use and faster to train, which enabled me to focus on a thorough analysis instead of the training process.

The primary embeddings used in this study were fastText (Bojanowski et al., 2017) and GloVe Pennington et al. (2014) embeddings.

One of the reasons for selecting fastText were that reference values for the WEAT and for word similarity tasks are available in literature for English and German embeddings. Furthermore, fastText handles out-of-vocabulary words better than W2V (see explanation in §2.1), which is especially relevant because the evaluation data I used contains some rare words like foreign first names.

Last but not least, pre-trained German fastText embeddings are readily available for download on the official fastText website[1]. The more commonly used version of these embeddings is trained on

[1] https://fasttext.cc/docs/en/pretrained-vectors.html

Common Crawl, but in this thesis I examine fastText embeddings trained on Wikipedia since they are more directly comparable to the GloVe embeddings I used. These fastText embeddings are 300-dimensional, their vocabulary size is 2,275,233 and they were trained using the skip-gram method with default parameters as specified in Bojanowski et al. (2017).

GloVe, on the other hand, I selected primarily because it is the only embedding model for which DD-GloVe (An et al., 2022), one of the debiasing methods I implemented (see §3.3.2), is available. Additionally, GloVe is an embedding model which generally performs well on standard benchmarks and is therefore widely used (see §2.1).

I obtained pre-trained German GloVe embeddings from Deepset[2], a private Berlin-based NLP solutions company who publish part of their products open-source. The embeddings were trained on a German Wikipedia dump of unspecified date. Based on Deepset's GitLab repository[3], the Wikipedia dump is likely to originate from 2018. The embeddings have 300 features and the vocabulary is 1,309,280 tokens large. Deepset do not explicitly specify further training details, but based on their model code appear to have trained their GloVe embeddings using a minimum vocabulary count of 20, no maximum vocabulary size, 30 iterations, and a symmetrical context window size of 15. Other parameters seem to be unchanged from Pennington et al. (2014).

In my experiments, I used these pre-trained GloVe embeddings only for validation purposes since I also trained GloVe embeddings from scratch, as described in the next subsection. Those are the main GloVe embeddings I analyse. I only report results for the pre-trained GloVe embeddings for experiments where it was necessary to validate my GloVe training.

Training GloVe Embeddings

For the DD-GloVe debiasing method introduced later (§3.3.2), it is necessary to train GloVe embeddings from scratch. To be able to accurately compare embeddings before and after debiasing, I also trained GloVe embeddings normally, i.e. without debiasing. For this purpose, I used the same training code and data as for DD-GloVe, only without the additional debiasing loss functions.

[2] https://www.deepset.ai/german-word-embeddings

[3] https://gitlab.com/deepset-ai/open-source/glove-embeddings-de

The code implementation of GloVe I utilised is that by An et al. (2022), available on GitHub[4]. It is largely based on the original GloVe code by Pennington et al. (2014), also published on GitHub [5]. An et al. provide an option to run their training code without debiasing, i.e. with original GloVe loss functions. In theory, this is equivalent to running Pennington et al.'s code since it excludes all modified objective functions.[6]

I ran GloVe training using a maximum vocabulary size of 400,000, minimum word count of 5, vector size of 300, symmetrical window size of 10, 40 iterations, $\alpha = 0.75$, and optimisation with AdaGrad (Duchi et al., 2011), with an initial learning rate of 0.05. These parameters are largely identical to those specified in Pennington et al. (2014) except for the smaller number of training iterations, which is based on 0raining specifications by An et al. (2022). Moreover, vector values are clipped to be in the range of $[-1, 1]$ as implemented by An et al. (2022) "to avoid numerical difficulties".

The training data for my German GloVe embeddings is a Wikipedia dump made available by HuggingFace[7] and dated 2022-03-01. It is comprised of 2,665,357 articles containing 1,147,061,829 words and 15,572,403 unique tokens. I selected a Wikipedia corpus due to its availability, manageable size compared to corpora such as CommonCrawl, and common usage as embedding training data and therefore comparability. As pre-processing steps, I lowercased the corpus and tokenised it using NLTK's `word_tokenize`[8], thereby also removing punctuation. This is based on pre-processing steps performed by Pennington et al. (2014) and An et al. (2022), except those works tokenised with the Stanford tokeniser[9].

To validate my self-trained embeddings, I performed a semantic evaluation task, which I will explain further in the next section, and compared their performance to values found in literature.

[4] https://github.com/haozhe-an/DD-GloVe

[5] https://github.com/stanfordnlp/GloVe

[6] I also briefly confirmed this empirically by calculating similarity scores for the WordSim-353 data set (see §2.2) for two sets of embeddings, one obtained from the Stanford NLP website (https://nlp.stanford.edu/projects/glove/) and one trained using An et al.'s code. I then compared the resulting similarity scores using Spearman's rank correlation coefficient. The correlation was very high with a coefficient ρ of approximately 0.8991 and an associated p-value of approximately 0, indicating that the two training methods result in similar embeddings. The difference can be explained by different training data used, since Pennington et al. (2014) use an older Wikipedia dump and slightly different training parameters such as a higher number of iterations than An et al. (2022) use.

[7] https://huggingface.co/datasets/wikipedia

[8] https://www.nltk.org/api/nltk.tokenize.word_tokenize.html

[9] https://stanfordnlp.github.io/CoreNLP/tokenize.html

3.2 Evaluation Metrics

In this section, I present the metrics with which I evaluated the embeddings introduced above. First, I briefly explain the word similarity task I used for semantic evaluation, and then I focus on my methods for bias evaluation. I expand on the theoretical background of the evaluation methods, the rationale for selecting them, and the changes I make to adapt them for German origin bias.

The selected metrics were applied to each set of embeddings before and after debiasing in order to answer the research questions of whether German embeddings contain origin bias and if the tested methods are effective in mitigating such bias.

3.2.1 Semantic Evaluation Metric

The purpose of including a semantic evaluation metric is to ensure that the embeddings I used are consistent with those used in other literature and that any debiasing algorithms applied to them affect only the bias components and not the general performance of the embeddings.

How best to measure the quality of embeddings is an open research question. As described in §2.2, a common test is word pair similarity, particularly using the WS353 data set (Finkelstein et al., 2001). Word pair similarity is a relevant metric in the context of debiasing because it focuses on the semantic relationships between word vectors. Debiasing methods often also utilise the semantic relationship between word pairs as a basis for debiasing (e.g. by defining a bias-direction) and could therefore potentially unintentionally impact these semantic properties in word embeddings in adverse ways. A decrease in performance after debiasing would indicate that this is the case.

Since I examine German embeddings, I used a German word similarity data set instead of WS353, namely the GUR350 data set[10] (Gurevych, 2006). Out of all available German data sets for word pair similarity, none are dominant over the others in terms of their use in relevant literature. An advantage of GUR350 is that the author provided reference test scores for embeddings that are comparable to the embeddings I used. Furthermore, with a size of 350 word pairs, it is one of the

[10] It can be found in full at https://github.com/dkpro/dkpro-similarity/blob/master/dkpro-similarity-experiments-wordpairs-asl/src/main/resources/datasets/wordpairs/de/wortpaare350.gold.pos.txt

two biggest German word pair similarity data sets. The other of those test sets is that created by Leviant and Reichart (2015)[11]. However, Leviant and Reichart (2015) take their word pairs directly from WS353 without reviewing whether they are suitable for embeddings trained on other languages. In comparison, the GUR350 data is derived with a corpus-based system using German language data and is therefore inherently based on the properties of German language. Besides, the multilingual WS353 translation by Leviant and Reichart (2015) contains some unaddressed spelling errors in its German variant ("Präzendensfall" instead of Präzendenzfall and "Palestinenser" instead of Palästinenser) which cast doubts upon its reliability.

To calculate an embedding's similarity score for a given word pair, I used cosine distance as implemented by the python library SciPy (Virtanen et al., 2020), which for two vectors $\vec{u}$ and $\vec{v}$ is defined as

$$d = 1 - \frac{u \cdot v}{||u||_2 ||v||_2}$$

For easier comparison to other works which measure similarity and not distance, I took the complement of d. I measured the correlation of the similarity calculated by the model with the judgement of humans using Spearman's rank correlation coefficient (Spearman, 1904), also implemented with SciPy[12].

For my experiments on GloVe embeddings, some words in GUR350 are out of vocabulary. In those cases, some existing works (Bojanowski et al., 2017) assign a similarity of zero to that word pair. This makes sense since usually word pair similarity tasks are used to measure an embedding's performance and the model should therefore be "punished" for words it does not know with the likely incorrect similarity score of zero. In this thesis, however, I primarily aim to measure the difference in performance before versus after debiasing. Out-of-vocabulary tokens are not included in the debiasing process since they are not part of the model. Because of this, the *difference* in performance can best measured by eliminating out-of-vocabulary words from the test. This puts a greater focus on the semantic performance for words which were affected by the debiasing algorithm.

I carried out three experiments: One validating my embeddings by comparing my embeddings' performance on the full GUR350 data set to values found in literature, one analysing the effects of reducing the GUR350 data set to exclude out-of-vocabulary tokens, and finally one with

[11] Data set available at https://leviants.com/multilingual-simlex999-and-wordsim353/.
[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

the reduced version of GUR350 comparing the embeddings' performance before versus after debiasing.

3.2.2 The WEAT Method

The metric I employed for bias measurement is the Word Embedding Association Test (Caliskan et al., 2017). The WEAT is a versatile metric which can be adapted for different languages and biases by exchanging the used seed lists.

Caliskan et al. (2017) compared WEAT values to the results of the original psychological IAT experiment and thereby backed the validity and expressiveness of the WEAT with a comparison to human biases. This gives the WEAT an advantage over other bias metrics which do not have a frame or reference for determining how well they are able to capture bias. Furthermore, the WEAT is easily interpretable because it is a statistical permutation test which has clear definitions for the significance of its results. Additionally, many reference values for WEAT experiments are available in peer-reviewed literature since it is the most common evaluation technique.

Flaws of the WEAT like seed instability are addressed in the next subsection, in which I explain the adjustments I make to the WEAT. First, however, I explain the methodology of the WEAT in more detail since this method holds great significance to my thesis and understanding it is necessary in order to understand the results of my experiments.

Each WEAT experiment always requires four seed lists, two attribute and two target lists. The association between these sets of words is measured using cosine similarity and a statistical permutation test. Note the terminology of *target words* versus *attribute words*, which are sometimes confused or swapped in literature. In my thesis, I will abide by the original terminology of Caliskan et al. (2017), in which *attribute words* are those pertaining to a certain human attribute for which the bias is to be tested, e.g. gender or race, and *target words* are those to which the attribute words are compared, e.g. professions or sentiments. While target and attribute words are mathematically indistinguishable and fully commutative (Maudslay et al., 2019), a clear terminology is still helpful for communication.

In her German WEAT translation, Kurpicz-Briki (2020) translated target and attribute words for WEAT 5, which is the WEAT experiment pertaining to race as per the original numbering in Caliskan et al. (2022). I list them in Table 1 for convenience. Kurpicz-Briki (2020)

Seed Type	Label	Terms
Attribute	Group 1 ("originally Swiss")	Peter, Daniel, Hans, Thomas, Andreas, Martin, Markus, Michael, Maria, Anna, Ursula, Ruth, Monika, Elisabeth, Verena, Sandra
Attribute	Group 2 ("different origin")	Ladina, Fatima, Fatma, Alma, Soraya, Svetlana, Elif, Vesna, Mehmet, Mustafa, Aleksandar, Mohamed, Ibrahim, Dragan, Hasan, Mohammad
Target	Pleasant	Spass, Liebe, Frieden, wunderbar, Freund, Freude, Lachen, glücklich
Target	Unpleasant	Qual, furchtbar, schrecklich, übel, böse, Krieg, scheusslich, Versagen

Table 1: Seed words translated by Kurpicz-Briki (2020) for the original WEAT 5 experiment (Caliskan et al., 2017). Associations are compared between the two different attribute lists in relation to the target lists. Each attribute list contains 8 female and 8 male first names, and the target lists 8 terms each.

reported results for fastText embeddings, which enables me to validate my WEAT implementation by comparing my scores for fastText embeddings with the values published by Kurpicz-Briki (2020).

The null hypothesis H_0 of the WEAT is that "there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words" (Caliskan et al., 2017). This hypothesis is tested as follows.

First, a function is defined which for a given target word w (e.g. "Liebe") measures whether this word is, on average, more similar to the words in attribute set A (e.g. German names) than the words in attribute set B (e.g. names of other origin). Caliskan et al. (2017) call this function $s(w, A, B)$, but since they overload s, I instead name it c for cosine and define it as

$$c(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

For unbiased embeddings, c would return a value close to 0 since the the two means would be approximately the same. A positive return would, in my example, mean a closer similarity of "Liebe" to names of German origin than to names of other origin, and a negative return would mean the opposite. By summing the returns of c over the entirety of a target set X, which could, for example, be a collection of pleasant words, one can estimate a tendency of whether one attribute group is more similar to pleasant words than the other group. The same can be repeated for the second target set Y (for example, the unpleasant words). This forms the test statistic s, which can be expressed as

$$s(X, Y, A, B) = \sum_{x \in X} c(x, A, B) - \sum_{y \in Y} c(y, A, B)$$

For perfectly unbiased embeddings, both summands should be close to zero. If the embeddings contained negative bias towards foreign names, the second sum would be positive and the first negative, resulting in a negative value s.

Using these function, it can be tested whether the null hypothesis can be rejected with statistical significance. To this end, a permutation test is performed in which all i possible equal-sized partitions A_i and B_i of the attribute sets A and B are formed.[13]. For example, one such permutation might be $A_{42} = \{$*Peter, Daniel, Hans, Thomas, Andreas, Martin, Markus, Michael, Maria, Anna, Ursula, Ruth, Ibrahim, Dragan, Hasan, Mohammad*$\}$.[14] For each of those permutations, the test statistic s is calculated and compared to the test statistic which results when using the original, i.e., non-permuted attribute sets. The percentage of permutations in which the measured association is greater than the original association, i.e., the percentage for which there was larger incidental bias measured than what the actual data shows, is the p-value of this test, formally given as:

$$Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B) | H_0]$$

I used the standard significance level $\alpha = 0.05$, meaning that for p-values below 0.05, the null hypothesis can be rejected. A p-value above 0.05 does not necessarily mean that the null hypothesis is true, i.e., that the embeddings are bias-free. The WEAT can only measure presence, not absence of bias.

In order to make the WEAT calculations computationally feasible, I followed Kurpicz-Briki (2020) and Chaloner and Maldonado (2019) and computed only 100,000 randomly selected permutations instead of all possible permutations. With $\binom{32}{16} = 601,080,390$ possible combinations

[13] Caliskan et al. (2017) talk about permuting the *attribute* sets, which makes sense contextually and in terms of the number of possible permutations, but then suddenly use X_i and Y_i in their notation for the permutations. This is likely an error, regrettably copied by subsequent publications, and meant to be A_i and B_i

[14] The partitioning is done with no consideration for gender, but for a large enough number of permutations, the law of large numbers suggests that on average, there will be an equal number of female and male names in A_i and B_i.

for an attribute set size of 16, it is highly improbable to encounter the same permutation twice.

Additionally to the p-value, I also reported the effect size of the test with Cohen's d (Cohen, 2013), defined as

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{{}_{w \in X \cup Y} s(w,A,B)}$$

It describes the difference between the average of two distributions — here, the difference between the association of *origin* with *pleasant* terms versus the association of *origin* with *unpleasant* terms. This distance is measured in standard deviation units. It can be interpreted as the strength of the observed effect, i.e. how "strong" the embedding bias is. A magnitude of $d < 0.5$ is commonly interpreted as a small effect, $0.5 \leq d < 0.8$ as a medium effect and values of $d \geq 0.8$ as a large effect (Sawilowsky, 2009).

3.2.3 WEAT for German Origin Bias

The WEAT as proposed by Caliskan et al. (2017) is affected by various limitations which I describe in §2.3. I want to ensure that it captures origin bias in German embeddings as accurately as possible. For this purpose, I created new WEAT name lists that address two main problems.

Addressed WEAT Problems First, the WEAT needs seeds that can concisely characterize the relevant bias attribute, which in my case is origin. In prior work, fairly small target and attribute lists have been used. Due to their size, they are more susceptible to distortion by infrequent or ambiguous terms. In this thesis, I expanded the attribute lists according to rigorous criteria to stabilise the WEAT measurements and make them more robust against artefacts in the data and manipulation with non-semantic information such as frequency. To further combat frequency issues (Ethayarajh et al. (2019), van Loon et al. (2022)), I made frequency in the training corpus a selection criterion for the names I included in my attribute sets.

Second, what I intend to analyse is the notion of origin bias, which is a multidimensional concept, i.e., there are more than just two manifestations of this attribute. One could define two attribute sets *German* and *Foreign*, as Kurpicz-Briki (2020) essentially does. However, this presents two issues. It frames Germans as having a special status and all other ethnic groups as being interchangeable, which is ethically questionable. It also oversimplifies the problem at hand: Different groups of minorities are discriminated against differently, and it is likely that word embeddings contain different biases for these groups of different national origin. In my thesis I therefore created multiple attribute groups with names of different national origin.

Name Selection Procedure To overcome these issues, I defined new attribute lists of first names following the selection and filtering process described below.

As a first step, I decided on the nationalities for which I want to evaluate bias. It is important for the debiasing and evaluation process that the selected names occur frequently enough in the training corpus. I expected that there would be more data available for larger minorities of the German population than for smaller minorities, so I considered the five biggest minorities in Germany, which are people from Türkiye, Poland, Syria, Romania and Italy (Bundeszentrale für politische Bildung, 2022). Out of those, I specifically focus on three in this thesis. To be able to clearly assign measured biases to nationalities, the names used for evaluation should be distinguishable between the different countries. Based on this, I selected Türkiye, Poland and Italy as countries since their cultures and languages differ more between each other than is the case for other combinations of the top nationalities. Additionally, these countries represent different historical waves of immigration to Germany[15], which might lead to different biases associated with these groups.

Sources for the name data I used differ for each country. I extracted names from sources listing the most common names in Germany, Türkiye, Poland and Italy respectively. Using most common names (instead of, e.g., a random selection) has the advantages of the names being representative for that country, the names having higher frequency in the training corpus, and simply the fact that such data is available at all. Ideally, since the training data is in German, it would be most appropriate to use the most common foreign names *in Germany* instead of in each respective country, but such data is not officially available in Germany and manually selecting foreign-sounding names would introduce subjective biases. For each country, I sourced an equal amount of female and male names to avoid unintended gender bias, and, wherever possible,

[15] See https://www.lpb-bw.de/anwerbeabkommen-tuerkei for Turkish immigration, https://www.bpb.de/themen/deutsche-einheit/migrantische-perspektiven/325312/migranten-aus-polen-im-wiedervereinigten-deutschland/ for Polish immigration and https://www.bpb.de/themen/deutschlandarchiv/259001/italienische-zuwanderung-nach-deutschland/ for Italian immigration.

used the most popular names across a wide time frame in order to prevent an age bias in the names.

Table 2 shows the precise sources used for each country, as well as comments on that source. "Popular names" sites hosted by private individuals like `beliebte-vornamen.de` or `behindthename.com` are useful resources because they aggregate data from various official sources like birth registries.

On `beliebte-vornamen.de`, the author Knud Bielefeld aggregates name statistics from sources such as registers of birth, newspapers, registers of university graduations, and more. It is the only such source available for Germany (Frank, 2013).

Mike Campbell operates `behindthename.com`, a website similar to `beliebte-vornamen.de` but including names from a variety of languages. The names are sourced from name dictionaries, personal collections, and official registries, among others.

"Wikipedia-Personensuche"[16] is a search engine for persons about whom an article exists in the German Wikipedia. It provides the functionality to sort by most common first names of a certain national origin, with an option to filter by gender. I used it on a supplementary basis to other sources because using resources other than Wikipedia provides an additional external judgement of the names' popularity outside of the embeddings' training corpus.

From these sources, I initially extracted the top 50 female and top 50 male names, as specified in Table 2. These names I then subjected to a filtering process, for which I defined criteria the names must meet. First, names should be unique to the country they are supposed to represent, especially in comparison to Germany. For example, Anna is a common Polish name, but it is also a common German name and thus should not be used. Second, the names should be frequent enough in the training data to produce meaningful embeddings and not be "automatically" associated with negative words (van Loon et al., 2022). Third, the names should be unambiguous with non-name tokens. For instance, the common Turkish name "Can", in a case-insensitive model, is ambiguous with the English word "can". Fourth, the names should be unambiguously assignable to one gender, i.e., not be gender-neutral.

In more concrete terms, I applied the following five step filtering process to ascertain that these criteria are met. The steps were applied consecutively, i.e., if a name was excluded after step one, it was not considered for step two or later steps.

[16] Person search: `https://persondata.toolforge.org/index.php`

Country	Sources	Comments
German	https://www.beliebte-vornamen.de/49519-erwachsene.htm	Most popular names in Germany for people born 1945-2000; only first names considered; phonetically identical names (e.g. Matthias/Mattias) treated as being the same; top 50 female and top 50 male names used
Türkiye	https://www.behindthename.com/top/extremes/turkey	Originally from official Turkish government statistics; top Turkish names used from 1980-2021; lists "most consistently popular", "top rises over two decades" and "top falls over two decades" used for age balance; 30 female/names in total
	https://www.beliebte-vornamen.de/1802-tuerkische.htm	Originally from Turkish government statistics for children born in 2022; top 5 female/male names
	https://persondata.toolforge.org/vorname/top/TUR	Top 15 female/male names used
Poland	https://www.behindthename.com/top/extremes/poland	Same methodology as for Turkish names; original name source not made entirely transparent (website's author thanks a site visitor for his contribution)
	https://persondata.toolforge.org/vorname/top/POL	Same methodology as for Turkish people; top 20 female/male names extracted
Italy	https://www.behindthename.com/top/extremes/italy	Same methodology as for Turkish and Polish; original source not entirely transparent
	https://www.beliebte-vornamen.de/562-italienische.htm	Top 10 female/male most commonly used names in Italy today; originally possibly from private blog post
	https://persondata.toolforge.org/vorname/top/ITA	Same methodology as for Turkish and Polish; top 10 female/male names used

Table 2: Online sources for name data used in my WEAT experiments per nationality, including commentary on the sources. behindthename.com and beliebtevornamen.de are privately operated name collection websites aggregating name data from official and unofficial sources alike. The Wikipedia person search is a search engine for German Wikipedia allowing to sort by popular names.

1. **Relevancy:** I checked whether a Wikipedia *name article* exists for the token in question. If not, the name was excluded. Wikipedia articles for the token which are not *name articles*, e.g., an article titled "Can" which describes beverage containers would, not fulfill this criterion, whereas an article titled "Can (Turkish first name)" would.

2. **Frequency:** I counted whether there were less than ten occurrences of the token in the Wikipedia corpus I use for training (see §3.1). If so, the name was excluded. This step is distinct from the fist since some names appear in the corpus but do not have their own Wikipedia article. The first step checks whether the term is relevant *as a name* whereas this step checks whether a meaningful embedding can be learned for this term.

3. **General ambiguity:** Wikipedia defines a *primary page* for tokens for which there exist multiple articles. For example, there is an article describing the first name "Alina", but also one for an opera, a film, an album and a ship all called "Alina". The primary page is the article which the user is directed to when searching for this token in the Wikipedia search bar without any further specifications. For "Alina", e.g., the primary page is the name page. The primary page can either be a proper article or a disambiguation page where the different possible articles are listed. Wikipedia editors determine a token's primary page according to various criteria such as article traffic statistics.[17] I discarded any tokens for which the primary page was an article other than the name page. In cases where the primary page was a disambiguation page, I qualitatively assessed whether it is likely that many occurrences of this token in the training corpus refer to a meaning other than the name. Often times, for example, the alternative articles cover ships, lesser known artistic works, or a variety of minor geographical objects, for which I deemed the ambiguity acceptable.

4. **Gender ambiguity:** I searched for the name in question using the Wikipedia persons search (https://persondata.toolforge.org/) introduced above. The tool allows the user to filter by gender. For each name, I retrieved the number of articles which treat a female versus a male person. If less than 95% of the articles were about a person of the gender which I intended to represent with the name in question, the name was excluded. For example, since only 61.32% of "Andrea" articles are about a woman, I did not include the name in the data set.

5. **National ambiguity:** Lastly, I again searched for the name in question using the Wikipedia persons search, similarly to step four. This time, I crawled the nationality metadata of the

[17] For more information on primary page criteria, see https://en.wikipedia.org/wiki/Wikipedia:Disambiguation#Is_there_a_primary_topic?.

> resulting articles and counted the occurrences of each country for the name in question. If any countries out of Turkish, Polish or Italian — except for the target country — were in the top five nationalities, I removed the name.[18] For example, for the name "Anna", most articles (35.53 % out of 3310) in the German Wikipedia were about a German Anna.[19] However, the fourth most common occurrence (5.46 %) was as a Russian name. Therefore, I excluded the name Anna from my experiments. I also removed names if under 50 % of the articles about bearers of that name were about a person of the target nationality. If that percentage was between 51 % and 75 %, I made my choice dependent on whether `behindthename.com` listed the target nationality as the most common usage in terms of nationality. If so, I kept the name, otherwise, I discarded it.

The experiments for which I used Wikipedia were conducted on the July 2023 version of Wikipedia. Due to technical restrictions, I did not distinguish between first, last, and middle names in my searches using the Wikipedia persons search. That tokens were lower-cased for all steps and represented accurately in their encoding, meaning that, e.g., Maria and María were not treated as the same name. The Wikipedia persons search also returns fictional characters, pen names and the like. These results were not filtered out.

After applying all inclusion and exclusion criteria according to this filtering algorithm, the smallest remaining name set were female Italian names with 14 remaining names. I found that this is due to female names generally being rarer than male names, and Italian names being less unique to Italy than German, Turkish or Polish names were to their countries. For better comparability, I reduced all other name sets to the size of 14 names per gender as well, cutting off those names with the lowest number of occurrences in the Wikipedia training corpus. This finally resulted in four name sets with 28 names each. The full list of names can be found in Table 3.

[18] If the target nationality of a name was one other than German and German was within the top five nationalities, I did not immediately exclude the name. Since I analysed name occurrences in the German Wikipedia, there were bound to be a certain number of German individuals for most names, especially since people might have dual citizenship. For example, only four out of the 200 gathered Turkish names had zero occurrences as a German name. I therefore only excluded names which had more bearers of German nationality than of the target nationality.

[19] An additional 7.51 % of articles were about Austrian Annas. I did not count Austrian or Swiss citizens as belonging to the same category as German citizens. I made this choice because there might be differences in the cultural perception of people from these three countries and combining them nationalities could therefore potentially skew results.

Nationality	Names
German	Katharina, Susanne, Karin, Ulrike, Renate, Birgit, Bettina, Jutta, Ute, Cornelia, Katja, Heike, Stefanie, Kerstin, Tanja, Hans, Carl, Wolfgang, Andreas, Werner, Christoph, Klaus, Philipp, Joachim, Jürgen, Dieter, Matthias, Manfred, Sebastian, Rainer
Turkish	Esra, Merve, Fatma, Sibel, Elif, Ayşe, Emine, Özlem, Zeynep, Hatice, Dilek, Ebru, Pınar, Hülya, Derya, Mustafa, Murat, Ahmet, Kemal, Orhan, Hüseyin, Bülent, Metin, Ömer, Emre, Halil, Erkan, Uğur, Burak, Volkan
Polish	Magdalena, Ewa, Zofia, Beata, Katarzyna, Krystyna, Małgorzata, Jadwiga, Danuta, Elżbieta, Urszula, Alicja, Aneta, Iwona, Edyta, Andrzej, Stanisław, Marek, Józef, Henryk, Krzysztof, Władysław, Tadeusz, Piotr, Janusz, Tomasz, Wojciech, Jakub, Marcin, Franciszek
Italian	Caterina, Francesca, Paola, Giulia, Chiara, Giovanna, Alessandra, Gioia, Antonella, Giuseppina, Azzurra, Antonietta, Ambra, Alessia, Giorgia, Giovanni, Carlo, Francesco, Giuseppe, Pietro, Luigi, Paolo, Alessandro, Angelo, Giorgio, Domenico, Enrico, Stefano, Vincenzo, Matteo

Table 3: The words, i.e. names, I use as attribute sets in my WEAT experiments. Each set contains an equal number of female (teal) and male (orange) names.

WEAT Experiment Setup With the new attribute sets defined above, I had devised three WEAT experiments, a German-Turkish, German-Polish, and German-Italian one. I ran the WEAT on the embeddings specified in §3.1 before debiasing, allowing me to analyse existing origin bias in German word embeddings and to validate my WEAT implementation against the scores published by Kurpicz-Briki (2020). In a second step, I compared the WEAT scores after debiasing with respect to the different nationalities, embeddings, and to the scores before debiasing.

Because of the four distinct attribute sets I used, I conducted multiple statistical tests at once, which leads to the*multiple comparisons problem*: Given a large enough number of samples, it becomes increasingly likely to eventually find statistically significant results due to chance. In my case, for one experiment with three tests with a significance level of 0.05, there would be a likelihood of up to 15% $(0.05 + 0.05 + 0.05)$ of a significant difference by chance. I utilised the Bonferroni correction (Bonferroni, 1936) to correct for this problem by dividing the significance level by the number of tests. Accordingly, the significance level in my WEAT experiments was $\alpha = \frac{0.05}{3} = 0.01\overline{6}$.

The target sets, i.e., the lists of pleasant and unpleasant terms I used in my experiments remained the same as in Kurpicz-Briki (2020). An exception is the word “scheußlich” which is replaced by “grausam” since “scheußlich” is missing from the vocabulary of the GloVe model.

3.3 Bias Mitigation Methods

For the purpose of mitigating bias, I selected two existing debiasing algorithms, Hard Debiasing (Bolukbasi et al., 2016) and DD-GloVe (An et al., 2022), and implemented, adapted and evaluated them for origin bias in German embeddings. In the following section, I present the reasoning for my selection, explain the methodology of the two methods and describe the process of adaptation.

For both algorithms it is the case that I attempted debiasing not for a general “German-Foreigner” bias-direction but instead for the specific nationalities Turkish, Polish, and Italian. This is analogous to the WEAT data sets I created for bias evaluation and has the aim of discovering the differences in bias between different countries of origin. Additionally, this decision facilitates the definition of seed words which are necessary for both algorithms. This is because analysing more specific bias categories results in a wider variety of available seed words since it is difficult to find seed words relating to the more general concept of origin.

3.3.1 Adapting Hard Debiasing for German Origin Bias

As introduced in §2.4, Hard Debiasing is a post-processing algorithm proposed by Bolukbasi et al. (2016) which aims to remove bias from pre-trained embeddings.

Rationale behind Choosing Hard Debiasing The work by Bolukbasi et al. (2016) constitutes a groundbreaking paper on bias in word embeddings. It is well-established, referenced many times in subsequent literature, and often used as a baseline for more sophisticated debiasing approaches. The linear subspace assumption Bolukbasi et al. (2016) make has been questioned by other researchers (Gonen and Goldberg, 2019) but can be mathematically proven (Vargas and Cotterell, 2020).

In this thesis, Hard Debiasing was implemented to serve as a reliable and easily transferable method. The execution of Hard Debiasing is resource-efficient since it is not necessary to train embeddings from scratch. Furthermore, since the algorithm is run post-hoc, it can be applied to already existing embeddings. It is therefore potentially a feasible and easily accessible way of improving the fairness of downstream NLP applications where pre-trained embeddings are already in use.

German embedding training methods do not differ from those for English embeddings and thus it is likely that the subspace assumption also holds for German origin bias. Because of this, I expected Hard Debiasing to be effective in reducing origin bias in German embeddings. No matter whether the debiasing would be successful or not, the outcome of this experiment could indicate a way forward for future research attempting to improve debiasing on German embeddings or debiasing for origin bias.

In the following, I present the three steps involved in the Hard Debiasing algorithm, first elaborating on the mechanics of each step and then presenting the adjustments I made.

Defining the Origin-Direction The first step Bolukbasi et al. (2016) perform is to define the bias-direction. As outlined in §2.3, they do this with a matrix of *defining word pairs* for which they compute the first principal component. This is supposed to extract embedding features which are characteristic for the difference between two concepts such as "national" versus "foreign". In particular, the authors utilise singular value decomposition to achieve this. Their definition is kept general to allow for a higher-dimensional bias subspace. However, in their work as well as most subsequent works including my thesis, the bias subspace is simply a two-dimensional direction, i.e., a vector of the same length as the normal word vectors. An accordingly slightly simplified bias-direction definition then becomes:

$$\text{SVD}\left(\sum_{i=1}^{n}\frac{1}{2}\sum_{w\in D}(\vec{w}-\mu_i)^T(\vec{w}-\mu_i)\right)$$

where D_i is a *defining pair* (e.g., `<Berlin, Istanbul>`) and μ_i is the mean embedding of D_i. This means that they center the two embeddings of each defining pair, construct the covariance matrix and then perform SVD. Part of an SVD calculation is factorising the covariance matrix into $U\Sigma V^T$. Bolukbasi et al. (2016) then use the first row of V^T (which has the dimensions 300×1 for 300-dimensional embeddings) as the bias-direction. The approach emphasises the features that differentiate each word pair, which in theory should be the bias components.

For the *defining sets* D which are needed for this step, I considered using first names as seen in the tutorial example mentioned above, e.g., one defining word pair could be `<Hans, Mustafa>`. However, this has two disadvantages. First, the debiasing algorithm performs a pair-wise centering of each word pair in the defining set as described above, but it is challenging to match pairs of first names together. For example, asking the question of "Which Turkish name is equivalent to the German name Sabine?" yields no fruitful answers. Frequency could be considered, but there are many more factors than just frequency influencing a name's embedding. Second, first names are already used in the WEAT evaluation. As described in §2.4, it has been criticised in literature (Gonen and Goldberg, 2019; Ethayarajh et al., 2019) that some debiasing algorithms only remove bias from words they have explicitly been told are biased. By *not* also using first names for defining the origin-direction, I can later use them in the WEAT to evaluate whether the capabilties of Hard Debiasing extend beyond just the words which it has explicitly been told are related to origin.

Instead of using first names, I manually created sets of defining pairs for German-Turkish, German-Polish and German-Italian debiasing. The pairs can be found in Table 4. Bolukbasi et al. (2016) define 10 pairs for their gender definition through a manual process they do not specify further. They then perform a human survey to confirm that their word pairs align with crowdworkers' idea of gender. I did not conduct such a survey, but the word pairs I selected are, for the most part, simply morphological variations of `<deutscher, türke>`, `<deutscher, pole>`, and `<deutscher, italiener>` and should therefore align well with the concept of "people of German origin versus people of Turkish/Polish/Italian origin". The issue, rather, lies in whether these defining sets might be too restrictive compared to Bolukbasi et al.'s gender seeds which contain a wider variety of words. I considered including word pairs such as `<berlin, instanbul>` but decided against it to keep the origin definition free of potential other biases that might be associated with such word pairs. More general word pairs were still included in the debiasing process, but in a later step in the form of *equalising word pairs*.

Removing Origin Component from Neutral Words After identifying the origin-direction, the next step in the Hard Debiasing algorithm is to "Neutralise". This happens by redefining a word's vector $\vec{w}$ to be $\vec{w}$'s projection onto the orthogonal origin-direction, thereby removing the component of $\vec{w}$ that lies in the origin-direction. Since I worked only with a bias-direction in the

Nationalities	**Word Pairs**
German ↔ Turkish	<deutscher, türke>, <deutsche, türkin>, <deutschen, türken>, <deutschen, türkinnen>, <deutschland, türkei>, <deutschlands, türkei>, <deutsch, türkisch>, <deutsches, türkisches>, <deutscher, türkischer>, <deutsche, türkische>, <germanisch, osmanisch>, <germane, osmane>, <deutschsprachig, türkischsprachig>, <deutschstämmig, türkischstämmig>
German ↔ Polish	<deutscher, pole>, <deutsche, polin>, <deutschland, polen>, <deutsch, polnisch>, <deutsches, polnisches>, <deutscher, polnischer>, <deutsche, polnische>,<deutschsprachig, polnischsprachig>, <deutschstämmig, polnischstämmig>
German ↔ Italian	<deutscher, italiener>, <deutsche, italienerin>, <deutschen, italienern>, <deutschen, italienerns>, <deutschen, italienerinnen>, <deutschland, italien>, <deutschlands, italiens>, <deutsch, italienisch>, <deutsches, italienisches>, <deutscher, italienischer>, <deutsche, italienische>, <deutschsprachig, italienischsprachig>, <deutschstämmig, italienischstämmig>

Table 4: Defining pairs chosen to calculate the origin-direction for Hard Debiasing on the German-Turkish, German-Polish, and German-Italian axis.

form of a vector and because all vectors are normalised, the projection of $\vec{w}$ onto the bias-direction b, called w_B can be expressed as $w_b = (\vec{w} \cdot b) \cdot b$, and the projection onto the orthogonal direction as $\vec{w} - w_b$. After neutralising, the embedding is again normalised. All together, this step can be denoted as

$$\vec{w} = \frac{\vec{w} - w_b}{||\vec{w} - w_b||}$$

Only words which do not actually relate to the origin concept should be debiased, e.g., "blau" should not contain any origin content, but "Migrant" should. I will call words which should not be related to origin *origin-neutral words*. Because this is the majority of words, it is easier to determine the complement of them, which I call *origin-specific words*.

Bolukbasi et al. (2016) derive these bias-specific words "using dictionary definitions", but do not elaborate on their methodology. Inspired by this idea, I derived origin-specific words by extracting entries from a dictionary that contain at least one of the defining words (specified above in Table 4).

The dictionary I use for this purpose is *Duden, Deutsches Universalwörterbuch* (abbreviated DDUW). The Duden, nowadays published by Cornelsen, was market-dominating among German dictionaries for a long time (Sauer, 1988) and is now still one of the most important reference books for the German language. I utilised the DDUW in its 2011 version, which is the seventh edition of the dictionary. It includes 172,663 entries spanning 124,944 unique tokens.

In the DDUW, I inspected for each entry if one of the defining words was present in that entry's definitional text. For this purpose, I capitalised the defining terms where appropriate, for instance, I checked for "Türkei" instead of "türkei", but "türkisch" remained as is. If any of the defining words were present, I added the current word to the list of origin-specific words. For example, the dictionary entry for "Bundeskanzlerin" contained the word "Deutschland", so it was included in the list of origin-specific words. I created three separate sets of origin-specific words: One for my German-Turkish experiments, one for German-Polish, and one for Italian-Polish. The German names were identical in all three cases.

The resulting word sets were 254 words large for German-Turkish, 270 words large for German-Polish, and 311 words large for German-Italian. To find the respective origin-neutral words, which is what is relevant for the "Neutralise" step, I then took the complement of the model's vocabulary with the origin-specific words.

Equalising Origin-Neutral Words The third step of Hard Debiasing is equalising origin-neutral words to be equidistant to a predefined set of *equality word pairs* $\mathcal{E}$, as mentioned in §2.4. Bolukbasi et al. (2016) again define this step for a more general case where the user would have multiple equality sets. Since I work with only one data set, I slightly simplify their equations.

For each equality pair (equality set in Bolukbasi et al.'s notation) $(e_1, e_2) \in \mathcal{E}$, its average $\mu = \dfrac{\vec{e_1} + \vec{e_2}}{2}$ is projected onto the orthogonal origin-direction: $\mu_B = \mu - \mu_B$.
The embeddings for e_1 and e_2 are then redefined as two components.

The first is the embedding's part which lies in the space orthogonal to the origin-direction. This part is simply equated to μ_B because the two words should be identical in their content except for their relationship to origin. For example, $\overrightarrow{Berlin}$ and $\overrightarrow{Istanbul}$ express the same concept of being the biggest cities of each country, except one relates to Germany and the other to Türkiye.

The second component of the equalised embedding is the part which lies in the bias-direction, called e_B. It is centered $(\vec{e_B} - \mu_B)$ to be symmetrically balanced across the origin-direction and then re-scaled to unit length, i.e., multiplied by $\frac{\sqrt{1-||\mu_B|^2}}{||\vec{e}_B-\mu_B||}$. This has the effect that the origin component of the two embeddings is now the same except for its direction.

All told, this results in the equation

$$\vec{e} = \mu_B + \sqrt{1 - ||\mu_B||^2} \cdot \frac{\vec{e_B} - \mu_B}{||\vec{e}_B - \mu_B||}$$

Bolukbasi et al. (2016) do not specify how they create their equality data set except that it is "crowdsourced". I utilised all word pairs I selected as defining word pairs (see Table 4). Additionally, I supplemented my equality sets with words from multiple nationality-specific categories, namely typical cuisine, currency, notable personalities, prominent landmarks, national dances, and biggest cities.

To collect words for each of these categories, I consulted Wikipedia's category listings in which for various categories such as "Italian cuisine" all articles which treat Italian cuisine are listed. I examined the Wikipedia entries for the seven categories listed above and then confirmed whether the items listed there have an entry in the Duden dictionary to confirm their relevancy and frequency of use in the German language.

These items then needed to be matched into pairs. I matched geographical entities to German equivalents by size; for example, I matched the (partly) Turkish river Euphrat to the (partly) German river Rhein. Famous persons I matched by occupation or position, for example "Merkel" is matched to "Erdoğan". Wherever possible, food was matched to other similar food, for example "Kebab" and "Stulle" were matched since both dishes are a type of filled bread. National dances were matched in the same way. As for currency, I equated "Euro" to "Złoty" for Poland and to "Lira" for Türkiye respectively, and omitted this equality pair for Italy since Germany and Italy use the same currency. The final list of equality sets can be found in Table 5.

Nationalities	Equalising Word Pairs
German ↔ Turkish	..., <berlin, istanbul>, <hamburg, ankara>, <münchen, izmir>, <köln, bursa>, <frankfurt, adana>, <stuttgart, gaziantep>, <düsseldorf, konya>, <leipzig, antalya>, <dortmund, kayseri>, <christlich, muslimisch>, <euro, lira>, <schnitzel, köfte>, <strudel, börek>, <sauermilch, kefir>, <flammkuchen, lahmacun>, <kohlrouladen, dolma>, <hackbällchen, köfte>, <brötchen, pide>, <stulle, kebab>, <walzer, hora>, <merkel, erdoğan>, <europäisch, asiatisch>, <rhein, euphrat>, <elbe, bosporus>, <alpen, ararat>, <ostseeküste, ägäis>, <bismarck, atatürk>
German ↔ Polish	..., <euro, złoty>, <berlin, warschau>, <hamburg, krakau>, <münchen, lodz>, <köln, breslau>, <frankfurt, posen>, <stuttgart, danzig>, <düsseldorf, stettin>, <leipzig, bromberg>, <dortmund, lublin>, <currywurst, bigos>, <grießsuppe, borschtsch>, <mohnkuchen, mazurek>, <maultausche, pirogge>, <walzer, mazurka>, <gardetanz, krakowiak>, <siebenschritt, polka>, <emsland, masuren>, <zugspitze, tatra>, <rügen, wollin>
German ↔ Italian	..., <berlin, rom>, <hamburg, mailand>, <münchen, neapel>, <köln, turin>, <frankfurt, palermo>, <stuttgart, genua>, <düsseldorf, bologna>, <leipzig, florenz>, <dortmund, bari>, <allgäu, toskana>, <ostseeküste, apulien>, <bratensoße, balsamico>, <jägermeister, amaretto>, <kloß, arancino>, <bratwurst, antipasto>, <stulle, focaccia>, <maultauschen, tortellini>

Table 5: Word pair sets used in the "Equalise" step of the Hard Debiasing algorithm. A separate set is used for the experiments concerning German-Turkish, German-Polish and German-Italian bias. Terms related to one of the nationalities are matched up to a similar term of the opposing country. Additionally to the terms shown here, all *defining word pairs* shown in Table 4 are also part of the *equalising word pair sets*. They are left out in this table for brevity, but indicated by an ellipsis.

After the creation of the origin-defining, origin-neutral and equality sets, Hard Debiasing is ready to be executed. To this end I utilised the code published by Bolukbasi et al. on GitHub[20] with no further changes.

[20] https://github.com/tolga-b/debiaswe

3.3.2 Adapting DD-GloVe for German Origin Bias

The second debiasing method I examined in this thesis was DD-GloVe as proposed by An et al. (2022). Different from Hard Debiasing, DD-GloVe is a train-time algorithm which adjusts GloVe's training objectives to produce bias-reduced embeddings. In the following, I am going to first explain the components involved in the DD-GloVe method and how I adapted them to origin bias in German embeddings, and then summarise my training setup including training data and hyperparameters.

Rationale behind Choosing DD-GloVe DD-GloVe is a novel approach to debiasing. It has not yet been reproduced or reviewed in subsequent works, but it provides a new perspective and promising results, especially with regards to mitigating indirect as well as direct bias.

It defines bias not only as the difference between certain defining word pairs, but additionally — and extensively — leverages dictionary definitions as an external source to define bias. Dictionary definitions are by no means perfectly bias-free. One problem they exhibit is that, since they are often written primarily by White authors and for White audiences (Murphy, 1991), they omit word senses used by non-White populations (Murphy, 1998). However, unlike other data sources such as Wikipedia, dictionaries are professionally proofread, deliberately worded neutrally, and deal with words detached from context, which means that the context of tokens has less potential to introduce bias. I therefore predicted the dictionary-debiasing approach to potentially be a valuable addition to debiasing research.

Another intriguing feature of DD-GloVe was that unlike for almost all other current debiasing methods, no extensive seed lists to define the bias concept are necessary. Only two initial seed words need to be given and the algorithm then automatically expands this to a longer list. This could prove to be useful for a difficult to define concept such as origin.

Additional potential strengths of DD-GloVe were that according to An et al. (2022), the model is not only successful on various bias metrics but even improves the embeddings' performance on semantic evaluation tasks due to the additional information learned from dictionary definitions. Last but not least, the authors make their vectors and code publicly available which makes reproduction decidedly easier.

In the next subsections, I explain the DD-GloVe model including its dictionary-guided loss functions, the data used and created by me, and my training setup.

Creating Dictionary Embeddings Fundamental to the DD-GloVe algorithm is the choice of dictionary. An et al. (2022) presumably used definitions from multiple Oxford dictionaries, in which case they had approximately 350,000 dictionary entries at their disposal. I am going to elaborate on the uncertainty of this assumption in §4.3.

For the German dictionary I considered multiple options.

With approximately 1,099,000 entries, the German version of Wikipedia is the largest free online resource available. However, raw Wiktionary data as published in Wiktionary dumps is structurally complex, and currently no parser for these files is publicly available. Furthermore, Wiktionary is a crowd-sourced resource and I hypothesised that it might be more susceptible to unconscious biases than a dictionary that was professionally proofread and edited by established publishing houses.

I therefore again utilised the Duden Universalwörterbuch, which I introduced in §3.3.1. To re-iterate, the edition of DDUW I utilise comprises 172,663 entries, which is at least 50.62% smaller than the dictionary used by An et al. (2022). I retrieved definitions for 59,721 tokens, out of which 17,957 tokens had multiple entries. It presents no immediate problem if no entry is found for a token since the DD-GloVe loss functions which require a dictionary entry are simply skipped in such cases.

An et al. (2022) calculate a word w's definition embedding $d(w)$ by taking the mean of the embeddings for all words in w's dictionary definition:

$$d(w) = \frac{1}{|D_w|} \sum_{w_i \in D} \vec{w_i}$$

where D_w is a list of all definition tokens of w and $w_i \in D_w$ is one token in this list. It is possible for tokens to occur multiple times in D. An et al. (2022) motivate this approach by virtue of it being "[simple] but empirically effective".

The authors emphasise that these definition embeddings are "trained from scratch" at the same time as all other embeddings in the model. This is necessarily the case because the definition embeddings are defined as being composed of regular word embeddings, and, as An et al. (2022)

state, using pre-trained embeddings to compose the definition embeddings would introduce bias contained in those pre-trained word embeddings to the dictionary embeddings. By training them at the same time, DD-GloVe aims to create a synergy where the definition embeddings start out bias-free (due to random initialisation) and remain as neutral as possible by continually influencing the word embeddings it is comprised of to orient themselves towards the definition embeddings.

Approximating the Origin-Direction Like Hard Debiasing, DD-GloVe too uses a bias-direction to determine how to realign word vectors. Unlike Hard Debiasing, in DD-GloVe the bias-direction is used only for one part of a multi-step debiasing process, as expanded upon in the next subsection. Nevertheless, the accuracy of approximating the bias-direction has a large effect on the overall success of the model, which I explore in later parts of this thesis (§4.3, §5.5).

For two sets of seed words A_1 and A_2 (named A for attribute), An et al. (2022) define the bias-direction vector b (in their notation g for gender) as

$$b = \frac{1}{|A_1|} \sum_{w \in A_1} \vec{w} - \frac{1}{|A_2|} \sum_{w \in A_2} \vec{w}$$

which is simply the difference vector between the means of the two seed word sets.

Part of DD-GloVe's functionality is an algorithm for automatically defining these seed word sets A based on two user-provided initial seed words s_1 and s_2. The algorithm first calculates the difference between the definition embeddings of s_1 and s_2, denoted as $b_{\text{initial}} = d(s_1) - d(s_2)$. Then, for all words w in the vocabulary their definition embedding $d(w)$ is projected onto b_{initial}, yielding a bias value $b(w)$. The words are sorted by their assigned bias value $b(w)$ and the words with the top n highest and lowest scores are added to A_1 or A_2 respectively. Both very high and very low (i.e., negative) resulting values of the projection onto the initial bias definition axis, will, in theory, indicate that these words are strongly associated with the bias concept.

The size of the sets n is determined empirically. An et al. (2022) recommend a smaller n for "attributes that have a smaller number of associated words, such as race", but do not specify an exact size. Guided by their choice of $n = 30$ for gender bias, I selected $n = 10$, in alignment with the authors' recommendation to choose a smaller n for concepts such as race.

Furthermore, due to the definition embeddings being learned along with all other embeddings, and the results of the seed word generation algorithm thus potentially changing with each iteration, the algorithm can be run multiple times throughout training. For racial debiasing, the authors run it in each of the first five iterations and then again every ten iterations, which I adopted for my experiments.

As with Hard Debiasing, I treated the origin setting as three separate debiasing scenarios and ran DD-GloVe with initial seed words for German–Turkish, German–Polish, and German–Italian bias. My initial seeds were `<deutscher, türke>`, `<deutscher, pole>`, and `<deutscher, italiener>`.

DD-GloVe Loss Functions Explained In order to gain insights into the results of debiasing with DD-GloVe, it is helpful to understand the loss functions involved in its training.

The overall training paradigm of the original GloVe model (Pennington et al., 2014) consists of minimising the difference between logarithmic word co-occurrence and the respective embeddings' similarity defined, with the optimisation typically being implemented as stochastic gradient descent.

An et al. (2022) re-defined the loss functions used in this optimisation process. They constructed three entirely new loss terms J_{ortho}, J_{proj} and J_{def}, and additionally altered the original GloVe loss function to be bias-aware (J_{G-bias}). They combined these four loss functions and defined DD-GloVe as

$$J = J_{G-bias} + \beta J_{ortho} + \gamma J_{proj} + \delta J_{def}$$

where the hyperparameters β, γ and δ control the influence of each loss term. These hyperparameters the authors typically set to be fairly small — between 0.0001 and 0.2 — leaving J_{G-bias} as the primary influence on the training process.

Bias-aware GloVe loss J_{G-bias} also constitutes the most complex out of the four loss functions. Its basic idea is to assign weights to individual co-occurrences in the co-occurrence matrix with the goal of balancing out unequal distributions in the training data. This way, the authors intend to achieve an effect similar to training on a corpus that has been balanced with data substitution methods (see §2.4).

The weight assigned to a co-occurrence pair depends on its bias. The bias $o(w)$[21] of a word w is calculated as the difference between w's projection $p(w)$ onto the difference vector $\vec{s_1} - \vec{s_2}$ and the projection $(d(w))$ of w's definition embedding onto $\vec{s_1} - \vec{s_2}$. Here, $\vec{s_1} - \vec{s_2}$ is a simple expression of the origin-direction defined by the initial seed pair (s_1, s_2). The bias definition $o(w)$ can be interpreted as the difference in origin content p in a word's embedding and its definition embedding, where a larger difference indicates bias since the definition embedding provides a neutral reference point. I formalise this as

$$o(w) = p(w) - p(d(w))$$

This bias o and origin association p is then calculated for each word pair $(w, \tilde{w})$ in the co-occurrence matrix and a weight $\omega(w, \tilde{w})$ is assigned based on the resulting values.

If the two co-occurring words are associated with opposite origin concepts, as would for example be the case for "Istanbul" and "German", a positive weight is assigned because the goal is to create a more balanced co-occurrence matrix. Conversely, for word pairs associated with the same origin-direction, a negative weight is assigned. This *direction* can be expressed with the signs of $p(w)$ and $p(w)$ as $1 - \text{sgn}(p(w)) \cdot \text{sgn}(p(\tilde{w}))$. The *magnitude* of the weight is set to whichever is the higher value out of $o(w)$ and $o(\tilde{w})$, i.e., $\max(o(w), o(\tilde{w}))$, meaning that the effect is weaker for bias-neutral words and stronger for bias-related words. Additionally, a constant α scales the weight function. This all comes together to

$$\omega(w, \tilde{w}) = 1 - \alpha \cdot \text{sgn}(p(w)) \cdot \text{sgn}(p(\tilde{w})) \cdot \max(o(w), o(\tilde{w}))$$

An et al. (2022) recommended α to be set to 0.4, which keeps $\omega(w, \tilde{w})$ in a range of about 0.9–1.1. I adopted this value for my own experiments.

[21] Named o for origin.

Unlike the other new loss terms An et al. (2022) define, bias-aware GloVe loss is not added to the overall GloVe loss as an additional summand, but integrated directly into the existing main GloVe function. There it can directly influence the original GloVe co-occurrence weight calculations. This results in a final "bias-aware GloVe loss" J_{G-bias} of

$$J_{G-bias} = \sum_{i,j=1}^{V} \omega(w_i, \tilde{w_j}) f(X_{ij}) \left(w_i^T j_j + b_i + \tilde{b_j} - \log(X_{ij})\right)^2$$

which is unchanged from the original GloVe loss except for minor notation details and the integration of ω.

Projection Loss While the bias-aware GloVe loss aims to achieve results reminiscent of Maudslay et al. (2019) or Lu et al. (2020), the projection loss shows similarities to the Hard Debiasing method. Instead of adjusting the projection of all neutral words onto the bias-direction to be zero, though, An et al. (2022) encouraged a word's projection onto the bias-direction to be similar to the projection of its definition embedding onto the bias-direction. They expressed this as

$$J_{proj}(w) = \left\| \frac{\vec{w} \cdot o}{o \cdot o} o - \frac{d(w) \cdot o}{o \cdot o} o \right\|$$

which captures the difference between the vector projections of word vector $\vec{w}$ onto the origin-direction o and w's definition embedding $d(w)$ onto the origin-direction o. The Euclidean norm of the difference vector then constitutes the projection loss.

For words without a dictionary definition, the authors set $d(w) \cdot o$ to be zero, assuming that the word does not contain bias content in such a case. Different than in Hard Debiasing, this function is applied to all words in the vocabulary and not just neutral or biased words because the authors hypothesised that the dictionary definition's embedding alone is able to indicate the whether a word is related to the bias concept or not.

Orthogonal Loss On top of defining a function for mitigating a specific bias such as origin bias, DD-GloVe also includes a loss function intended for debiasing general biases. The "orthogonality" refers to the component of a word embedding $\vec{w}$ which lies in the subspace orthogonal to $\vec{w}$'s definition embedding $d(w)$ and which is encouraged to be dropped by the model. Mathematically, the authors define this as

$$\phi(w, (d(w)) = \vec{w} - \frac{\vec{w} \cdot d(w)}{d(w) \cdot d(w)} d(w)$$

which is the vector projection of $\vec{w}$ onto $d(w)$ removed from $\vec{w}$. This function $\phi(w, (d(w))$ is expected to express undesired and likely biased information in $\vec{w}$ because An et al. (2022) hypothesise that any information not contained in dictionary embeddings is unnecessary information. The final orthogonal loss function formulated is

$$J_{ortho}(w) = (\phi(w, (d(w)) \cdot \vec{w})^2$$

which minimises the squared dot product between $\vec{w}$ and $\phi(w, d(w))$ to encourage dissimilarity between the two vectors. The penalty term would be higher when there is a larger projection of $\vec{w}$ onto the orthogonal subspace, thus promoting the model to reduce the impact of information not aligned with $d(w)$.

If a word does not have a dictionary definition assigned, this loss term is ignored.

Definition Loss Lastly, An et al. (2022) declare a fourth loss function which also leverages dictionary definitions, but targets improvement in the embeddings' semantic performance instead of bias mitigation. Where *orthogonal loss* encourages the model to remove embedding components orthogonal to their respective dictionary definitions, *definition loss* promotes the model to generally increase similarity between an embedding and its definition embedding. This is defined simply as the L_1 norm between the difference of a word embedding $\vec{w}$ and its definition embedding $d(w)$:

$$J_{def}(w) = ||\vec{w} - d(w)||_1$$

This difference is lower when the word embedding in question and its definition embedding are more similar.

Training Setup For training my own DD-GloVe models, I set the parameters which weight the different loss function to be $\beta = 0.0001, \gamma = 0.05$, and $\delta = 0.001$. These values were taken from An et al.'s experiments concerning racial bias and *not* tuned specifically for my data. I assumed that even without fine-tuning, it should become clear whether the method is reproducible and generally viable for German origin bias or not.

All other general GloVe settings such as number of features or vocabulary size are identical to the specifications I made in §3.1 regarding the regular (i.e., non-debiased) GloVe embeddings I trained. A description of the Wikipedia training data I use can be found at the same place. I elaborated on my dictionary source in §3.3.2.

In their basic version, my experiments regarding DD-GloVe consisted of training regular GloVe embeddings as well as "debiased" embeddings using DD-GloVe and comparing their performance on semantic and bias evaluation tasks as specified in this chapter. As will become evident in the next chapter, more extensive insights into the performance of DD-GloVe became necessary to understand the results of these experiments. For this reason, some additional experimental setups are going to be presented in §4.3.

4 Results

In this chapter, I present the results of my experiments as defined in chapter 3. I organise my findings by analysis method. First I show results of my semantic evaluation (explained in §3.2.1) on the original embeddings (determined in §3.1) as well on the embeddings after debiasing (see §3.3). Then I present an analysis regarding bias before and after debiasing using the WEAT (see §3.2.3). Finally, in order to better understand the results of my attempts at debiasing with DD-GloVe, I dedicate a section of this chapter to carrying out and reporting on additional post-hoc experiments regarding this model.

4.1 Semantic Evaluation with Word Pair Similarity

The first part of my evaluation concerns the semantic performance of the two different embeddings I examined.

Validation Experiments

Before analysing the effects of debiasing on the semantic performance of the embeddings, I first confirmed that the performance of the unchanged embeddings is in line with values reported in literature. To do so, I report word pair similarity scores using the original GUR350 data set (Gurevych, 2006) introduced in §3.2.1. Out-of-vocabulary pairs were here assumed to have a similarity of zero.

Embedding	**Measured** ρ	**Previously Reported** ρ
Pre-trained GloVe	0.45	0.49 (Forthmann and Doebler, 2022)
Self-trained GloVe	0.44	0.49 (Forthmann and Doebler, 2022)
Pre-trained fastText	0.72	0.70 (Bojanowski et al., 2017)

Table 6: Spearman's rank correlation coefficients ρ for word pair similarity on the original GUR350 data (Gurevych, 2006) set measured using different word embeddings. Details on the used embeddings can be found in §3.1. The table compares the measured ρ values with previously reported values from the literature.

Table 6 shows the results for the original word embeddings in comparison to values reported in literature. It can be seen that the values were generally within the same range as previously published results. This indicates that my implementation of the word pair similarity measurement, the GUR350 data set, and the embeddings were all fundamentally working as intended.

The 0.02 points difference for the fastText embeddings might be explained by the fastText embeddings published on `https://fasttext.cc` being more recent than those analysed in Bojanowski et al. (2017) and therefore possibly being trained on a larger Wikipedia dump.

The values reported by Forthmann and Doebler (2022) are, to the best of my knowledge, the only published values for a word pair similarity task on German GloVe embeddings. However, the authors trained their GloVe model on the deWaC corpus (Baroni et al., 2009), a collection of German web text, instead of Wikipedia and the word pair similarity data set they use appears to be a translated version of WordSim-353 (Finkelstein et al., 2001). The comparison of the GloVe values should therefore be taken with a grain of salt and can only serve as an approximate guideline.

Notably, the performance of both of the German GloVe embeddings I examined was significantly lower than what is usually reported for English GloVe models. Pennington et al. (2014), for example, report a coefficient ρ of 0.66 on the WordSim-353 data set. However, since the pre-training embeddings by Deepset (see §3.1), the embeddings trained by Forthmann and Doebler (2022) and the embeddings I trained all perform in this much lower range, I assume that this poor performance is not a flaw in my methods but due to the model itself.

Modifying GUR350

The GUR350 word pair similarity set created by Gurevych (2006) contains 49 words that were outside the vocabulary of the GloVe model I trained as described in §3.1. The 49 tokens are listed in Table 7. As explained in §3.2.1, for the purpose of analysing the difference in performance before and after debiasing, I excluded these word pairs from the data set instead of assuming their similarity to be zero. Excluding all word pairs in which one of the words is out-of-vocabulary leaves one with 283 remaining word pairs.

A possible concern with removing a significant part of the data set might be that the task becomes easier or harder for the model because, for example, harder words might be more likely to be

stellenanzeige	gepäckkontrolle	makake	kopfairbag
weißblau	dieselversion	drehfreudig	gehaltsunterschied
flaschenöffner	frühlingssonne	frustrieren	leidensgenosse
geschirrdurcheinander	urwaldhaus	herausstreichen	hirnsignal
kaffeetasse	tv-kamera	hinaufklettern	krebserkennung
premium-hersteller	niederschmetternd	reiseschutzpass	plätschernd
prozentzeichen	sandwich-konzept	betrugshandlung	küchenschrank
frontalkollision	suchmaschinenbetreiber	suchstrategie	topmanagement
arbeitssuchender	entwicklungschef	gepäcknetz	inaugurationsmesse
portokosten	beziehungsarbeit	heimgang	lebensbedürnis
quartalsumfrage	flachlegen	sports-tourer	volierenzelt
wegrennen	berlin-kreuzberg	rot-weiß	a-säule

Table 7: Words in the GUR-350 word pair similarity data set (Gurevych, 2006) which are out of vocabulary for the GloVe model I train on German Wikipedia text.

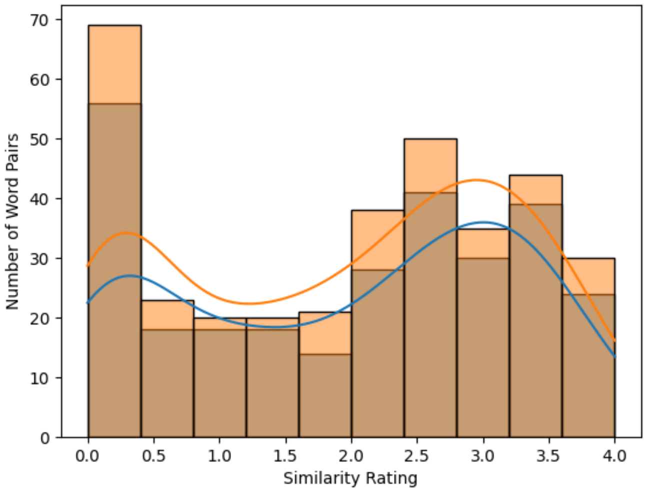

Figure 4: The number of word pairs in the GUR-350 word pair similarity data (Gurevych, 2006) set with a certain similarity ranking. Word pairs can have rankings in 0.5 step increments ranging $[0.0 - 4.0]$. Light orange bars show the number in the original data set and dark orange bars after removing the words specified in Table 7. The orange (original) versus blue (after removing) lines are Kernel Density Estimate plots showing that the distribution remains approximately the same.

		fastText	GloVe (self-trained)
original		0.7011	0.5697
after HD	German ↔ Turkish	0.7073 (Δ0.0062)	0.5806 (Δ0.0109)
	German ↔ Polish	0.7027 (Δ0.0016)	0.5805 (Δ0.0108)
	German ↔ Italian	0.7069 (Δ0.0058)	0.5803 (Δ0.0106)
after DD	German ↔ Turkish	n/a	0.5807 (Δ0.0110)
	German ↔ Polish	n/a	0.5806 (Δ0.0109)
	German ↔ Italian	n/a	0.5802 (Δ0.0105)

Table 8: Results for the word pair similarity task before and after debiasing. Each cell shows the Spearman rank correlation coefficient ρ calculated for the GUR283 data set. Columns indicate the embedding technique and rows indicate whether the original embedding or the embedding after debiasing was used. For the debiasing rows, the debiased nationality is additionally indicated. HD refers to Hard Debiasing (Bolukbasi et al., 2016) and DD to Dictionary Debiasing (An et al., 2022). Higher values are better. The values in brackets indicate the change after debiasing compared to the original embedding.

outside the model's vocabulary and therefore more likely to be removed. To analyse this, I measured the distribution of the word pairs' similarity ratings in the original GUR350 data set and in the reduced version. Figure 4 shows that removing these word pairs did not change the distribution of the data, suggesting that decreasing the data set size in this matter does not make the task easier or harder. It does still decrease the stability of the results since the data set is smaller.

I used the reduced version of GUR350 in the experiments concerning the performance difference before versus after debiasing and refer to it as GUR283. Table 8 shows the effect debiasing had on the word pair similarity task.

Effect of Debiasing

The embeddings' performance did not decrease after debiasing compared to the original value in any of the cases. This indicates that no semantic features other than those related to origin bias were removed during the debiasing process. In fact, all scores were marginally better after debiasing. The trend towards improvement might be explained when one considers the biased information contained in embeddings to be redundant information which sometimes obscures more meaningful components. This would be in line with the GloVe scores showing a greater improvement on the word pair similarity task than the fastText scores, since the WEAT analysis shows that the GloVe embeddings contain greater origin bias than the fastText embeddings (see §4.2). Overall, the debiasing algorithms did not have a strong effect on the embeddings' performance.

The scores for the original embeddings, displayed in Table 8, can also be compared to the scores the embeddings achieved on the full GUR350 data set, as reported in Table 6. It can be seen that the score for the self-trained GloVe embeddings was 0.1297 points higher after removing out-of-vocabulary words. This was to be expected since the model is no longer punished for out-of-vocabulary words with a default similarity value of zero, as explained in §3.2.1. Since fastText embeddings are capable of handling out-of-vocabulary words anyway, no improvement can be noted for them. Instead, they scored 0.0189 points lower, perhaps since the removed word pairs happened to be ones which the model had previously judged well.

4.2 Bias Analysis with the WEAT

After assessing the basic semantic performance of the embeddings used in my experiments, I now turn towards the main bias analysis of this thesis and present the results of my WEAT experiments.

Bias Measured in Original Embeddings

As explained in §3.2.2, I created three WEAT data sets: One with German and Turkish names, one with German and Polish names, and one with German and Italian names (see Table 3). Table 9 reports the origin bias measured on the original (i.e. non-debiased) embeddings using these different WEAT attribute lists. The mean μ of the Turkish, Polish and Italian WEAT scores is included to approximate an understanding of the origin bias contained in German word embeddings across different nationalities.

Validation Experiments I first compared my WEAT implementation to the results published in Kurpicz-Briki (2020). In particular, I compared Kurpicz-Briki's WEAT results for her translated WEAT 5 experiment, i.e., her measurement of origin bias using general "Swiss names" versus "Foreign names" to a reproduction of the same experiment. In this particular case, I did not use my adapted WEAT name lists, but the ones given in Kurpicz-Briki (2020). The author did not measure bias on GloVe embeddings, so only values for fastText are shown in the respective

Test Data	fastText		GloVe (self-trained)	
	d	p	d	p
Kurpicz-Briki (2020) WEAT 5	1.1340	$< 10^{-3} *$	n/a	n/a
Kurpicz-Briki (2020) Reproduction	0.9800	$0.0042*$	1.7889	$< 10^{-4} *$
German ↔ Turkish	1.3571	$0.0109*$	1.8321	$< 10^{-4} *$
German ↔ Polish	0.2829	0.5185	1.7301	$< 10^{-4} *$
German ↔ Italian	1.0331	0.1082	1.4650	$0.0040*$
μ {Turkish, Polish, Italian}	0.8910	0.2125	1.6757	$0.0013*$

Table 9: WEAT results for German fastText and GloVe embeddings before debiasing. The first column indicates which WEAT data set was used, with the last row showing an average of the previous three. Absolute effect size (Cohen's d) and p-value of the WEAT permutation test are reported. Smaller values of d and higher values of p mean less origin bias. Statistically significant bias ($p < 0.01\bar{6}$) is marked with an asterisk.

row in Table 9. The fastText version used by Kurpicz-Briki (2020) is different to the fastText embeddings I used. Kurpicz-Briki (2020) used embeddings trained on CommonCrawl whereas I used a version trained only on Wikipedia. This difference explains why my measured values were slightly different than those reported by Kurpicz-Briki (2020). In both cases, however, a significant bias was found.

I extended Kurpicz-Briki's results by also measuring origin bias in GloVe using the same attribute lists, i.e., the translation by Kurpicz-Briki (2020). While both fastText and GloVe embeddings showed significant origin bias in this experiment, the results for the GloVe embeddings were much more pronounced. The effect size for GloVe is almost twice as large and the p-value is close to zero compared to 0.0042 for fastText.

Effect of Adapted WEAT Seeds Looking at the rest of Table 9, I analysed the effect that my changed attribute lists, which differentiate between different nationalities, had on the WEAT results. A continuing stark difference between fastText and GloVe embeddings can be observed. Considering the mean μ of the three nationalities, it is apparent that while the values for GloVe were similar to those in the Kurpicz-Briki (2020) reproduction, overall no significant bias could be measured for the fastText embeddings. In fact, the fastText measurements were considerably far apart from the significance level $1.\bar{6}\%$.

As concerns the different nationalities, it can be seen that in the case of the fastText embeddings, significance bias was measuring using the German-Turkish name lists but not with the German-Polish or German-Italian ones. The GloVe results mirror this in a less extreme fashion. Here,

Embeddings	Turkish		Polish		Italian	
	$d \downarrow$	$p \uparrow$	$d \downarrow$	$p \uparrow$	$d \downarrow$	$p \uparrow$
fastText original	1.3571	0.0109*	0.2829	0.5185	1.0331	0.1082
fastText HD Turkish	1.1332	0.0573	0.2123	0.5639	0.9480	0.1469
fastText HD Polish	1.3612	0.0103*	0.1786	0.5699	1.0477	0.1018
fastText HD Italian	1.3299	0.0139*	0.1901	0.5747	0.5896	0.3510
GloVe original	1.8321	$< 10^{-4}*$	1.7301	$< 10^{-4}*$	1.4650	0.0004*
GloVe HD Turkish	1.5220	0.0010*	1.5315	0.0003*	1.1688	0.0115*
GloVe HD Polish	1.7833	$< 10^{-4}*$	1.3190	0.0032*	1.2579	0.0046*
GloVe HD Italian	1.7943	$< 10^{-4}*$	1.6358	$< 10^{-4}*$	0.8406	0.0538
GloVe DD Turkish	1.8209	$< 10^{-4}*$	1.6582	$< 10^{-4}*$	1.4522	$< 10^{-4}*$
GloVe DD Polish	1.7717	$< 10^{-4}*$	1.5567	$< 10^{-4}*$	1.3303	$< 10^{-4}*$
GloVe DD Italian	1.7891	$< 10^{-4}*$	1.7005	$< 10^{-4}*$	1.3057	0.0001*

Table 10: Cohen's d and p-value for WEAT permutation test before and after running debiasing algorithms. Hard Debiasing (HD) and Dictionary Debiasing (DD) results are shown for German-Turkish, German-Polish, and German-Italian debiasing attempts, with WEAT data sets for German-Turkish, German-Polish, and German-Italian names. Light green fields highlight matching embeddings and WEAT tests in regards to nationality. Statistically significant bias ($p < 0.01\bar{6}$) is marked with an asterisk.

significant origin bias could be shown for all three nationalities, but the measured bias effect size decreases slightly from Turkish over Polish to Italian. The difference between fastText and GloVe embeddings was less pronounced for the Turkish WEAT compared to the other two tests, in which the p-values were very high for fastText but very low for GloVe. The Polish measurement stands out among the fastText results because of its unusually high p-value of 0.5185 and its small effect size. For the other two nationalities, the fastText embeddings still produced a large effect size despite their high p-values.

Bias Measured After Debiasing Attempts

The second part of my WEAT experiments compared the bias measured before versus after debiasing, i.e., the success of the two debiasing methods. Table 10 shows these results.

Both Hard Debiasing and Dictionary Debiasing were carried out for specific nationalities and not general origin bias, as explained in §3.3. The various embeddings were all measured on all three WEAT data sets — Turkish, Polish, and Italian — even though this means that for some embeddings, the debiasing direction (e.g. German-Turkish) does not match up with the bias measured in the WEAT (e.g. German-Italian). These values were still included in the results in order to analyse whether the debiasing algorithms were perhaps able to mitigate even more than just the narrow sense of origin they were instructed on. For better comprehensibility, Table 10 highlights in light green those WEAT experiments and debiased embeddings where the treated nationality is the same.

Hard Debiasing with fastText Embeddings In the fastText embeddings, a significant bias was only measured for Turkish names, as mentioned above. In the German-Turkish WEAT experiments on the fastText embeddings after Hard Debiasing, the effect size was reduced by a magnitude of 0.2239 and the p-value increased by 0.0464 points. This resulted in a p-value comfortably above the significance level of $0.01\bar{6}$, meaning that no significant bias could be measured after debiasing.

Since no statistically significant bias for fastText embeddings was measured in the Polish or Italian WEAT experiments before debiasing, I ran the Hard Debiasing algorithm for these cases mainly in order to provide complete results. Interestingly, it can still be noted that effect sizes further decreased and p-values further increased after debiasing.

Hard Debiasing with GloVe Embeddings Analysing the WEAT scores for the GloVe embeddings, which initially all showed statistically significant bias, it can be observed that Hard Debiasing was only partly able to mitigate this bias. For the Turkish and Polish case, the algorithm reduced the effect sizes by 0.3101 and 0.4111 points respectively. The p-value was approximately 0.001 and 0.0032 points higher after debiasing. Both of those values are still below the significance level of $0.01\bar{6}$.

For Italian origin bias, however, the algorithm significantly reduced the bias measured by the WEAT. The effect size was reduced by a magnitude of 0.6244 and the p-value was increased from 0.004 to 0.0538, which exceeds the specified significance level. Still, however, this debiased version of the GloVe embeddings performs worse in the WEAT than the original version of the fastText embeddings, which has a p-value approximately twice as high. Notably, the Italian WEAT experiment was also the one with the lowest origin bias measured in the original GloVe embeddings in comparison to the Turkish or Polish test.

Hard Debiasing Across Nationalities In terms of cross-national Hard Debiasing, no success could be reported. Debiasing with German-Polish or German-Italian seed words did not have a consistent effect on German-Turkish bias. The effect size and p-value variably either slightly increased or decreased, and no significant reduction in bias was achieved. None the GloVe biases could be significantly mitigated by debiasing for a different nationality than the one measured for by the WEAT. However, debiasing with Turkish seed words slightly improved Polish and Italian bias, the Italian bias even considerably more so than Turkish bias.

Dictionary Debiasing Finally turning to the results of my Dictionary Debiasing approach, it becomes apparent that while Hard Debiasing might not have fully satisfied my goal of removing origin bias, Dictionary Debiasing fully failed in doing so. For all WEAT runs and all DD-GloVe embeddings, the p-values are 0.0001 or lower and the lowest effect size is still very large at a magnitude of 1.3057. The debiasing training process evidently had almost no noticeable effect on the WEAT's p-values. The effect sizes marginally decreased for the nationality for which the embeddings were trained, but only by magnitudes of 0.0112 to 0.1734 points. While An et al. (2022) claim that DD-GloVe has the capability to remove "general biases", the model in some cases even increased the bias measured by the WEAT, such as when training the German-Turkish model and then evaluating for German-Italian bias. I investigate this behaviour more closely in the next section.

4.3 Additional Attempts at Reproducing DD-GloVe

In response to the unexpectedly ineffective results of the DD-GloVe algorithm, I carried out further experiments with the aim of determining the error source. Since DD-GloVe is a novel approach, an additional goal is to critically review the work by An et al. (2022), which unlike for Bolukbasi et al. (2016) has not been done so far.

Testing Assumption of Bias-Free Dictionaries

Since the semantic evaluation of the DD-GloVe embeddings did not produce considerably worse results than for the pre-trained GloVe embeddings by Deepset, I assume that the basic GloVe training is not an issue. Instead, it is likely that the model's loss functions do not operate as

Word Set	**Contents of Word Set**
"German" Occupations	Brenner, Verkoster, Anlagenführer, Orientalist, Altphilologe, Fotograf, Kameramann, Richter, Verfassungsrichter, Tierarzt, Führungskraft, Aufseher, Projektleiter, Politiker, Schornsteinfeger, Bestatter, Sozialarbeiter, Orthopäde, Lobbyist, Stenotypist
"Foreigner" Occupations	Tänzer, Koch, Dachdecker, Trockenbaumonteur, Bauarbeiter, Fleischer, Bäcker, Kellner, Betonbauer, Sänger, Fischwirt, Musiker, Fliesenleger, Barkeeper, Metallograf, Bodenleger, Gebäudereinigung, Putzkraft, Reinigungskraft, Stuckateur
Origin-Characterising Words	`ausland, ausländ, migrant, exot, pol, italien, türk, fremd, einwander, zuwander, gast, deutsch, inland, inländ, german, bundesrepublik, tradition, heimat, ddr, brd, osteurop, österreich, schweiz, asia, südeurop, asie, immigrant, flüchtling, geflüchtet`

Table 11: Collection of word sets aggregated for verifying the origin bias content of German dictionaries. Contains stereotypical German and Non-German occupations as determined by employment statistics. Origin-characterising words are such words which might appear in a term's dictionary definition and indicate that this term is related to the concept of origin in some way.

intended or that the training data is not suitable for the task at hand. One of the basic assumptions the DD-GloVe loss functions all make use of is that dictionary data can serve as a relatively bias-free external training source. An et al. (2022) carried out an experiment testing whether this assumption is true. I replicated this test for German origin bias.

To do so, An et al. (2022) derived a list of gender-biased occupations, i.e., occupations that are stereotypically associated with one gender, from Zhao et al. (2018b). For each of these terms, they examined whether the term's dictionary definition contained any *gendered words* such as "he", "women" or "female". They utilised a list of 1,441 gendered words first compiled by Wang et al. (2020). The authors found that 39 out of 40 dictionary definitions for gender-biased occupations did not contain gendered words. Based on this, they concluded that dictionary definitions are "almost bias-free"[1].

I replicated this test by first compiling a list of German origin-biased occupations. I based this list on a statistic[2] first published by the German employment office ("Agentur für Arbeit") which

[1] according to a somewhat narrow definition of "bias-free dictionaries", as discussed in §3.3.2

[2] `https://mediendienst-integration.de/integration/arbeitsmarkt.html`

breaks down the citizenship of employees in all recorded occupations. In line with Zhao et al. (2018b), I extracted the top 20 occupations with the largest proportion of German employees and the top 20 with the largest non-German citizenship proportion. The resulting list can be found in Table 11. This data can only serve as an approximated of origin-biased occupations since occupations with a large proportion of non-German employees are not necessarily the same as those which are socially seen as the most stereotypical occupations for foreigners. However, no reliable data on the most stereotypical occupations for different nationalities is currently available.

The equivalent to the list of *gendered words* necessary for this test I created manually by defining a list of 30 word stems which could appear in the definition of an origin-related term, e.g., `ausland`, `heimat`, `immigrant`, or `türk`. I name this list *origin-characterising words*. The full list can also be found in Table 11. With these lists I then inspected how often origin-characterising words appear in dictionary definitions for statistically "German" versus "non-German" occupation. The result of this test was that in zero out of 40 occupation definitions, an origin-characterising word appears, confirming An et al.'s findings.

In a second part of this test, the authors confirmed that gender-specific words, i.e., words which *should* relate to gender such as "queen", do contain gendered words in their definitions. To replicate this test, I utilised the list of origin-specific words I created for the Hard Debiasing algorithm (see §3.3.1) and checked for the presence of origin-characterising words in the dictionary definitions of the origin-specific words. My results were that only 69 out of the 105 words considered contained origin-characterising words, which constitutes 65.71%. This is a lower percentage than the 86.25% (327 out of 379) reported by An et al. (2022). This likely either means that German dictionary entries do not indicate their relation to origin topics as clearly as is the case for English dictionaries and gender, or that my origin-characterising word list was not extensive enough.

Reproducing Results for English and Gender

In order to investigate whether the unsuccessful results of my DD-GloVe debiasing were due to changes I made while adapting the method to German origin bias or whether there are flaws inherent to the method itself, I attempted to reproduce the results published in An et al. (2022). This means that in this case, I utilised English instead of German data and examined gender instead of origin. The steps necessary to achieve a reproduction are to ensure that the training code is identical to the code published by the authors, that hyperparameters match those specified in their paper, and that the same data is used.

Investigating Dictionary Source Not only the training data itself, but also the dictionary data used is essential for DD-GloVe training. However, it does not become entirely clear in An et al. (2022) which dictionary the authors use.

In their paper's text, they claim to derive definitions from the "Oxford online dictionary" and link to https://www.lexico.com/, which in fact used to host a collection of Oxford dictionary sources[3], but now redirects the user to https://www.dictionary.com/ instead. Assuming that the authors accessed the site before the redirection, which took place in August 2022[4], they would have been using a combination of the *Oxford Dictionary of English*, the *New Oxford American Dictionary*, and parts of the *Oxford English Dictionary*.[5] However, comparing example definitions given in An et al. (2022) to definitions from those sources does not yield matching results.

Examining then the code published by An et al. (2022) leads one to the dictionary API resource "Free Dictionary API"[6]. Upon searching the GitHub repository for the project, one can find a comment[7] by the project's creator dated August 2021 stating that the API's data source will be migrated to Wiktionary. It is unclear what the data source was before this date.

The definitions provided by the English version of Wiktionary are relatively similar to the example definitions shown in An et al. (2022), but still do not match up exactly. For example, the Wiktionary definition for "mistress" reads "A woman, specifically one with great control, authority or ownership"[8], while An et al. (2022) claim that a mistress is defined as "[a] woman in a position of authority or control". It is conceivable that the Wiktionary article was edited in between the point in time at which An et al. (2022) accessed it and the time of writing this thesis. It is also possible that An et al. (2022) carried out their experiments before August 2021. Ultimately, it cannot be determined with certainty which dictionary the authors used in their experiments.

[3] https://web.archive.org/web/20190616173800/https://languages.oup.com/lexico-faqs

[4] https://web.archive.org/web/20220813074549/https://www.lexico.com/

[5] https://web.archive.org/web/20140122061925/http://www.oxforddictionaries.com/words/content-help

[6] https://dictionaryapi.dev/

[7] https://github.com/meetDeveloper/freeDictionaryAPI/issues/102

[8] https://en.wiktionary.org/wiki/mistress

Training Setup For my reproduction, I decided to use the dictionary source specified in An et al. (2022), the website dictionary.com. This is likely not the resource An et al. (2022) used, since the website now houses content from the *Random House Unabridged Dictionary*, but unlike the *Oxford English Dictionary*, it is a freely accessible dictionary website. While the results might then not be identical to those published in An et al. (2022), the process should still be similar enough to determine whether the method is reproducible at all. I crawled dictionary.com for definitions of all tokens in the model's vocabulary and retrieved definitions for 116,380 out of 400,000 tokens. If a word had multiple definitions, I simply concatenated all definitions, as An et al. (2022) did.

For the model's training corpus, I used a Wikipedia dump from HuggingFace[9], in accordance with An et al. (2022). The data set I downloaded is dated 2022-03-01. It is possible that An et al. (2022) used an earlier version of Wikipedia.

The training code was taken directly from the authors' GitHub page[10]. Most training parameters stayed the same as in my German experiments except that the hyperparameter γ was set to 0.2 instead of 0.05, the number of seed words generated was $n = 30$, and seed words were generated only in the first iteration. These settings were taken from the specifications for gender debiasing in An et al. (2022). Lastly, as initial seed words I chose the default "he" and "she" which An et al. (2022) used.

Reproduction Evaluation I compared scores for word pair similarity and the WEAT between the vectors produced through my reproduction and the values published in An et al. (2022). Additionally, I ran my evaluation on the vectors that An et al. (2022) have published on their GitHub repository.

Unlike in my experiments for German embeddings, I used the WordSim-353 data set (Finkelstein et al., 2001) for the word pair similarity task. Word pairs with out-of-vocabulary words were treated as having zero similarity. For the WEAT I use the original WEAT 6 values from Caliskan et al. (2017). This test consists of female versus male names as the attribute words, and career versus family terms as the target words.

Table 12 compares the values An et al. (2022) reported after debiasing for gender bias, the vectors the authors published on GitHub, and the embeddings I recreated.

[9] https://huggingface.co/datasets/wikipedia
[10] https://github.com/haozhe-an/DD-GloVe

	Paper	GitHub	Reproduced
WordSim-353	n/a	$\rho = 0.5718,\ p < 10^{-4}$	$\rho = 0.6136,\ p < 10^{-4}$
WEAT 6	$d = 1.25,\ p = 0.0029$	$d = 1.2941,\ p = 0.0004$	$d = 1.882,\ p < 10^{-4}$

Table 12: Comparison of semantic similarity and bias metrics between different version of DD-GloVe embeddings. Embeddings from An et al. (2022) are compared with the vectors the authors have published and my reproduction of their methods. For WordSim-353, Spearman's rank correlation coefficient ρ is specified and for the WEAT permutation test, p-value and Cohen's d are given.

Overall, it can be seen that the results are incongruous. The gender bias measured with the WEAT was much higher in my version of the embeddings than what was reported by An et al. (2022). According to the authors, the p-value of 0.0029 that DD-GloVe achieves on the WEAT for gender bias is the best result out of a range of other debiasing methods such as Double Hard Debias (Wang et al., 2020) and GN-GloVe (Zhao et al., 2018a). In my version of the embeddings, however, the gender bias contained in them was as prevalent as if I had not been utilising the debiasing loss functions at all. For comparison, I also trained a DD-GloVe model on English data without any debiasing and ran the WEAT, resulting in a p-value of $p < 10^{-4}$ and an effect size of $d = 1.831$, which is as high as for the "debiased" version I trained.

The vectors published by the authors on GitHub achieved a WEAT score similar but not identical to what is reported in An et al. (2022). It is possible that the published embeddings were trained using slightly different parameters than those for which An et al. (2022) reported results in their paper, and therefore produced different results. Nevertheless, one would expect the published vectors to score closer to the published values.

The authors did not measure their embeddings' performance for any word pair similarity task, so only my reproduction and the published vectors could be compared. It can be seen that the embeddings I trained show a slightly higher correlation for the WordSim-353 data set. This might be due to me possibly using a newer Wikipedia dump than An et al. (2022). However, it is not certain whether this factor alone would cause a difference of 0.0418 points.

Examining Seed Word Generation

An integral part of most debiasing algorithms is determining the bias direction with the help of seeds words. In DD-GloVe, these seed words are automatically generated from a pair of initial seed words. If this generation were unsuccessful, this would make it impossible for the algorithm to succeed in debiasing embeddings. Because of this, I analysed the seed words generated by DD-GloVe under different training setups.

There are multiple parameters which influence the generated seed words. One such parameter is the number n of seed words generated, which An et al. (2022) set to 30 for their gender experiments, and for which I chose $n = 10$ for origin debiasing. Furthermore, the content of the seed words' definitions is a deciding factor. An et al. (2022) manipulate this in their code by cutting off the definitional text for "he" after the first 11 words and thereby including only the dominant dictionary gloss. Last but not least, the choice of the initial seed word pair of course heavily influences the outcome of the seed word generation algorithm.

I experimented with different combinations of these parameters and qualitatively assessed the resulting seed words. In Table 13, I present an excerpt of these experiments. All training settings were the same as for my usual DD-GloVe training unless otherwise specified.

Generally, it became apparent that many of the generated results were entirely unrelated to the concept that the initial seed words expressed. This was especially the case for the seed words "Türke" and "Ausländer": Evaluating the main training setup I used throughout this thesis (first row in Table 13), it can be seen that the seed words generated for "Türke" were all unrelated to Türkiye or any other origin concept. The only exception was the seed word "Türke" itself, which was added to the list automatically since it was the initial seed word. For "Deutscher", the rough orientation of the generated seeds were more in line with the desired concept, although some terms (e.g., "Rotwelsch", "Holländisch") appeared far-fetched or even unfitting considering that more obvious choices like "Deutschland" or "deutschstämmig" would have been available.

Using only the first gloss (second row in Table 13) did — at least in this case — not improve the quality of the results. The relatedness of the generated seed words to the origin concept was similar to before.

Changing the initial seed words to "Deutscher" and "Ausländer" produced the most relevant seed words. The seed words for "Deutscher" were all clearly related to the concept of being German and while the seeds generated for "Ausländer" might not necessarily all have been related to the concept of being a "foreigner", they at least were all clearly related to the origin concept.

Increasing the number of generated seed words lead to worse results. One might wonder why the previous top ten terms were not included in these top 30 terms. This is because the seed word

Setting	Seed Words
`<deutscher, türke>`	**“Türke”:** Ziehharmonika, Nachbau, Blutzucker, Konstrukt, Türke, Versteifung, Paradoxie, Provisorisch **“Deutscher”:** Biedermeier, Hitlerdeutschland, Bundesbank, Deutscher, Germania, Hochdeutsche, Holländische, Schweizerdeutsche, Rotwelsch, Lufthansa
`<deutscher, türke>`, only first gloss	**“Türke”:** Bienenstock, Informationsveranstaltung, Wandschmuck, Absperrgitter, Türke, Selbstinszenierung, Orientierungshilfe, Wärmedämmung, Herrschaftsinstrument, Funktionär **“Deutscher”:** Nationalsozialismus, Volksherrschaft, Landesversicherungsanstalt, Tamile, Deutschtum, deutsch-französisch, Deutsche, Deutscher, Bundesbank, Hochdeutsche
`<deutscher, ausländer>`	**“Ausländer”:** Vokabel, Lehnwort, Fremdwort, Inländer, Staatswappen, Dolmetscher, Ausländer, fremdländisch, Staatsfinanzen, Staatsschutz **“Deutscher”:** deutschnational, Großdeutschland, Deutschlandtour, Deutschtum, deutsch-französisch, Hitlerdeutschland, Deutsche, Deutschrock, Deutscher, Sudetendeutsche
`<deutscher, ausländer>`, $n = 30$	**“Ausländer”:** Mondsonde, Schießer, Harfner, Mundschenk, Sterbekasse, Datenschützer, Minima, Adoptivsohn, Inländer, Ger, Dragqueen, Thermostat, Guano, Raumkapsel, Ausländer, Junggeselle, Fremdarbeiter, Staatsbürger, [...] **“Deutscher”:** Deutschtum, deutschstämmig, deutsch-französisch, Achtundvierziger, Westmark, Deutschkunde, Schwabenspiegel, Deutschlandlied, Novemberrevolution, Landesversicherungsanstalt, Deutschsprachig, Hitlerdeutschland, Deutschlandweit, Nachkriegsdeutschland, Biedermeier, Ostpreußen, Gesamtdeutschland, Deutsch, [...]
An et al. (2022) Reproduction	**“he”:** he, he/she, H.E **“she”:** she, she/he, she/her
An et al. (2022)	**“he”:** he, son, brother, brothers, boys, sons, boy, businessman, yang, gentleman, wizard, headmaster, statesman, nobleman, policeman, salesman, bahadur, stallion, fiance, manny, [...] **“she”:** ex-wife, girl, jane, woman, wife, witch, women, she, pilipinas, heroine, maids, hens, dona, wives, fiancee, goddess, bint, sheila, hostess, hen, [...]

Table 13: Seed words generated by the DD-GloVe algorithm for different training setups. Parameters are the initial seed words, number of seed words generated, dictionary content used (entire entry or only the dominant gloss), and basic training scenario (English gender or German origin). Additionally, the seed words reported by An et al. (2022) are shown. Seed words in each row except the last are ordered according to the calculated bias value $b(w)$. For experiments with 30 generated seed words, only the first 20 are shown due to lack of space.

generation happens at train time and the seed words determined in the first iteration can change in later iterations — depending on which seed words were chosen initially. In other words, in each training iteration, the seed word sets influence the embeddings, which changes the definition embeddings, and in turn this changes the seed words selected in the next iteration. Therefore, if 20 new seed words are introduced in the first iteration, this can lead to an output of 30 completely different seed words in the last iteration.

My reproduction of An et al. (2022), i.e., gender debiasing for English embeddings as described above, resulted in only six seed words total being identified. A lower number of seed words being found than originally specified is possible whenever the algorithm determines multiple words to have the same bias value $b(w)$ (see §3.3.2). Due to the particular way An et al. (2022) implemented their algorithm, only the top n values *bigger* than other values are included. If, for example, the five highest "gender values" were $[0.8, 0.8, 0.85, 0.9, 0.95]$, then for $n = 5$ only $[0.85, 0.9, 0.95]$ would be included since the first two values are the same. This suggests that the three terms found for "he" and "she" were particularly high in their genderedness, but the next-highest values were all too similar to each other, perhaps all zero. During the course of carrying out these seed word experiments, I was able to note that in many cases, the vast majority of bias values calculated was zero or close to zero.

It can further be seen that the seed words generated by my reproduction attempt greatly diverged from those specified in An et al. (2022). Because dictionary definitions are a core part of the seed word generation algorithm, the difference in the dictionaries used could potentially explain this stark difference.

Seed Pair	**WEAT Turkish**		**WEAT Polish**		**WEAT Italian**	
	$\downarrow d$	$\uparrow p$	$\downarrow d$	$\uparrow p$	$\downarrow d$	$\uparrow p$
`<deutscher, türke>`	1.821	$< 10^{-4}$	1.658	$< 10^{-4}$	1.452	$< 10^{-4}$
`<deutscher, pole>`	1.772	$< 10^{-4}$	1.557	$< 10^{-4}$	1.330	$< 10^{-4}$
`<deutscher, italiener>`	1.789	$< 10^{-4}$	1.701	$< 10^{-4}$	1.306	0.0001
`<deutscher, ausländer>`	1.790	$< 10^{-4}$	1.716	0.0008	1.442	0.0008

Table 14: WEAT Scores for embedding bias comparison across different nationality pairs. Each row represents a seed word pair, and columns show the effect size (Cohen's d) and statistical significance (p-value) for WEAT using Turkish, Polish, and Italian first names as attributes. Lower effect size and higher p-values indicate improved performance in mitigating bias.

Since the seed pair `<deutscher, ausländer>` produced the most promising results, I again perform a WEAT evaluation for the embeddings resulting from these seed words. Table 14 shows that while the seed words generated with the initial pair <deutscher, ausländer> may have seemed more coherent than those for <deutscher, türke>, this did not necessarily reflect in better debiasing results. For the WEAT with Turkish names, the performance of both embeddings was approximately the same. For the Polish and Italian WEAT, the <deutscher, ausländer> embeddings achieved slightly better results than <deutscher, pole> and <deutscher, italiener> respectively. The p-value was at least 0.0008 points higher in both cases. However, this still resulted in p-values below the significance level of $1.\bar{6}\%$.

5 Discussion

In this chapter, I am going to discuss the results presented in chapter 4 and answer the three research questions posed in chapter 1. In particular, I am going to focus on challenges encountered in my experiments, unexpected results, and the implications of my findings.

5.1 How Do You Measure a Measure?

The first of my leading research questions was how origin bias can be measured in German word embeddings. I chose to employ the WEAT as the bias metric of this thesis. Kurpicz-Briki (2020) had already shown that the seed words used in English WEAT experiments can be translated and applied to German embeddings and I was able to confirm these findings. This result was as expected since the basic methodology of the WEAT — calculating word associations based on similarity — also works for German embeddings since apart form training data, there is no significant difference between the training process of GloVe and fastText embeddings for English versus for German. It follows that the basic properties of the embeddings are the same and calculations like cosine similarity transcend the embedding language. I therefore find that the WEAT could theoretically applied to any scenario in which two distinct groups of attribute and target words can be defined.

Apart from the WEAT, I also presented other options for measuring bias in §2.3 like the neighbourhood metric or word analogy tests. It is difficult to determine whether the WEAT is an appropriate tool to measure origin bias or not. In order to do so, one would need a gold measure to compare this metric to, i.e., a metric for which it could be said with certainty that it accurately quantifies the bias contained in word embeddings. However, no such gold measure exists, since all presented bias metrics already *are* attempts at creating a measure of bias. In other words, no ground truth exists for the question of "How biased is an embedding?", only different theoretical approaches. This leads to the question of how best to determine how well a metric captures embedding bias.

Bolukbasi et al. (2016) in this regard state that "the difficulty of evaluating embedding quality [...] parallels the difficulty of defining bias in an embedding", meaning that both measuring the semantic performance of embeddings as well as measuring embedding bias are challenging tasks. One possible method to resolve this could be to compare the results produced by a human study to those measured with embeddings. Since Caliskan et al. (2017) based the WEAT on the psychological IAT, this was one of the arguments I presented for selecting the WEAT as a metric. Still, a direct comparison of embedding bias to human bias is not possible since humans cannot judge the content of 300-dimensional vectors. Only a validation via a proxy is possible, e.g., by employing methods such as calculating the associations between word embeddings and then having humans rate these words in combination as well.

Another approach to appraising different bias metrics could be to assess the strengths and weaknesses of a metric in comparison to other metrics, which I have done in §2.3. One of the main issues researchers have pointed out for the WEAT is its dependency on token frequency. Van Loon et al. (2022) criticised that the bias measured with the WEAT can be explained solely by rare and negative terms being clustered together in the embedding space. Their research was placed in the context of sociology, where embeddings are used as predictors of anti-Black sentiment. In this context, their criticisms are valid, since the frequency dependency means that embeddings can not be used as an independent indicator of anti-Black sentiment. However, I would argue that the issue they criticised lies with the embeddings themselves and not the metric. Calculating cosine scores, which is how the WEAT measures association, is an integral part of many downstream NLP applications which do not "control for relative name frequency", as van Loon et al. (2022) suggest. If biased outputs are produced by these applications, they are still problematic even if they can be explained by frequency. Therefore, measuring this "frequency bias" can be a valuable part of the process of measuring bias in word embeddings. This view then rewards debiasing methods such as Double-Hard Debiasing (Wang et al., 2020) which remove frequency information from the embeddings.

As explained in chapter 2, the step of determining a bias metric should not be equated with defining what bias fundamentally is. As part of my research questions, I made the assumption that bias can be captured using mathematical methods. To discuss this, it could be explored whether the WEAT can be said to detect embeddings which fulfill the initial definitions of bias proposed in chapter 1.

I assess that especially the second definition by Friedman and Nissenbaum (1996) is well-addressed by the WEAT. This definition asks whether computer systems "systematically and unfairly discriminate" against certain people. The WEAT shows that, e.g., Turkish names are more associated with unpleasant terms and German names are more associated with pleasant terms. This could lead to *discriminatory* behaviour of downstream applications in which these embeddings are utilised. The WEAT measures this discrepancy in a *systematic* way with a permutation test over lists of dozens of seed words.

The first bias definition by the Brookings Institution (2023) asks whether algorithms are predicting their target inaccurately or inequitably. This would probably best be measured by a method which generates predictions from language models or which employs embeddings in downstream tasks for which the performance can be measured. However, accurate predictions can in part also be measured with the semantic evaluation I carried out and which showed slight improvements in performance after debiasing. This indicates that perhaps the bias contained in the embeddings did indeed hinder accurate predictions. The equity of predictions is again captured by the WEAT, since I used the WEAT to compare predictions of similarity between two nationalities and found them, at least in part, to not be equal.

Considering the factors discussed here, I estimate the WEAT to be an appropriate tool to measure origin bias in German word embeddings. Nevertheless, employing additional metrics, especially ones which examine indirect bias, could provide an even more accurate insight into origin bias in German embeddings. Future work should thus ideally include a comparison of a broad selection of bias metrics.

5.2 Seed Lists Introduce Subjectivity

Antoniak and Mimno (2021) stated that seed words, which are used in most bias-related methods, often pose problems due to instability. I have encountered this issue in various parts of my own research. In particular, I can confirm Antoniak and Mimno's observations that results of bias analysis and debiasing methods heavily depend on the choices made regarding the selection of seed words.

For my WEAT measurements, I defined a process for creating new attribute sets, which I am going to reflect upon in §5.3. The name lists resulting from these changes produced measurements different from those reported by Kurpicz-Briki (2020). As discussed above, it is difficult to determine which of these WEAT variants "better" captures bias. It is clear, however, that the differences in output are due to the different seed words chosen. The selection of these seed words posed multiple challenges.

For the primary step of gathering name data, no official records were available for Germany. For the other three countries, official records were sometimes inaccessible due to language barriers, or sources contradicted each other. Other researchers have occasionally used medical data such as cancer registries in which names and ethnic origin of patients are documented (e.g., Razum et al., 2001). However, this type of data is classified as sensitive information and therefore not easily accessible. I therefore utilised privately operated websites which aggregate name information. It could be questioned whether the names extracted from these sources truly represent the most common names of each country. I attempted to mitigate this potential issue by including name prevalence, frequency and nationality uniqueness in my filter criteria.

The details of my filter criteria were chosen with efficiency, reproducibilty, and stability in mind. However, there are some limitations to my methodology. The thresholds for some of the criteria, e.g., gender distribution, was chosen empirically based on the number and type of names included at different thresholds. Furthermore, it was assumed that if a name is used predominantly as, e.g., German in German texts, it is a typically German name, even though it might be that it is just used this way in *German* texts, not in general. Lastly, the process of excluding names with ambiguous origin could potentially have resulted in an age bias in the names since there could be differences in the typical origin of names between different generations.

Since there is no benchmark for correct choices in this case, the validity of my selection procedure can ultimately only be assured in the sense that more comprehensive attribute lists should, in theory, lead to more reliable results. This holds true as long as the names of the list are fitting for the origin categories, which I ensured with the inclusion criteria I specified. In any case, my adapted WEAT attribute lists present a more extensive, methodically selected, and differentiated version of the name lists first presented in Caliskan et al. (2017) and translated by Kurpicz-Briki (2020).

Both debiasing algorithms I examined were influenced by the seed words chosen for them. For DD-GloVe, I chose only two initial seed words and the model automatically generated additional seed words. While this might seem like a solution to the subjective nature of seed word choices, in practice, this method only lead to a higher influence of the initial two seed words chosen and less possibilities to adapt the algorithm to specific circumstances, as can be seen in the analysis in §4.3.

Especially for the Hard Debiasing algorithm, I defined many seed words through a process of qualitative evaluation. The seed words in this case are the primary factor in determining which information is to be removed from the embeddings. Since the semantic performance was not decreased after debiasing, it is unlikely that the word lists I created were too extensive. However, it is possible that they were not extensive enough or contained unsuitable words, and that this decreased the debiasing performance. This could explain why the algorithm was able to remove some of the measured biases but not others.

Overall, there is a great level of subjectivity involved in the creation of seed lists. There exist barely any recognized and extensively reviewed standard data sets, and virtually none at all for cases other than English gender bias. Seed sets are a decisive factor in the success of debiasing methods; the failure of the DD-GloVe algorithm in this thesis is a striking example of this. It would therefore be a worthwhile endeavour for future work to critically review and rethink existing word lists and compare the outcome of debiasing algorithms using different seeds.

5.3 Characteristics of Origin Bias in German Embeddings

Origin Bias Differs Between Nationalities The second research question I posed in chapter 1 was whether German word embeddings contain origin bias. I expected to confirm the findings of Kurpicz-Briki (2020), who found origin bias in German embeddings similar to what has been observed for English embeddings using the WEAT. Overall, I was able to show that German fastText as well as GloVe embeddings do contain origin bias. However, my results in this regard were mixed and a significant bias was not measured in all examined cases. This is surprising because prior literature has been unanimous in its discovery of gender and racial bias in English embeddings.

This contrast is likely due to the adapted WEAT seeds I employed, which differentiate between different nationalities. Kurpicz-Briki (2020) in her experiments created name lists which, by happenstance, mainly included names of Turkish or Arabic origin (see Table 1 in §3.2.2). The bias I measured with the WEAT for my Turkish name lists was indeed comparable to the values published by Kurpicz-Briki (2020); it is only for the Polish and Italian experiments that my values differed considerably. It can therefore be assumed that, if the nationalities in the data set used by Kurpicz-Briki (2020) were more diverse, the bias measured by the author would also be lower or even non-significant. A similar hypothesis can be raised for works on English embeddings which measured racial bias on a Black-White axis. It is possible that, if racial bias in English embeddings were differentiated between, e.g., White, Black, Hispanic, and Asian groups, the results might differ to previously reported biases too.

This raises the question of what is preferable — the analysis of general origin bias or examining different individual ethnic origins? According to the research motivation presented in chapter §1, the aim of bias research can be understood as raising awareness of biased patterns in word embeddings and mitigating the discriminatory effect these patterns can have in downstream applications. With these goals in mind, a more differentiated view of bias with specific nationalities or races can provide more thorough information to users of embeddings. This additional knowledge could then potentially lead to the development of more specialised debiasing methods, which more effectively reduce origin bias from embeddings. I therefore encourage future bias research to further explore multi-dimensional bias analysis and bias mitigation methods.

One insight the differentiation between nationalities allows is a comparison of the WEAT values with the discrimination these minorities experience in real-world scenarios. In a survey by the research team *Deutsches Zentrum für Integrations- und Migrationsforschung*, the authors reported that people of Turkish origin experienced discrimination more often than other groups (Brinkmann et al., 2023). This matches the bias I measured in German word embeddings. In fastText embeddings, Turkish names were the only ones for which a significant bias was measured, and in GloVe embeddings, the Turkish bias was accompanied by the largest effect size. Furthermore, Brinkmann et al. (2023) reported that people of Polish or South European origin experienced discrimination considerably less often than people from other ethnic origins such as Africa or the Middle East. Again, this agrees with my WEAT results, which indicated a lower bias for the Polish and Italian tests. This is especially the case for the fastText embeddings, in which I was not able to measure a statistically significant bias against people of Polish or Italian origin. In light of these results and the statistics reported by Brinkmann et al. (2023), future work might focus primarily on measuring and mitigating bias against African, Middle Eastern, and Turkish descent.

Origin Bias Differs Between Embeddings Next to the differences between nationalities, my adapted WEAT experiments results also included unexpected differences between fastText and GloVe embeddings. While no statistically significant bias against Polish or Italian names could be measured in fastText embeddings, the GloVe embeddings contained Polish and Italian biases almost as high as for Turkish names. This raises the question of where this difference stems from. The two sets of embeddings were both trained on similar Wikipedia data and evaluated using the same WEAT attribute lists. Therefore, it is likely that the difference lies in the training methodology of the two embeddings.

One explanatory approach arises from comparing the effect sizes and p-values of the respective tests. While the average p-value for the fastText embeddings (0.8910) is considerably above the significance level and much higher than its GloVe equivalent (0.0013), the average effect size for the fastText embeddings is still large with a magnitude of 0.8910. These values can be interpreted as there being a large practical difference between the fastText embeddings for German versus other names, but this difference not necessarily being a biased one, i.e., not being associated with pleasant versus unpleasant terms in a statistically significant way. In practice, this might mean that a *difference* between, e.g., Italian and German names exists in both embedding models, but only GloVe optimises the embeddings in such a way that this difference expresses itself as Italian names being more *negative* than German names.

A possible reason for such behaviour can be found in how the two models treat infrequent terms. In §2.3, I presented research on how embedding models tend to group infrequent and negative terms together and how this affect bias metrics. It is possible that GloVe embeddings exhibit this behaviour more intensely than fastText embeddings, which are trained on subwords and therefore able to handle rare words more accurately. If this is the case, the GloVe model might associate Polish and Italian names with unpleasant terms because of their shared infrequency, but the fastText model might be able to circumvent this association.

This hypothesis in turn leads to the question of why, then, the fastText embeddings contain significant bias against Turkish origin. Frequency alone does not determine the embeddings that a model produces. For Turkish origin, it might be the case that there are clear semantic data patterns relating Turkish and negative terms to each other, such as texts about Turkish people written in a negative tone of voice. According to the semantic content of the data, fastText would then accurately learn a bias against people of Turkish origin.

In order to test this hypothesis, future work might analyse where in the corpus bias stems from and compare this between fastText and GloVe embeddings. It could then be seen whether the same corpus documents lead to different outcomes in terms of bias in the two models. Brunet et al. (2019) developed a method which could be employed to this end. They analysed how removing certain documents from the corpus affected bias in word embeddings. Furthermore, in future work the WEAT could be carried out and compared between fastText and Word2Vec embeddings, which are similar to fastText except that they are not trained on subwords. From this, conclusions could be drawn about the effect that fastText's special handling of rare words has on origin bias.

This explanatory approach does not explain all differences between the two models to full satisfaction. In particular, it does not provide a reason as to why the effect size measured for Polish bias in fastText embeddings was as low as it was. To reiterate, the training corpus used was similar and the WEAT attribute lists were identical between GloVe and fastText. Because of this, the discrepancy cannot stem from, e.g., German and Polish names perhaps being more similar etymologically, or any other property of the data. Instead, the difference should also be explainable by the different architectures of the two embeddings models. When counting the frequency of names in the training corpus, I found that on average, the Polish WEAT names appeared 3081 times in the corpus whereas the Italian names appeared 9155 times on average. It is conceivable that due to the lower number of Polish occurrences, the model used more subword information for constructing the Polish name embeddings than was the case for the Italian name embeddings. These subword embeddings might share features with the embeddings for the German name embeddings, leading to a greater similarity between the two sets of name embeddings and hence a smaller effect size. Again, Word2Vec embeddings could be evaluated to assess whether this hypothesis has merit.

Finally, it should be noted that the, in part, low bias measured in fastText embeddings does not necessarily mean that these embeddings are not biased against people of Polish or Italian origin. The WEAT can only measure the presence of bias, not its absence, since it can only disprove the null hypothesis of there not being any bias, not prove it. In future work, further bias metrics could be applied to these embeddings to determine with more certainty whether they are truly free of Polish and Italian bias.

5.4 Difficulties in Adapting Debiasing for German Embeddings

The third research question I formulated was how origin bias in German embeddings can be mitigated. My sub-questions under this were whether existing methods can be adapted to the German origin scenario, how different methods compare to each other, and what reasons for the success or failure of these methods might be.

As described in §2.4, a variety of approaches to debiasing exist. While many authors state that their methods can be adapted to other languages or bias attributes, in practice, the adaptation of such methods to German and origin bias was not without challenges. German is a resource-rich language, but nevertheless, data specific to bias research is scarcely available, so it became necessary to create new data sets. During the process of creating seed words, own biases or instability can be introduced to the data, as explained in §5.2.

Particularly for the Hard Debiasing algorithm, the creation of extensive seed word sets was necessary. The resulting performance of Hard Debiasing might have been hindered by limitations in my methodology in this regard. Since I did not carry out any human studies to confirm my selected origin-defining word pairs, it is possible that they did not capture the concept of origin in an ideal way.

Furthermore, the set o origin-specific words defined in this work was smaller in size than the one used in Bolukbasi et al. (2016). However, this should, in theory, not have lead to worse debiasing results, only possibly to a decrease in semantic performance. This is due to the fact that the origin-specific seed words define the words *not* to be debiased and therefore a small set size might mean that too many words were debiased. Since the semantic evaluation did not show a decrease in performance after Hard Debiasing, it can be assumed that this was not a prevalent issue.

Part of my initial research question was whether the employed methods are able to mitigate bias according to the bias definitions given in chapter 1. In this thesis, the success of debiasing was measured using the WEAT. As discussed in §5.1, the WEAT can capture bias according to the bias definitions proposed in chapter 1. Therefore, if the bias contained in embeddings was mitigated according to the WEAT, the methods were able to reduce bias so that the initial definitions of biased computer systems were no longer fulfilled for these embeddings.

The results of my WEAT experiments showed that this was not the case for the DD-GloVe algorithm. Possible reasons for this are going to be explored in §5.5. The Hard Debiasing algorithm, on the other hand, was able to reduce the measured bias above the significance level in two out of four cases where statistically significant bias was initially measured: In fastText embeddings, Hard Debiasing reduced Turkish origin bias below statistical significance. Since this was the only nationality for which bias was detected in the original embeddings, the algorithm appears to be well-suited for fastText embeddings. In the GloVe embeddings, only Italian origin bias was removed to a statistically significant degree, even though the WEAT also indicated Turkish and Polish origin bias in the original embeddings.

One conceivable explanation of this result is that the German-Italian seed words used for debiasing better expressed the origin concept than the German-Turkish or German-Polish seed words.

However, this is unlikely because the same seed words were also used to successfully reduce Turkish origin bias in the fastText embeddings. This suggests that the difference lies in how the bias is structured, i.e., represented as features, in the embeddings. As described in §2.4, researchers have criticised that Hard Debiasing is not able to remove indirect biases well (Gonen and Goldberg, 2019). It stands to reason that GloVe embeddings might contain more indirect bias for Turkish and Polish origin than for Italian origin and the debiasing algorithm is therefore more successful in removing Italian origin bias.

It should be noted that while the Hard Debiasing algorithm was not as successful for Turkish and Polish origin bias as for the Italian case, the bias measured by the WEAT still *was* decreased after debiasing. This becomes especially relevant when comparing the WEAT scores for these results to values published in prior literature. According to baselines evaluated by An et al. (2022), even methods for which researchers have reported debiasing success sometimes still only produce p-values close to zero after debiasing. However, the values reported by An et al. (2022) differ from those published in the original papers which An et al. (2022) reference. For example, An et al. (2022) claim that Double-Hard Debiasing (Wang et al., 2020) results in a p-value of 0.0014 on the WEAT 1 (which measures gender bias) whereas Wang et al. (2020) themselves state that the algorithm achieves a p-value of 0.0366. In both cases, that value would still be below the significance level $\alpha = 0.05$ specified by the authors. In comparison to other literature, it might therefore be reasonable not to assume failure of the method if the WEAT still measures significant bias after debiasing, but instead to strive mainly for a general trend towards mitigation. Under this assumption, Hard Debiasing can be said to overall have successfully mitigated origin bias in German word embeddings, whereas the DD-GloVe algorithm failed in doing so.

5.5 Insights from Failed DD-GloVe Reproduction

The experiments I carried out in response to the lack of success of the DD-GloVe method allow some conclusions to be drawn as to why the model did not perform as reported by An et al. (2022).

The fundamental assumption of dictionary-based debiasing methods such as DD-GloVe is that dictionary definitions can be leveraged as neutral reference points. These reference points are then supposed to indicate which words should or should not be related to the bias attribute. Due to this important role, the dictionary data used has a large impact on the resulting embeddings. As noted in §3.3.2, the dictionary used in my experiments contained only approximately half as many definitions as the one used by An et al. (2022). In my implementation for German embeddings and origin bias, I retrieved definitions for 59,721 tokens, whereas in my reproduction of An et al. (2022), the English dictionary yielded definitions for 116,380 tokens. Because of this, there was less data available for the German DD-GloVe model to determine bias values for its vocbulary tokens.

Moreover, the experiment described in §4.3 showed that even for words which should be related to origin, the German dictionary definitions did not contain origin-characterising terms. This might have prevented the DD-GloVe model from successfully differentiating which words should and should not be related to the bias dimension: The algorithm assumes that definition embedding can be used to determine whether a word's embedding is biased or not. This assumption is especially important for the bias-aware GloVe loss function J_{G-bias}, which due to hyperparameter settings was the primary influence on the overall loss. The function depends on the existence of a significant number of definitions which indicate whether a token is related to origin. However, my experiments showed that many dictionary definitions in fact did not contain information about the token's relation to origin. This might have led the algorithm to judge a large number of words to be similarly biased — or rather, similarly unbiased. For most parts of the DD-GloVe algorithm, this behaviour is not detrimental. Particularly the definition loss J_{def} will still enhance the embeddings with general information from the dictionary definitions, unrelated to origin bias. The effect of this can be observed in the increased semantic performance of the embeddings after debiasing, despite the bias not having been removed. In terms of origin-debiasing, however, missing dictionary definitions and dictionary definitions which erroneously do not contain origin-characterising words will lead to the model not being able to accurately identify biased embeddings. The debiasing therefore cannot be successful.

My experiments further showed that the seed word generation algorithm employed by DD-GloVe did not produce the desired results. The reason for this could again be related to the issue of dictionary definitions not containing enough information about origin: In the seed word generation algorithm, vocabulary tokens are sorted according to their bias value to identify the most origin-specific tokens. This bias value is calculated using definition embeddings. It is assumed that a word related to origin will contain origin words in its definition and therefore its projection onto the initial seed words' difference vector will have a large absolute value. Since the dictionary definitions, however, are in large parts not just bias-free, but also generally origin-free, this assumption in practice was not fulfilled. Furthermore, it is assumed that the difference between the two initial seed words chosen by the user clearly indicates the origin direction. It is possible that the initial seed words selected by me were not indicative enough of the origin concept. The result of these two faulty assumptions could then have been that arbitrary words were found by the algorithm because the projection vector was not meaningfully connected to origin, and there were not enough words with a large bias value.

Additional aspects which might have had a detrimental effect on the outcome of the seed word generation algorithm are the calculations the algorithm performs. While An et al. (2022) state that they calculated the bias value of a given token as the projection of the tokens' vector onto the difference vector between the initial seed words, this is not what is specified in their code or mathematical expression. There, they defined the bias value as the difference between the cosine similarities of the token to the two initial seed words. This is not equal to the scalar projection they claim to use in their explanation of the algorithm. It is possible that actually using vector projection instead of cosine similarity in practice might have lead to more meaningful results, since this would have made the bias values more dependent on the *difference* between the two initial seed words instead of just their individual associations with the token in question.

Apart from these difficulties, there are also features inherent to the design of the DD-GloVe method which could be called into question: An et al. (2022) chose to calculate definition embeddings by averaging all embeddings in the definition embeddings. This method will include any content present in the definition irrespective of whether it is relevant to the semantic meaning of the word. In cases where there is an origin-characterising word present in the definition, but the definition also contains many other words unrelated to origin, the origin content might not get across in the average of all definitional embeddings. This is especially relevant since the authors also chose to include all definitional glosses of a word into its definition and not just the most dominant one, even though some words may be homonyms or otherwise ambiguous. A more suitable approach to calculating the definition embeddings could be to apply a weighting function to the various glosses of a word, or to calculate the definition embeddings with methods designed for calculating singular vectors from multiple-word texts, such as document embeddings.

Finally, I attempted to reproduce the results published by An et al. (2022) by directly replicating their experiments, but did not succeed in doing so. A primary possible reason for this is the training data used. In §4.3, I detailed my procedure of selecting a dictionary for the reproduction attempt. It is likely that the dictionary data used in my reproduction was not the same as that used by An et al. (2022). In addition, it is possible that the authors used a different version of

Wikipedia data than the one used in this thesis. There can also be some training parameters found in the code published by An et al. (2022) which are not explained in their paper and for which it is therefore uncertain how they should be set. This includes a cap c which restricts the choice of seed words to the first c words in the vocabulary, and the restriction of the initial seed words' dictionary definitions to the first gloss. All of these factors potentially influenced the reproduction in this thesis and might have led to the difference in outcomes.

The failure of this reproduction demonstrates the importance of documentation of the precise data and hyperparameters used in scientific works. The incongruence of the work by An et al. (2022) and their training code make parts of any reproduction attempt a matter of best guesses. On top of this, the results reported by An et al. (2022) for their reproductions of various other debiasing methods were not consistent with the values reported in the corresponding original works, as mentioned in §5.4. In particular, An et al. (2022) reported a p-value of 0.0029 for the WEAT 1 experiment measuring gender bias after debiasing with DD-GloVe, and claimed that this result is state-of-the-art compared to other methods. However, some of the works they compared to their method in actuality report higher p-values than 0.0029.

Overall, these inconsistencies point towards the broader context of the reproducibility crisis ongoing in NLP research. In a work reviewing current efforts to increase reproducibility in NLP and machine learning research, Belz et al. (2021) found that only approximately 14% of reproduction studies obtained the same results as the original study. They also reported that reproductions usually yield worse results than what was claimed by the original works and that "worryingly small differences in code have been found to result in big differences in performance". The results of my experiments concerning the DD-GloVe algorithm were in line with these concerning findings.

6 Conclusion

Researchers in recent years have shown that embeddings contain undesired content in regards to protected attributes such as gender, which has real-world consequences when algorithms using such embeddings output discriminatory judgements. Previous studies have made progress in assessing and addressing such bias but have focused mainly on English embeddings and gender bias. In this present study, my leading research questions were how origin bias in German word embeddings can be measured, to what extent German embeddings are origin-biased, and how such bias can be mitigated. In order to answer these questions, I defined a bias metric and adapted and applied two debiasing algorithms on fastText and GloVe embeddings.

I employed the WEAT as a bias metric and found that it is generally well-adaptable to German and fulfills the purpose of quantifying bias as I defined it. Expanding upon work by Kurpicz-Briki (2020), I created new name seeds to measure bias relating to Turkish, Polish, and Italian origin. This differentiation allowed a more fine-grained analysis, but was challenging in terms of gathering reliable data and defining appropriate filter criteria.

An analysis of German embeddings with this metric partly confirmed the existence of origin bias as previously found by other researchers, but partly also resulted in lower bias values than expected: In fastText embeddings, no significant bias was measured for Turkish and Polish origin. GloVe embeddings, on the other hand, showed significant bias for all three nationalities. The highest bias in both sets of embeddings was measured for Turkish origin. Especially for the GloVe embeddings analysed, the bias captured by the WEAT corresponded to experiences of discrimination made by people with a migration background in Germany.

The performances of the debiasing algorithms were heterogeneous: The Hard Debiasing approach proposed by Bolukbasi et al. (2016) achieved somewhat successful bias mitigation, whereas the DD-GloVe model propsed by An et al. (2022) did not produce improved results.

Hard Debiasing was able to raise p-values above the significance level for Turkish bias in fastText embeddings and Italian bias in GloVe embeddings, and otherwise also reduced the measured bias, but not to such a great extent that no statistically significant bias was measured afterwards. Paralleling my WEAT adaption, a challenge in adapting this method for German embeddings and origin bias was the creation of fitting seed lists, which necessarily introduce subjectivity into the process. The results of the algorithm might be improved by expanding and validating the seed lists I created with a human study.

DD-GloVe achieved barely any reduction in bias across all experiments I carried out. Further investigative experiments showed that the dictionary data needed for the method was likely not suitable to the task of identifying origin-related versus origin-neutral words and therefore the model was not able to learn which parts of the embeddings should be debiased. This along with other discovered weaknesses such as debatable methods of calculating definition embeddings might have led to the undesired outputs of the seed word generation algorithm of DD-GloVe. An attempted reproduction of the experiments in An et al. (2022) under the same settings was not successful, possibly due to differences in training data and hyperparameters.

The failure of my DD-GloVe reproduction demonstrates the importance of the reproducibility of scientific studies. In addition to training code and hyperparameters used, authors in bias research should also consistently specify the seed words used for their algorithms and the exact procedures of how these seed words were procured. This is especially important to avoid various issues of "bad seeds" as discovered by Antoniak and Mimno (2021).

Many works on bias mitigation claim that their methods should, in theory, be applicable to other bias attributes and languages even when they carried out experiments only on English embeddings and for gender bias. An et al. (2022) and Bolukbasi et al. (2016) are among such works. In view of my results, these claims do not hold up. I therefore encourage future work to treat bias attributes other than gender not just as a theoretical matter, but as a real issue with large potential societal impact.

In this thesis, I first explored which kind of debiasing might be promising for origin bias in German embeddings and where potential problems lie. I thereby was able to provide some guidance for future work which might focus more extensively on optimizing debiasing performance. In particular, my findings showed that dictionary-based methods are challenging to adapt for this purpose and that the Hard Debiasing approach might be a promising avenue. Since frequency plays a significant role in the measurement and mitigation of bias, the Double-Hard Debiasing method (Wang et al., 2020), which accounts for frequency information in embeddings, might be an interesting approach to consider.

Lastly, the partly disparate bias values I measured for different nationalities demonstrate that it is important to consider exactly which bias attribute and dimensions researchers should analyse and mitigate. Gender and race, for example, have been assumed as biased attributes, but might be too general categories. It would be worthwhile for future research to examine more closely the motivations for debiasing and, if necessary, redefine bias categories accordingly.

All together, the strides already made towards a fairer NLP world are commendable. However, in the context of the rising influence of natural language technologies, researchers should not settle for simple solutions to a complex problem. Bias is multi-faceted and should be measured from multiple points of view; different kinds of bias must be treated differently; and the challenges of languages other than English should not be neglected in bias research, since bias differs across cultures and languages.

Bibliography

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Haozhe An, Xiaojiang Liu, and Donald Zhang. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, 2022.

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94 (4):991–1013, 2004.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, October 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 2023-11-28.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678. Association for Computational Linguistics, July 2019. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, page 231–232. Association for Computing Machinery, 2020. doi: 10.1145/3372923.3404804. URL https://doi.org/10.1145/3372923.3404804.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463, 2015.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*, 2019.

Ifeoma Ajunwa. The paradox of automation as anti-bias intervention. *Cardozo L. Rev.*, 41:1671, 2019.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

Brookings Institution. To stop algorithmic bias, we first have to define it, 2023. URL https://www.brookings.edu/articles/to-stop-algorithmic-bias-we-first-have-to-define-it/. Accessed 2023-07-13.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347, 1996.

Merriam-Webster. Bias, 2023. URL https://www.merriam-webster.com/dictionary/bias?utm_campaign=sd&utm_medium=serp&utm_source=jsonld. Accessed 2023-12-11.

Calvin K. Lai, Kelly M. Hoffman, and Brian A. Nosek. Reducing implicit prejudice. *Social and Personality Psychology Compass*, 7(5):315–330, 2013. doi: 10.1111/spc3.12023. URL https://compass.onlinelibrary.wiley.com/.

Richard Fletcher, Daniel Frey, Mike Teodorescu, Amit Ghandi, and Audace Nakeshimana. Exploring fairness in machine learning for international development, 2020. URL https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spri pages/module-three-framework/protected-attributes/. Accessed 2023-11-29.

James L. Hilton and William von Hippel. Stereotypes. *Annual Review of Psychology*, 47(1): 237–271, 1996. doi: 10.1146/annurev.psych.47.1.237. URL https://doi.org/10.1146/annurev.psych.47.1.237.

Gordon W. Allport, Kenneth Clark, and Thomas Pettigrew. *The Nature of Prejudice*. Addison-Wesley, 1954.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. How good is nlp? a sober look at nlp tasks through the lens of social impact. *arXiv preprint arXiv:2106.02359*, 2021.

Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Mascha Kurpicz-Briki. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. 2020.

Statista Research Department. Diskriminierung in deutschland nach diskriminierungsmerkmal 2022, July 2023. URL https://de.statista.com/statistik/daten/studie/1123809/umfrage/diskriminierung-in-deutschland-nach-diskriminierungsmerkmal/. (Accessed 2023-11-29).

Florian Pfisterer. Democratizing machine learning, October 2022. URL http://nbn-resolving.de/urn:nbn:de:bvb:19-309477.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003.

Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.431. URL https://aclanthology.org/2020.acl-main.431.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Julia El Zini and Mariette Awad. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31, 2022.

Chuan Li. Openai's gpt-3 language model: A technical overview, 2020. URL https://lambdalabs.com/blog/demystifying-gpt-3. Accessed 2023-11-29.

Philipp Dufter, Nora Kassner, and Hinrich Schütze. Static embeddings as efficient knowledge bases? *arXiv preprint arXiv:2104.07094*, 2021.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146, 2017.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Amir Bakarov. A survey of word embeddings evaluation methods. *CoRR*, abs/1801.09536, 2018. URL http://arxiv.org/abs/1801.09536.

Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414, 2001.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. 2009.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. URL https://doi.org/10.2307/1412159. Accessed 29 Nov. 2023.

Ira Leviant and Roi Reichart. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*, 2015.

Iryna Gurevych. Computing semantic relatedness across parts of speech. Technical report, Darmstadt University of Technology, Germany, Department of Computer Science, Telecooperation, 2006.

Francisco Vargas and Ryan Cotterell. Exploring the linear subspace hypothesis in gender bias mitigation. *arXiv preprint arXiv:2009.09435*, 2020.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018a.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

Haswanth Aekula, Sugam Garg, and Animesh Gupta. [re] double-hard debias: Tailoring word embeddings for gender bias mitigation. *arXiv preprint arXiv:2104.06973*, 2021.

Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, 2020.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. Araweat: Multidimensional analysis of biases in arabic word embeddings. *arXiv preprint arXiv:2011.01575*, 2020.

Davide Biasion, Alessandro Fabris, Gianmaria Silvello, and Gian Antonio Susto. Gender bias in italian word embeddings. In *CLiC-it*, 2020.

Chunlin Qin, Xin Zhang, Chaoran Zhou, and Yan Liu. An interactive method for measuring gender bias and evaluating bias in chinese word embeddings. In *International Conference on Computer Vision, Application, and Algorithm (CVAA 2022)*, volume 12613, pages 216–221. SPIE, 2023.

Meichun Jiao. Investigating gender bias in word embeddings for chinese, 2021.

Austin van Loon, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. Negative associations in word embeddings predict anti-black bias across regions–but only via name frequency. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1419–1424, 2022.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Frage: Frequency-agnostic word representation. *Advances in neural information processing systems*, 31, 2018.

Maria Antoniak and David Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, 2021.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170, 2022.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 446–457, 2020.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. Double-hard debias: Tailoring word embeddings for gender bias mitigation. *arXiv preprint arXiv:2005.00965*, 2020.

Masahiro Kaneko and Danushka Bollegala. Dictionary-based debiasing of pre-trained word embeddings. *arXiv preprint arXiv:2101.09525*, 2021.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202, 2020.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*, 2019.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Kaytlin Chaloner and Alfredo Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, 2019.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

Shlomo S Sawilowsky. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8(2):26, 2009.

Bundeszentrale für politische Bildung. Ausländische bevölkerung nach staatsangehörigkeit, 2022. URL https://www.bpb.de/kurz-knapp/zahlen-und-fakten/soziale-situation-in-deutschland/61631/auslaendische-bevoelkerung-nach-staatsangehoerigkeit/. Accessed 2023-12-05.

Charlotte Frank. Beliebte kindernamen: Der namenjäger, 3 2013. URL https://www.sueddeutsche.de/leben/beliebte-kindernamen-der-namenjaeger-1.1631646. Published in Süddeutsche Zeitung. Accessed 2023-12-06.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Wolfgang Werner Sauer. *Der »DUDEN«. Geschichte und Aktualität eines »Volkswörterbuchs«*. Springer-Verlag (originally J. B. Metzlersche Verlagsbuchhandlung and Carl Ernst Poeschel Verlag), Stuttgart, 1988. ISBN 978-3-476-00638-7. Limited preview available in the Google Book Search.

M Lynne Murphy. Defining racial labels: Problems and promise in american dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 13(1):43–64, 1991.

M Lynne Murphy. Defining people: race and ethnicity in south african english dictionaries. *International Journal of Lexicography*, 11(1):1–33, 1998.

Boris Forthmann and Philipp Doebler. Fifty years later and still working: Rediscovering paulus et al.'s (1970) automated scoring of divergent thinking tests. 2022.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226, 2009.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018b.

Oliver Razum, Hajo Zeeb, and Seval Akgün. How useful is a name-based algorithm in health research among turkish migrants in germany? *Tropical Medicine & International Health*, 6 (8):654–661, 2001. doi: https://doi.org/10.1046/j.1365-3156.2001.00760.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-3156.2001.00760.x.

Marvin M. Brinkmann, Frederic Gerdon, and Simon Kühne. Diskriminierungswahrnehmung und herkunftsregion. *DeZIMinutes*, (13), January 2023. URL https://www.dezim-institut.de/fileadmin/user_upload/Demo_FIS/publikation_pdf/FA-5504.pdf. Accessed on: 2024-01-21.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. *arXiv preprint arXiv:2103.07929*, 2021.