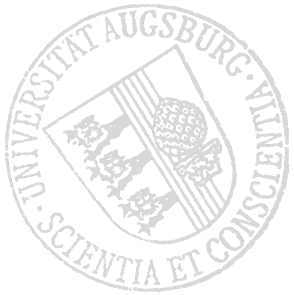
Universität Augsburg
Fakultät für Angewandte Informatik

Bachelorarbeit
für die Prüfung zum Bachelor of Science
im Studiengang Informatik und Multimedia

Erkennung von rassistischem Bias in Deutschen Word Embeddings mithilfe Analogie-Findungsaufgaben

Jarl Hengstmengel



Jarl Hengstmengel

Erkennung von rassistischem Bias in Deutschen Word Embeddings mithilfe Analogie-Findungsaufgaben

Erstprüferin:	**Prof. Dr. Elisabeth André**, Fakultät für Angewandte Informatik, Universität Augsburg
Zweitprüfer:	**Prof. Dr.-Ing. habil. Björn Schuller**, Fakultät für Angewandte Informatik, Universität Augsburg
Betreuerin:	**Stina Klein**, Fakultät für Angewandte Informatik, Universität Augsburg
Matrikelnummer:	1530788
Abgabedatum:	22. September 2023

Zusammenfassung

Zusammenfassung

Es wurde wiederholt gezeigt das Word2Vec Embeddings sozialen Bias übernehmen und reproduzieren. Bisherige Forschung fokusiert sich aber auf englischsprachige Word Embeddings oder untersucht insbesondere Gender Bias. Deshalb untersucht diese Arbeit Methoden zur Identifizierung und Analyse von rassistischen Bias in deutschen Word Embeddings. Dabei wird auf Analogie-Findungs-Eigenschaften zurückgegriffen und ein fastText Embedding genauer analysiert. Mittels einer Umfrage und Bias definierenden Begriffen, wird in Analogien mit Bildungsabschlüssen ein rassistisscher Bias festgestellt. Von 36 automatisch generierter Analogien, werden 26 als rassistisch bewertet. Über eine Hauptkomonentenanalyse wird ein Bias Subspace ermittelt und ein direkter Bias in Schulabschlüssen gemessen. Dabei erreichen diese einen direkten Bias Wert von 0,08. Dies bedeutet, dass die Schulabschlüsse einen nicht unerheblichen Bias Anteil in ihrer Vektorrepräsentation besitzen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

NLP Natural Language Processing

NaDiRa Nationaler Diskriminierungs- und Rassismusmonitor

CBOW Continuous Bag-of-Words

PCA Principal Component Analysis

1 Einleitung

1.1 Motivation

Die Integreat-App wurde im Zuge der hohen Fluchtbewegung 2015 entwickelt mit dem Ziel, geflüchteten Menschen möglichst niedrigschwellig alle relevanten Informationen zukommen zu lassen. Herausgekommen ist eine App, welche von Kommunen genutzt wird, um Informationen an zugewanderte Menschen zu verbreiten. Die Entwicklung verfolgt dabei einen Open-Source-Ansatz. Ein zentraler Teil der App sind Übersetzungen von Inhalten, welche durch Kommunen zur Verfügung gestellt werden. Dabei werden die Inhalte durch externe Übersetzungsbüros übersetzt. Zentrale Ziele sind dabei Barrierefreiheit, Inklusivität und Diskriminierungsfreiheit.

Um die Veröffentlichung schnelllebiger Inhalte zu ermöglichen, bietet die App seit Juli 2023 die Funktion, Inhalte direkt zu veröffentlichen. Um den zeitintensiven Redaktions- und Übersetzungsprozess zu verkürzen, wird zur Übersetzung die Übersetzungssoftware DeepL verwendet. Damit ein gewisser Qualitätsstandard eingehalten wird, wird der von der Kommune verfasste Inhalt mittels Hohenheimer-Verständlichkeitsindex auf die Verständlichkeit hin bewertet. Bei diesem Vorgehen fehlt aber eine Überprüfung auf Diskriminierungsfreiheit. Dies ist insbesondere deshalb ein Problem für eine Anwendung wie die Integreat-App, da ihre Zielgruppe kürzlich nach Deutschland migrierte Menschen ist, welche nicht selten Teil rassistisch diskriminierter Gruppen sind.

Für eine Bewertung auf Diskriminierungsfreiheit muss darauf geachtet werden, dass das Ziel insbesondere ist, dass auch automatisch übersetzte Inhalte diese Anforderung erfüllen müssen. Es ist also relevant, wie ein Übersetzungsmodel die Eingaben interpretiert. Eine Überprüfung auf Diskriminierung eines Inhaltes sollte an einer Stelle stattfinden, die offenlegt, wie das Übersetzungsmodel den Inhalt interpretieren bzw. assoziieren könnte.

1.2 Problembeschreibung und Zielstellung

Viele Anwendungen aus dem Bereich von Natural Language Processing (NLP) wie Übersetzungsprogramme arbeiten mit Word Embeddings. Dies sind mehrdimensionale Vektoren, welche in Wörtern enthaltene Informationen oft mittels eine Co-Occurrence Matrix lernen, sodass diese Informationen für weitere NLP-Anwendungen interpretierbar werden (Jurafsky und Martin, 2023). Embeddings werden mittels Machine Learning trainiert. Sie übernehmen den im Trainingsdatensatz enthaltenen Bias, also

Stereotypen und Vorurteile, welche die Gefahr für Diskriminierung bestimmter Gruppen birgt (Caliskan et al., 2017).

Es existiert Forschung dazu, wie Bias in Embeddings gefunden und negative Auswirkungen verringert werden können. Diese fokussiert sich auf Embeddings, welche auf amerikanisches Englisch ausgerichtet sind. Diese sind nicht optimal ins Deutsche übertragbar. Es gibt Embeddings, welche deutlich geeigneter für die Verwendung mit deutscher Sprache sind (Grave et al., 2018). Es existieren relativ wenige Arbeiten dazu, inwiefern diese diskriminierende Assoziationen übernehmen. Hinzu kommt, dass das Thema Rassismus in Embeddings – wenn überhaupt – bisher meist nur am Rande untersucht wurde. Die vorhandenen Arbeiten fokussieren sich überwiegend auf Assoziationen mit Geschlecht (Field et al., 2021).

Das Ziel dieser Arbeit ist die Untersuchung, wie rassistische Assoziationen in deutschen Word Embeddings festgestellt und analysiert werden können. Dabei wird auf die Fähigkeit von Embeddings zurückgegriffen, Beziehungen zwischen Wörtern mit hoher Genauigkeit darzustellen und wird dem Embedding Analogiefindungsaufgaben gestellt, welche mithilfe einer Online-Umfrage auf einen rassistischen Bias hin untersucht werden. Anschließend wird das Embedding auf rassistischen Bias untersucht und geprüft, inwiefern Begriffe in Zusammenhang mit Bildungsabschlüssen eine rassistische Konnotation im Embedding erhalten.

1.3 Vergleichbare Arbeiten

Allen vorweg muss hier die Arbeit von Bolukbasi et al. (2016) erwähnt werden, als erste Veröffentlichung in dem Bias in englischsprachigen Word2Vec Embeddings (vgl. Abschnitt 2.2.1) untersucht wird. Weitere erwähnenswerte Arbeiten die direkt daran anschließen sind Caliskan et al. (2017) welche in Anlehnung an den Implicit Association Test den Word Embedding Association Test entwickeln und Manzini et al. (2019) die die Methodik von Caliskan et al. ausweiten um Multiklassenbias zu finden. Garg et al. (2018) untersuchen Bias in Embeddings mit historische Daten um die Entwicklung von Bias in Sprache über 100 Jahre hinweg zu untersuchen.

Auch existieren Arbeiten welche Bias deutschsprachige Embeddings untersuchen. Chen et al. (2021) untersuchen Gender Bias in deutschsprachigen Word2Vec Embeddings, Papakyriakopoulos et al. (2020) betrachten Bias in Embeddings nachgelagerten Anwendungen, die Masterthesis von Kraft, 2021 und die Bachelorthesis von Gerard (2017).

Wichtige Arbeiten welche sich mit Bias und insbesondere auch mit Rassismus in NLP generell beschäftigen, sind Hanna et al. (2020), Field et al. (2021) und Bender et al. (2021).

2 Theoretische Grundlagen

Im Folgenden werden theoretische Grundlagen zum Finden und Analysieren von Rassismus in Embeddings eingeführt. Zuerst werden einige relevante theoretische Hintergründe zu Rassismus und der Diskurs in der deutschen Rassismusforschung betrachtet. Darauffolgend werden Zusammenhänge zwischen Rassismus und Bildung erläutert, um darzustellen, warum der Bereich für eine genauere Untersuchung bezüglich Rassismus interessant ist. Anschließend folgen Ausführungen zum Zusammenwirken von Sprache und Rassismus. Die Grundlagen zu Rassismus werden geschlossen mit Erläuterungen zur Wirkung von Rassismus in NLP und Möglichkeiten dagegen vorzugehen.

Für ein besseres Verständnis von Embeddings werden Grundlagen von Embeddings und Word2Vec-Modellen eingeführt, um zu verdeutlichen, wie diese Informationen aufgenommen, repräsentiert und schlussendlich wieder reproduziert werden. Darauf aufbauend wird näher darauf eingegangen, wie rassistische Diskriminierung in Embeddings dargestellt wird und wie diese erkannt und analysiert werden kann. Es werden für diese Analyse benötigte Bias definierende Begriffe definiert. Zum Abschluss des Kapitels zu theoretischen Grundlagen werden Bildungsabschlüsse als Kategorien festgelegt, um den in Embeddings enthaltenen rassistischen Bias in Abschnitt 2.3.3 weiter zu untersuchen.

2.1 Grundlagen zu Rassismus

2.1.1 Begriffsklärung

Für die Untersuchung von Rassismus werden zuerst einige Begriffe festgelegt und erläutert. Rassismus ist ein aktuelles Thema, das auch in der Forschung einem konstanten Wandel und Diskurs unterliegt. Aus diesem Grund ist das Definieren der wichtigsten Begriffe relevant.

Rassismus und Rassifizierung

Die Definition von Rassismus orientiert sich an Arndt (2021) und Memmi (1987). Ihrem Verständnis zufolge entsteht Rassismus um die Konstruktion von körperlichen Unterschieden aus ökonomischen und politischen Machtbestrebungen. Aus diesen vermeintlich „natürlichen“ Unterschieden werden statische und vermeintlich objektive Kriterien zur Unterscheidung von Menschen aufgestellt. Diesen Kriterien werden soziale, kulturelle und religiöse Eigenschaften sowie Verhaltensmuster zugewiesen und dann verallgemeinert, verabsolutiert und hierarchisiert. Rassifizierung beschreibt wiederum den Prozess, in Zuge dessen einer Person bzw. einer Gruppe Eigenschaften

zugeschrieben, sie verallgemeinert und hierarchisiert werden (Banton, 1977 und Miles, 2004).

Intersektionalität und Mehrfachdiskriminierung

Intersektionalität wurde insbesondere durch die Arbeit von Crenshaw (1989) geprägt. Der Begriff beschreibt das Ineinandergreifen mehrer Diskriminierungsformen zu einer Diskriminierungserfahrung, welche anders wirkt, als die jeweiligen Diskriminierungsformen für sich genommen. Ein klassisches Beispiel sind hier Diskriminierungserfahrungen von Schwarzen Frauen mit Rassismus und Sexismus, welche beim Zusammenfallen eine Diskriminierungserfahrung bilden, die durch keine der beiden Diskriminierungsformen vollständig beschreibbar ist. Die Notwendigkeit dieser konkreten Erfassung der intersektionalen Betroffenheit durch mehrere Diskriminierungsformen wird im Fehlen von entsprechenden Erfassungsmöglichkeiten, wie beispielsweise juristischer Terminologie, gesehen.

Der Diskurs endet aber nicht beim Begriff der Intersektionalen Diskriminierung. Makkonen (2002) unterscheidet neben der intersektionalen Diskriminierung auch zwischen mehrfacher und verbundener Diskriminierung. Mehrfache Diskriminierung beschreibt dabei eine Form, bei der sich die generelle Betroffenheit von mehreren Diskriminierungsformen zu einer Gesamtlage für eine Person akkumuliert. Verbundene Diskriminierung wirkt eher additiv von mehreren Diskriminierungsformen, von denen eine Person zum gleichen Zeitpunkt betroffen ist. Im deutschen Sprachraum setzt sich der Begriff der mehrdimensionalen Diskriminierung im rechtswissenschaftlichen Diskurs durch (Marten und Walgennbach, 2017). In den Sozialwissenschaften wird viel auf Begriffe wie Verschränkung, Schnittpunkte, Durchkreuzungen, Überschneidungen und Achsen zurückgegriffen. Dabei werden vor allem die Wechselbeziehungen verschiedener Formen von Diskriminierung zueinander erforscht (Marten und Walgennbach, 2017). In englischer Literatur, findet sich oft der Begriff der Mehrdimensionalität.

Race und Mehrdimensionalität

Im englischsprachigen Diskurs findet sich oft der Begriff Race. Die direkte Übersetzung des Begriffs ins Deutsche wäre der Begriff Rasse. Durch den historischen Kontext des Nationalsozialismus und die daraus resultierende negative Besetzung des Begriffs, folgt ein Ausbleiben eines ausgeweiteten Diskurses und teilweise Umdeutung für die Beschreibung der sozialen Konstruktion, wie es der Begriff Race tut (Arndt und Hornscheidt, 2004). Da es dadurch kein äquivalentes Konzept im Deutschen gibt, welches uns eine gleichwertige Übersetzung bietet, wird in dieser Arbeit Stellenweise Race verwendet. Race ist nicht ein eindimensionales Konzept. Der Begriff beschreibt eine mehrdimensionale soziale Konstruktion, bei der je nach Kontext bestimmte Dimensionen wichtiger sind, zu beleuchten bzw. in denen es angemessen ist, eine Dimension zu untersuchen und abzufragen. Die verwendete Dimension kann bei der Untersuchung von Rassismus einen signifikanten Einfluss auf die Ergebnisse haben (Roth, 2016). Die folgende Auflistung beschreibt die Dimensionen Racial Identity, Racial Self-Classification,

Observed Race, Reflected Race und Phenotype nach Hanna et al. (2020). Sie exkludieren dabei die von Roth noch mitaufgenommene Dimension Racial Ancestry aufgrund der Kritik an der Verwendung von genetischer Geschichte in der biomedizinischen Forschung.

1. *Racial Identity* - Eigene subjektive Identität. Abfrage über Self-ID, also die freiwillige und offene Abfrage der Identität. Angemessene Untersuchungskontexte sind z.B.: Politische Mobilisation, Assimilation, soziale Netzwerke, Wahlen, lokale Entscheidungsfindung, Meinungsabfragen.

2. *Racial self-classification* - Selbstgewählte Klassifikation bei vorgegebenen Optionen wie auf Formularen. Abfrage mit geschlossenen Antwortoptionen. Angemessene Untersuchungskontexte sind z.B.: demographische Veränderungen, Gesundheitsstatistik, Sterbe- und Krankheitsraten.

3. *Observed Race* - Race, wie Dritte dies wahrnehmen. Klassifikation im Rahmen von Interviews. Angemessene Untersuchungskontexte sind z.B.: Diskriminierung, sozioökonomische Ungleichheiten, Segregation, Justiz, Verfügbarkeit von Gesundheitsleistungen.
    a) *Appearance-Based* - Observed Race basierend auf äußerlichen Charakteristiken. Klassifizierung im Rahmen von Interviews, basierend auf dem ersten Eindruck. Angemessene Untersuchungskontexte sind z.B.: Racial Profiling, Diskriminierung im öffentlichem Raum.
    b) *Interaction-based* - Wahrnehmung von Race auf Basis von Charakteristiken in der Interaktion, wie Sprachgebrauch, Akzent oder Name. Klassifizierung im Rahmen von Interviews mit Einschätzung nach der Interaktion. Angemessene Untersuchungskontexte sind z.B.: Diskriminierung bei Wohnungssuche oder am Arbeitsplatz, sprachbasierte Diskriminierung.

4. *Reflected race* - Race, wie eine Person selbst denkt, dass sie durch andere wahrgenommen wird. Abfrage über Eigeneinschätzung der befragten Person. Angemessene Untersuchungskontexte sind z.B.: Self-ID Prozesse, wahrgenommene Diskriminierung.

5. *Phenotype* - Phänotyp. Klassifizierung durch Interviewer*in, dabei oft basierend auf „objektiven“ Charakteristiken. Angemessene Untersuchungskontexte sind z.B.: Diskriminierung, sozioökonomische Ungleichheiten, Justiz, Verfügbarkeit von Gesundheitsleistungen.

2.1.2 Analysekategorien

In diesem Absatz werden einige Analysekategorien vorgestellt, welche die Grundlage für Parameter für die Analyse eines Embeddings in Kapitel 3 bilden.

Verschiede Formen des Rassismus

Nachfolgend werden verschiedene Rassismusformen bzw. von Rassismus betroffene Gruppen beschrieben. Dabei werden, wenn nicht anders gekennzeichnet, die Definitionen des Nationalen Diskriminierungs- und Rassismusmonitors (NaDiRa) von Foroutan et al. (2022) verwendet, welche eine Vielzahl an Quellen zusammentragen, um ihre Definitionen zu formulieren. Die in Abschnitt 2.3.2 aufgestellten Bias definierenden Begriffe basieren auf diesen Definitionen.

Antischwarzer Rassismus
Antischwarzer Rassismus geht zurück auf Zeiten der Kolonialisierung und Versklavung. Rassistische Zuschreibungen wurden dazu genutzt, Kolonialisierung, Versklavung und Ausbeutung des afrikanischen Kontinents zu legitimieren. Verbrechen wie der Genozid an den Herero und Nama wurden durch bedrohliche Zuschreibungen mit dem Auflösen dieser angeblichen Bedrohung gerechtfertigt. Die aus dieser Zeit entspringenden Zuschreibungen wirken bis heute nach und manifestieren sich im gegenwärtigen antischwarzen Rassismus. Schwarze dienen außerdem als Projektionsfläche für negative Eigenschaften. Weiterhin existierende Zuschreibungen sind z.B. die Herabwürdigung von Schwarzer Körperlichkeit, Exotisierung, Hypersexualisierung und die Zuordnung bedrohlicher Eigenschaften.

Antisemitismus
Im Gegensatz zu anderen Rassismen gehen mit dem Antisemitismus Machtzuschreibungen gegenüber Jüdinnen*Juden einher. Deshalb wird Antisemitismus oft abgegrenzt von anderen Formen rassistischer Diskriminierung. Geschichtliche Zuschreibungen lassen sich grob in zwei Phasen aufteilen. Einerseits eine christlich motivierte Judenfeindschaft, aus der einige Mythen hervorgingen wie Hostienschändung, Brunnenvergiftung und Ritualmord. Andererseits der völkisch motivierte Antisemitismus des 19. Jahrhunderts, welcher u.a. als Gegenbewegung zu Emanzipationsbestrebungen aufgefasst wird. Jüdinnen*Juden erhielten die Zuschreibung, gleichermaßen für Kapitalismus, Kommunismus, Feminismus und Liberalismus verantwortlich zu sein und als Minderheit die Mehrheit dominieren zu wollen. Ein Element, welches sich seit der Vormoderne konstant mitzieht, ist die biologistische Vorstellung, Jüdinnen*Juden seien von „Natur“ aus nicht „bekehrbar“.

Antimuslimischer Rassismus
Diese Form des Rassismus konstruiert eine Gruppe, der alle Personen zugeordnet werden, die, unabhängig von der tatsächlichen Religionszugehörigkeit, mit einem dem Islam verbundenen Bild assoziiert werden. Dieser Fokus resultiert aus einer Wahrnehmungsverschiebung, bei der früher als „Gastarbeiter*innen“ oder „Ausländer*innen“ wahrgenommenen Personen, als Muslim*innen eingeordnet werden. Dabei gelten sie als „unintegrierbar“ und als Abstammungsgemeinschaft. Die Abstammung wird als entscheidender Faktor herangezogen, ob Personen als muslimisch gelten oder nicht. Es folgen Zuschreibungen wie „Gefahr“ oder „Bedrohung“ auf Personen, welche auf-

grund ihres Aussehens oder Namens dieser angeblichen Abstammungsgemeinschaft zugeordnet werden. Dabei wird ein Feindbild von „westlicher“ bzw. „abendländisch-christlicher“ Kultur gegen „islamischer“ Kultur konstruiert. „Islamische“ Kultur erhält in diesem Prozess Zuschreibungen wie „Rückständig“, „Unwandelbar“, „Irrational“, „Barbarisch“ und „Nicht demokratiekompatibel“.

Antiasiatischer Rassismus
Historisch begründet sich antiasiatischer Rassismus in Deutschland an verschiedene Ereignissen wie der deutschen Kolonialpolitik in China, Umgang mit Chines*innen im Nationalsozialismus, Arbeitsmigration in den 1950er Jahren, Fluchtmigration aus Vietnam und rassistische Gewalt in den 1990er Jahren. Dabei verschleiern vermeintlich positive Zuschreibungen des „Model-Minority-Mythos“ Diskriminierung. Während der Corona-Pandemie nahmen Rassismuserfahrungen zu.

Antislawischer Rassismus
In Deutschland geht antislawischer Rassismus zurück auf Vorstellungen eines monolithischen, geschichts- und kulturlosen Slawentums im 19. Jahrhundert. Der Nationalsozialismus verknüpfte dies dann mit Vorstellungen von „rassischer Minderwertigkeit“ um Expansionsvorstellungen zu legitimieren. Dies führte zur rassistischen Okkupations- und Germanisierungspolitik, der große Teile der Bevölkerung in Osteuropa zum Opfer fielen. Bis heute findet antislawischer Rassismus wenig Beachtung in Debatten um Rassismus.

Rassistische Diskriminierung von Sinti und Roma
Rassistische Diskriminierung von Sinti und Roma reicht mindestens zurück in das 15. Jahrhundert und wird seither durch den strukturellen Ausschluss dieser Gruppe aus der Gesellschaft gekennzeichnet, was im Nationalsozialismus in der gewaltsamen Verfolgung und dem Völkermord, dem Porajmos, gipfelte. Scherr (2017) beschreibt die Konstruktion dieser vermeintlich homogenen Gruppierung als geschichtlich fokussiert auf die fehlende Ortsansässigkeit von Personen. Dabei wurden die Ursachen von niedrigem sozioökonomischen Status u.a. in dem gleichen seit dem 18. Jahrhundert existierenden Vorurteil eines sog. „Wandertriebs”’ gesehen. Basierend auf diesem Vorurteil wurde fehlende bis nicht-mögliche Assimilation zum grundlegenden Hindernis zur Lösung der sozialen Frage erklärt, ungeachtet vom tatsächlichen Wahrheitsgehalt dieser Annahme. Viele Maßnahmen seit den 1950er Jahren fokussieren sich deshalb darauf, einen ambulanten Lebensstil zu unterbinden und so die Situation auf Arbeits- und Wohnungsmarkt zu verbessern. Dieses Muster zieht sich bis in die 1990er Jahre durch. Eine Folge dieser vermeintlichen Unterstützungsmaßnahmen, aber auch von offen rassistisch agierenden Behörden, ist ein oft niedriges Vertrauen in staatliche Institutionen (Scherr (2017) und Foroutan et al. (2022)). Mit dem Jugoslawienkrieg flüchten vermehrt Roma vom Balkan in die Bundesrepublik. Zu dieser Zeit nehmen auch Ressentiments in der Gesellschaft zu.

Weißsein als dominierende Gruppe
Die konkrete Benennung von durch Diskriminierung betroffenen Gruppen erfolgt in der Abgrenzung zur dominierenden Gruppe in der Gesellschaft. Die dominierende Gruppe nimmt sich selbst als Norm wahr. In der Folge findet keine Benennung durch sich selbst statt. Um das Verhältnis von diskriminierten Gruppen zur dominierenden Gruppe zu untersuchen und zu hinterfragen, muss die dominierende Gruppe auch benannt werden als das, was sie ist. Im Fall von Rassismus muss hier also das Weißsein auch als solches betitelt werden (Hornscheidt, 2005).

Diskriminierung und Rassismus im Bildungssystem

Viel Forschungsarbeit im Bereich von Rassismus in NLP findet beispielsweise bezogen auf Namen, Berufe oder Hate Speech statt. In dieser Arbeit wird der Bereich Bildung bzw. Bildungsabschlüsse untersucht. Deshalb werden im Folgenden Grundlagen zu Rassismus in der Bildung beleuchtet.

Hummrich (2017) beschreibt den geschichtlichen Rahmen des deutschen Bildungssystems. Das dreigliedrige deutsche Schulsystem geht auf die Klassenlogik des 19. Jahrhunderts zurück. In den 1920er Jahren wurden Ausländer*innen gesetzlich vom Bildungssystem ausgeschlossen. Die Bundesrepublik Deutschland drehte diese Schritte erst in den 1950er und 60er Jahren zurück und ein Großteil westdeutscher Bundesländer führte eine Schulpflicht für Kinder von sogenannten Gastarbeiter*innen ein. Ein gesetzlich abgesicherter Bildungszugang für Kinder von Asylbewerber*innen existiert flächendeckend erst seit 2012. Diese historisch strukturellen Voraussetzungen legen den Grundstein für Probleme mit verschiedenen Formen von Diskriminierung wie Rassismus im Zusammenhang mit dem deutschen Bildungssystem. Hummrich stellt fest, dass in der deutschen Antidiskriminierungsforschung insbesondere auch ein Zusammenhang zwischen Klassenwiederholungen und Rückstellungen zur sozialen Stellung und somit auch zum Bildungserfolg gesehen wird. Auch der Bildungshintergrund der Eltern wird als ein starker Faktor ausgemacht. Insgesamt gelten Herkunftseffekte als entscheidend dafür, welche Abschlüsse erzielt werden. Die folgenden Faktoren werden als maßgeblich dafür gesehen, wie der Bildungsweg verläuft:

1. Bildungsaspirationen der Eltern
2. Zusammenspiel von interaktiver und institutioneller Ausgrenzungserfahrung
3. Zurückweisungserfahrung bei nicht anschlussfähigem Habitus
4. Ausschluss von nicht anschlussfähigen Jugendlichen
5. Schulen präferieren bestimmte Milieus
6. Annahmen zu der Leistungsfähigkeit migrantisierter Menschen

Zusätzliche Faktoren, welche Hummrich ausmacht, sind personenspezifische Merkmale, die Unterscheidung nach Aufenthaltsstatus und das Nichtvorhandensein eines

direkten Migrationshintergrundes einer migrantisierten Minderheit. Die Vornamensstudie von Kaiser (2010) legt einen Zusammenhang nahe zwischen Annahmen von Lehrer*innen zu Schüler*innen aufgrund von Namen. Aber auch milieu-spezifischer Habitus beeinflusst Annahmen von Lehrkräften. Begriffe wie „Migrationshintergrund“, die Zuschreibung von Arbeitsmigrant*innen über Generationen hinweg, verfestigen Diskriminierung und Rassismus auf lange Sicht. Die Unterscheidung nach Aufenthaltsstatus beginnt beim vom Bildungssystem separierten Bildungszugang von Geflüchteten. Eine mögliche Folge ist Isolation. Beschulungsdauer in Kombination mit Alter beeinflusst den Zugang zu Folgeausbildungen. Kettenduldungen führen zu einer Unplanbarkeit des Bildungswegs. Zusätzlich leiden migrantisierte Gruppen ohne tatsächlichen Migrationshintergrund unter Diskriminierungsproblemen des Bildungssystems. Frühes Ausscheiden aus dem Bildungssystem von Roma und Sinti wird oft kulturell verklärt und die eigentlichen Probleme, wie der sozioökonomische Hintergrund und die erfahrene Diskriminierung, werden ignoriert. (Scherr, 2017).

Des weiteren konkretisiert Quehl (2010) die Probleme der Diskriminierung im Bildungssystem im Kontext Rassismus. Rassistische Normalität zeigt sich in einem Zusammenspiel der drei Ebenen: subjektiver Denk- und Handelsweisen, der sozialen Bedeutung und der gesellschaftlich-strukturellen Bedingungen. Auf der ersten Ebene vermitteln Lehrkräfte rassistisches Wissen als „soziale Erkenntnis“ an Schüler*innen. Die zweite Ebene vermittelt Dominanzverhältnisse der Migrationsgesellschaft und deren Diskurs und damit auch den zugehörigen rassistischen Diskurs. Auf der dritten Ebene zeigen sich gesellschaftlich-strukturelle Bedingungen. Darunter fällt die seit Jahren gleichbleibende Überrepräsentation von Schüler*innen mit Migrationshintergrund auf Sonderschulen mit Förderschwerpunkt Lernen, was auf eine strukturelle Problematik hindeutet.

2.1.3 Rassismus in Sprache

Es gibt die Tendenz Begriffe, welche als diskriminierend gekennzeichnet sind, nicht mehr zu verwenden. Neben vielen sexistischen Ausdrücken betrifft das insbesondere Wörter, die mit dem Nationalsozialismus in Verbindung gebracht werden. Oft wird in der Auseinandersetzung mit diesen Wörtern ihre Bedeutung hauptsächlich im Kontext des Nationalsozialismus betrachtet. Viele Wörter haben jedoch bereits eine diskriminierende Deutungsgeschichte davor - nicht selten liegt der Ursprung schon im Kolonialismus. Die reine Nicht-Verwendung von Begriffen führt aber nicht direkt zu einer Schwächung der mit ihnen transportierten Ideologie. Entsprechende Prozesse müssen breit angegangen werden. Die kritische Auseinandersetzung mit Sprache ist hierbei ein Werkzeug von vielen. Im Gegensatz zum Nationalsozialismus, findet in der Gesellschaft die Aufarbeitung des deutschen Kolonialismus kaum statt. Dieser Teil der Geschichte und seine Wirkmacht spiegeln sich in der deutschen Sprache meist unreflektiert wider. (Arndt und Hornscheidt, 2004)

Wann ist Sprache rassistisch?

Um zu bestimmen, ob bzw. inwiefern Wörter einen rassistischen Gehalt haben, schlagen Arndt und Hornscheidt (2004) folgende Strategien vor:

1. Suche nach Zitaten zur Illustration der historischen und aktuellen Verwendung des betreffenden Begriffes in gängigen Wörterbüchern, die u.a. die Kontinuität von diskriminierenden Erklärungsmustern demonstrieren können.
2. Verwendungsgeschichte des Wortes untersuchen.
3. Aktuelle Konnotation des Wortes untersuchen.
4. Interpretation von Wortzusammensetzungen und Redewendungen zur Bewusstmachung von Konnotationen des Begriffs sowie zur Illustration, wie rassistische Begriffe in Komposita und Redewendungen breite Verwendung finden.
5. Assoziationen der sprachnutzenden Personen betrachten.
6. Analogietest: Wäre Übertragung auf Weiße möglich?
7. Asymmetrie der Begriffsverwendungen untersuchen.
8. Derzeit mögliche und nach Kontexten differenzierte Alternativvorschläge zusammenstellen.

Rassismus ohne rassistische Wörter

Rassismus in Sprache steckt nicht nur in direkt rassistischen Wörtern. Oft bekommen Wörter erst eine rassistische Konnotation durch den Kontext, in dem sie vorkommen. Durch eine Übertragung auf marginalisierte Gruppen, ohne dass diese Begriffe für dominante Gruppen verwendet werden, entsteht beispielsweise eine Asymmetrie der Verhältnisse zwischen der dominanten und der marginalisierten Gruppe (Arndt und Hornscheidt, 2004). Hierfür sind Begriffe wie „Hütte“ oder „Dialekt“ gute Beispiele. Auch Begriffe, welche in kolonialistischen und rassistischen Konstruktionen eine Rolle spielen, transportieren rassistische Ideen durch den verwendeten Kontext weiter, z.B. „primitiv“ oder „traditionell“. Reisigl (2017) beschreibt die folgenden Elemente mit denen diskriminierende und rassistische Ideen durch Sprache transportiert werden:

Diskriminierung über Sprechakte
Eine diskriminierende Gesellschaftsordnung bildet und erhält sich über verschiedene Sprechakte. Es existieren folgende Sprechakte: Assertive (z.B. Behauptungen), Fragende, Direktive (z.B. Aufforderungen), Kommissive (z.B. Versprechen), Expressive (z.B. Beschimpfungen) und Deklarative (z.B. Anweisungen). Über Assertive Sprechakte werden z.B. negative Stereotype und Vorurteile geäußert. Über Direktive Sprechakte reproduziert man die diskriminierende Gesellschaftsordnung über negierende und mobilisierende Sprechakte. Negierung findet über entsprechende Verbote statt. Eine

Mobilisierung wird über Aufforderungen verwirklicht. Aufforderungen regeln in einer Gesellschaft welche Personengruppe welchen Zugang zu welchen Orten und Räume hat. Orte und Räume können hier als physisch abgegrenzte Orte, aber auch als gesellschaftlich konstruierte Räume verstanden werden. Deklarative Sprechakte schaffen und erhalten so eine diskriminierende Gesellschaft. So wird die Gesellschaftsordnung über Regeln und Gesetze zementiert.

Diskriminierende Nominationen
Diskriminierende Personenbezeichnungen gelten als die mitunter offensichtlichsten Mittel, um Personengruppen zu diskriminieren. Bei Nominationen müssen indessen einige Nuancen beachtet werden: Es muss festgestellt werden, ob eine Überindividualisierung stattfindet, also ob eine Stigmatisierung auf eine ganze Gruppe bezogen erfolgt. Auch relevant ist, inwiefern diese in der Breite der Gesellschaft verankert ist, ob eine Asymmetrie oder eine Depersonalisierung vorhanden sind.

Diskriminierende Prädikationen
Diskriminierende Prädikationen sind diskriminierende Zuschreibungen, z.B. wie die Zuschreibung von Eigenschaften, Merkmalen und Qualitäten. Diese werden mit vielen sprachlichen Instrumenten direkt oder indirekt transportiert. So kommen verschiedene Formen von Nominationen zum Einsatz wie Vergleiche, Attribute, Metaphern und andere rhetorische Figuren und Mitteln.

Diskriminierende Argumentationen
Diskriminierende Argumentationen sind ein zentrales Mittel zur Stützung von Diskriminierung in Diskursen, insbesondere durch Topoi, also wiederkehrende Argumentationsmuster. Diskriminierende Topoi dienen u.a. zur Überindividualisierung bestimmter Gruppen. Dadurch kommt es oft zu Fehlern in der Argumentation und daraus resultierenden Trugschlüssen.

2.1.4 Rassismus in Natural Language Processing

NLP gewinnt zunehmend an Bedeutung in Bereichen in denen der ethische Anspruch an NLP-Systemen zunehmend steigt. Der Anteil an NLP-Anwendungen die direkt mit Nutzer*innen interagieren, welche keine Expert*innen in Linguistik bzw. Computerlinguistik sind, steigt immer weiter. Ebenso erreicht mit Chat GPT eine solche Anwendung erstmals eine breitere Öffentlichkeit. Dadurch wird die Frage danach, inwiefern solche Anwendungen Diskriminierung reproduzieren oder sogar verstärken, weiter in den Mittelpunkt gerückt. Dabei ist Chat GPT nicht die erste Anwendung mit breiter Nutzer*innen-Basis, welche Teil einer kritischen Diskussion über sozialen Bias in NLP-Systemen wird. Google-Translate etwa gerät immer wieder in den Blick von Forscher*innen, ob dessen Übersetzungsvorschläge einen gefährlichen Bias übernehmen und reproduzieren (Prates et al., 2020, Sólmundsdóttir et al., 2022). Im Jahr 2016 rufen Hovy und Spruit (2016) dazu auf, die sozialen Auswirkungen von NLP-Anwendungen zu erforschen und ethische Grundprinzipien bzw. Leitlinien für NLP

zu entwickeln. Sie stellen drei Problemfelder fest: *Exklusion und Falschdarstellung*, *Übergeneralisierung* und *Unter- bzw. Überexposition*.

Im Folgenden werden zuerst Probleme von NLP mit Rassismus betrachtet anhand der klassischen NLP-Pipeline. Danach wird eine kritische Methodik zum Umgang mit Rassistischen Bias eingeführt.

Rassismus in der NLP-Pipeline

Dieser Abschnitt folgt hauptsächlich den Ausführungen von Field et al. (2021). Sie orientieren sich in ihrer Analyse an einer klassischen NLP-Pipeline und betrachten, an welchen Stellen Probleme mit Bias auftauchen können. Eine klassische NLP-Pipeline besteht aus folgenden Komponenten:

- Daten
- Datenannotation
- Modell
- Output
- Analyse des Outputs

Daten und -Annotation

Daten - NLP-Systeme können sozialen Bias aus Daten übernehmen und auch verstärken. Insbesondere zwei Fragestellungen zu den verwendeten Daten tun sich auf:

- *Wie werden Minderheiten durch die Trainingsdaten beschrieben?* Vor allem Embeddings sind hier ein Untersuchungsgegenstand. Wie in dieser Arbeit auch diskutiert wird (vgl. 2.3), neigen Embeddings dazu sozialen Bias und damit auch rassistischen Bias aus den ihnen vorhanden Daten zu übernehmen. Die Darstellung von marginalisierten Gruppen reproduziert sich in der Darstellung durch Embeddings (Bolukbasi et al., 2016, Chen et al., 2021, Caliskan et al., 2017; Manzini et al., 2019). Auf Embeddings aufbauende Anwendungen können diesen Bias wiederum übernehmen (Kiritchenko und Mohammad, 2018, Davidson et al., 2019).
- *Wer hat die Daten erstellt?* Personen, welche die genutzten Datensätze zusammenstellen, sind selbst nicht frei von sozialen Bias. Dies können sie auch auf die Daten übertragen. Ein interessantes Beispiel, was auch für diese Arbeit relevant ist, ist Wikipedia. Viele Modelle greifen auf Wikipedia als Datengrundlage zurück. Der Großteil der Verfasser*innen ist oft Weiß bzw. stammt aus Ländern in denen Weiß die dominierende Gruppe ist. In der Folge hat Wikipedia einen systematischen rassistischen Bias in der inhaltlichen Abdeckung (Adams et al., 2019, Field et al., 2022).

Datenannotation - Auch der Annotationsprozess wird von sozialem Bias beeinflusst. Schemata die zur Annotation aufgestellt und verwendet werden, neigen immer wieder dazu Sprachvariationen nicht miteinzubeziehen (Blodgett et al., 2018). Anweisungen zur Annotation beeinflussen wie Annotator*innen Daten labeln. Rassistischer Bias in Instruktionen, auch in Form der Nichtbeachtung von Variation, führt zu einen rassitischen Bias in den Labeln (Sap et al., 2019). Und schlussendlich ist die Auswahl von Annotator*innen selbst nicht frei von Bias. Aktivist*innen gegen Rassismus labeln oft anders als Arbeiter*innen (Waseem, 2016).

Exklusion und Falschdarstellung - Eine Folge von sozialen Bias in Daten und deren Annotationsprozess, sind Exklusion und Falschdarstellung von Minderheiten (Hovy und Spruit, 2016). Der Bias entsteht in den Datensätzen sowie im Annotationsprozess ähnlich, sie werden hauptsächlich von und aus der Perspektive von westlich geprägten, gut ausgebildeten, industrialisierten, eher reichen und demokratisch geprägten Personen erstellt bzw. gestaltet. Die Konsequenz ist in beiden Fällen ähnlich. Es entsteht ein Overfitting auf die Perspektiven dieser Personen. Für Minderheiten schneiden NLP-Anwendungen dann qualitativ schlechter ab, stellen sie falsch dar oder exkludieren sie (Hovy und Spruit, 2016).

Modelle

Modelle - Die Repräsentation von marginalisierten Gruppen durch Modelle wird stark davon beeinflusst, welche Personen diese entwickeln, aber auch von Personen in Entscheidungspositionen (Field et al., 2021). Dazu kommt aber auch eine große Abhängigkeit von verwendeten Parametern. Zufällig generierte Seeds für die Initialisierung von Variablen können hier einen Einfluss haben (Sommerauer und Fokkens, 2019). Eine Untersuchung von rassistischen und Gender Bias in 200 Sentiment Analyses systems mit standardisierten Daten stellte in allen Modellen ein ähnliches Level an Bias fest (Kiritchenko und Mohammad, 2018).

Übergeneralisierung - Ein Problem mit sozialem Bias in der Architektur der Modelle, ist eine mögliche Übergeneralisierung von marginalisierten Gruppen. Hieraus folgen immer wieder False-Positives, bei dem durch Generalisierung über Attribute immer wieder falsche Annahmen über Nutzer*innen getroffen werden. Hier muss die Frage gestellt werden, ob keine Antwort in manchen Fällen besser ist, als eine falsche Antwort (Hovy und Spruit, 2016).

Output und Analyse

Output - Was am Ende als Output einer NLP-Anwendung steht und welche Auswirkungen dieser Output auf Nutzer*innen hat, ist ein zu debattierendes Thema. Diverse Systeme neigen zu diskriminierendem Output. Klassifizierer für Hate Speech bewerten Nicht-Standard-Englisch schneller als Hate Speech (Davidson et al., 2019, Sap et al., 2019). Identitätsterminologie wie „Black“ wird schnell als verletzend bewertet (Dixon et al., 2018). GPT generiert Textoutput mit negativen Sentiment für afroamerikanisches Englisch (Groenwold et al., 2020).

Soziale Analyse des Outputs - NLP-Systeme finden auch Anwendung in der Analyse von Rassismus und Gesellschaft. Auch diese Arbeiten sind anfällig für rassistischen Bias in den Systemen. Alles vorherig Genannte findet auch hier Anwendung. Eine Strategie dies transparent zu machen, ist die Verwendung von Einordnungen der eigenen Perspektive (Field et al., 2021).

Weitere Problemfelder

Neben Probleme bezogen auf eine klassische NLP-Pipeline, existieren noch einige weitere generelle Probleme.

Fehlende Breite der Daten - NLP-Forschung, wie der NLP-Bereich generell, ist fixiert auf einige wenige Datensätze. Diese werden immer wieder mit wenig kritischer Auseinandersetzung übernommen. Field et al. (2021) identifizieren insbesondere drei Datenquellen, welche immer wieder verwendet werden: Labels basiernd auf U.S. Zensus bzgl. Rassismus und Ethnizität (Blodgett et al., 2016), Namenslisten von Sweeney (2013), Caliskan et al. (2017) oder Garg et al. (2018), und explizite Begriffe wie „Black Woman“ für Tests der Performance von Modellen. Durch die Verwendung der immergleichen Datengrundlagen entsteht ein verengtes Bild der multidimensionalen Aspekte von Rassismus und Diskriminierung (Field et al., 2021).

Klassifizierung basierend auf festen, eindimensionalen und U.S.-Zentrierten Label - Nicht selten werden Gruppen auf Basis des U.S.-Zensus definiert. Es findet kaum eine Auseinandersetzung mit durch rassifizierte Gruppen selbst gewählte Bezeichnungen statt. Zudem wird ignoriert, dass Bezeichnungen sich je Land und Sprache unterscheiden können. Datensätze werden oft unhinterfragt weiterverwendet, ohne eine größere Auseinandersetzung mit deren Limitierungen zu führen. Caliskan et al. (2017) entfernen z.B. einige als African American gelabelte Namen aufgrund ihres niedrigen Vorkommens. Dies kann die Effektivität von Debiasing beeinflussen. Hierauf aufbauende Arbeiten diskutieren dies kaum (Field et al., 2021). Ein drittes Problem bei Klassifizierung ist, dass fast jegliche Forschung von Race als eindimensionale Kategorie ausgeht. Wie in Abschnitt 2.1.1 beschrieben, handelt es sich bei Race um eine komplexe mehrdimensionale Konstruktion.

Spezifizierte Ausrichtung von NLP-Forschung - Die Erforschung von Rassismus in NLP-Systemen fokussiert sich auf einen limitierten Umfang von Anwendungsgebieten. Systeme im Bereich von Übersetzungen, Zusammenfassungen oder Question-Answer werden kaum genauer untersucht. Hate Speech, Soziale Medien oder Embeddings sind häufiger Gegenstand von Forschungsbemühungen (Field et al., 2021).

Kritische antirassistische Methodik

Für den Umgang mit Rassismus in dieser Arbeit wird eine Methodik benötigt. Orientierend an der Mehrdimensionalität von Race (vgl. 2.1.1), schlagen Hanna et al. (2020) eine Methodik zum kritischen Umgang mit Race in Algorithmen vor. Die erste Frage, die sich für die Methodik stellt, ist inwiefern eine Kategorisierung überhaupt notwendig ist. Um dies festzustellen, muss zuerst der Anwendungskontext analysiert

werden. Die Frage die sich dann stellt ist, wie kann Race denaturalisiert werden, ohne zu dematerialisieren (vgl. M'charek, 2013)? Dazu muss Race als multidimensionales, zusammenhängendes und sozial situiertes Konstrukt betrachtet werden. Dafür muss die Konzeptionalisierung und Operationalisierung von Race fokussiert werden, eine kritische Auseinandersetzung mit Kategorisierung und Bewertungsschema und transparente Kommunikation dieser stattfinden. (Hanna et al., 2020)

Hanna et al. (2020) machen einige Vorschläge zur Operationalisierung von Race, wenn diese als Variable in algorithmischen Kontexten verwendet wird. Bestehende Racial Schemata sollten kritisch bewertet werden bevor sie zum Einsatz kommen. Insbesondere auf öffentliche Zensus basierende Daten unterliegen instabilen und auf Ungleichheit basierenden Entscheidungen an den erstellenden Stellen. Probleme in der Modellierung der Messgrößen sollten ernst genommen und beachtet werden. Race ist Gegenstand einer fortwährenden Diskussion mit diversen Dimensionen und lässt sich kaum auf eine einzige Lösung reduzieren. Die Entscheidungen die zur Operationalisierung getroffen werden, können die Ergebnisse maßgeblich beeinflussen. Daher sollte viel Wert auf eine transparente und deutliche Kommunikation der zugrunde liegenden Entscheidungen erfolgen. Nach Möglichkeit sollten mehrere Messungen von Race stattfinden und als eigenes empirisches Problem wahrgenommen werden. Wenn eine Kategorisierung auf Basis von äußerlich wahrnehmbaren Aspekten vorgenommen wird, sollte diese sich auch an den Phänotypen orientieren und nicht an Kategorien von Race. Dabei sollte beachtet werden, dass dadurch rassistische (biologistische) Zuschreibungen nicht befördert werden. Race und Phänotyp dürfen nicht gleichgesetzt werden.

2.2 Grundlagen zu Embeddings

Um Embeddings im praktischen Teil untersuchen zu können, muss erst erläutert werden, was Embeddings sind, wie sie trainiert werden und Informationen aufnehmen und wie mit ihnen gerechnet werden kann. Außerdem werden in den folgenden Unterkapiteln Möglichkeiten zur Analyse von Embeddings auf sozialen Bias aufgezeigt.

2.2.1 Grundlagen zu Embeddings

Die Abschnitte Word Vectors und Embeddings, Dense Vectors und Cosine Similarity orientieren sich an Kapitel 6 in Speech and Language Processing (Third Edition Draft) von Jurafsky und Martin (2023).

Word Vectors und Embeddings

Unter einem Word Vector wird ein Vektor verstanden, der die Beziehung eines Wortes gegenüber seiner Umgebung beschreibt. Die Umgebung besteht meistens aus einem Vokabular an Wörtern. Genutzt hierfür werden oft distributive Modelle. Diese Modelle basieren auf einer Co-Occurrence Matrix, einer Matrix, welche das Auftreten von

Wörtern zu einem Kontext wie andere Wörter oder einem Dokument anzeigt. Erste Modelle in diese Richtung sind Modelle wie eine Term-Document-Matrix oder eine Term-Term-Matrix. Einfache Word Vectors nehmen die Größe bzw. Dimensionalität der Größe des Vokabulars V, also $|V|$, an. Die zugehörige Matrix schlüsselt diese dann anhand des Vorkommens im Kontext auf. Falls das Vorkommen von Wörtern in Dokumenten betrachtet wird, bedeutet das, dass dies Matrix eine Dimensionalität von $|V|$ x $|Dokumente|$ annimmt. Wenn das Vorkommen von Wörtern im Kontext von anderen Wörtern betrachtet wird, bedeutet dies eine Dimensionalität von $|V|$ x $|V|$. Solche Matrizen enthalten viele Null-Einträge, da viele Wörter nie zusammen mit Wörtern oder auch Dokumenten auftauchen. Diese Art von Word Vectors werden deshalb auch als Sparse Vectors bezeichnet. Oft werden Vektoren, die Wörter repräsentieren, auch als Embeddings benannt. (Jurafsky und Martin, 2023)

Dense Vectors

Sparse Vectors besitzen eine hohe Dimensionalität. Wenn das Vokabular klein ist, z.B. im Bereich von einigen 10.000 Wörtern, ist dies bzgl. des Rechenaufwands noch einigermaßen beherrschbar, jenseits der 50.000 Wörter aber gestaltet sich der Trainingsprozess aufwendig. Um dieses Problem entgegenzuwirken, können Wörter auch durch eine deutlich dichtere Repräsentation beschrieben werden. Die Dimensionalität eines Word Vectors liegt dann im Bereich von 50 bis 1.000 für die gesamte Matrix $|V|$ x 50 bis 1.000. Diese Art von Vektoren werden als Dense Vectors bezeichnet. In Dense Vectors haben die einzelnen Dimensionen keine klare Interpretation mehr. Dies erleichtert den Lernprozess, da durch die niedrigere Dimensionalität deutlich weniger Parameter gelernt werden müssen. Die Modelle erreichen eine deutlich bessere Generalisierung und eine geringere Anfälligkeit für Overfitting. Außerdem, was für diese Arbeit wichtig ist, erfassen Dense Vectors Ähnlichkeiten zwischen Wörtern besser. Die wichtigsten Modelle in dieser Arbeit basieren alle auf Dense Vectors. (Jurafsky und Martin, 2023)

Word2Vec Model

Das grundlegende Modell für Embeddings auf das in der Analyse in Kapitel 3 zurückgegriffen wird, ist Word2Vec, vorgeschlagen durch Mikolov, Chen et al. (2013). Word2Vec ändert den oben erläuterten Ansatz ab. Vektoren beschreiben nicht mehr die gezählten Vorkommen von Wörtern, sondern die Wahrscheinlichkeit p, dass ein Wort ein anderes Wort benachbart. Die Nachbarschaft eines Wortes erstreckt sich über ein Fenster von Wörtern vor und nach dem Wort. Dieses Fenster wird Kontextfenster genannt. Es wird nicht nur die Wahrscheinlichkeit des Vorkommens eines Wortes, sondern auch das Nicht-Vorkommens des Wortes beschrieben. Ziel des Modells ist die Maximierung der logarithmischen Wahrscheinlichkeit. Für Word2Vec schlagen Mikolov, Chen et al. (2013) zwei Modellvarianten zum Erlernen des Embeddings vor: Continuous Bag-of-Words (CBOW) und Continuous Skip-Gram.

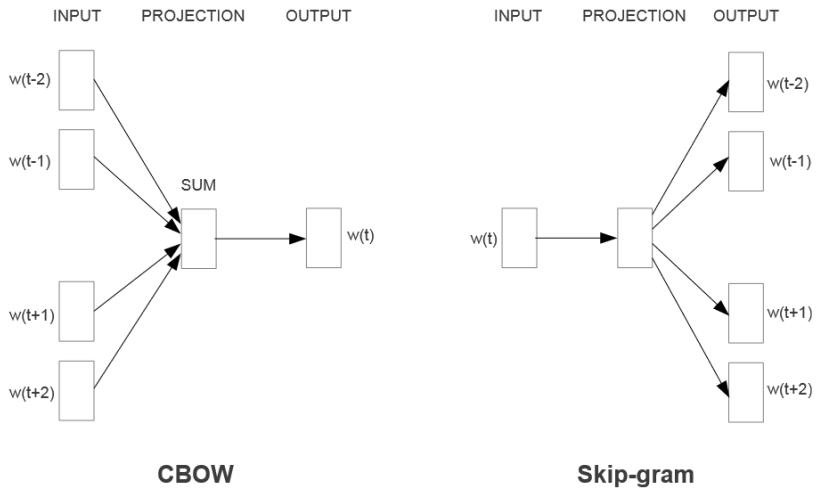


Abbildung 2.1: CBOW sagt aus den Kontextwörtern w_{t-2} bis w_{t+2} das Zielwort w_t vorher, Skip-gram sagt auf Basis von w_t die Kontextwörter w_{t-2} bis w_{t+2} vorher. Diese Abbildung stammt von Mikolov, Chen et al. (2013).

Continuous Bag-of-Words Model

Mikolov et al. (2017) stellen das Trainingsziel von CBOW vor, als die Optimierung der Vorhersage des Zielworts w_t aus einem Kontextfenster um das Zielwort herum (vgl. Abbildung 2.1). Dafür werden alle Wörter $w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}$ aus dem Kontextfenster C_t auf eine zentrale Stelle t projiziert. t ist dabei die Stelle des Zielworts w. Um die Projektion zu erreichen, werden die Vektorrepräsentationen aller Wörter aus dem Kontextfenster gemittelt. Die Reihenfolge der Wörter wird dadurch irrelevant. Wörter, die eigentlich in der Reihenfolge in der Zukunft liegen, werden so miteinbezogen. CBOW maximiert die logarithmische Wahrscheinlichkeit p eines Zielworts w_t in Abhängigkeit ihres Kontextes C_t an der Stelle t mit:

$$\sum_{t=1}^{T} \log p(w_t|C_t) \tag{2.1}$$

t ist dabei der Index über eine Sequenz an Trainingswörter T. Um p zu erhalten, gibt es unterschiedliche Möglichkeiten. Softmax bietet sich an, um eine bedingte Wahrscheinlichkeit über den Kontext und Zielwort zu erhalten. Ein Nachteil ist dabei die Ineffizienz für ein großes Vokabular. Eine Möglichkeit, dies zu lösen, ist eine Annäherung durch eine binäre Klassifikation. Die Wahrscheinlichkeit p kann dann wie folgt bestimmt werden:

$$\log\left(1 + e^{-s(w,C)}\right) + \sum_{n \epsilon N_{C_t}} \log\left(1 + e^{s(n,C)}\right) \tag{2.2}$$

N_{C_t} sind dabei Negativbeispiele, also zufällig aus dem Vokabular ausgewählte Wörter, welche keine Überlappung mit dem Kontext C_t haben, also sicher nicht im Kontext des Zielworts w liegen (Negative Sampling). Die Funktion s übernimmt hier eine Bewertung auf Ähnlichkeit zwischen dem Wort w und dem Kontext C. Dabei ist $s(w, C)$ wie das innere Produkt zwischen dem Word Vector von w und den zu in C gemittelten Vektoren der Kontextwörter (vgl. zu den Ausführungen zu Gleichungen 2.5 und 2.7). Gleichung 2.2 eingesetzt in Gleichung 2.1 ergibt:

$$\sum_{t=1}^{T} \left[\log\left(1 + e^{-s(w_t, C_t)}\right) + \sum_{n \epsilon N_C} \log\left(1 + e^{s(n, C_t)}\right) \right] \tag{2.3}$$

Mikolov et al. (2017) beschreiben noch einige Erweiterungen und Verbesserungen für CBOW. In 2.2.2 wird so eine Erweiterung, nämlich fastText (Bojanowski et al., 2017), beschrieben.

Continuous skip-gram model

Das *skip-gram model* dreht das Vorgehen von CBOW um. Das Ziel ist anhand eines Wortes die Kontextwörter in der Umgebung des Wortes korrekt vorherzusagen (vgl. Abbildung 2.1). Skip-gram maximiert dabei die durchschnittliche Wahrscheinlichkeit der Kontextwörter $w_{j-c}, ...w_{j-1}, w_{j+1}, ..., w_{j+c}$ im Fenster von $-c$ bis c um die Position t herum in Abhängigkeit vom Wort w_t:

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \tag{2.4}$$

Ähnlich wie bei CBOW, würde sich hier Softmax als Funktion für p anbieten und auch hier existiert das Problem der Ineffizienz von Softmax für ein großes Vokabular. Mikolov et al. lösen dies für Skip-gram durch eine Annäherung an den Softmax mit dem Hierarchical Softmax als binärer Klassifikator (Mikolov, Sutskever et al., 2013). Auf eine genauere Beschreibung dieses Klassifikators wird an dieser Stelle verzichtet, da für die Analyse in Kapitel 3 ein Modell basierend auf CBOW verwendet wird.

Cosine Similarity

Die Repräsentation von Wörtern als Vektoren ermöglicht es auch, mit diesen zu rechnen. Beispielsweise lassen sich Vektoren miteinander auf Ähnlichkeit vergleichen. Eine Metrik dafür ist die Cosine Similarity. Diese beschreibt den Winkel zwischen zwei Vektoren. Der Winkel zwischen zwei Vektoren gibt an, in welche Richtung die Vektoren sich jeweils im Vergleich zueinander zeigen. Je kleiner dieser Winkel ist, desto ähnlicher die Richtung der Vektoren und damit desto ähnlicher die Vektoren selbst. Um mit der Cosine Similarity rechnen zu können, braucht es das innere Produkt zweier Vektoren:

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{N} v_i w_i \tag{2.5}$$

Dieses Produkt multipliziert jede Komponente eines Vektors mit der jeweils gleichen Komponente des anderen Vektors und summiert diese dann auf. Eine Beobachtung ist, dass längere Vektoren einen größeren Einfluss auf das Produkt haben wie kürzere. Öfter vorkommende Wörter haben Vektorrepräsentationen einen längeren Vektor wie weniger oft vorkommende Wörter und hätten so einen überproportionalen Einfluss. Um das entgegenzuwirken, müssen die Vektoren normalisiert werden mit ihrer Länge:

$$\|\vec{v}\| = \sum_{i=1}^{N} v_i^2 \tag{2.6}$$

Das innere Produkt wird normalisiert, indem es durch das Produkt der normalisierten Vektoren geteilt wird:

$$\frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\|\|\vec{w}\|} = \cos(\vec{v}, \vec{w}) \tag{2.7}$$

Diese Formel entspricht dem Cosinus des Winkels zwischen zwei Vektoren. In vielen Anwendungen werden Vektoren mit Einheitslänge verwendet. Dies vereinfacht die Formel auf das innere Produkt und simplifiziert die Implementierung.

Semantische Eigenschaft: Analogie-Findung

Eine für diese Arbeit wichtige semantische Eigenschaft von Embeddings ist die Fähigkeit, relationale Bedeutungen aufzugreifen. Diese Eigenschaft lässt sich dafür nutzen, einen möglichen sozialen Bias im Embedding zu untersuchen. Die genaue Beschreibung der Umsetzung im Rahmen dieser Arbeit folgt in Abschnitt 2.3.1. Embeddings greifen Assoziationen und Relationen in Trainingsdaten auf. Wiederfinden lassen sich diese Assoziationen über eine Methodik aufbauend auf dem Parallelogramm Modell, vorgeschlagen durch Rumelhart und Abrahamson (1973). Grundlage dafür ist die Frage *„a is to x as b is to y?“*, wobei a, b und x vorgegebene Wörter sind. Ziel ist es, das optimale y zu finden. Das optimale y soll dabei eine Analogie zu x mit dem Kontext von a, b darstellen. Dafür werden nach Turney (2012) folgende Gleichungen und Ungleichungen als Maßstab für hochwertige Analogien festgelegt:

$$sim_r(a : x, b : y) = sim_r(x : a, y : b) \tag{2.8}$$

$$sim_r(a : x, b : y) = sim_r(b : y, a : x) \tag{2.9}$$

$$sim_r(a : x, b : y) \neq sim_r(a : x, y : b) \tag{2.10}$$

$$sim_r(a : x, b : y) \neq sim_r(a : y, b : x) \tag{2.11}$$

Dabei sind $(a : x)$ und $(b : y)$ jeweils ein Wortpaar. sim_r ist eine Funktion für Ähnlichkeit. Diese Gleichungen und Ungleichungen beschreiben jeweils die Verhältnis-

se zwischen den Wörtern der Analogie. Gleichung 2.8 besagt, dass wenn a mit x und b mit y vertauscht werden, die gleiche Ähnlichkeit zwischen den Wortpaaren ergeben soll. Gleichung 2.9 besagt, dass die Ähnlichkeit beim Vertauschen der Wortpaare auch gleich bleiben sollte. Die Ungleichung 2.10 besagt, dass wenn nur die Wörter eines der Wortpaare vertauscht werden, nicht die gleiche Ähnlichkeit gegeben sein darf. Ungleichung 2.11 besagt, dass die Ähnlichkeit nicht gleich bleiben darf, wenn y mit x vertauscht wird. Eine der Nicht-Erfüllung der Ungleichungen bedeutet das y eher ein Synonym als wie eine Analogie ist. Für *„a is to x as b is to y“* wird ab hier die Notation $a : x :: b : y$ verwendet.

2.2.2 Embeddings und der Spezialfall Deutsch

Die Optimierung von Embeddings auf Deutsch ist für uns relevant. Wie bereits in Abschnitt 2.1.4 aufgezeigt, zentrieren viele Arbeiten und Anwendungen, aber auch Analysen von Embeddings, Englisch. Ihre Performance ist auch entsprechend optimiert. Deutsch ist eine morphologisch komplexere Sprache wie Englisch. Eine Schwierigkeit ist die Groß- und Kleinschreibung. Viele Tokenizer haben die Eigenschaft von Lower Casing, also das Ersetzen von Großbuchstaben durch Kleinbuchstaben, um Komplexität zu reduzieren. Tokenizer normalisieren die Textdaten, mit denen Embeddings trainiert werden, indem sie Texte in Wörter segmentieren. Groß- und Kleinschreibung führt im Deutschen einiges an Informationsgehalt mit sich. Verschiedene Wörter haben unterschiedliche konkrete Bedeutungen in Abhängigkeit von ihrer Groß- und Kleinschreibung bei sonst gleicher Schreibweise. Dieses Problem lässt sich einigermaßen leicht umgehen durch Verwendung geeigneter Tokenizer wie SoMaJo (Proisl und Uhrig, 2016) und ist nicht unbedingt abhängig vom verwendeten Modell. Deutlich schwieriger gestaltet sich aber der Umgang mit Wortzusammensetzungen und eine große Anzahl an Wortformen (Bojanowski et al., 2017). Diese führen zu einem sehr großen Pool an Wörtern. Viele dieser Wörter werden aber bereits in der Vorverarbeitung wegen ihres seltenen Vorkommens durch viele Modelle aussortiert. In den nächsten Abschnitten wird dieses Problem präziser in Word2Vec beleuchtet und mit fastText ein Modell vorgestellt, welches diese Problematik für diverse Sprachen verbessert. Um dieses Problem auch mit Zahlen aufzeigen zu können, wird zuerst die Word Analogy Task betrachtet, welche als Metrik dient, um die Performance zwischen verschiedene Parametern und Modellen vergleichen zu können.

Word Analogy Task

Die Eigenschaft von Embeddings, gut abzuschneiden beim Aufgreifen relationaler Beziehungen, lässt sich dazu nutzen, ein allgemeines Maß für die Qualität von Embeddings festzulegen. Vorgreifend muss aber erwähnt werden, dass die Qualität von Embeddings sehr schwer zu messen ist bzw. die Performance in einzelnen Aufgabengebieten mehr über die Eignung eines Embeddings für eine bestimmte Aufgabe aussagt. Allgemeine Aussagen lassen sich so nicht gut treffen. Ein Embedding kann schlecht in einer allgemeineren Aufgabe abschneiden, aber sehr gut für spezifische Aufgaben

geeignet sein. Oft wird aber trotzdem eine Methodik benötigt, um festzustellen, ob ein Embedding für allgemeinere Aufgaben gut abschneidet. Hier wird oft auf die Word Analogy Task zurückgegriffen (Mikolov, Yih et al., 2013, Mikolov, Chen et al., 2013). Ähnlich wie in 2.2.1 werden Aufgaben an das Embedding gestellt in der Form *„a is to x as b is to y“*. Über einen standardisierten Aufgabensatz wird dann mit der erreichten Accuracy, also der Anzahl der erfüllten Analogie-Aufgaben, ein Vergleich zwischen verschiedenen Embeddings möglich. Der Aufgabensatz enthält dabei semantische und syntaktische Aufgaben. Syntaktische Aufgaben testen Basisformen, vergleichende und superlative Formen von Adjektiven; Singular und Plural von Nomen; Possessiv Nomen sowie Gegenwarts- und Vergangenheitsformen von Verben. Semantische Aufgaben beinhalten Tests, um festzustellen, inwiefern Beziehungen zwischen Wörtern aufgegriffen werden. Ein Aufgabensatz für Englisch stammt von Mikolov, Yih et al. (2013), einer für Deutsch von Köper et al. (2015). In 2.3.1 wird die Generierung von Analogien ausführlicher diskutiert.

Word2Vec und deutsch

Der Word2Vec Ansatz wie ihn Mikolov, Chen et al. (2013) beschreiben, gilt vor allem als großer Sprung in der Trainierbarkeit von Neural Net Language Models, insbesondere durch die Reduzierung der Komplexität. Dabei erreicht Word2Vec auch eine höhere Accuracy als andere Modelle bis zur dessen Vorstellung in 2013. In ihrem Benchmarktest erreichen Bojanowski et al. (2017) in Englisch mit Word2Vec eine Accuracy von 78,5% für semantische Aufgaben und 70,1% für syntaktische Aufgaben für Skip-gram. Für CBOW erreichen sie 78,2% für semantische und 69,9% für syntaktische Aufgaben. In Deutsch erreichen sie mit Skip-gram 66,5% für semantische und 44,5% für syntaktische Aufgaben. Mit CBOW werden 66,8% für semantische und 45,0% für syntaktische Aufgaben erreicht. Hier fällt auf das Word2Vec insbesondere für syntaktische Aufgaben deutlich schlechter abschneidet. Dies gilt nicht nur für deutsch, sondern auch für andere Sprachen, welche morphologisch komplexer sind.

fastText als Lösung

Um das Problem des deutlich schlechteren Abschneidens von Word2Vec in morphologisch komplexere Sprachen zu begegnen, schlagen Bojanowski et al. (2017) fastText vor. Dafür wird zusätzlich zu den Word Vectors, wie sie in Word2Vec vorkommen, ein Subword Model eingeführt. Dieses zerlegt die Wörter in jeweilige Teile einer bestimmten Länge. Die einzelnen Wortteile werden Character N-grams genannt. Die ermöglichen den Austausch von Informationen, sogenannter Subword Information, zwischen den einzelnen Teilen von Wörtern, welches verschiedenen Wortformen und Wortzusammensetzungen zugutekommt. Ein zusätzlicher Effekt dieses Subword Models ist die Fähigkeit von fastText, Wörter, welche nicht im Vokabular des Embeddings vorkommen, darzustellen. Die zusammengesetzten Wortteile besitzen immer eine Vektorrepräsentation, auch wenn das Wort nicht Teil des Vokabulars ist. Aufbauend auf dem Skip-gram Modell von Mikolov, Chen et al. (2013) erreichen sie so eine Accuracy von

77,8% für semantische und 74,9% für syntaktische Aufgaben in Englisch. Bezeichnend ist die in Deutsch erreichte Steigerung der Accuracy für syntaktische Aufgaben von 44,5% Accuracy für Word2Vec auf 56,4% für fastText. Für semantische Aufgaben wird eine Accuracy von 62,3% in fastText erreicht.

In einer Erweiterung von fastText schlagen Grave et al. (2018) die Verwendung von Positionsgewichtung nach Mnih und Kavukcuoglu (2013) mit CBOW vor und erreichen z.B. für Deutsch noch mal eine Steigerung der Accuracy für die Word Analogy Task. Sie erreichen so eine gesamte Accuracy von 71,7% auf semantische und syntaktische Aufgaben. Mit Optimierung des Negative Sampling und der Anzahl der Epochen wird dies noch gesteigert auf 73,9%. So schneidet fastText mit CBOW für Deutsch deutlich besser ab wie Word2Vec.

2.3 Bias in Embeddings

Die folgenden Abschnitte behandeln, wie diskriminierende Assoziationen in Embeddings gefunden werden könne. Dabei wird sich insbesondere an Bolukbasi et al. (2016) und den darauf aufbauenden Arbeiten von Beispielsweise Manzini et al. (2019) und Chen et al. (2021) orientiert.

2.3.1 Diskriminierende Assoziationen

Geometrie von Rassismus

Wie vorhin bereits ausgeführt haben Embeddings die Eigenschaft, relationale Bedeutungen gut aufzugreifen (vgl. 2.2.1). Das legt auch die Vermutung nahe, dass Embeddings auch Assoziationen aufgreifen können, die diskriminierend sind. Einige Arbeiten untersuchen dies insbesondere mit Fokus auf Gender. Startpunkt bildet hier Bolukbasi et al. (2016), welche als erste in Embeddings nach einem Gender Bias suchen und Methoden vorschlagen, diese zu finden. Wörter können mehr oder weniger mit einem Bias assoziiert werden. Der Bias selbst kommt aus den vorhandenen Daten. Dieser lässt sich abgrenzen durch die Beschreibung mithilfe eines Sets an definierenden Begriffen. Wörter werden wiederum mit dem Bias assoziiert aufgrund dessen, dass sie in den Trainingsdaten oft im Kontextfenster der Bias definierenden Begriffen auftauchen oder weil Bias definierende Begriffe oft in deren Kontextfenster vorkommen. Das Word2Vec Modell lernt dann entsprechend höhere Wahrscheinlichkeiten für das gemeinsame Vorkommen. Diese Dynamik führt dann zur stärkeren Assoziation durch das Embedding. In den folgenden Abschnitten werden einige Möglichkeiten genauer ausgeführt, wie sich diese Assoziationen feststellen und untersuchen lassen.

Analogiefindung

In Anlehnung an Bolukbasi et al. (2016) weißen wird in Kapitel 3 die Existenz eines rassistischen Bias in einem vortrainierten fastText Embedding nachgewiesen. Dafür wird eine u. a. eine Umfrage durchgeführt, bei dem Teilnehmende beurteilen sollen,

ob nach dem Muster *„a is to x as b is to y“* generierte Analogien rassistisch sind. Um diese Analogien zu generieren, wird das Konzept genutzt, welches in 2.2.1 beschrieben ist. Der klassische Vorgang, so wie er auch in der Word Analogy Task in Abschnitt 2.2.2 verwendet wird, erreicht dafür bereits gute Ergebnisse. Die Generierung erfolgt dabei über folgende Methode (Mikolov, Yih et al., 2013):

$$y = \vec{x} - \vec{a} + \vec{b} \tag{2.12}$$

y entspricht nicht immer einem mit dem Vokabular korrespondierenden Vektor. In dem Fall muss der Wort w mit dem korrespondierenden Word Vector w mit der höchsten Cosine Similarity gefunden werden. Die Cosine Similarity (2.7) wird dafür maximiert mit dem Word Vector $\vec{w}$ und y aus 2.12:

$$w = argmax_w \frac{\vec{w} \cdot y}{\|\vec{w}\| \cdot \|y\|} \tag{2.13}$$

Ein Problem, das auch in dieser Arbeit beobachtet und u. a. durch Bolukbasi et al. (2016) diskutiert wird, ist, dass die in Abschnitt 2.2.1 aufgestellten Gleichungen und Ungleichungen immer wieder gebrochen werden durch Ergebnisse der Gleichungen 2.12 und 2.13. Ein Grund dafür ist, dass oft der am nächsten liegende Word-Vector für $y = x$ ist. Dies wird oft gelöst, indem x als Lösung ausgeschlossen wird. Trotzdem liegt y sehr nahe an x. In diesem Fall entspricht $a : x :: b : y$ oft auch $a : y :: b : x$, was eine Verletzung von der Gleichung 2.11 ist. Bolukbasi et al. (2016) unterstreicht in Anlehnung an Levy und Goldberg (2014) die erstrebenswerte Eigenschaft, dass die Winkel zwischen $\vec{a} - \vec{b}$ und $\vec{x} - \vec{y}$ möglichst ähnlich sein sollten. Dabei stellen sie fest, dass $cos(\vec{a} - \vec{b}, \vec{x} - \vec{y})$ oft mit y als Extrem maximiert wird. Hier wird das Beispiel genannt, indem für $a = he$ und $b = she$ oft *her* als Ergebnis für y auftaucht, unabhängig von x. Um dieses Verhalten zu begrenzen, schlagen sie die Einschränkung $|\vec{x} - \vec{y}| \leq \delta$ vor. Dies entspricht einer Schranke für Ähnlichkeit und verhindert zu starke Extreme von y zu x. Den für sie optimalen Wert für $\delta = 1$ bestimmen sie experimentell. In Abschnitt 3.3.2 wird festgestellt, das $\delta = 1$ für die Analyse in Abschnitt 3.3 nicht optimal ist und wird da diskutiert.

Aus den genannten Ausführungen leiten Bolukbasi et al. (2016) die folgende Metrik für das finden von Analogien ab:

$$S_{(a,b)}(x, y) = \begin{cases} cos(\vec{a} - \vec{b}, \vec{x} - \vec{y}) & \text{wenn } \| \vec{x} - \vec{y} \| \leq \delta \\ 0 & \text{sonst} \end{cases} \tag{2.14}$$

Die Differenzen $\vec{a} - \vec{b}$ und $\vec{x} - \vec{y}$ stellen die Relationen von a mit b und x mit y dar. S beschreibt über die Cosine Similarity den Winkel zwischen der Relation a und b und der Relation von x und y. S wird maximiert für y und daraufhin als y in 2.13 eingesetzt, um die finale Analogie zu erhalten. Durch das Zusammenspiel von Maximieren von S und der Schranke für Ähnlichkeit von x und y sollte S also das y finden, welches einerseits ähnlich zu x ist, andererseits eine möglichst parallele Beziehung zu a und b wie x besitzt. In Abschnitt 3.3.2 zeigt sich, dass diese vorgeschlagene Metrik die

Problematik, die die Metrik erst motiviert hat, auch nicht vollständig löst. Eine im praktischen Teil generierte Analogie mit $x = Abitur$ ergibt $y = abitur$. Dies wird auch in 3.3.2 genauer ausgeführt.

Bias Subspace

Beim Betrachten der Methodik zur Generierung von Analogien lässt sich ein Prinzip beobachten, welches dazu genutzt werden kann, einen Bias Subspace oder eine Bias-Richtung zu finden. Bei der Generierung von Analogien wird über die Differenz die Beziehung zwischen zwei Wörter erfasst. Sind diese Wörter sehr deutlich in ihrer Bedeutung bzw. welchem übergeordneten Konzept sie beschreiben, so lässt sich mutmaßen, dass die Differenz der zwei Vektoren einem Vektor entspricht, der hauptsächlich von einem gemeinsamen Konzept der Word Vectors dominiert wird. Die Differenz beschreibt dann eine Art Bias Subspace. Word Vectors enthalten oft nicht nur Informationen über das gesuchte Konzept. Um den Subspace besser anzunähern, können die Differenzen mehrerer Word Vectors berechnet und diese Differenz gemittelt werden. Bolukbasi et al. (2016) schlagen diese Methodik vor, um nach einen Gender Subspace zu suchen. Sie verwenden dafür 10 Begriffspaare, welche sie als besonders definierend für einen Gender Bias feststellen. Chen et al. (2021) konkretisieren dieses Vorgehen, indem sie zuerst über den Durchschnitt des Wortpaars P_i deren Zentrum μ_i berechnen. Im Anschluss wird die Differenz jedes Word Vectors des Wortpaars zu deren Zentrum berechnet. Mit dem kompletten Set an Wortpaaren ergibt sich so eine Matrix. Auf Basis dieser Matrix wird nun eine Hauptkomponentenanalyse bzw. Principal Component Analysis (PCA) durchgeführt. Eine PCA komprimiert die Dimensionalität von den Word Vectors, um herauszufinden, welche Komponenten der Word Vectors maßgeblich an deren Form beteiligt sind. Dadurch das die Bias definierenden Begriffe stark mit dem gesuchten Bias assoziiert werden, definieren die dominierenden Komponenten dann den Bias Subspace B. Die beschriebe Methodik lässt sich wie folgt zusammenfassen:

$$B = PCA\left(\bigcup_{i=1}^{n}\bigcup_{w \epsilon P_i} \vec{w} - \mu_i\right) \tag{2.15}$$

n ist die Anzahl Wortpaare der Bias definierenden Begriffen. Manzini et al. (2019) entwickeln den Ansatz von Bolukbasi et al. (2016) weiter für die Untersuchung eines Multiklassen-Bias. Sie schlagen eine Methodik vor, einen Subspace zu finden, welcher mehrere Biase erfasst. Sie berechnen zuerst das Zentrum μ_i des gesamten Defining Sets D_i. Für jeweils alle Wörter des Defining Sets wird die Differenz des Word Vectors vom Zentrum berechnet und eine Matrix gebildet aus allen Differenzen. Dies wird mit allen Defining Sets durchgeführt und aus den Ergebnissen eine Matrix gebildet. Anschließend wird die PCA auf diese Matrix durchgeführt. Der aus der PCA resultierenden Bias Subspace B kann anschließend verwendet werden, um Bias im Embedding weiter zu untersuchen.

Direkter Bias

Unter direktem Bias wird verstanden, wie stark Wörter mit dem Bias assoziiert wird. Bolukbasi et al. (2016) schlagen hier erstmals eine Metrik vor, um diese Assoziation zu messen. Ähnlich wie bei der Generierung von Analogien beschreibt die Metrik den Winkel zwischen einem Word-Vector w und dem Bias Subspace B mit der Cosine Similarity. Die Metrik betrachtet also, wie ähnlich ein Word-Vector und der Bias Subspace sind. Diese Metrik wird dann über ein Set an Wörtern gemittelt, sodass das folgende Gleichung ergibt:

$$DB = \frac{1}{|N|} \sum_{w \epsilon N} |cos(\vec{w}, B)|^c \tag{2.16}$$

c dient als Maßgabe zur Striktheit. Für $c = 0$ nimmt $|cos(\vec{w}, B)|^c$ nur den Wert 0 an, wenn es keine Überschneidung von $\vec{w}$ mit B existiert. Sobald eine Überschneidung existiert, würde der Wert 1 rauskommen. Mit welchen konkreten Wert c belegt wird, ist abhängig von der Situation, indem Bias untersucht wird. Beim Einsatz von NLP-Systemen bei z.B. Bewerbungsverfahren sollte eine stärkere Striktheit verwendet werden. N besteht aus dem Set an Wörtern, welche untersucht werden sollen. Bolukbasi et al. (2016) verwenden hier z.B. Berufe. Tendenziell werden hier Begriffe untersucht, welche neutral sein sollten. In dieser Arbeit werden Bildungsabschlüsse untersucht. In 2.3.3 wird näher darauf eingegangen.

2.3.2 Bias definierende Begriffe für Rassismus

Um den Bias Subspace finden zu können, werden Begriffe benötigt, von denen vermutet wird, dass diese den Bias im Embedding repräsentieren könnten. Bei der Untersuchung von Gender Bias in Word2Vec Modelle erstellen Bolukbasi et al. (2016) über Ccrowd Sourcing und experimentelle Verifizierung mittels Umfragen auf Mechanical Turk eine Liste mit 10 Wortpaaren, von denen sie vermuten, dass diese den Subspace repräsentieren könnten. Die Paare bestehen dabei aus Tupeln wie *(she, he)* und stellen jeweils gegenüberstehende Extreme des Subspaces dar. Mittels Umfrage verifizieren sie das mit diesen Paaren generierte Analogien einen Gender Bias enthalten. Manzini et al. (2019)) basieren ihre Bias definierende Begriffe für Race auf Literatur über Race und Caucasians, African Americans und Asian Americans. Der Fokus liegt hier auf den amerikanischen Diskurs. Dabei rücken sie ab vom Paaransatz und rechnen mit Differenzen von Sets zum Zentrum der Liste (vgl. Abschnitt zu Bias Subspace 2.3.1). Chen et al. (2021) untersuchen Gender Bias in einem mit Wikipedia trainierten deutschsprachigen Word2Vec Modell. Für die Erstellung von Bias definierenden Begriffen orientieren sie sich in erster Instanz an den von Bolukbasi et al. erstellten Begriffen und übersetzen diese u. a. ins Deutsche. Dabei stoßen sie auf einige Herausforderungen wie uneindeutige Bedeutungen bei Wörtern wie „sie“ oder „ihrer“. Schlussendlich stellen sie eine Liste mit sieben Wortpaaren für neun Sprachen auf.

Für Bias definierende Wörter für Rassismus im deutschen Diskurskontext wird auf den Ausführungen des NaDiRa zurückgegriffen (Foroutan et al., 2022). Die dort be-

schriebenen Gruppen dienen als Grundlage für Bias definierende Begriffe. Im Kontrast zu Manzini et al., wird das Problem in dieser Arbeit binär betrachten. Wie bereits in 2.1.2 angesprochen, ist es wichtig, bei der Untersuchung von Diskriminierung und einen sozialen Bias die dominierende Gruppe zu benennen und abzugrenzen zu den diskriminierten Gruppen. Im Kontext vom deutschen Diskurs bzgl. Rassismus bedeutet dies, Weißsein explizit zu benennen und als Konstrukt von rassifizierten Gruppen abzugrenzen. Aus diesem Grund heraus wird Weißsein als ein Bias-Extrem festgelegt. Die rassifizierten Gruppen, wie sie im NaDiRa und in 2.1.2 aufgeführt sind, werden als weitere Extreme festgelegt. Ein Element, welches untersucht wird, ist, ob der Bias Subspace auch gefunden werden kann, wenn eher abstraktere Begrifflichkeiten genutzt werden. Für die schlussendlich verwendeten Begriffe wird festgelegt, die Konzepte, als Adjektive zu verwenden. Hieraus ergibt sich folgende Liste an Bias definierende Begriffen:

- Weiß
- Schwarz
- Jüdisch
- Muslimisch
- Asiatisch
- Osteuropäisch
- Roma

Zwei abschließende Anmerkungen zu dieser Liste, Osteuropäisch wird als Bezeichnung für antislawischen Rassismus ausgewählt, um die ganze Bandbreite der Betroffenen abdecken zu können. Für Rassismus gegen Roma und Sinti wird Roma festgelegt vor dem Hintergrund, dass es sich bei Roma um eine Bezeichnung für eine Gruppe an Personen handelt, welche insbesondere ab den 1950er-Jahren nach Deutschland migrierten oder flüchteten. Unter Sinti wird eher die Gruppe verstanden, welche bereits vor den 1930er-Jahren in Deutschland ansässig war. Den Begriff Roma wird deshalb eine höhere geschichtliche Aktualität zugestanden. Experimentelles ausprobieren deutet aber darauf hin, dass ähnliche Ergebnisse mit dem Begriff Sinti erzielt werden können. Eine andere Möglichkeit wäre gewesen, die zwei Word Vectors von Sinti und Roma zu mitteln und den resultierenden Word-Vector weiter zu verwenden.

2.3.3 Bildungsabschlüsse als Prüfkategorie

In dieser Arbeit wird untersucht, wie ein Bias sich überträgt in den Ausgaben eines Embeddings. Es ist sinnvoll, Bereiche zu untersuchen, bei denen ein diskriminierender Bias besonders kritisch wirken kann. So untersuchen Bolukbasi et al. (2016) Beschäftigungen auf einen Gender Bias, bewerten diese mit der in 2.3.1 vorgestellte Metrik und

stellen dabei einen Bias-Anteil fest. Ebenso untersuchen Chen et al. (2021) Berufsbezeichnungen und stellen für ein Word2Vec Embedding, welches mit Wikipedia-Corpus trainiert wurde, auch einen Gender Bias fest. Auch existieren Arbeiten, welche Namen in Zusammenhang von Ethnizität untersuchen (Gerard, 2017). Zum Zeitpunkt des Verfassens dieser Arbeit sind uns aber keine Arbeiten bekannt, welche den Bereich Bildung näher beleuchten. Wie bereits in 2.1.2 beleuchtet, ist Bildung aber ein Feld, indem rassifizierte Gruppen Diskriminierung ausgesetzt sind. In abschnitt 3 werden deshalb Bildungsbegriffe im Embedding analysiert.

Für die Begriffe, welche untersucht werden, orientiert diese Arbeit sich in erster Linie an Schulabschlüsse, wie sie das Statistische Bundesamt verwendet (2022). Da einzelne Wörter in einem Embedding untersucht werden, werden alle Bezeichnungen wie „Ohne allgemeinen Schulabschluss“ entfernt. „Abschlüsse der polytechnischen Oberschule“ werden aufgrund der fehlenden Aktualität entfernt. Nach diesen Schritten ergibt sich so eine Liste mit den Schulabschlüssen „Hauptschulabschluss“, „Realschulabschluss“ und „Abitur“, deren Vorkommen im Embedding genauer untersucht wird in Abschnitt 3.3.5.

3 Analyse eines Embeddings

Nachdem Kapitel 2 theoretische Grundlagen und Hintergründe zu Rassismus und NLP behandelt hat, wird im Folgenden nun ein konkretes Embedding aufbauend auf diesen Grundlagen auf rassistischen Bias hin untersucht und analysiert werden. Dafür wird zuerst kritische antirassistische Methodik in unseren Kontext eingeordnet. Ein ursprüngliches Ziel war es, ein fastText Embedding mit aktuellen Daten zu trainieren und zu analysieren. Dabei traten Schwierigkeiten auf, welche hier beschreiben werden. Aus den aufgetretenen Problemen ergab sich die Änderung, ein vortrainiertes fastText Embedding zu verwenden. Um die Existenz eines rassistischen Bias nachzuweisen, wurde eine Umfrage mit durch das Embedding generierten Analogien durchgeführt, welche in diesem Teil ausgewertet wird. Anschließend wird der Bias Subspace festgestellt mittels einer PCA. Schlussendlich werden Bildungsabschlüsse im Embedding genauer beleuchtet und der Anteil an direkten Bias gemessen mit der in 2.3.1 vorgestellten Metrik.

3.1 Einordnung antirassistischer Methodik

Da im praktischen Teil rassifizierte Gruppenbezeichnungen als einen der zentralen Parameter verwendet werden, wird zuerst eingeordnet, wie die in Abschnitt 2.1.4 erläuterten Punkte zur Operationalisierung von Race in unserer Methodik beachtet werden. Für die Gruppenbezeichnungen werden die durch den NaDiRa definierten Gruppen und in Abschnitt 2.3.2 aufgeführten Begriffe verwendet. Diese Begriffe stützen sich auf neuere Forschungsarbeit in dem deutschen Diskurs der Rassismusforschung wie den NaDiRa. Der NaDiRa verwendet diese als Selbstbezeichnung in ihren Umfragen. In der Umfrage in dieser Arbeit wird dies ähnlich gehandhabt und wird eine Option für Angaben hinzugefügt, falls teilnehmende Personen sich nicht durch die vorgegebenen Bezeichnungen repräsentiert sehen. Wir orientieren uns damit an dem in Abschnitt 2.1.1 aufgeführten Vorschlag zur Verwendung von Racial Self Classification. Sofern Angaben von teilnehmenden zu durch einer der Definitionen der Gruppen eingeschlossen werden, werden diese der entsprechenden Gruppe zugeordnet. Auch wird nach Rassismuserfahrung gefragt. Diese Frage orientiert sich auch an der eigenen persönlichen Einschätzung.

3.2 Training eines fastText Embeddings

Eine breitere gesellschaftliche Debatte zu Rassismus in Deutschland bekam 2020 mit den Protesten rund um die „Black Lives Matter“-Bewegung einen Auftrieb. Gesell-

schaftliche Debatten schlagen sich nieder in Beiträgen und Artikeln von Nachrichtenplattformen und auch in wissenschaftlichen Schwerpunktsetzungen. Aus dieser Motivation war diese Arbeit ursprünglich darauf ausgelegt, ein fastText Embedding mit aktuellen Datensätzen zu trainieren, um auch die Debatten der letzten Jahre reflektieren zu können. Im Folgenden wird der Aufbau des Trainings des Embeddings beschrieben und dargelegt, warum letztlich kein selbsttrainiertes Embedding für die spätere Analyse verwendet wird.

3.2.1 Daten

Nachrichtenartikel genießen oft den Ruf, einen gewissen Grad an Neutralität beziehungsweise Objektivität zu reflektieren. Anknüpfend an Bolukbasi et al. (2016) kann die Analyse eines Embeddings, welches mit Nachrichtenartikeln trainiert wurde, auf sozialen Bias hin interessant sein, um festzustellen, ob auch vermeintlich objektive Datengrundlagen diskriminierende Assoziationen aufgreifen und reproduzieren. Deshalb werden Artikel großer deutscher Nachrichtenplattformen als Datengrundlage für das Training eines Embeddings verwendet. Einen ausführlichen und jährlich aktualisierten Datensatz stellt Wortschatz Leipzig zur Verfügung (Goldhahn et al., 2012).

3.2.2 Modell

Als Modell wird fastText verwendet. fastText steht in der Python-Bibliothek `Gensim` mit diversen Methoden zur Verfügung. Die Parameter, welche verwendet werden, orientieren sich an den verwendeten Parametern von Grave et al. (2018). Wie Grave et al. wird auf CBOW mit Subword Information und Positionsgewichtung wie in Abschnitt 2.2.2 beschrieben zurückgegriffen. Als Tokenizer wird SoMaJo (Proisl und Uhrig, 2016) als ein explizit für Deutsch entwickelter Tokenizer verwendet.

3.2.3 Training

Das Training findet mit einem Datensatz mit ungefähr 10 Millionen Sätzen statt. Nach zwölf Epochen wird das Training abgebrochen aufgrund einer niedrigen Accuracy im Verhältnis zur Zeitdauer.

3.2.4 Evaluierung

Die Qualität des Embeddings wird mit der Word Analogy Task wie in 2.2.2 beschrieben evaluiert. Als Aufgaben-Datensatz werden die von Köper et al. (2015) zusammengestellten Analogie-Aufgaben genutzt. Für die Evaluation werden die 200.000 im Datensatz am häufigsten vorkommenden Wörter verwendet. Die Evaluation wird in zwei Konstellationen durchgeführt, einmal inklusive Wörter aus dem Evaluationsdatensatz, welche nicht Teil dieser 200.000 Wörter sind und einmal exklusive dieser Wörter. Mit

dem beschriebenen Training wird auf den kompletten Evaluationsdatensatz eine Accuracy von 27% erreicht. Ohne die Wörter, die nicht in den 200.000 Wörtern vorkommen, wird eine Accuracyy von 35% erreicht.

3.2.5 Fehlende Rechnerkapazität

Die nach den Trainingsdurchläufen niedrige Accuracy kann dadurch erklärt werden, dass größere Datensätze für das Training benötigt werden. Diese stehen bei Wortschatz Leipzig durchaus zur Verfügung, das Nadelöhr hier ist eher die zur Verfügung stehende Rechnerkapazität. Mit ausreichender Kapazität und/oder Zeit lässt sich das Modell gut trainieren. Dies würde aber den Rahmen dieser Bachelorarbeit sprengen.

3.3 Analyse auf rassistischen Bias

3.3.1 Vortrainiertes fastText Embedding

Parameter

Um mit einem qualitativ hochwertigen Embedding zu arbeiten, wird ein durch Grave et al. (2018) vortrainiertes deutsches fastText Embedding genutzt. Bei diesem Embedding handelt es sich um ein CBOW Word2Vec Modell mit
Subword Information und Positionsgewichtung wie in Abschnitt 2.2.2 beschrieben. Die Dimensionalität beträgt 300, die Character N-grams besitzen eine Länge von 5, das Kontextfenster hat eine Größe von 5 und das Negative Sampling wird mit jeweils 10 Negativbeispielen durchgeführt. Als Datengrundlage verwenden Grave et al. Common Crawl und Wikipedia. Es wird der Tokenizer der Europarl Preprocessing Tools verwendet (Koehn, 2005).

Common Crawl und Wikipedia

Grave et al. verwenden für das Training ihrer fastText Embeddings Daten von Common Crawl und Wikipedia. Common Crawl ist eine Non-Profit Organisation, welche das Internet durchsucht, Ergebnisse filtert sowie vorverarbeitet und daraus HTML- oder Textdatensätze erstellt. Diese werden dann frei zur Verfügung gestellt. Grave et al. verwenden von Common Crawl den Textdatensatz von Mai 2017 in UTF-8. Nach Vorverarbeitung entsteht aus diesem Datensatz ein deutsches Vokabular mit über 19,7 Millionen Wörtern und über 65 Milliarden Tokens.

Wikipedia bietet eine große Anzahl an kuratierten und hochqualitativen Texten und wird in diversen Anwendungen verwendet. Auch ist Wikipedia immer wieder Gegenstand von Arbeiten bei der Untersuchung von sozialem Bias (vgl. Abschnitt 2.1.4). Die Texte sind frei verfügbar und können frei heruntergeladen werden. Der durch Grave et al. verwendete Wikipedia-Datensatz stammt vom 11. September 2017. Nach Vorverarbeitung ergibt sich ein deutsches Vokabular mit etwas über 3 Millionen Wörtern und über 1,38 Milliarden Tokens.

Für das Training des deutschsprachigen Embeddings benutzen Grave et al. diesen Wikipedia-Datensatz. Dieses erreicht eine höhere Accuracy von 73,9% in der Analogy-Task im Verhältnis zu 72,9% durch Training mit Common-Crawl Daten. Für die Analogy-Task verwenden Grave et al. den Analogiedatensatz von Köper et al. (2015).

3.3.2 Generierung von Analogien mit fastText

Bei der Umfrage, bei der Teilnehmende Analogien darauf bewerten, ob sie rassistisch sind, werden Aussagen mithilfe des vortrainierten Embeddings generiert. Dafür wird die in 2.3.1 vorgestellte Metrik in Python implementiert.

Es wird in erster Linie auf die Parameter von Bolukbasi et al. (2016) zurückgegriffen, $\delta = 1$ gesetzt und die Vektoren von den in Abschnitt 2.3.2 aufgestellten Begriffen für $\vec{a}$ und $\vec{b}$ genutzt. Für $\vec{x}$ werden Begriffe aus der Liste von Schulabschlüssen aus Abschnitt 2.3.3 eingefügt. In der Tabelle 3.1 sind drei Beispiele für generierte Analogien aufgelistet. Ein Problem, welches an den ersten zwei Beispielen deutlich wird, ist die Nähe der resultierenden Wörtern für $\vec{y}$. Die Richtung der gestellten Aufgabe ist annähernd austauschbar. Ein zweites Problem, das auffällt, lässt sich anhand der dritten Analogie in der Tabelle 3.1 erkennen. Zwar macht die Analogie von „Matura“ zu „Abitur“ als Abschlüsse grundsätzlich Sinn, der Strich am Ende von „Matura-“ ist aber Überflüssig. Eine Erklärung ist, dass die Schranke für Ähnlichkeit δ nicht optimal ist. Eine zu lose Schranke kann dazu führen, dass Analogien generiert werden, welche zu wenig Bezug zum gegebenen Kontext haben. Aus diesem Grund wird δ auf 0,9 herabgesetzt, um so Ergebnisse zu erhalten, welche näher an dem gegebenen Kontext liegen.

Tabelle 3.1: Beispielergebnisse für die Generierung von Analogien nach der Gleichung 2.14 mit $\delta = 1$

$\delta = 1$
Muslimisch : Hauptschulabschluss :: Weiß : Schulabgang
Weiß : Hauptschulabschluss :: Muslimisch : SchulabgängerInnen
Weiß : Abitur :: Jüdisch : Matura-

Mit dem Herabsetzen von δ auf 0,9 werden jetzt Analogien generiert, welche näher am vorgegebenen Kontext liegen. Drei Beispiele sind in Tabelle 3.2 dargestellt. In den ersten zwei Beispielen sind „Weiß“ und „Muslimisch“ nicht mehr annähernd austauschbar. Ein neues Problem ist die Generierung von zwei dem dritten Beispiel entsprechenden Analogien, bei denen lediglich die Groß- und Kleinschreibung von „Abitur“ der Unterschied zwischen x und y ist. Da aber insgesamt über alle 36 Analogien die Ergebnisse näher am gesuchten Konzept von Schulabschlüssen liegen als bei einer Schranke für Ähnlichkeit von $\delta = 1$, werden diese Ergebnisse für die Umfrage in Abschnitt 3.3.3 weiter verwendet. Die generierten Analogien sind in Tabelle 3.4 aufgelistet.

Tabelle 3.2: Beispielergebnisse für die Generierung von Analogien nach der Gleichung 2.14 mit $\delta = 0{,}9$

$\delta = 0{,}9$
Muslimisch : Hauptschulabschluss :: Weiß : Bildungsabschluss
Weiß : Hauptschulabschluss :: Muslimisch : Hauptschulabsolventen
Weiß : Abitur :: Jüdisch : abitur

3.3.3 Umfrage

Aufbau

Um grundsätzlich festzustellen, ob das untersuchte Embedding einen rassistischen Bias aufweist, wird eine Umfrage mit den Analogien aus dem vorherigen Abschnitt durchgeführt. Diese wird mithilfe eines Limesurvey-Online-Fragebogens umgesetzt. Teilnehmende beantworten Fragen der Form „Ist die Aussage „*Schwarz* und *Weiß* stehen im gleichen Verhältnis zueinander wie *Hauptschulabschluss* und *Realschulabschluss*“ rassistisch?“. Dabei sind „Schwarz“, „Weiß“, „Hauptschulabschluss“ und „Realschulabschluss“ der generierten Analogie *Schwarz : Hauptschulabschluss :: Weiß : Realschulabschluss* entnommen. „a is to x as b is to y“ wird nicht direkt als „a verhält sich zu x wie b zu y“ übersetzt, um für die Teilnehmenden eine präzisere Formulierung der Fragen zu erhalten. Außerdem findet eine Korrektur der Groß- und Kleinschreibung der generierten Analogien statt und werden alle Begriffe der Analogien großgeschrieben, damit der Faktor Rechtschreibung keinen Einfluss auf die Bewertungen durch die Teilnehmenden hat. Bei den Fragen handelt es sich um Ja-Nein-Fragen. Hiermit die Methodik ähnlich wie die Umfrage von Bolukbasi et al. (2016) umgesetzt. Die Teilnehmenden sollen bewerten, ob sie eine gegebene Aussage als rassistisch einschätzen. Es wird das Set an generierten Analogien aus Abschnitt 3.3.2 verwendet. Dieses Set umfasst insgesamt 36 Analogien, sechs pro Bias definierende Gruppe immer in Kontrast zu „Weiß“. Für jede Gruppe ist pro verwendeten Schulabschluss eine Analogie in jede Richtung generiert. Richtung bedeutet hier, ob $x =$ *Weiß* oder $y =$ *Weiß* ist. Um die Umfrage für die Teilnehmenden übersichtlich zu halten, müssen sie lediglich 18 von 36 Analogien bewerten. Dafür werden aus jeder der sechs Gruppen drei Analogien randomisiert ausgewählt.

Um besser abschätzen zu können, wie die Zusammensetzung der Teilnehmenden ist, wird nach einigen demografischen Daten gefragt. Es werden Alter, Geschlecht, Rassismuserfahrung, Gruppenzugehörigkeit und höchster schulischer Abschluss abgefragt. Eine Vermutung ist, dass Alter einen Unterschied bei der Beantwortung der Fragen machen könnte. Ähnlich wird die Vermutung aufgestellt, dass auch Geschlecht einen Unterschied machen könnte. Bei Geschlecht sind „Weiblich“, „Männlich“, „Nichtbinär“, „Inter*“ und ein Feld für andere Eingaben vorhanden. Da es bei der Umfrage um Rassismus geht, wird nach Rassismuserfahrung gefragt. Hier stehen die Optionen „Ja“ und „Nein“ zur Verfügung. Die Optionen bei Gruppenzugehörigkeit orientieren

sich an den im Nationaler Diskriminierungs- und Rassismusmonitor (NaDiRa) aufgeführten Gruppen. Außerdem wird die Gruppe „Weiß“ sowie ein optionales Textfeld ergänzt. Das optionale Textfeld dient dazu, Personen, welche sich durch keine der vorgegebenen Gruppen repräsentiert sehen, die Möglichkeit zu geben, eine Angabe zu machen. Die Frage nach eigener Betroffenheit ist interessant, da diese Gruppen ein Untersuchungsgegenstand dieser Arbeit sind und eigene Betroffenheit ein interessanter Faktor ist bei der Bewertung der Analogien. Als letztes wird nach dem höchsten Schulabschluss gefragt und dafür die in Abschnitt 2.3.3 festgelegten Schulabschlüsse verwendet. Außerdem werden die beiden Optionen „Noch in Ausbildung“ und „Keinen Abschluss“ hinzugefügt. Die Fragen nach Geschlecht und Gruppenzugehörigkeit sind bewusst mit einem „Sonstiges“-Feld ausgestattet, um hier genutzter Selbstidentifikation ausreichend Rechnung tragen zu können. Bei Gruppenzugehörigkeit wird die Auswahl mehrerer Optionen zugelassen, da diese sich nicht gegenseitig ausschließen. Eine Person kann sich z.B. als muslimisch und asiatisch gleichzeitig bezeichnen.

Durchführung

Die Durchführung fand vom 24.08.2023 bis zum 06.09.2023 statt. Verbreitet wurde die Umfrage auf diversen Plattformen von gemeinnützigen Organisationen, Unternehmen sowie intern bei Integreat, über Soziale Medien (Instagram) und unter Studierenden. Während der laufenden Umfrage wurden zweimal kleinere Rechtschreibfehler ausgebessert, sonst aber keine Änderungen vorgenommen. Insgesamt wurde die Umfrage 171 Mal begonnen, davon wurde sie 107 Mal vollständig ausgefüllt. Von den 107 vollständigen Einträgen wurde ein Eintrag gelöscht, da dieser offensichtlich nicht ernsthaft ausgefüllt wurde. Schlussendlich ergibt das 106 vollständig ausgefüllte Einträge, welche im Folgenden ausgewertet werden.

Auswertung

Als Kriterium dafür, ob eine Analogie insgesamt als rassistisch bewertet wurde, wird festgelegt, dass zumindest die Hälfte der Teilnehmenden eine Analogie als rassistisch bewertet haben. Das Kriterium entspricht dem der Umfrage von Bolukbasi et al. Insgesamt werden 106 vollständige Einträge ausgewertet. Zuerst findet eine Aufschlüsselung der Angaben bei den demografischen Fragen statt. Die demografischen Angaben sind in Tabelle 3.3 zusammengefasst. Von 106 Teilnehmenden gaben 57 (53,77%) an, weiblich zu sein, 43 (40,57%) männlich und 6 (5,66%) nicht-binär. Andere Angaben gab es keine. Es haben also deutlich mehr Frauen als Personen mit anderem Geschlecht teilgenommen. Bei der Frage nach Rassismuserfahrung gaben 18 Personen (16,98%) „Ja“ und 88 (83,02%) „Nein“ an. Dies liegt 5,22 Prozentpunkte unter dem Wert, welcher im NaDiRa ermittelt wurde. Bei der Frage nach der eigenen Gruppenzuordnung geben zwei Personen „Schwarz“ an (1,72%), eine Person „Jüdisch“ (0,86%), acht Personen „Muslimisch“ (6,90%), vier Personen „Asiatisch“ (3,45%), zwei Personen „Sinti und Roma“ (1,72%), fünf Personen „Osteuropäisch“ (4,31%) und 87 Personen „Weiß“ (75,00%). Neben den gegebenen Optionen gibt es sieben Angaben unter Sonstiges:

„Balkan“, „Arabisch“, „Halbjüdisch“, zwei Mal „Westeuropäisch“, einmal „Atheist/-Nihilist“ und einmal „Migrantisch“. Einige Teilnehmenden wählten mehrere Optionen aus bei der Beantwortung dieser Frage. Insgesamt geben 18 Personen an (16,98%), Teil einer der von uns vorgegebenen von Rassismus betroffenen Gruppen zu sein. Zum Vergleich: beim NaDiRa beträgt dieser Anteil 13,2%. Bei Schulbildung geben drei Personen an, noch in Ausbildung zu sein (2,83%), eine Person gibt an, einen Hauptschulabschluss zu haben (0,94%), fünf geben Realschulabschluss an (4,72%) und 97 Personen geben an, die Fach- oder Hochschulreife zu haben (91,51%). Es hat keine Person ohne Schulabschluss teilgenommen.

Tabelle 3.3: Aufschlüsselung der demografischen Angaben von Teilnehmenden

	Absolut	Anteilig
Insgesamt vollständig teilgenommen	106	100%
Weiblich	57	53,77%
Männlich	43	40,57%
Nicht-binär	6	5,66%
Rasssismuserfahrung	18	16,98%
Schwarz	2	1,72%
Jüdisch	1	0,86%
Muslimisch	8	6,90%
Asiatisch	4	3,45%
Roma und Sinti	2	1,72%
Osteuropäisch	5	4,31%
Weiß	87	75,00%
Sonstige Gruppen	7	6,03%
Noch in Ausbildung	3	2,83%
Haupt- /Mittelschulabschluss	1	0,94%
Realschulabschluss	5	4,72%
Fach- /Hochschulreife	97	91,51%

Eine Analogie gilt als rassistisch wenn mindestens 50% der ausfüllenden Personen diese als rassistisch bewerten. Insgesamt werden so 26 von 36 Analogien als rassistisch bewertet. Die Analogie mit der höchsten Zustimmung erreicht 87,27%, die niedrigste Zustimmungsrate ist 9,62%. Die Zustimmungsraten sind pro Analogie in Tabelle 3.4 aufgelistet. Für eine genauere Aufschlüsselung nach den Gruppen aufgrund der demografischen Angaben werden Gruppen mit einer Größe von zumindest zehn Personen betrachtet, um ausreichend begründete Aussagen treffen zu können. Mit kleineren Gruppen existiert das Risiko, das nicht alle Analogien durch diese Gruppe bewertet wurden. Fokus fällt insbesondere auf die Kategorien Alter, Geschlecht, Rassismuserfahrung und Gruppenzuordnung. Die Anzahl von als rassistisch bewerteten Analogien per Gruppe, ausgenommen Alter, findet sich in Abbildung 3.1. Die Aufteilung nach

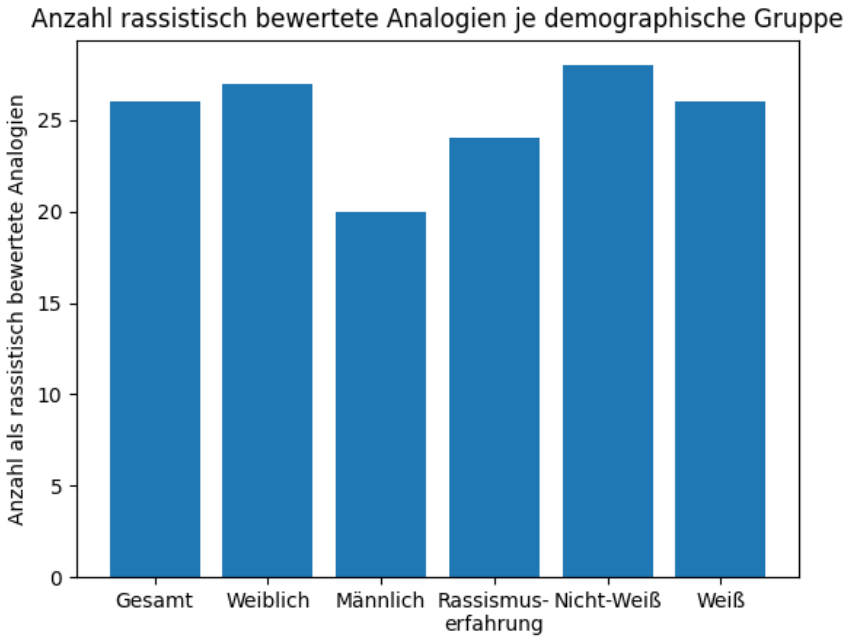


Abbildung 3.1: Anzahl der 36 Analogien, welche durch mehr als 50% der ausfüllenden Personen als rassistisch bewertet werden, aufgeschlüsselt nach demografischen Gruppen. „Nicht-Weiß“ umfasst alle Personen welche sich einer nicht-weißen Gruppe zugeordnet haben bzw. aufgrund ihrer Angabe unter sonstiges unter der Definition einer nicht-weißen Gruppe fallen.

Alter ist in Abbildung 3.3 dargestellt.

Bei einer genaueren Betrachtung von Geschlecht ergibt sich folgendes Bild: Frauen geben an, dass 27 von 36 Analogien rassistisch sind, Männer geben an, dass 20 von 36 Analogien rassistisch sind und nicht-binäre Teilnehmende geben an, dass 24 von 36 Analogien rassistisch sind. Bei nicht-binären Personen sei darauf hingewiesen, dass es sich hier um Angaben von sechs Personen handelt, hieraus also wenig Schlüsse gezogen werden können. Auffällig ist der starke Unterschied zwischen den Bewertungen von Männern und Frauen, welcher auch definitiv der stärkste Effekt zwischen verschiedenen Gruppen in dieser Umfrage ist. Das ist auch in Abbildung 3.1 visuell deutlich erkennbar. Dazu kommt, dass Frauen viele Analogien deutlich sicherer als rassistisch bewerten als Männer. 15 Analogien werden von 80% oder mehr Frauen als rassistisch bewertet und auch mit über 90% oder mehr werden immer noch fünf Analogien als rassistisch bewertet. 80% oder mehr der Männer beurteilen hingegen nur zwei Analogien als rassistisch. Diese Zahlen sind auch noch mal in Abbildung 3.2 dargestellt. Die blaue Linie zeigt die Anzahl der durch Frauen als rassistisch bewertete Analogien und den erreichten Prozentsatz der Bewertungen. Die orange Linie zeigt die Bewertungen durch Männer. Einen Effekt zu beobachten ist nicht überraschend, dass dieser so stark

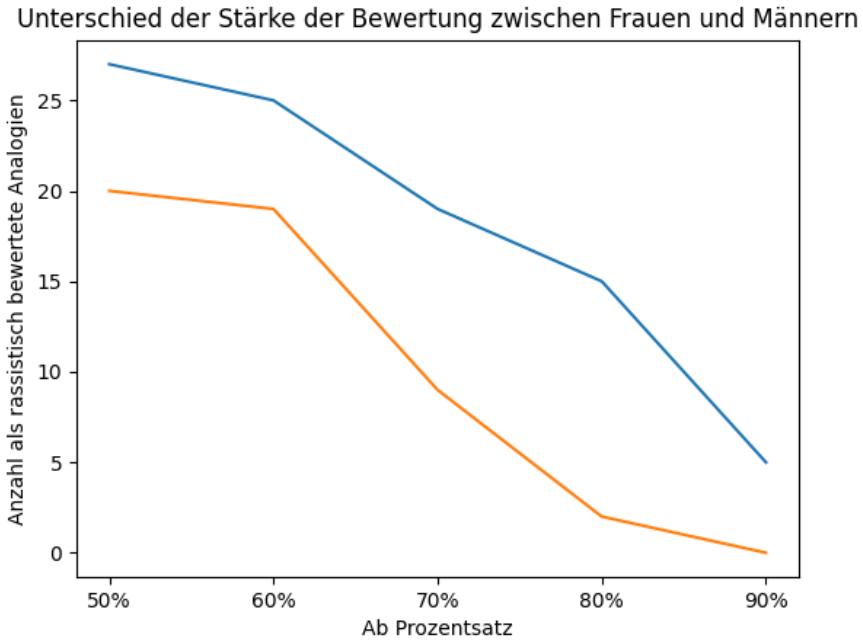


Abbildung 3.2: Anzahl der Analogien Welche ab entsprechenden Prozentsatz der Frauen (blaue Linie) oder Männer (orange Linie) als rassistisch bewertet wurden.

ausfällt, ist auffällig.

Bei der Aufschlüsselung nach Alter wird zwischen Gruppen von Teilnehmenden unter 30 mit insgesamt 73 Personen, Teilnehmenden zwischen 30 und 40 mit insgesamt 20 Personen sowie Teilnehmenden über 40 Jahren mit insgesamt 13 Personen unterschieden. Bei den Teilnehmenden unter 30 Jahren werden 26 Analogien als rassistisch bewertet, bei Teilnehmenden ab 30 bis 39 werden 25 Analogien als rassistisch bewertet und bei Teilnehmenden ab 40 Jahren werden 24 Analogien als rassistisch bewertet. Es lässt sich zwar ein Abnehmen der als rassistisch bewerteten Analogien mit steigender Altersgruppe beobachten, der Effekt fällt aber deutlich kleiner aus als der beim Faktor Geschlecht. Diese Abnahme ist in Abbildung 3.3 dargestellt.

Als Nächstes wird die Frage nach Rassismuserfahrung aufgeschlüsselt. Personen, welche Rassismuserfahrung angegeben haben, bewerten 24 Analogien als rassistisch. Auffällig ist hier die Differenz zu der Gesamtbewertung. Beim genaueren Betrachten der Angaben fallen einige Einträge auf, bei denen Personen Rassismuserfahrung und gleichzeitig bei Gruppenzugehörigkeit „Westeuropäisch“ oder „Weiß“ angegeben haben. Diese Angaben stehen im Widerspruch zu den verwendeten Definitionen. Eine Erklärung ist, dass die Frage nach Rassismuserfahrung sehr subjektiv ist und schwer Schlüsse zulässt.

Die letzte demografische Kategorie, die näher betrachtet wird, ist die Gruppenzugehörigkeit. Da abgesehen von „Weiß“ keine Kategorie von mehr als acht Teilnehmende

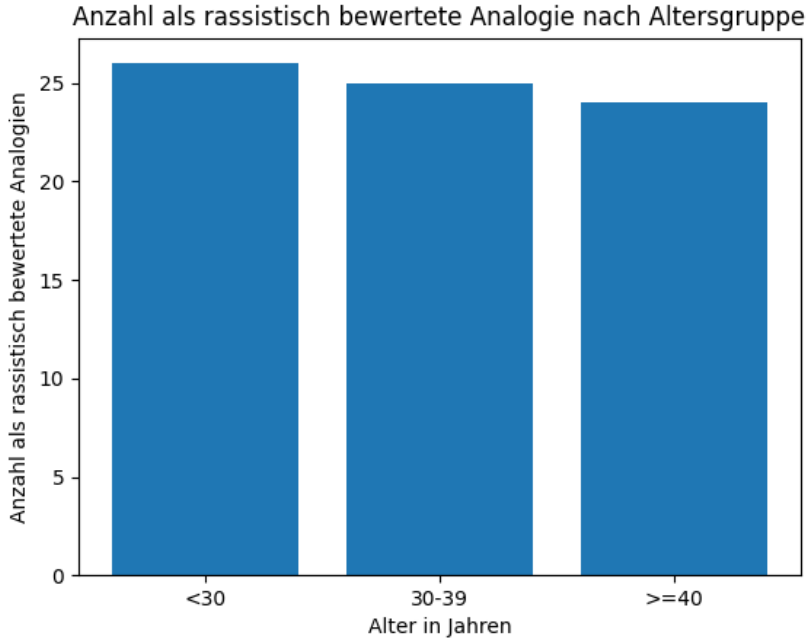


Abbildung 3.3: Anzahl als rassistisch bewerteter Analogien nach Altersgruppen.

angegeben wurde, werden diese im Folgenden als „Nicht-Weiß“ zusammengefasst. Darin werden alle inkludiert, die zusätzlich „Weiß“ angegeben haben, da diese Personen trotzdem Teil einer von Rassismus betroffenen Gruppe sind. Bei sonstigen Angaben werden für die Kategorie „Nicht-Weiß“ die Angaben selektiert, welche sich zu von Rassismus betroffenen Gruppen zuordnen lassen. Das umfasst die Angaben „Balkan“, „Arabisch“, „Halbjüdisch“ und „Migrantisch“. Bei mehrfachen Angaben durch eine Person wird die Person bei den Bewertungen nur einfach gezählt. Mit diesem Rahmen werten Personen der Gruppe „Nicht-Weiß“ 28 Analogien als rassistisch. Dies liegt zwei Analogien über dem Gesamtdurchschnitt.

Als Nächstes wird betrachtet, wie jede Analogiegruppe abschneidet. Eine Analogiegruppe umfasst alle Analogien welche eine von Rassismus betroffenen Gruppe Korrespondieren. Hier wird im Vorhinein die Vermutung aufgestellt, dass für Analogien mit „Jüdisch“ und „Muslimisch“ weniger als rassistisch bewertet werden. Eine Begründung dafür ist, dass Teilnehmende diese als Religionszuschreibungen interpretieren könnten und diese deshalb losgelöst von der Frage nach Rassismus sehen. Bei „Jüdisch“ kommt noch der Diskurs zum Verhältnis von Antisemitismus zu Rassismus als Faktor hinzu, welcher einen Einfluss auf die Bewertung nehmen könnte. Jede Gruppe von Analogien enthält sechs Analogien. Analogien mit „Schwarz“ werden viermal als rassistisch bewertet, „Jüdisch“ fünfmal, „Muslimisch“ viermal, „Osteuropäisch“ viermal, „Roma“ viermal und „Asiatisch“ fünfmal. Ein größerer Unterschied zeigt sich erst, wenn betrachtet wird, wie viele Teilnehmende die einzelnen Analogien als rassistisch bewerten. Keine Analogie mit „Jüdisch“ wird durch 70% oder mehr der Teilnehmenden als

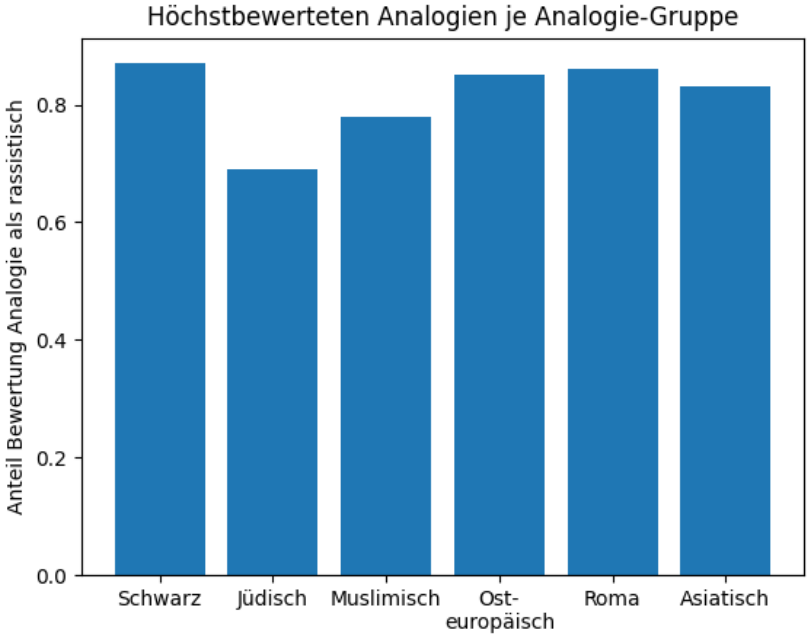


Abbildung 3.4: Anteil Bewertungen mit „Ja“ von Analogien mit maximalem Wert als rassistisch pro Analogiegruppe.

rassistisch bewertet. Zwei Analogien mit „Muslimisch“ werden noch durch 70% oder mehr der Teilnehmenden als rassistisch gewertet. Alle anderen Gruppen erreichen hier noch drei oder mehr Analogien. Zumindest eine Bewertung als rassistisch durch über 80% oder mehr der Teilnehmenden erreichen alle außer „Jüdisch“ und „Muslimisch“. Um den Effekt zu verdeutlichen, werden in Abbildung 3.4 der maximale Anteil an „Ja“-Bewertungen für eine Analogie der jeweiligen Analogiegruppen dargestellt. Es ist erkennbar, dass der maximale Wert für die jüdische Analogiegruppe niedriger ist wie der Wert der anderen Gruppen. Auch die muslimische Analogiegruppe hat einen etwas niedrigeren maximalen Wert. Hier ist eine Tendenz in Richtung Anfangsvermutung erkennbar, die weiter untersucht werden könnte.

Der letzte Punkt untersucht die Relevanz der Richtung in die Analogiefindungsaufgaben gestellt werden. Von den 18 Analogien, welche mit $a = \textit{Weiß}$ erstellt wurden, werden 11 als rassistisch bewertet. Von den 18 Analogien, bei denen *Weiß* als Eingabe für b diente, werden 15 als rassistisch bewertet. Es existiert also ein Unterschied bzgl. der Richtung der Analogien. Eine mögliche Erklärung ist die Art und weiße, wie die Begriffe zusammen in den Trainingsdaten stehen. „Weiß“ könnte öfter zusammen mit höheren Bildungsabschlüssen auftauchen, die anderen Gruppen könnten hingegen mehr mit allen Abschlüssen verknüpft werden. Dies kann dazu führen das "´Weiß“ oft eine „Aufwertung“ des Abschlusses in der Analogie erfährt, wenn ein Abschluss für „Weiß“ gesucht wird. In die andere Richtung könnte durch eine breitere Assoziation Analogien generiert werden, welche für y mit x vergleichbare Begriffe beinhalten. Dies

müsste aber genauer untersucht werden, um andere Effekte wie Fragebogengestaltung auszuschließen.

Zusammenfassung

Den wichtigsten Schluss, welcher aus der Auswertung der Umfrage gezogen wird, ist, dass 72,23% der Analogien als rassistisch bewertet werden. Zum Vergleich: in der Arbeit von Bolukbasi et al. (2016) werden 29 von 150 Analogien als mit einem problematischen Gender Bias bewertet, was einem Prozentsatz von 19,34% entspricht. Dies begründet die Feststellung, dass das in dieser Arbeit betrachtete fastText Embedding einen rassistischen Bias aufweist, der Auswirkungen auf die dem Embedding gestellten Aufgaben haben kann. Eine zusätzliche Beobachtung ist, dass die Reihenfolge der Parameter bei der Generierung von Analogien einen Einfluss haben können.

Eine Feststellung neben den Beobachtungen, welche das Embedding direkt betreffen ist, dass das Geschlecht der teilnehmenden Personen einen großen Effekt auf die Bewertung hat. Kleinere Auswirkungen, dennoch aber erkennbare Tendenzen, haben die Kategorien Alter und Gruppenzugehörigkeit. Andere Gruppen sind zu klein, um sie näher zu betrachten. Generell gilt, dass Schlüsse aus den demografischen Daten schwierig zu ziehen sind, sich aber Tendenzen zwischen verschiedenen Kategorien erkennen lassen, welche damit zusammenpassen, das die Perspektiven von Männern sich eher widerspiegeln in Wikipedia-Datensätzen (vgl. Adams et al. (2019), Field et al. (2022)).

3.3.4 Rassistischer Bias Subspace in fastText

Zusätzlich dazu, dass die Umfrage das Vorhandensein eines rassistischen Biases im Embedding bestätigt, bestätigt diese auch, dass die Bias definierenden Begriffe, welche in Abschnitt 2.3.2 festgelegt werden, dazu geeignet sind, einen rassistischen Bias im Embedding zu finden. In diesem Abschnitt wird mithilfe dieser Begriffe der Bias Subspace ausfindig gemacht. Dafür wird das in Abschnitt 2.3.1 erläuterte Verfahren in Python mithilfe von den Bibliotheken `Numpy` und `scikit-learn` implementiert.

Für die Bias definierenden Begriffe werden hier die ersten fünf Hauptkomponenten des Embeddings analysiert. Dabei hat die erste Hauptkomponente (PC1) einen Varianzanteil von 63,5%, die zweite (PC2) 12,2%, die dritte (PC3) 8,7%, die vierte (PC4) 6,6% und die fünfte (PC5) 4,8%. In Abbildung 3.5 sind die Varianzverhältnisse als Balkendiagramm visualisiert. Es lässt sich deutlich erkennen, dass die erste Hauptkomponente den eindeutig größten Anteil an die Richtung der Bias definierenden Begriffe hat. Die Differenz zwischen der ersten und zweiten Hauptkomponente beträgt 51,3 Prozentpunkte. Zum Vergleich: Chen et al. (2021) erhalten beim Bestimmen eines Gender Bias Subspace eine Differenz zwischen erster und zweiter Hauptkomponente von 17 Prozentpunkte, bei Bolukbasi et al. (2016) eine Differenz von ca. 50 Prozentpunkten. Bei Bolukbasi et al. erreicht die erste Hauptkomponente einen ähnlichen Anteil wie bei unserer PCA mit ca. 60%. Zusammenfassend lässt sich schlussfolgern, dass die erste Hauptkomponente den Bias Subspace für Rassismus ausreichend reprä-

sentiert. Daher wird dieser als Bias Subspace für die Analyse von Bildungsabschlüssen im nächsten Abschnitt verwendet.

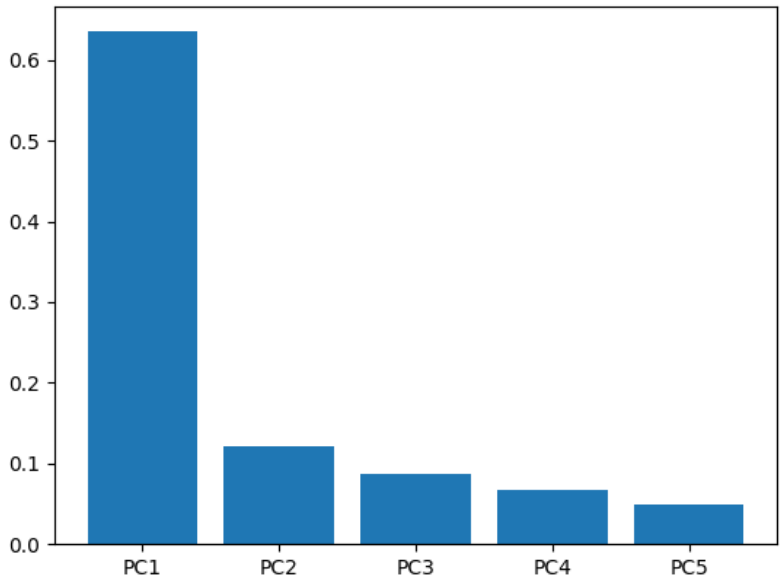


Abbildung 3.5: Varianzanteil der Hauptkomponenten der Bias definierenden Begriffe im verwendeten fastText-Embedding.

3.3.5 Rassistischer Bias in Schulabschlüssen

Nachdem im vorherigen Abschnitt der Bias Subspace bestimmt wurde, werden mit der in Abschnitt 2.3.1 beschriebenen Metrik für direkten Bias, Begriffe für Schulabschlüsse auf ihren Bias-Gehalt untersucht. Dafür wird die Gleichung 2.16 in Python implementiert.

Als Parameter dienen die Word Vectors der drei in 2.3.3 festgelegten Schulabschlüsse und der in Abschnitt 3.3.4 festgestellte Bias Subspace. Als Maßgabe für Striktheit wird wie durch Bolukbasi et al. (2016) und Chen et al. (2021) $c = 1$ verwendet. Mit diesen Parametern resultiert das in einem Wert von $DB = 0{,}08$. Die Metrik lässt sich folgendermaßen deuten, ein Wert für Cosine Similarity von 0 würde bedeuten, dass ein Word Vector senkrecht auf dem Bias Subspace steht. Es existiert dann keine Ähnlichkeit zwischen dem Word Vektor und dem Bias Subspace und sie teilen sich keine Komponente. Jeder Wert, welcher nicht 0 entspricht, bedeutet, dass ein geteilter Anteil existiert. Ein Wert von 0,08 bedeutet also, dass die Begriffe einen Bias beinhalten. Bolukbasi et al. erhalten mit ihrem untersuchten Embedding ein Wert für deren Set von Beschäftigungen von 0,08 für einen Gender Bias. Chen et al. erhalten in dem von ihnen untersuchten deutschen Embedding einen Wert von 0,04 für einen männlichen

Gender Bias in deren Set von Beschäftigungen. Manzini et al. (2019) erhalten mit ihrer Metrik, dem Mean Average Cosnie Distance, einen Wert von 0,892 für rassistischen Bias. Da die Mean Average Cosnie Distance statt der Cosine Similarity die cosine Distance verwendet, kann dieser Wert mit unserem verglichen werden, indem das Komplemnt unseres Wertes berechnet wird: $1 - 0{,}08 = 0{,}92$. Zusammenfassend begründet das, dass ein Bias bei den Bildungsabschlüssen im untersuchten Embedding festgestellt wird.

Tabelle 3.4: Liste der generierten Analogien und der Anteil der Personen welche diese als rassistisch bewerten

Generierte Analogie	Bewertung
Schwarz : Hauptschulabschluss :: Weiß : realschulabschluss	87,27%
Weiß : Hauptschulabschluss :: Schwarz : Lehrabschluss	56,36%
Schwarz : Realschulabschluss :: Weiß : realschulabschluss	14,29%
Weiß : Realschulabschluss :: Schwarz : Volksschulabschluss	83,67%
Schwarz : Abitur :: Weiß : Abiturdurchschnitt	29,79%
Weiß : Abitur :: Schwarz : Volksschulabschluss	85,71%
Jüdisch : Hauptschulabschluss :: Weiß : Bildungsabschluss	66,67%
Weiß : Hauptschulabschluss :: Jüdisch : Hochschulreife	69,09%
Jüdisch : Realschulabschluss :: Weiß : Schulabschluss	65,57%
Weiß : Realschulabschluss :: Jüdisch : Hochschulreife	65,31%
Jüdisch : Abitur :: Weiß : Vorbereitungsstudium	55,93%
Weiß : Abitur :: Jüdisch : abitur	9,62%
Muslimisch : Hauptschulabschluss :: Weiß :Bildungsabschluss	71,43%
Weiß : Hauptschulabschluss :: Muslimisch : Hauptschulabsolventen	27,27%
Muslimisch : Realschulabschluss :: Weiß : Schulabschluss	64,15%
Weiß : Realschulabschluss :: Muslimisch : Hauptschulabsolventen	43,14%
Muslimisch : Abitur :: Weiß : Mathematik-Studium	50,85%
Weiß : Abitur :: Muslimisch : Turbo-Abitur	78,18%
Asiatisch : Hauptschulabschluss :: Weiß : Abitur	83,33%
Weiß : Hauptschulabschluss :: Asiatisch : Pflichtschulabschluss	57,63%
Asiatisch : Realschulabschluss :: Weiß : Abitur	77,78%
Weiß : Realschulabschluss :: Asiatisch : Pflichtschulabschluss	77,55%
Asiatisch : Abitur :: Weiß : Mathematik-Studium	72,73%
Weiß : Abitur :: Asiatisch : abitur	19,15%
Osteuropäisch : Hauptschulabschluss :: Weiß : Abitur	85,19%
Weiß : Hauptschulabschluss :: Osteuropäisch : Pflichtschulabschluss	60,47%
Osteuropäisch : Realschulabschlus :: Weiß : Abitur	26,67%
Weiß : Realschulabschluss :: Osteuropäisch : Pflichtschulabschluss	74,58%
Osteuropäisch : Abitur :: Weiß :Abiturexamen	84,31%
Weiß : Abitur :: Osteuropäisch : Fremdsprachenstudium	43,14%
Roma : Hauptschulabschluss :: Weiß :Abitur	86,21%
Weiß : Hauptschulabschluss :: Roma : Hauptschulabsolventen	20,75%
Roma : Realschulabschluss :: Weiß : Fachabi	80%
Weiß : Realschulabschluss :: Roma : Hauptschulabsolventen	83,67%
Roma : Abitur :: Weiß : Vordiplom	78%
Weiß : Abitur :: Roma : Hochschulreife	27,08%

4 Diskussion und Ausblick

Zusammenfassung der Ergebnisse

Diese Arbeit untersucht Methoden zur Identifizierung und Analyse von rassistischen Bias in deutschen Word Embeddings. Zuerst werden Grundlagen zu Rassismus und dem deutschen Diskurs eingeführt sowie Grundlagen zu Embeddings erläutert. Es werden Bias definierende Begriffe festgelegt und aufgezeigt, warum Bildung und Bildungsabschlüsse relevant sind zu untersuchen. Mit den eingeführten Methoden zur Generierung von Analogien wird eine Liste an Analogien erzeugt welche die Beziehung zwischen Race und Bildungsabschlüssen in einem fastText Embedding repräsentieren.

Mittels einer Umfrage mit 106 Teilnehmenden stellen wir fest, dass 26 der generierten Analogien einen rassistischen Bias beinhalten und somit das Embedding einen rassistischen Bias reproduziert. Außerdem bestätigt dies, dass die Bias definierenden Begriffe dazu geeignet sind, um den Bias Subspace für Race bzw. Rassismus genauer zu bestimmen. Eine Auffälligkeit der Umfrage ist, dass Geschlecht einen erheblichen Einfluss auf die Beurteilung von Analogien als rassistisch hat. Durch Männer wurden 20 von 36 Analogien als rassistisch bewertet, durch Frauen 27 von 36. Dies ist zwar keine völlig unerwartete Beobachtung da NLP-Modelle die mit Datensätze von Wikipedia trainiert wurden dazu neigen hauptsächlich die Perspektive von Männern zu übernehmen (Field et al., 2021), nichtsdestotrotz ist dies aber auffällig.

Mittels PCA wird festgestellt, dass bereits die erste Hauptkomponente der Bias definierenden Begriffe ausreicht, um den Bias Subspace zu beschreiben. Die erste Hauptkomponente besitzt einen Varianzanteil von 63,5%. Das Prüfen von Schulabschlüssen auf direkten Bias ergibt einen Wert von 0,08. Dieser Wert deutet auf einen rassistischen Bias in den Begriffen hin. Auch wenn dieser klein erscheint, so zeigt sich in den generierten Analogien und der Umfrage, dass dies ausreicht, um entsprechende Ergebnisse zu produzieren.

Bedeutung dieser Arbeit

Die Ergebnisse entsprechen Erwartungen, welche sich aufgrund vergleichbarer Arbeiten ergeben. In englischsprachigen Embeddings hat Manzini et al. (2019) einen rassistischen Bias festgestellt und mit einer vergleichbaren Metrik für direkten Bias ein ähnliches Ergebnis für die durch ihn verwendete Begriffe erhalten. Bolukbasi et al. (2016) erhalten auch einen Bias Subspace mit der ersten Hauptkomponente mit einem Varianzanteil von ca. 60%. In ihrer Umfrage werden 29 von 150 Analogien bewertet, dass sie einen problematischen Gender Bias beinhalten. Wir übertreffen das Verhältnis deutlich mit 26 von 36 Analogien. Für direkten Bias erhalten sie auch einen Wert von

0,08. Chen et al. (2021) benötigen für ihrem Gender Bias in einem deutschsprachigen Embedding mehr als eine Hauptkomponente, erhalten bei ihren direkten Bias für Gender einen Wert von 0,04. Die Ergebnisse für direkten Bias sind also in vergleichbaren Größenordnungen. Die Umfrage in dieser Arbeit übertrifft die Bewertung von Bolukbasi et al. um einiges.

Die Beiträge dieser Arbeit lassen sich wie folgt zusammenfassen: Es wird ein Set an deutschen Bias definierender Wörter für Rassismus festgelegt, nachgewiesen, dass diese in Analogiefindungsaufgaben rassistische Analogien ergeben, ein Bias Subspace im analysierten Embedding bestimmt und einen direkten rassistischen Bias in Schulabschlüssen gemessen.

Limitationen

Einige Punkte müssen aber kritisch betrachtet werden bzw. sind noch offen. Die Teilnehmenden der Umfrage sind kein repräsentatives Abbild der Bevölkerung. Einige Bereiche können nicht tiefergehender analysiert werden da zu wenige Personen, auf die bestimmte Kategorien zutreffen, teilgenommen haben. Ein Beispiel ist, dass wir zwar feststellen, dass Geschlecht einen maßgeblichen Einfluss auf die Bewertung hat, können dies aber nur mit den binären Kategorien Mann und Frau untersuchen. Für Menschen außerhalb dieser beiden Kategorien haben zu wenige teilgenommen, um hier klare Aussagen zu treffen. Zudem wird beobachtet, dass fragwürdig ist, wie gut eine Abfrage nach „Rassismuserfahrung“ eine sinnvoll zusammenfassbare Gruppe repräsentiert aufgrund der sehr subjektiven Wahrnehmung Begriffs. Möglicherweise müssten andere Wege gefunden werden, dies abzufragen oder klarer formuliert werden, was gefragt ist. Hinzu kommt auch die Überrepräsentation von Menschen mit Abitur. Aussagen zu Menschen ohne Abitur lassen sich so nicht treffen.

Ein Kritikpunkt, der immer zu nennen ist, sobald Race mit Kategorien operationalisiert wird, inwiefern eine Kategorisierung notwendig ist. Es werden diverse von Rassismus betroffene Gruppen aus dem deutschen Kontext verwendet, dessen Definitionen selbst Teil eines andauernden Diskurses sind. Außerdem stellt sich die Frage, inwiefern die Auffassung eines binären Problems zwischen Weiß und Nicht-Weiß zu untersuchen, wirklich zutrifft oder ob das Problem besser als Multiklassenproblem zu behandeln wäre.

Ein starker Kritikpunkt durch Field et al. (2021), der generell an auf Bolukbasi et al. (2016) aufbauende Arbeit gerichtet ist, ist dass die Analyse stark auf den Analogieeigenschaften und der Cosine Similarity beruht. Es werden dadurch vor allem Strukturen in den Embeddings im Verhältnis der Parallelität zueinander untersucht. Ob und wie sich ein Bias in Embeddings clustert, wird kaum betrachtet. Es ist also nicht sicher, wie vollständig dieses Vorgehen einen Bias beschreibt.

Ausblick und Möglichkeiten für Integreat

Ein Ansatzpunkt, der weiter untersucht werden könnte, wäre eine Ausweitung auf mehr Begriffe als die drei untersuchten Bildungsabschlüsse. Es gibt viele Bereiche,

in denen Rassismus eine Rolle spielt und für die es sich lohnen könnte, sie zu analysieren. Bolukbasi et al. (2016) schlagen außerdem noch eine Metrik für indirekten Bias vor, um zu bestimmen, wie ein Bias in der Beziehung zweier Wörter zueinander wirkt. Dieser könnte auch noch weiter untersucht werden. Ein weiterer Punkt der noch betrachtet werden könnte, wäre die Zusammenwirkung verschiedener Arten von Bias. Wie in Abschnitt 2.1.1 dargelegt, ist Rassismus nicht eindimensional und können verschiedene Formen von Diskriminierung unterschiedlich zusammenwirken. Manzini et al. (2019) setzen hier bereits für englischsprachige Embeddings mit einer Methodik für Multiklassen-Bias an.

Für Integreat bedeutet diese Arbeit einen Aufschlag für die Auseinandersetzung mit Rassismus. Neben einem ersten grundsätzlichen befassen mit Rassismus in Sprachsystemen gibt es einige interessante Ansatzpunkte mit der existierenden Methodik. Sie kann möglicherweise dazu genutzt werden, in Texten anzuzeigen, welche Wörter eine größere Assoziation mit Race bzw. Rassismus haben. Diese Anzeige könnte möglicherweise als Indikator dafür dienen, ob ein Übersetzungsprogramm wie DeepL dazu neigen würde, diese Begriffe zu einem rassistischen Ergebnis zu übersetzen. Hier würde sich weitere Forschung anbieten, um dies genauer zu untersuchen.

Literatur

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing 3rd edition draft.

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

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. *arXiv preprint arXiv:2106.11410*.

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

Manzini, T., Yao Chong, L., Black, A. W., & Tsvetkov, Y. (2019). Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 615–621. https://doi.org/10.18653/v1/N19-1062

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M., & Matthews, J. (2021). Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 24–34. https://doi.org/10.1145/3461702.3462530

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 446–457.

Kraft, A. (2021). *Triggering models: Measuring and mitigating bias in german language generation* (Diss.). Universität Hamburg.

Gerard, S. (2017). Examining societal biases in word vector models trained on German language corpora.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 501–512.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of*

the 2021 ACM conference on fairness, accountability, and transparency, 610–623.

Arndt, S. (2021). Rassismus. In *Wie Rassismus aus Wörten spricht: (K)Erben des Kolonialismus im Wissensarchiv deutsche Sprache; ein kritisches Nachschlagwerk*. Arndt, Susan; Ofuatey-Alazard, Nadja.

Memmi, A. (1987). *Rassismus*. Europäische Verlagsanstalt.

Banton, M. (1977). The Idea of Race, London.

Miles, R. (2004). *Racism*. Routledge.

Crenshaw, K. W. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine. *The University of Chicago Legal Forum, 139*.

Makkonen, T. (2002). Multiple, compoud and intersectional discrimination: bringing the experiences of the most marginalized to the fore.

Marten, E., & Walgennbach, K. (2017). Intersektionale Diskriminierung. In *Handbuch Diskriminierung*. Scherr, Albert; El-Mafaalani, Aladin; Yüksel, Gökçen.

Arndt, S., & Hornscheidt, A. (2004). *Afrika und die deutsche Sprache: Ein kritisches Nachschlagewerk*. Unrast.

Roth, W. D. (2016). The multiple dimensions of race. *Ethnic and Racial Studies*, *39*(8), 1310–1338.

Foroutan, N., Ha, N., Kalter, F., Shooman, Y., & Sinanoglu, C. (2022). Rassistische Realitäten: Wie setzt sich Deutschland mit Rassismus auseinander?

Scherr, A. (2017). Diskirminierung von Roma und Sinti. In *Handbuch Diskriminierung*.

Hornscheidt, A. (2005). (Nicht)Benneungen: Critical Whiteness Studies und Linguistik. In *Mythen, Masken und Subjekte*. Eggers, Maureen Maisha; Kilomba, Grada; Piesche, Peggy; Arndt, Susan.

Hummrich, M. (2017). Diskriminierung im Erziehungssystem. In *Handbuch Diskriminierung*. Scherr, Albert; El-Mafaalani, Aladin; Yüksel, Gökçen.

Kaiser, A. (2010). Vornamen: Nomen est omen? Vorerwartungen und Vorurteile in der Grundschule. *Schulverwaltung. Zeitschrift für Schulleitung und Schulaufsicht*, *21*(2), 58–59.

Quehl, T. (2010). Immer noch die anderen? Ein rassismuskritischer Blick auf die Normalität schulischer Bildungsbenachteiligung. In *Rassismus Bildet*. Broden, Anne; Mecheril, Paul.

Reisigl, M. (2017). Sprachwissenschaftliche Diskriminierungsforschung. In *Handbuch Diskriminierung*. Scherr, Albert; El-Mafaalani, Aladin; Yüksel, Gökçen.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*, 6363–6381.

Sólmundsdóttir, A., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. (2022). Mean Machine Translations: On Gender Bias in Icelandic Machine Translations. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3113–3121.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598.

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Adams, J., Brückner, H., & Naslund, C. (2019). Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test". *Socius*, *5*, 2378023118823946.

Field, A., Park, C. Y., Lin, K. Z., & Tsvetkov, Y. (2022). Controlled analyses of social biases in wikipedia bios. *Proceedings of the ACM Web Conference 2022*, 2624–2635.

Blodgett, S. L., Wei, J., & O'Connor, B. (2018). Twitter universal dependency parsing for African-American and mainstream American English. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1415–1425.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Sommerauer, P., & Fokkens, A. (2019). Conceptual change and distributional semantic models: An exploratory study on pitfalls and possibilities. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 223–233.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Groenwold, S., Ou, L., Parekh, A., Honnavalli, S., Levy, S., Mirza, D., & Wang, W. Y. (2020). Investigating African-American Vernacular English in transformer-based text generation. *arXiv preprint arXiv:2010.02510*.

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868*.

Sweeney, L. (2013). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, *11*(3), 10–29.

M'charek, A. (2013). Beyond fact or fiction: On the materiality of race in practice. *Cultural anthropology*, *28*(3), 420–442.

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

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135–146.

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

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of artificial intelligence research*, *44*, 533–585.

Proisl, T., & Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. *Proceedings of the 10th Web as Corpus Workshop*, 57–62.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Köper, M., Scheible, C., & im Walde, S. S. (2015). Multilingual reliability and "semantic" structure of continuous word spaces. *Proceedings of the 11th international conference on computational semantics*, 40–45.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, *26*.

Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. *Proceedings of the eighteenth conference on computational natural language learning*, 171–180.

Bildungsstand: Verteilung der Bevölkerung in Deutschland nach höchstem Schulabschluss im Jahr 2022. (2022). https://de.statista.com/statistik/daten/studie/1988/umfrage/bildungsabschluesse-in-deutschland/

Goldhahn, D., Eckart, T., Quasthoff, U., et al. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. *LREC*, *29*, 31–43.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of machine translation summit x: papers*, 79–86.

Eidesstattliche Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe, und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat.

Alle Ausführungen der Arbeit, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Augsburg, 22. September 2023

(Jarl Hengstmengel)