JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG

Philosophische Fakultät

Institut für deutsche Philologie

Masterarbeit

Identifikation von Hassrede in Tweets mit Deep Learning

Studienfach Digital Humanities

angefertigt von

Thora Hagen

Würzburg, 2019

Masterarbeit in den Digital Humanities

Titel: **Identifikation von Hassrede in Tweets mit Deep Learning**

Betreuer:	Prof. Dr. Fotis Jannidis
Semester:	6
Nachname:	Hagen
Vorname:	Thora
Matrikelnummer:	1957145
Geburtsdatum:	27.05.1995

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CNN	Convolutional Neural Network
GRU	Gated Recurrent Units
KI	Künstliche Intelligenz
LSTM	Long Short-Term Memory
NetzDG	Netzwerkdurchsetzungsgesetz
NLP	Natural Language Processing
PCA	Principle Component Analysis
RNN	Recurrent Neural Network
SVM	Support Vector Machines
TFIDF	Term frequency-inverse document frequency

1. Einleitung

Informationszugang, Kommunikation, politisches Engagement – alle diese Punkte sowie vieles mehr wurde durch den Zugang der breiten Gesellschaft zum Internet erheblich erleichtert. Damit einher gehen einige unschöne Aspekte des Internets, gegen welche versucht wird anzugehen. Dazu gehören unter anderem die durch das Internet ebenso erleichterte Verbreitungen von Extremismus und Des- bzw. Falschinformation („Fake News"), sowie auch Hassrede (vgl. Baldauf et. al. 2018, 8). Dies sind beileibe keine neuen Mittel zur Interessensdurchsetzung, doch haben sie durch soziale Netzwerke einen derartigen Aufschwung genommen, sodass sie aktuell präsenter denn je sind. Einige Maßnahmen wurden bereits gegen jene nicht erwünschten Arten der Meinungsmache ergriffen, allerdings ohne genau zu wissen, wie genau Hassrede und Radikalisierung funktioniert. Denn die Forschung zum Thema Hassrede im Internet steht hauptsächlich vor dem Problem, dass das Thema und die Technologie weitaus schneller voranschreiten, als was in derselben Zeit erforscht werden kann (vgl. Neumann 2018, 5f.). Zudem ist das Thema sehr jung. Erst seit 2015 ist der Begriff „Hassrede" in der Öffentlichkeit in Deutschland präsent (vgl. Rafael und Ritzmann 2018, 11).

Besonders aufgrund der Flüchtlingskrise im Jahr 2015 kam es, dass der öffentliche Hass zugenommen hat und das Thema Hassrede somit überhaupt erst in das allgemeine Interesse gerückt ist. Hassrede im Internet wird dazu zunehmend häufiger unter bürgerlichem Namen verbreitet. Dies weist darauf hin, dass Hass zurzeit eine „Normalisierung" durchläuft, was bedeuten kann, dass aktuell eine „Grenzverschiebung dessen, was öffentlich sagbar sei" geschieht (Meßmer und Krause 2018, 5). Umso mehr verbreitet sich auch der Wunsch nach Deradikalisierung in der Gesellschaft. Hinsichtlich des Tempos und des Ausmaßes des Themas hinken die Lösungsansätze allerdings, wie bereits angedeutet, hinterher (vgl. Baldauf et. al. 2018, 8).

Die fortschreitende Digitalisierung hat es möglich gemacht, auf immer mehr Daten digital zugreifen und diese somit auch computergestützt auswerten zu können. Insbesondere maschinelle Lernverfahren aus dem Bereich der künstlichen Intelligenz (KI) konnten so zunehmend häufiger als Lösungen für Probleme angewandt werden. Deep Learning ist einer von vielen Ansätzen im Bereich des überwachten maschinellen Lernens. Dieser Ansatz ist insofern besonders, als dass er erst seit wenigen Jahren Anwendung findet, da zuvor die notwendigen Ressourcen wie genügend Daten aber auch Rechenleistung fehlten. Bei klassischen, überwachten maschinellen Lernverfahren werden die Eigenschaften, die die

Kriterien für die Kategorisierung ausmachen, vom Menschen bestimmt und an die Maschine weitergegeben. Bei Deep Learning werden jene Kriterien ebenfalls durch die Maschine definiert. Es wird also ein sonst unvermeidbarerer Schritt eingespart; dafür werden eben deutlich mehr Rechenarbeit und Daten benötigt, um diese Eigenschaften ausfindig zu machen (vgl. Couto 2018). Mit Hilfe von Deep Learning konnten in vielen Feldern revolutionäre Fortschritte getätigt werden; dazu gehören automatische, ärztliche Diagnosen, Gesichtserkennung, oder auch selbstfahrende Autos (vgl. Wang und Siau 2018, 1).

Es sind inzwischen einige Anwendungsfälle für Deep Learning feststellbar, wie beispielsweise das Generieren von Untertiteln für Bilder, automatisches Zusammenfassen von Texten, Wettervorhersage, automatisches Übersetzen oder aber auch das schlichte Zuordnen einer Kategorie zu einem Textabschnitt. 2018 wurden insbesondere zwei Bereiche weiterentwickelt: Natural Language Processing (NLP) und Video Generierung. Da der Aufbau von Welten für Computerspiele mit einer Grafik-Engine viel Programmieraufwand erfordert, haben Forscher von NVIDIA es sich zum Ziel gesetzt, diese Welten nicht mehr über eine Grafik-Engine zu erstellen, sondern fotorealistische Äquivalente mittels schemenhafter Videos und Deep Learning automatisch zu erschaffen. Aufgrund der überzeugenden Ergebnisse könnte dieser Ansatz schon bald für Computerspiele, aber auch für Filme, zum Einsatz kommen (vgl. Couto 2018). Was NLP betrifft, so wurde hauptsächlich die Art und Weise wie Sprache für NLP Aufgaben modelliert wird, revolutioniert. Dazu gehört hauptsächlich BERT (Devlin et. al. 2018), ein von Google entwickeltes Sprachmodell, welches aktuell für 11 NLP Bereiche State-of-the-Art Ergebnisse erzielt (vgl. Couto 2018).

In dieser Arbeit treffen damit also zwei hochaktuelle Themen, Hassrede und Deep Learning, aufeinander. Genauer gesagt geht es um die automatische Identifikation von Hassrede in Tweets. Im Laufe der Arbeit sollen vor allem folgende Fragen beantwortet werden:

- Wie gut kann ein Deep Learning Modell Hassrede, ausgehend von unterschiedlichen Textrepräsentationen, insgesamt abbilden?
- Welche detaillierteren Möglichkeiten zur Verarbeitung von Text (im Hinblick auf Deep Learning) stehen zur Verfügung und welche Gründe sprechen für oder gegen die Anwendung der jeweiligen Möglichkeiten, wenn es um die Identifikation von Hassrede geht?
- Welche konkreten Probleme können bei den in dieser Arbeit erstellten Modellen auftreten?

Der Hauptfokus dieser Arbeit besteht damit darin, die verschiedenen notwendigen Schritte vor dem eigentlichen Deep Learning Algorithmus genauestens zu beleuchten. Natürlich ist die Erstellung des neuronalen Netzes ein genauso zentraler Schritt. Allerdings ist es nicht Ziel der vorliegenden Thesis, ein ideales neuronale Netz für die jeweiligen Repräsentationsarten zu bestimmen, da dies nochmals einen erheblichen Zeitaufwand für minimal bessere Ergebnisse erfordern würde. Weil es auf diesen Unterschied in dieser Arbeit nicht ankommt, sind die hier erstellten neuronalen Netze trotzdem gut geeignet.

Zunächst soll es darum gehen, den Begriff „Hassrede“ genauer zu untersuchen. Dabei werden vor allem die unterschiedlichen Definitionen des Worts beleuchtet sowie die Rolle von Hassrede im Internet dargelegt. Zusätzlich werden unterschiedliche Kategorisierungsversuche von Hassrede dargestellt. Des Weiteren geht es in diesem Abschnitt um die Art und Weise, wie bislang mit Hassrede in Deutschland umgegangen wird; daran anknüpfend wird auch die Debatte um die Zensur von Internet-Posts kurz erörtert.

Im darauffolgenden Kapitel wird das zusammengestellte Korpus beleuchtet. Es wird vor allem erklärt, wie die Datensätze, aus denen das Korpus letzten Endes besteht, erstellt worden sind und aus welchen Gründen die einzelnen Datensätze in der Form vereinigt worden sind. Dargestellt werden zusätzlich einige deskriptive Statistiken sowie eine kurze, qualitative Analyse zu dem Korpus, um einen Überblick über die Daten zu geben.

Dann wird der Stand der Forschung zum Thema Erkennung von Hassrede bzw. aggressiven Inhalten im Internet beleuchtet. Dabei geht es zunächst um die Studien, welche mit zum Teil gleichen Daten geforscht haben. Dies dient der Etablierung einer Vergleichsgrundlage. Danach werden einige Studien aufgeführt, in denen zwar nicht dieselben Daten verwendet wurden, grundsätzlich aber mit Deep Learning nutzergenerierte, kurze Texte aus dem Internet in aggressive und nicht-aggressive Inhalte klassifiziert werden. Diese Studien sollen zu dem eigenen Vorgehen, vor allem was die Vorverarbeitungsschritte und die Architektur des neuronalen Netzes angeht, beitragen. Dieser Abschnitt des Kapitels ist dabei nochmals in drei Teile untergliedert, in welchen es um eine grundlegende Entscheidung innerhalb der Vorverarbeitung geht: die Wahl der Repräsentation der Texte. Für diese Arbeit wurde sich dazu entschieden, den Einfluss der Repräsentation auf Wortebene und Buchstabenebene auf die Ergebnisse zu untersuchen. Auch gemischte Varianten wurden einbezogen.

Im darauffolgenden Kapitel geht es um die Durchführung. Nach einer kurzen, allgemeinen Beschreibung des grundsätzlichen Arbeitsablaufs von Projekten im Bereich Deep Learning

folgen die in dieser Arbeit getroffenen Entscheidungen hinsichtlich Vorverarbeitung und Modellerstellung sowie deren Erläuterungen zu allen drei oben genannten Ansätzen. Für jeden der Ansätze werden die Ergebnisse verglichen und schließlich werden zwei Modelle mit der besten Performanz ausgewählt.

Anschließend werden die zwei finalen Modelle mit neuen, vermeintlich problematischen Daten getestet, um die Ergebnisse hinsichtlich der möglichen Schwachstellen der beiden Modelle zu erweitern. Dazu gehören hauptsächlich Untersuchungen im Bezug auf mögliche Vorurteile gegenüber weniger verbreiteten Arten von Hass sowie schwarzen Humor. Im Anschluss werden alle im Laufe der Arbeit aufgetretenen Probleme angesprochen und diskutiert, gegliedert nach den einzelnen Arbeitsschritten. Im Zuge dessen werden noch Wege und Möglichkeiten dargestellt, welche diese Arbeit erweitern könnten. Ein abschließendes Kapitel fasst die wichtigsten Erkenntnisse der Arbeit kritisch zusammen und gibt kurze, mögliche Ausblicke auf die Themen Hassrede und Deep Learning.

2. Definition von Hassrede

Bevor von einer automatischen Klassifikation von Hassrede gesprochen werden kann, sollte zuerst deutlich gemacht werden, was der Begriff „Hassrede“ eigentlich enthält. Laut Mendel et. al. in Saleem et. al. (2017) lässt sich Hassrede definieren als eine Ausdrucksweise, welche Hass gegenüber einer Person oder einer Personengruppe ausdrückt. Dieser Hass wird verursacht durch die Angehörigkeit der Zielperson zu einer Personengruppe beziehungsweise durch ein Merkmal der Gruppe im Allgemeinen (Geschlecht, Sexualität, Ethnizität etc.). Hassrede ist laut dieser Definition von Mobbing abzugrenzen. Zwar ist Mobbing oft von Hass geleitet, allerdings wird dieser Hass meist nicht durch die Zugehörigkeit des Opfers zu einer Gruppe geschürt (vgl. Saleem et. al. 2017).

Dies ist jedoch nur eine von vielen möglichen Definitionen. Laut Barlett et. al. (2014, 11) gibt es für den Begriff keinen klaren Rahmen, auch wenn viele Definitionen der oben beschriebenen Variante ähneln. Weitere, enger gefasste Definitionen betonen eher die Auswirkungen von Hassrede. Laut Gagliardone et. al. (2015, 54) kann Hassrede auch eine Ausdrucksweise sein, welche vor allem Gewalt gegenüber den angesprochenen Gruppen anstiftet. Die Verbindung zwischen Hassrede und physischer Gewalt ist allerdings komplex und bislang gibt es noch keine systematische Studie, die diesen Zusammenhang durch Messungen hergestellt hat.

Geschichtliche Vorgänge wie der Aufstieg des Dritten Reichs zeigen jedoch deutlich, wie Hassrede Gewalt verstärken kann (vgl. Shaw 2012, 282). Eine alternative Definition, welche ebenfalls die Auswirkungen von Hassrede in den Fokus stellt, bezieht sich auf die Verletzung der Würde des Menschen (vgl. Bakircioglu 2008, 4). Es zeigt sich, dass bei der Identifikation von Hassrede unterschiedliche Perspektiven wahrgenommen werden sollten.

Eine wichtige Feststellung ist allerdings auch, dass Hassrede durch vorherrschende soziale Normen, Kontext und individuelle sowie kollektive Interpretation bestimmt wird (vgl. Saleem et. al. 2017). So werden manche Verunglimpfungen (engl. „slur"), welche in der Vergangenheit generell akzeptiert waren, heute von der breiten Gesellschaft nicht mehr gebilligt. Die Feststellung impliziert auch, dass zur selben Zeit in unterschiedlichen sozialen Kontexten Begriffe toleriert werden, welche insgesamt nicht als akzeptabel gelten (vgl. Barlett et. al. 2014, 11). Und letztlich kann auch dieselbe Aussage von unterschiedlichen Personen als Hassrede oder eben nicht wahrgenommen werden (vgl. Saleem et. al. 2017).

Durch die unterschiedlichen Definitionen und Wahrnehmungen ist der Begriff „Hassrede" auch zu einem Instrument geworden. Um beispielsweise eine Meinung zu zensieren, kann diese von einer Partei als Hassrede betitelt werden. Der Begriff kann also selbst zum Verunglimpfen einer Person oder Personengruppe verwendet werden, da eine Partei die Definitionen von Hassrede zu ihrem Gunsten auslegen kann (vgl. Gagliardone et. al. 2015, 55). Alexander Brown (2018) beschreibt diese Beobachtung als eine Disparität zwischen Hassrede als Rechtsbegriff und Hassrede als gebräuchliches Konzept – wie der Begriff also tatsächlich von der Gesellschaft eingesetzt wird. Besonders wird die Verwendung dieses Konzepts mit der linken Seite der Politik in Verbindung gebracht. Denn in den Augen der politisch Konservativen wird oftmals beispielsweise Kritik an einer offenen Grenzpolitik durch den Begriff „Hassrede" niedergeschlagen, da für die Liberalen diese Kritik in die Kategorien „Rassistisch" oder „Islamophob" fällt. So kann es sein, dass „Hassrede" hauptsächlich für „einer anderen Meinung sein" steht (vgl. Brown 2017, 425ff.).

Speziell in Deutschland (im Unterschied zu den Niederlanden oder Norwegen) gibt es keine juristische Definition von Hassrede. So handelt es sich eher um einen politischen bzw. alltäglichen Begriff. Er umfasst zwar einige justiziable Delikte wie beispielsweise Volksverhetzung, Androhung von Straftaten oder Bedrohungen, allerdings beinhaltet das Konzept von Hassrede, wie bereits dargelegt, weitaus mehr als nur diese Fälle (vgl. Meßmer und Krause 2018, 6ff.).

Prinzipiell lassen sich vier unterschiedliche Kategorien von Hassrede laut Yong (2011, 394ff.) feststellen. Diese sind auch aus der erstgenannten Definition ableitbar. Die erste Kategorie ist die gezielte Diskriminierung. Bei dieser Art von Hassrede gibt es genau eine Zielperson oder eine sehr kleine Zielgruppe. Die Hauptmotivation ist dabei, das Ziel mit möglichst verletzenden Wörtern zu schaden. Diffuse Diskriminierung ist im Gegensatz dazu an ein großes Publikum gerichtet. Dabei ist das Ziel dasselbe wie bei der gezielten Diskriminierung, allerdings gibt es hier einen großen Teil an Zuschauern. Per Definition hat die Zielgruppe der Diskriminierung die Möglichkeit, dieser aus dem Weg zu gehen, was bei dem gerichteten Gegenstück nicht so leicht der Fall ist. In die dritte Kategorie fallen Verordnungen, die den Ausschluss oder die Eliminierung von Gruppen bewirken. In die letzte Kategorie fallen Tatsachenbehauptungen und Meinungen mit Werturteil, welche sich nachteilig auf Personengruppen auswirken.

Neben der Unterscheidung zwischen unterschiedlichen Hassrede-Kategorien aufgrund der möglichen Auswirkungen gibt es eine weitere Unterscheidungsmöglichkeit: explizite und implizite Hassrede. Dabei steht die Direktheit des Angriffs im Vordergrund. Bei expliziter Hassrede handelt es sich um Aussagen, die eindeutig das Potential haben, missbräuchlich zu sein. Diese beinhalten oftmals zum Beispiel rassistische oder homophobe Verunglimpfungen. Die implizite Variante dagegen besitzt nicht direkt solche Anzeichen von Hass. Es fehlen für Hassrede typische Anzeichen wie Verunglimpfungen oder Schimpfwörter; stattdessen wird mit Mehrdeutigkeit oder Sarkasmus gearbeitet, aber auch vermeintlich „rationale“ oder „objektive“ Beweisführungen zählen dazu. Dazu gehört auch Hassrede, die möglicherweise unbewusst geäußert wurde. Ein Beispiel für solche implizite Hassrede wäre etwa *most of them come north and are good at just mowing lawns*, wobei *them* für eine ethnische Gruppe steht (vgl. Waseem et. al. 2017).

Wesentlicher Bestandteil von Hassrede sind Verunglimpfungen, wie bereits erwähnt. Aufgrund der Wesentlichkeit muss dies hier nochmals genauer erklärt werden. Verunglimpfungen sind Wörter, welche alleinstehend eine Abwertung einer Personengruppe ausdrücken. Zu diesen Wörtern gibt es immer ein neutrales Synonym. Beispielsweise kann das Wort *f*ggot* als *homosexual and despicable because of it* paraphrasiert werden (vgl. Bianchi 2014, 36). Aus der Sicht der Pragmalinguistik wird diese unterschwellige Abwertung allerdings lediglich durch den Kontext vermittelt, in dem das Wort verwendet wird. Von jenem Standpunkt aus bedeutet das, dass diese Abwertung auch ausbleiben kann. Hierbei hat sich herausgestellt, dass es innerhalb der entsprechenden Personengruppen akzeptiert wird, die mit ihnen verbundenen Verunglimpfungen zu verwenden, um unter anderem eine innere Solidarität auszudrücken (vgl.

ebd., 36f.). Essenziell bei dem nicht-anstößigen Gebrauch ist dabei die Abgrenzung von einer politischen oder sozialen Absicht, was eben vor allem durch die Zugehörigkeit zu der jeweiligen Gruppe gesichert sein kann (vgl. ebd., 40). Ein Beispiel dafür ist das Wort *n*gger* und dessen akzeptierter Gebrauch innerhalb der afroamerikanischen Gemeinschaft. Dieser nicht-anstößige Gebrauch kann dazu beitragen, graduell die unterschwellige Abwertung vom eigentlichen Wort innerhalb der Gruppen loszulösen. Im Idealfall könnte diese Loslösung auch auf die gesamte Gesellschaft übergreifen, so wie es für das Wort *gay* schon beobachtet werden konnte (vgl. ebd., 43).

Vor allem mit dem Aufstieg des Internets hat Hassrede eine starke Verbreitung gefunden. Dies liegt hauptsächlich an der Anonymität, welche das Internet bietet, sowie an der psychologischen Distanzierung der Sprecher von ihrem Publikum. Einerseits werden Nutzer also eher dazu verleitet, sich hasserfüllter Sprache zu bedienen, da sie sich nicht mit den Reaktionen anderer auseinandersetzen müssen, weil sie diese nicht zwingend sehen (vgl. Brown 2018, 300). Andererseits stammt die Verbreitung auch von der weitreichenden und dauerhaften Natur des Internets (vgl. Shaw 2012, 280). Somit ist es allerdings auch einfacher geworden, Instanzen von Hassrede zu finden und auszuwerten. Was sich hierbei gezeigt hat ist, dass größere Mengen an Hassrede oft die Folge eines bedeutsamen Ereignisses sind (vgl. Barlett et. al. 2014, 12). Das bedeutet im Allgemeinen, dass verschiedene Ausprägungen von Hassrede zu unterschiedlichen Zeitpunkten auch unterschiedlich häufig auftreten. Um ein Beispiel für unsere heutige Gesellschaft zu nennen: in den United States ist zum Beispiel Hassrede zum Thema „Rasse“ stärker repräsentiert als Hassrede zum Thema „sozialer Stand“ (vgl. Mondal et. al. 2017).

Die Studie von Mondal et. al. (2017) zeigt auch, dass verschiedene Hassrede-Themen in verschiedenen Ländern unterschiedlich stark vertreten sind. Dies korreliert vor allem mit den beiden Aussagen, dass Hassrede durch vorherrschende soziale Normen geprägt wird und durch bedeutsame Ereignisse ausgelöst werden kann. Denn so wie die sozialen Normen verschiedener Länder unterschiedlich sein können, so können sich parallel dazu die Problematiken innerhalb der Länder unterscheiden, welche wiederum zu den Ereignissen führen. Zudem finden manche Ereignisse nur auf nationaler Ebene Gehör. So kommt es, dass die häufigsten Ziele von Hass in den United States Schwarze sind, während es im Vereinigten Königreich hauptsächlich Dicke und Homosexuelle sind (vgl. ebd., 2017).

Die Gefahr von Hassrede auf den Einzelnen besteht darin, dass die Opfer psychischen Schaden erleiden können. Außerdem können insgesamt im Ansatz bereits bestehende, gefährliche

Überzeugungen in der breiten Zuhörerschaft verstärkt werden. Je weitverbreiteter diese Falschdarstellungen und Vorurteile in der Gesellschaft sind, desto wahrscheinlicher sind auch physische Angriffe auf die betroffenen Gruppen (vgl. Shaw 2012, 281ff.). So kann es zum Beispiel auch eine Folge von Hassrede sein, dass einzelne Opfer aus der Gesellschaft ausgeschlossen werden. Ganze Gruppen können ihre kulturelle Identität sowie ihre Stellung in der Gesellschaft verlieren, womit Ungleichheit erzeugt wird. Zusätzlich werden die Opfer durch die verbreitete Angst zum Schweigen gebracht und können sich deshalb oftmals nicht gegen die Angriffe wehren (vgl. Bakircioglu 2008, 4f.). Auch die Überzeugungen der Täter selbst können durch hasserfüllte Äußerungen intensiviert werden (vgl. Shaw 2012, 281ff.).

Was das Internet angeht, so werden deshalb immer mehr Anbieter von Webseiten von unterschiedlichen Akteuren dazu aufgerufen, Hassrede zu regulieren. Diese Akteure können einzelne Nutzer oder Nutzergruppen sein; aber auch Regierungen können Druck auf die Webseiten ausüben (vgl. Gagliardone et. al. 2015, 55). In Deutschland gibt es bislang das Netzwerkdurchsetzungsgesetz (NetzDG), welches im Januar 2018 eingeführt wurde. Dieses besagt, dass rechtswidrige Inhalte in sozialen Netzwerken vom Betreiber gelöscht werden müssen, wenn sie ihm gemeldet werden (vgl. Rafael und Ritzmann 2018, 16). Aber „auf der Ebene der Europäischen Union existiert […] bisher keine gesetzliche Regelung gegen ‚Hate Speech' im Internet", lediglich einen „EU-Verhaltenskodex zur Bekämpfung von Hetze im Internet" (Deutscher Bundestag 2018, 8) gibt es seit 2016, zu welchem sich einige große IT-Unternehmen wie Facebook und Twitter verpflichtet haben (vgl. ebd., 8). Die Webseiten nehmen dabei aber eben nur eine passive Rolle ein, denn sie sind auf Meldungen von Nutzern oder Moderatoren angewiesen, um gegen Hassrede vorgehen zu können. Bei der Entscheidung, welcher der Aussagen als Hassrede einzustufen ist, wird meist eine von der Webseite eigens erstellte Definition verwendet (vgl. Gagliardone et. al. 2015, 55).

Facebook zum Beispiel definiert Hassrede als einen „direkten Angriff auf Personen aufgrund geschützter Eigenschaften". Ein direkter Angriff wird definiert als „gewalttätige oder entmenschlichende Sprache, Aussagen über Minderwertigkeit oder Aufrufe, Personen auszuschließen oder zu isolieren". Humor und Gesellschaftskritik in dieser Verbindung wird allerdings explizit zugelassen. Facebook macht hierbei explizit auf die Deutlichmachung der Absicht aufmerksam.[1]

[1] https://www.facebook.com/communitystandards/objectionable_content (Stand 12.03.2019).

Auf Twitter hingegen wird hassschürendes Verhalten als eine Förderung von Gewalt gegenüber geschützten Kategorien (ähnlich derer zu Facebook) verstanden. Dazu gehören gezielte Gewaltandrohungen, Ausdrücke, die einer geschützten Person oder Personengruppe Schaden wünschen, gezielte Entmenschlichungen und Diskriminierungen, Verbreitungen schädlicher Klischees und Belästigungen von Personen durch Gewaltdarstellungen, bei denen geschützte Personengruppen primäres Opfer sind. Auch hier wird insbesondere auf die Absicht der Nutzer geachtet.[2]

Wenn auch beide Definitionen sehr ähnlich sind, unterscheiden sie sich bereits in wesentlichen Punkten. Was die Kategorien bzw. Personengruppen angeht, so erwähnt Facebook auch „Kaste" und „Einwanderungsstatus", während Twitter „Alter" mitzählt und die Kategorie „Krankheiten" durch ein vorangestelltes „ernsthafte" eingrenzt. Auch der Ausschluss von Humor lässt sich so wie bei Facebook in der Definition von Twitter nicht explizit finden. Dies zeigt, dass die verschiedenen Definitionen der Webseiten und deren Anwendung die Sichtweise der Nutzer auf das Thema Hassrede auch unterschiedlich prägen können.

Der Aufruf nach weiterer Eindämmung von Hassrede, insbesondere im Internet, ist stetig. Allerdings bestehen einige Sorgen vor allem was die Überregulierung von Hassrede betrifft. Eine Regulierung von Hassrede durch die Regierung kann die Demokratie gefährden. Robert Post (1991) beispielsweise argumentiert, dass ein freier, öffentlicher Diskurs entscheidend für die Entwicklung eines demokratischen Kollektivs ist. Diese Limitierung der Meinungsfreiheit wirkt sich nicht nur negativ auf das Kollektiv, sondern auch auf die betroffenen Individuen aus, da diese sich nicht mehr frei äußern dürfen. Auch die persönliche Entwicklung ist damit gefährdet (vgl. ebd., 290ff.). Auf der einen Seite ist die Meinungsfreiheit ein Menschenrecht und sollte deshalb immer garantiert werden. Auf der anderen Seite greift Hassrede ebenso einige Grundprinzipien der Menschenrechte an. Bei dieser Debatte stehen sich also zwei Definitionen der Menschenwürde gegenüber: Würde als freie Meinungsäußerung und Würde als eine Einschränkung der Meinungsfreiheit, welche wiederum dem Schutz der Menschenwürde und der persönlichen Entwicklung eines jeden dient. Eine übereifrige Zensur stellt damit genauso eine Bedrohung dar wie auch die Verbreitung von hassschürenden Inhalten (vgl. Shaw 2012, 283ff.).

Die Problematik zur Regulierung von Hassrede wird unter anderem auch von Bakircioglu (2008) diskutiert, vor allem im Bezug auf die Aufrechterhaltung der Demokratie. Es muss eine

[2] https://help.twitter.com/de/rules-and-policies/hateful-conduct-policy (Stand 12.03.2019).

Balance aus beiden Ansätzen gefunden werden. Die meisten Demokratien besitzen bereits eine Form von Regulierung von Sprache, wie beispielsweise Nazi-Propaganda in Deutschland. Eine Gesellschaft, welche die Meinungsfreiheit respektiert, muss damit nicht unbedingt keine Beschränkungen festlegen. Um die persönliche Entwicklung jedes Individuums zu schützen ist es gerechtfertigt, Sprache, die dem entgegenwirkt, zu unterbinden. Denn in einer ungleichen Gesellschaft werden die Meinungen und Informationen der Stärkeren die der Schwächeren überschatten und somit wird das Ungleichgewicht an Macht aufrechterhalten (vgl. ebd., 2f.). Zu viel Kontrolle kann allerdings, wie im vorherigen Absatz erläutert, ebenso viel Schaden anrichten. Durch die Tatsache, dass es nur vage und unterschiedliche Definitionen von Hassrede gibt, können allerdings auch nur vage Limitierungen verordnet werden. Aufgrund der daraus entstehenden, möglichen Unsicherheit kann es passieren, dass harmlose Aussagen unterdrückt werden, da Individuen nicht einschätzen können, ob diese akzeptiert sind oder nicht (vgl. ebd., 44f.).

Zusammenfassend lässt sich sagen, dass Hassrede ein Konzept ist, welches sich nicht einfach definieren lässt. Zwar gibt es Überschneidungen in den verschiedenen Ansätzen, Hassrede zu definieren, doch eine allgemein akzeptierte Definition gibt es nicht. Soweit lässt sich klar herausstellen, dass Hassrede einerseits

- justiziable Delikte (Volksverhetzung, Androhungen von Gewalt) und andererseits
- Ausdrücke, welche geschützte Personen oder Personengruppen herabsetzen,

beinhaltet. Dennoch können sich die Definitionen unterscheiden. Denn es ist oftmals nicht klar, welche Personengruppen genau betroffen sein können und welche Ausdrücke als Herabsetzung gezählt werden können. Dies kann von Person zu Person unterschiedlich wahrgenommen werden. Auch Zeit und Ort können die Wahrnehmung einer Gesellschaft auf diesen zweiten Punkt der Definition maßgeblich prägen.

Dadurch ist es sehr schwierig einzuschätzen, wann genau ein Ausdruck als Hassrede einzustufen ist. Manche Ausdrücke mögen zwar beleidigend sein und Vorurteile gegen beispielsweise Frauen fördern; diese Ausdrücke müssen dadurch allerdings nicht unbedingt Hassrede sein (vgl. Bakircioglu, 2008, 4). Genau dieser Unterschied zwischen „nur“ beleidigenden und hasserfüllten Ausdrücken sollte für das vorliegende Projekt auch relevant sein, weswegen genaue diese beiden Klassen zusätzlich zu neutralen Ausdrücken unterschieden werden. Zunächst wird deshalb die für diese Arbeit zusammengestellte Datengrundlage vorgestellt; einschließlich einer kurzen qualitativen und quantitativen Analyse.

3. Datengrundlage

Der erste Schritt für das Maschinelle Lernen ist die Zusammenstellung eines Korpus. Dieses Korpus muss ausreichend sogenannte Tweets (kurze, auf dem sozialen Netzwerk Twitter veröffentlichte Mitteilungen) aus mindestens zwei Klassen („Hassrede“ und „Neutral“) beinhalten. Jedes Tweet muss mit dessen Klasse gekennzeichnet sein. Es ist beispielsweise möglich, mittels der Twitter API aktuelle Tweets aus hassanfälligen Themenkreisen, wie beispielsweise der Politik, auszugeben und eigenständig zu klassifizieren. Das kostet allerdings nicht nur viel Zeit; auch die Klassifizierung durch nur eine Person kann sehr subjektiv ausfallen.

Für diese Arbeit wurde sich aus Gründen der Zeiteffizienz darauf konzentriert, bereits gekennzeichnete Daten zu finden und zu verwenden. Zunächst wurde das englische Twitter Korpus von Thomas Davidson[3] benutzt. Gerade dieses Korpus erwies sich als besonders vielversprechend, da hier sogar in drei Klassen („Hassrede“, „Beleidigung“ und „Neutral“) unterschieden wird. Etwa 24.000 Tweets, welche gewisse Schlüsselwörter enthalten, wurden aus Twitter ausgelesen. Bei der Klassifizierung wurden Nutzer der Webseite *CrowdFlower* (jetzt *Figure Eight*) herangezogen, welche mit entsprechenden Informationen zum Thema Hassrede ausgestattet wurden. Aus dem Datensatz geht hervor, wie viele der Nutzer (mindestens drei je Tweet) sich für welche Klassen entschieden haben. Nur bei deutlicher Mehrheit wurde ein Tweet in das Korpus aufgenommen. Somit wurde eine gewisse Objektivität bei der Klassifizierung gewährleistet (vgl. Davidson et. al. 2017, 513).

Dieser Datensatz allein reicht allerdings nicht aus. Aus dem Paper (vgl. ebd., 514) geht bereits hervor, dass die automatische Klassifikation mit jenen Trainingsdaten Probleme aufgezeigt hat. Denn die Mehrheit der Tweets (76%) wurde als „Beleidigung“ eingestuft, während nur 5% als „Hassrede“ klassifiziert wurden. Gerade die wichtigste Klasse wurde somit in der Testphase nur zu 61% korrekt wiedererkannt, was vermutlich an der Klassenverteilung lag.[4] Deswegen sollten vor allem weitere Tweets der Klasse „Hassrede“ und „Neutral“ dieses Korpus ergänzen. Hinzugezogen wurden deswegen die Datensätze von Mai ElSherief[5] und Zeerak Waseem[6]. Da

[3] https://github.com/t-davidson/hate-speech-and-offensive-language (Stand 12.02.2019).
[4] Diese Vermutung wurde in der vorliegenden Arbeit überprüft. Bei einer Größe des Trainingssets von zufällig ausgesuchten, 5.000 Tweets wurden nur 12% der Klasse „Hassrede“ korrekt zugeordnet. Sobald von allen Klassen gleich viele Tweets für das Training verwendet wurden, stieg die Genauigkeit für diese Klasse an, auch wenn aufgrund des kleineren Trainingssets die allgemeine Genauigkeit sank.
[5] https://github.com/mayelsherif/hate_speech_icwsm18 (Stand 12.02.2019).
[6] https://github.com/ZeerakW/hatespeech (Stand 12.02.2019).

hier nur die Tweet IDs gegeben sind, wurden diese mithilfe der Twitter API in die eigentlichen Tweets umgewandelt.

In dem Datensatz von Waseem wurden die Tweets in die Klassen „Rassismus“, „Seximus“ und „Neutral“ eingeteilt. Es haben dabei die Autoren selbst die Tweets annotiert; überprüft wurden die Annotationen von zwei Experten. Die Tweets wurden ebenso zuvor nach beinhalteten Schlüsselwörtern ausgesucht (vgl. Wassem und Hovy 2016, 89). Die beiden ersten Klassen hätten unter „Hassrede“ zusammengefasst werden können – allerdings sind nicht nur diese beiden Klassen bereits recht eingrenzend was den Überbegriff Hassrede angeht, auch besteht die Mehrheit dieser Tweets aus zu spezifischen Themen (zum Beispiel *#mkr* für eine Kochsendung). Nur der neutrale Anteil der Daten wurde deshalb in das Korpus aufgenommen. Denn mit den zusätzlichen Daten von ElSherief konnte die Klasse „Hassrede“ genügend vergrößert werden. Die Tweets aus diesem Datensatz wurden nach zwei Vorgehensweisen ausgesucht: Hashtagbasiert und Schlüsselwortbasiert. Mit speziellen Hashtags (Schlagwörter, welche vom Verfasser eines Tweets angegeben werden können, um diesen auffindbar zu machen) und Schlüsselwörtern wurden hier ebenfalls Tweets aus dem Twitter-Stream herausgefiltert. Da für diese Studie nur Tweets mit Hassrede-Inhalten relevant waren, wurden Tweets, die zwar diese Wörter enthalten, aber nicht Hassrede sind, durch ein qualitatives Vorgehen heraussortiert (vgl. ElSherief et. al. 2018).

Die kombinierten Daten werden innerhalb der Vorverarbeitungsschritte nach Duplikaten durchsucht, womit 47 Tweets aussortiert werden. Insgesamt umfasst das Korpus damit 37.555 Tweets. Tabelle 3.1 zeigt die Klassenverteilung. Da die Klassen im Korpus gleichverteilt sein sollten, wurden einmalig 6.000 Tweets aus allen drei Klassen zufällig entnommen. Das endgültige Korpus beinhaltet damit 18.000 Tweets.

	Hassrede	Beleidigung	Neutral
Davidson	1.430	19.190	4.163
ElSherief	9.755	-	-
Waseem	-	-	3.017
Gesamt	**11.185**	**19.190**	**7.180**

Tabelle 3.1: Übersicht über die Klassenverteilung innerhalb des vorläufigen Korpus sowie die Quelle der Tweets.

Bei der Zusammenstellung des Korpus in dieser Art und Weise tauchten bereits erste Probleme auf. Beim Umwandeln der Tweet IDs in die Tweets wurde beobachtet, dass ein Großteil der IDs nicht mehr zugeordnet werden konnte. Vermutlich wurden diese Tweets notwendigerweise

aufgrund ihres Inhalts gelöscht. Damit sind die mutmaßlich interessantesten Tweets der Datensätze von ElSherief und Waseem nicht mehr zugänglich. Dies ist ein großer Nachteil der Nachnutzung bereits zusammengestellter Daten, wenn es um ein solches Thema geht. Es hätte sich also durchaus ein Vorteil ergeben, die Daten selbst aus Twitter abzufangen und als Klartext abzuspeichern. Dann wären möglicherweise extremere und somit repräsentativere Tweets im Korpus enthalten.

In Tabelle 3.2 wurden einige deskriptive Daten zum Korpus festgehalten. Hier zeigt sich, dass Tweets aus der Klasse „Hassrede“ im Durchschnitt etwas kürzer im Vergleich zu denen aus den anderen beiden Klassen sind. Auch lässt sich ein vergleichsweise größerer Wortschatz innerhalb der Klasse „Neutral“ feststellen.

	Durchschn. Länge (Buchstaben)	Durchschn. Länge (Wörter)	Anzahl an Types	Anzahl an Tokens	Type-Token-Relation
Hassrede	76,9	11,3	8.424	68.341	0,123
Beleidigung	83,1	13,5	9.105	81.080	0,112
Neutral	88,8	13,3	13.148	80.372	0,163
Gesamt	**82,9**	**12,7**	**21.543**	**229.793**	**0.093**

Tabelle 3.2: Durchschnittliche Länge der Tweets (in Buchstaben und Wörtern gemessen) sowie die Anzahl an Types und Tokens je Klasse. Zu „Token“ werden alle Vorkommnisse der Wörter gezählt, während zu „Type“ nur ein Vorkommnis je Worttyp gezählt wird. Die Type-Token Relation ist ein einfaches Maß zur Errechnung der Komplexität eines Texts.

An dieser Stelle sollte deutlich gemacht werden, dass die Tweets aller Klassen durch die oben beschriebenen Auswahlverfahren aus Hass-anfälligen Themenkreisen wie zum Beispiel politischen Diskussionen stammen. Ein Problem bei automatischen Klassifizierungsverfahren stellt sich, wenn bei der Klassenerkennung zufällige Eigenschaften fälschlicherweise einer Klasse zugeordnet werden. Falls beispielsweise in jedem Trainings-Tweet der Klasse „Hassrede“ das Wort *Trump* fällt (aber eben nicht in der Klasse „Neutral“), so wird der Algorithmus dieses Wort als eine Eigenschaft für diese Klasse festhalten. Dadurch, dass die Korpora, besonders von Davidson und Waseem, in erster Linie aus einer Schlagwort-basierten Vorauswahl an Tweets bestehen, und die Tweets erst anschließend klassifiziert wurden, ist diesem Problem zum Großteil entgegengekommen worden.

In diesem Fall würde ein Problem nur dann entstehen, wenn alle drei Korpora durch zu unterschiedliche Schlagwörter zustande gekommen wären. Dann gäbe es in der hier erstellten Zusammenstellung vermutlich keine „Gegenbeispiele“ für Tweets in den Korpora von Waseem

und ElSherief. Es wurde deshalb ein Topic Model mit dem TopicsExplorer[7] von DARIAH-DE für das Korpus erstellt, um mögliche Probleme angesichts der Themen zu identifizieren. Die Tweets innerhalb der Klassen wurden dabei durchmischt, um gleiche Bedingungen wie bei der Klassifikation zu schaffen. Abbildung 3.1 zeigt das Ergebnis als Heatmap.

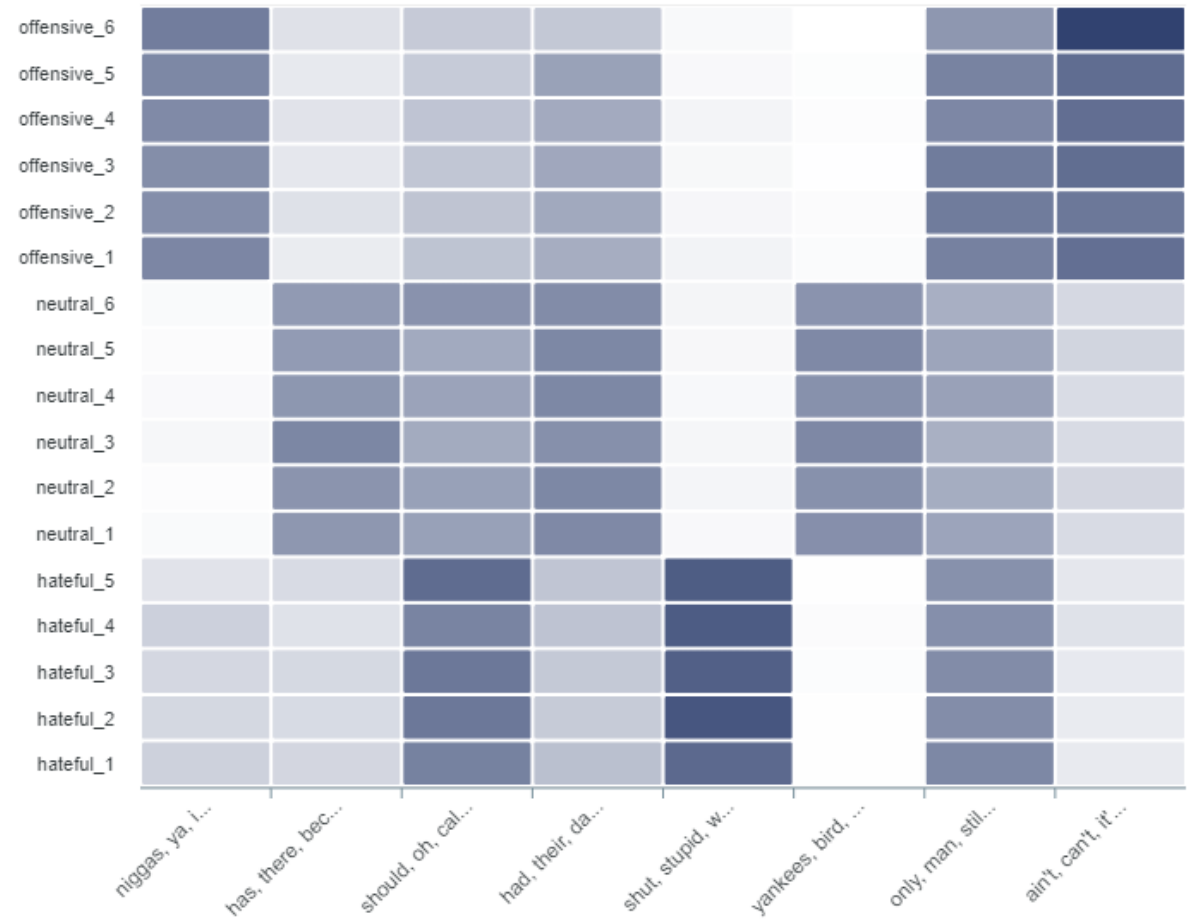

Abbildung 3.1: Topic-Verteilung über die Segmente der drei Klassen „hasserfüllt“, „beleidigend“ und „neutral“ als Heatmap. Auf der X-Achse werden die Topics durch ihre repräsentativsten Wörter dargestellt.

Innerhalb der Klassen sind die Topics auf den ersten Blick unterschiedlich stark ausgeprägt. Bei einer genaueren Betrachtung wird allerdings deutlich, dass die Schlagwörter innerhalb der Topics überwiegend keine themenspezifische Aussagekraft besitzen, mit Ausnahme des sechsten Topics. Hier finden sich Keys wie *yankees*, *bird* und *charlie*, welche vermutlich alle mit dem Baseball Team „New York Yankees“ zusammenhängen. Die Tweets mit diesen

[7] https://github.com/DARIAH-DE/TopicsExplorer (Stand 21.05.2019).

Schlüsselwörtern stammen unerwarteterweise aus dem Korpus von Davidson. Vermutlich wurde das Schlüsselwort *yankees* ausgesucht, um Tweets zu filtern, in welchen dieses Wort als Beleidigung verwendet wird. Da der Begriff allerdings zwei Bedeutungen hat, wurden überwiegend nur neutrale Tweets mit diesem Wort gefunden – mit weniger Gegenbeispielen aus der Klasse „Hassrede". Wie damit umgegangen wird, wird in Kapitel 5 erläutert. Abgesehen von diesen Wörtern finden sich keine weiteren Indikatoren für Themen; dies kann auch anhand der 200 häufigsten Wörter der Klassen beobachtet werden. Es finden sich weder dort noch in den wichtigsten Keys der anderen Topics Wörter, welche zu einem klaren Themengebiet zugeordnet werden können.

Es lassen sich von diesen Untersuchungen bereits erste Erkenntnisse ableiten. Die Heatmap lässt vermuten, dass Topics den Klassen unterschiedlich stark zugeordnet sind. Von einigen interessanten Keys der Topics wurden deshalb die Häufigkeitsverteilungen herausgesucht. Was hierbei zunächst auffällt ist, dass das Wort *n*gger* hauptsächlich in der Klasse „Hassrede" vorkommt. Die alternativen Schreibweisen allerdings, wie zum Beispiel *n*gga*, tauchen meist in der Klasse „Beleidigung" auf. Aber nicht alle Tweets, die das Wort *n*gger* enthalten, gehören zwangsläufig zur Klasse „Hassrede". Dies zeigt, dass beim Annotieren der Tweets zum Teil auf den Kontext geachtet wurde. Denn solche Verunglimpfungen müssen, wie schon in Kapitel 2 erläutert, nicht zwangsläufig in einer hasserfüllten Art und Weise benutzt werden. In der zugehörigen Gemeinschaft verwendet, fällt die negative Konnotation der Verunglimpfung weg (vgl. Anderson und Lepore 2013, 350). Anhand eines Tweets zu identifizieren, ob der Verfasser einer bestimmten Gemeinschaft angehört oder nicht, ist grundsätzlich schwer realisierbar. Vermutlich ist daher auch nur ein Tweet, welcher eines der beiden Wörter enthält, als „Neutral" eingestuft worden. Trotzdem ist eine Abstufung der Verunglimpfungen durch die Klasse „Beleidigung" abgebildet worden.

Aufgrund dieses Befunds wurde nach Wörtern gesucht, welche nur in einer Klasse vorkommen – sozusagen stellvertretend für eine Klasse stehen. Für „Hassrede" fallen tatsächlich mehrere Wörter aber auch viele Emojis auf. Die häufigsten Wörter sind *istandwithhatespeech*, *whitepower* und *trump*. Die ersten Beiden sind zwei der Hashtags, die von ElSherief et. al. (2018) ausgesucht worden sind, um nach Hassrede zu filtern. Weiterhin lassen sich vor allem Emojis finden, welche eine Hautfarbe darstellen.

Für die Klasse „Beleidigung" fallen keine besonderen Wörter auf. Für die Klasse „Neutral" sticht hauptsächlich das Wort *gamergate* heraus. Dies ist vermutlich ein Schlüsselwort, welches

für das Korpus von Waseem verwendet wurde und nicht in den anderen Klassen repräsentiert ist, da keine Tweets aus den Klassen „Sexismus“ und „Rassismus“ in dieses Korpus integriert wurden. Natürlich steht *gamergate* grundsätzlich nicht für ein neutrales Tweet, da es ein sehr aufgeladenes Thema ist, mit welchem Hass grundsätzlich assoziiert wird. Da es allerdings nur insgesamt 49 Vorkommnisse des Worts im Korpus gibt, wurde dies nicht weiter beachtet. Denn andere, augenscheinlich zufällige Verbindungen von Wörtern zum Beispiel mit der Klasse „Beleidigung“ kommen ähnlich häufig im gesamten Korpus vor.

Soweit der Überblick über das zusammengestellte Korpus, mit dem die Folgeschritte für das maschinelle Lernen angegangen werden. Im Folgenden werden die Studien, welche diese Daten (wenn auch in unterschiedlichen Varianten) verwenden, um mit maschinellen Lernverfahren Hassrede automatisch klassifizieren zu können, behandelt. Im Anschluss werden weitere Studien, welche neuronale Netze für dasselbe Ziel mit etwas unterschiedlichen Daten gebrauchen, vorgestellt.

4. Stand der Forschung

4.1 Datensatz

Es gibt bereits Forschungsarbeiten, welche das im letzten Kapitel beschriebene Korpus in unterschiedlichen Ausprägungen zur computergestützten Klassifikation von Tweets benutzen. Bereits angesprochen ist das zugehörige Paper von Davidson et. al. (2017); hier ist zwar ein F1 Score von 0,90 mittels logistischer Regression (Walker und Duncan 1967) erreicht worden, allerdings wurde aufgrund der mangelnden Daten für die Klasse „Hassrede“ bei den Testdaten für diese Klasse nur eine Erkennungsrate von 61% erzielt.

Eine weitere Arbeit stammt von Gaydhani et. al. (2018), in welcher die Korpora von Davidson und Waseem ebenfalls verwendet wurden. Zusätzlich wurde allerdings ein älteres Korpus von Davidson hinzugezogen, welches lediglich eine Teilmenge des neueren Korpus darstellt. Dies bedeutet, dass etwa 15.000 Duplikate in den Daten vorhanden waren, welche nicht behandelt wurden. Die Daten wurden nur zufällig durchmischt und dann in Trainings- und Testset aufgeteilt. Ebenfalls mit logistischer Regression sowie TFIDF Normalisierung und der Verwendung von N-Grammen als Features wurde eine Klassifikationsgenauigkeit von 95,6% erzielt. Durch die Duplikate ist es allerdings sehr wahrscheinlich, dass teilweise dieselben Tweets in den Trainings- und Testdaten vorhanden waren. Das Modell konnte die doppelten

Tweets in der Testphase korrekt klassifizieren, allerdings nicht durch Erbringen einer Transferleistung, sondern durch Wiedererkennung – oder schlichtem „Auswendiglernen“. Dies verfälscht die Ergebnisse der Arbeit.

Es wurde außerdem bereits erforscht, wie gut Deep Learning im Vergleich zu klassischen, maschinellen Lernverfahren abschneidet (Robinson et. al. 2018 und Zhang et. al. 2018). In zwei Arbeiten wurden verschiedene Support Vector Machine (SVM) Modelle (Cortes und Vapnik 1995) und Modelle mit CNN (LeCun et. al. 1998) und GRU (Cho et. al. 2014) auf sieben unterschiedlichen Datensätzen getestet, darunter „DT“ (Davidson) und „WZ-L“ (Waseem). Es zeigte sich, dass selbst mit aufwändig betriebener Feature Selektion keines der SVM Modelle die F1 Scores der Deep Learning Modelle übertreffen konnte. Die höchsten erreichten F1 Scores liegen bei 0,82 (Waseem) und 0,94 (Davidson). Allerdings wurden im Falle von letzterem Korpus die Klassen „Beleidigung“ und „Neutral“ zu einer Klasse zusammengefasst (Zhang et. al. 2018, 754). Daher ist das Ergebnis nicht mit dem von Davidson et. al. (2017) vergleichbar.

Diese Befunde stellen somit eine Richtlinie für die zu erwartenden Ergebnisse der vorliegenden Arbeit dar (unter Berücksichtigung der Zusammenstellung und Art und Weise der Behandlung der Daten). Im folgenden Abschnitt soll nun der Stand der Forschung bezüglich der bereits erprobten Vorgehensweisen zum Thema Klassifikation von Tweets bezüglich Hassrede oder Sentiment vorgestellt werden.

4.2 Methodik

Da Deep Learning der Fokus dieser Arbeit ist, sollen in diesem Unterkapitel hauptsächlich die Forschungsarbeiten erwähnt werden, in welchen diese Methode für eine ähnliche Fragestellung ebenfalls verwendet, beziehungsweise mit weiteren Feature-basierten Methoden, wie zum Beispiel Support Vector Machines, verglichen wurde. Eine recht grundsätzliche Frage, welche sich bei der Klassifikation von Text stellt, ist, wie genau der Text eigentlich repräsentiert wird. Für diese Arbeit stehen besonders Wörter, Buchstaben N-Gramme sowie Buchstaben im Vordergrund.

4.2.1 Repräsentation von Text durch Wörter

In der Arbeit von Raiyani et. al. (2018) wurde versucht, Aggression in Daten von Facebook und Twitter mit Deep Learning zu identifizieren. Es hat sich gezeigt, dass ein dichtes neuronales Netz (auch „Dense“) mit drei Layern und einer One-Hot Kodierung der Wörter die besten Ergebnisse liefert. Dabei wurden aufwändige Vorverarbeitungsschritte angewendet; insbesondere wurden alle Abkürzungen und Rechtschreibfehler in den Daten durch vordefinierte Regeln ausgebessert, Emojis durch Wortrepräsentationen ersetzt und Stoppwörter entfernt. Das Modell erreichte einen F1 Score von 0,60 auf englischen Twitter Daten bei einer Anzahl von drei Klassen („offen aggressiv“, „versteckt aggressiv“ und „nicht aggressiv“).

Dieselben drei Klassen sollten auch in der Arbeit von Aroyehun und Gelbukh (2018) automatisch unterschieden werden, dabei dienten Facebook Posts als Grundlage. Die Vorverarbeitungsschritte setzten sich zusammen aus: Entfernen von Satzzeichen, Zahlen, URLs, Benutzernamen, Umwandlung von Großbuchstaben und Hashtags, Dekodierung von Emojis und Rechtschreibkorrektur. Die Daten wurden außerdem durch zweimaliges Übersetzen der Posts augmentiert. Die Wörter wurden durch 300-dimensionale FastText Embeddings kodiert und jeweils mit einem zusätzlichen, zufällig initialisierten, 50-dimensionalen Vektor konkateniert. Insgesamt wurden sieben Deep Learning Architekturen getestet. Davon haben LSTM (Hochreiter und Schmidhuber 1997) und CNN-LSTM mit F1 Scores von 0,64 und 0,60 am besten abgeschnitten.

Eine weitere Arbeit von Yenala et. al. (2018) untersuchte die Identifikation von unangemessenen Inhalten in Suchanfragen. Die Daten wurden der Aufgabenstellung nach in zwei Klassen aufgeteilt. Die einzelnen Wörter werden durch DSSM Embeddings (Huang et. al. 2013) repräsentiert. Diese werden erst an drei hintereinandergeschaltete CNN Layer weitergeleitet. Der resultierende Output wird durch eine Bi-LSTM (Graves et. al. 2005) Layer und schließlich in eine weitere Dense Layer vor der Ausgabelayer geleitet. Von den insgesamt neun getesteten Ansätzen konnte der oben beschriebene Ansatz mit einem F1 Score von 0,87 überzeugen. Auch erwähnenswert ist die Tatsache, dass alle vier neuronalen Netze deutlich besser abschnitten als die vier Feature-basierten Lernmethoden. Im Schnitt liegt der Unterschied bei 0,33.

Um die Identifikation von beleidigenden Nutzerkommentaren geht es in der Studie von Pavlopoulos et. al. (2017). Dabei wurden verschiedene Netzarchitekturen jeweils mit Word2Vec Embeddings (Mikolov et. al. 2015) untersucht. Getestet wurden die Modelle unter

anderem auf englischen Wikipedia „Talk Page“ Kommentaren. Zwei Korpora sind entstanden, jeweils annotiert nach Angriff (positiv und negativ) und Schädlichkeit (positiv und negativ). Die Modelle, welche aus RNNs (Cho et. al. 2014) aufgebaut wurden, konnten die CNN-basierten Modelle übertreffen. Es hat sich auch bewiesen, dass ein Attention Mechanismus die Performanz weiter verbessern konnte. Es wurden Genauigkeiten von 0,97 bzw. 0,98 erzielt.

Auch beschäftigten sich Ahluwalia et. al. (2018) mit der Erkennung von Frauenfeindlichkeit in Tweets. Die Tweets wurden mit NLTK tokenisiert und Wörter, die in 60% aller Tweets oder insgesamt weniger als vier Mal vorkommen, entfernt. Die Wörter wurden mit GloVe Embeddings (Pennington et. al. 2014) ausgestattet. Unter anderem wurde ein neuronales Netz trainiert, dass Tweets in „frauenfeindlich“ und „nicht frauenfeindlich“ klassifizieren sollte. Dieses besteht aus einer Bi-LSTM Layer und einer Dense Layer; zusätzlich zur Ausgabelayer. Interessanterweise konnte dieses Modell nicht gegen Feature-basierte Lernmethoden ankommen. Gepaart mit einem weiteren Task allerdings (Erlernen der Art der Frauenfeindlichkeit und des Vorhandenseins einer Zielperson), bei welchem die dann klassifizierten Tweets weiterverwendet wurden, konnte das neuronale Netz die besten Ergebnisse liefern.

Es gibt eine Studie, in welcher unterschiedliche N-Gramm Ansätze gegeneinander verglichen wurden. In der Arbeit von Mehdad und Tetreault (2016) wurden für die Identifikation von beleidigenden Kommentaren unter anderem ein neuronales Netz trainiert. Für die Repräsentationen wurden Wort N-Gramme und Buchstaben N-Gramme getestet. Das neuronale Netz besteht aus RNNs und es wurden Embeddings auf Grundlage des Skip-Bigram Models (vgl. Mikolov et al. 2013) selbst trainiert und eingesetzt. Das Modell, welches die Buchstaben N-Gramme benutzt, konnte bessere Ergebnisse als das Wort N-Gramm basierte Modell erzielen. Es wurde festgestellt, dass die bessere Performance daran lag, dass Wörter, in welchen Buchstaben durch Zahlen oder andere Zeichen ersetzt worden sind, besser durch die Buchstaben N-Gramme abgebildet wurden und somit vom Modell besser verarbeitet werden konnten.

4.2.2 Repräsentation von Text durch Buchstaben

Zhang et. al. (2017) schlugen ein Buchstaben-basiertes neuronales Netz vor, um vor allem mit demselben Modell über verschiedene Sprachen hinweg Text (in diesem Fall Tweets) klassifizieren zu können. Je nach Klassifikationsaufgabe wurden Benutzernamen, Hashtags und

URLs entweder standardisiert oder entfernt. Es wurde eine Embedding Layer in das Modell eingebaut, wobei die Embeddings zufällig initialisiert wurden und trainierbar waren. Auf die Embedding Layer folgten vier CNN Layer gefolgt von zwei Dense Layern. Zum Vergleich wurde ein Wort-basiertes CNN Modell sowie ein SVM Modell herangezogen. Hierbei hat sich gezeigt, dass der Buchstaben-basierte Ansatz für Tweets in Sprachen abgesehen von Englisch besser funktioniert, als diese erst maschinell zu übersetzen und dann zu klassifizieren. Auch kann damit die Klassifikation von Daten, in welchen verschiedene Sprachen auftauchen, verbessert werden.

Auch Zhang et. al. (2015) arbeiteten mit CNNs und Buchstaben, um Text vor allem hinsichtlich des Sentiments und der Topics zu klassifizieren. Genauer gesagt wurden zwei Modelle aufgebaut – ein Größeres und ein Kleines. Beide bestanden aus sechs CNN und drei Dense Layern. Lediglich die Länge des Inputs (maximale Länge des Texts) variierte. Das Alphabet wurde auf 70 Zeichen standardisiert und es wurden nur Kleinbuchstaben verwendet. Dieses Modell wurde bei gleichen Hyperparametern mit einem Wort-basiertem LSTM Modell, einem Wort-basiertem CNN Modell sowie mehreren Feature-basierten Modellen verglichen. Bei den Wort-basierten Modellen wurden 300-dimensionale Word2Vec Embeddings verwendet. Die Modelle wurden auf acht verschiedenen Datensätzen getestet. Es stellte sich heraus, dass der Buchstaben-basierte Ansatz mit den anderen Ansätzen mithalten konnte. Welches Modell das beste Ergebnis lieferte, hing von dem jeweiligen Datensatz ab. Aufgrund der Ergebnisse wurde die Hypothese, dass Buchstaben-basierte Modelle für rohen, benutzergenerierten Text besser geeignet seien, aufgestellt.

Eine weitere Arbeit stammt von Serrà et. al. (2017). Um Hassrede zu erkennen, werden zwei hintereinandergeschaltete neuronale Netze trainiert. Zunächst wird für jede Klasse ein Sprachmodell trainiert, welches dafür gedacht ist, nachfolgende Buchstaben vorherzusagen. Dieses Netz nimmt zunächst One-Hot kodierte Buchstaben als Input und verarbeitet diese in einer TimeDistributed Dense Layer. Drauf folgt eine GRU Layer und eine weitere TimeDistributed Dense Layer für die Ausgabe. Die Performanz dieser Modelle wird wiederum als normalisierte, sequenzielle Messung an das nächste neuronale Netz weitergegeben. Dieses besteht aus zwei Layern sowie Dropout (Srivastava et. al. 2014) und Ausgabelayer. Der Ansatz ist dafür gedacht, besonders gut mit Out-of-Vocabulary Wörtern umgehen zu können, da so der Fokus „von der Beschreibung gesehener Daten zu der Vorhersage ungesehener Daten verlagert wird“ (Serrà et. al. 2017). Getestet wurde das Modell deswegen auf Daten von Twitter. Verglichen wurde der Ansatz mit Naive Bayes, Naive Bayes gepaart mit logistischer

Regression und einem einfachen buchstabenbasierten neuronalen Netz, welches aus einer TimeDistributed Embedding Layer, GRU Layer, Dense Layer und anschließender Ausgabelayer besteht. Diese konnten alle von der vorgeschlagenen zwei-Schritt Klassifikation übertroffen werden.

4.2.3 Gemischte Ansätze

Statt sich für nur eine Repräsentation zu entscheiden kann auch nach einem Weg gesucht werden, mehrere Repräsentationen gleichzeitig in die Modellarchitektur aufzunehmen. Die Fusion mehrerer Eingaben kann auf unterschiedliche Weisen geschehen.

In der Arbeit von Park und Fung (2017) wurden drei verschiedene, auf CNNs basierende Modelle getestet. Das Ziel war die Klassifikation von Twitter Daten von Waseem und Hovy (2016), welche mit den Klassen „Sexismus“, „Rassismus“ und „Neutral“ annotiert wurden. Nur die Eingabevektoren unterscheiden diese Modelle. Als Eingabe wurden Wörter, Buchstaben sowie eine Hybrid-Version getestet. Das Buchstaben-CNN-Modell wurde von Zhang et. al. (2015) und das Wort-CNN-Modell von Kim (2014) übernommen. In der Arbeit von Kim wurden die Texte nach dem Segmentieren mit Word2Vec Embeddings ausgestattet. Zusätzlich haben Park und Jung einige, nicht im Vokabular von Word2Vec enthaltene Wörter (oftmals Hashtags) in kleinere Einheiten segmentiert, sodass diese auch erfasst werden können. Die Hybrid-Version der beiden Modelle verarbeitet beide Repräsentationen separat mit drei CNN Layern und konkateniert die Ausgabe dieser zu einem Vektor, welcher direkt an die Ausgabelayer weitergeleitet wird. Es hat sich gezeigt, dass Buchstaben als Eingabe am schlechtesten abschneiden. Die Hybrid-Version konnte die Testdaten am besten klassifizieren.

Auch Dos Santos und Gatti (2014) haben sich unter anderem mit der Sentiment Klassifikation von Tweets beschäftigt. Sie haben dabei ein neuronales Netz (CharSCNN) entwickelt, welches gleichzeitig Features von Buchstaben bis hin zu Sätzen verarbeiten kann. Das Netz erhält zunächst die Wörter eines Satzes als Eingabe. Die Wörter werden dann durch Embeddings kodiert. Die Embeddings bestehen dabei aus zwei untergeordneten Vektoren: Wort und Buchstaben Embeddings. Die untergeordneten Wort Embeddings werden anhand der englischen Wikipedia mit Word2Vec erlernt. Die Buchstaben je Wort werden zunächst One-Hot kodiert. Diese werden dann an die erste CNN Layer in CharSCNN weitergeleitet und die Ausgabe aller Buchstaben wird mit einer „Max“-Operation zu einem Vektor zusammengeführt. Eine zweite CNN Layer in CharSCNN extrahiert Satz-Embeddings. Diese werden ebenfalls mit

einer „Max“-Operation über die vorher erstellten, übergeordneten Wort-Embeddings generiert. Auf den Daten von Twitter erreichte CharSCNN eine Genauigkeit 0,86 für eine binäre Klassifikation.

In der Studie von Risch und Krestel (2018) ging es um die Identifikation von Aggression in Nutzerposts auf den Sozialen Medien im Allgemeinen. Dafür wurden nicht nur verschiedene Repräsentationsebenen, sondern auch ganz verschiedene Modellansätze vereint. Aus einem Post wurden verschiedene Repräsentationen extrahiert: Wörter und deren Embeddings, Buchstaben N-Gramme, Wort N-Gramme sowie einige syntaktische Eigenschaften. Nur auf die erstgenannte Repräsentation wurde ein neuronales Netz angewandt. Die drei anderen Repräsentationen wurden mittels logistischer Regression untersucht. Die Ergebnisse aller vier Repräsentationen wurden mit Entscheidungsbäumen und Gradiant Boosting vereint. Wie bei Park und Fung (2017) auch wurde innerhalb der Vorverarbeitung insbesondere darauf geachtet, dass Twitter Hashtags in ihre einzelnen Wörter aufgetrennt werden, da diese oftmals aus mehreren, zusammengeschriebenen Wörtern bestehen. Es wurden vortrainierte FastText Embeddings ausgewählt. Das neuronale Netz besteht aus einer Dropout Layer von 0,1 jeweils nach der Eingabelayer und vor dem Ausgabelayer. Dazwischen befindet sich eine bi-direktionale GRU Layer, dessen Ausgabe separat von einer Pooling Layer und einer Lambda Layer verarbeitet wird, deren Ausgaben wiederum konkateniert an den zweiten Dropout Layer und dann an den Ausgabelayer weitergereicht werden. Auf Daten von Twitter und Facebook erzielte dieses Modellensemble jeweils einen F1 Score von 0,60.

Letztlich haben auch Vijayaraghavan et. al. (2016) neuronale Netze entwickelt, welche die Haltung („für“ oder „gegen“) eines Verfassers gegenüber einem Thema erkennen. Aufgrund der unterschiedlichen Themenkreise wurden zwei Modelle sowie eine kombinierte Version dieser geschaffen. Das buchstabenbasierte Modell orientierte sich an Zhang und LeCun (2015), welches hinsichtlich der Größe des Alphabets angepasst wurde. Die Texte wurden One-Hot kodiert, welche dann an vier hintereinander gereihte CNNs sowie mehrere Dense Layer weitergereicht werden. Ein Dropout von 0,5 wurde außerdem angewandt. Innerhalb des wortbasierten Modells werden zunächst zufällige Embeddings initialisiert, die dann erlernt werden. Die Embeddings werden an mehrere, parallel geschaltete CNNs mit unterschiedlichen Schrittgrößen weitergereicht. Die Ausgabevektoren dieser werden zu einem einzigen Vektor konkateniert, welcher nach einer Kompression an den Ausgabelayer weitergegeben wird. Für die Auswertung wurden die F1 Scores beider Klassen und beider Modelle separat für drei unterschiedliche Themenkreise verglichen. Da für ein Thema der beste F1 Score der Klassen

drastisch für die Modelle verschieden war, wurde das Modell nur für diesen Anwendungsfall durch eine Heuristik kombiniert: wenn das buchstabenbasierte Modell „gegen“ vorhersagt, bleibe bei dieser Klassifikation, ansonsten verwende die Klassifikation des wortbasierten Modells. Grundsätzlich hat diese Arbeit allerdings gezeigt, dass das buchstabenbasierte Modell das wortbasierte Modell übertrifft.

In der vorliegenden Thesis werden einige verschiedene Ansätze aus diesen Arbeiten verglichen. Der Schwerpunkt liegt dabei auf den unterschiedlichen Arten Text zu repräsentieren. Die vorgestellten verschiedenen Möglichkeiten, Text (speziell im Hinblick auf Hassrede) vorzuverarbeiten, werden dabei überdacht oder getestet und daraufhin ausgewählt sowie teilweise verbessert. Dasselbe wird für die Wahl der Wort- und Buchstaben Embeddings durchgeführt. Ein weiterer wesentlicher Punkt ist, dass diese Ansätze mit dem neuesten Werkzeug aus dem NLP Bereich, BERT, verglichen werden, welches bislang für dieses Thema noch nicht eingesetzt wurde.

5. Vorstellung der unterschiedlichen Arbeitsabläufe

Der Stand der Forschung zeigt nicht nur, wie divers die Ansätze, was die Repräsentationsarten von Text angeht, sein können, sondern auch, wie vielfältig die Vorverarbeitungsschritte und Modellarchitekturen sein können. Zusammenfassend lässt sich damit sagen, dass es keine universellen Vorverarbeitungsschritte und auch keine universelle Modellarchitektur für alle Anwendungen von automatischer Textklassifikation gibt. Trotzdem können die verschiedenen Ansätze durch die bereits geleisteten Arbeiten eingegrenzt werden. Auch die Tatsache, dass es sich bei den Daten um sehr kurze, benutzergenerierte Texte handelt, hilft bei der Wahl der Vorgehensweise.

Ein typischer Deep Learning Workflow besteht immer aus denselben Schritten. Erst muss sich für eine Repräsentation der Eingabedaten entschieden werden. Speziell für Textdaten wird damit festgelegt, ob der Algorithmus Text als eine Anreihung von Buchstaben, Wörtern, Sätzen oder sonstigen Elementen (N-Gramme, Paragrafen…) verstehen soll. Dann müssen die Daten in eine für den Algorithmus notwendige Form gebracht werden. Je nach Repräsentationsart müssen die einzelnen Elemente des Textes in Vektoren umgewandelt werden. Dafür gibt es grundsätzlich zwei Möglichkeiten: One-Hot Encoding und Embeddings (vgl. Chollet 2018, 69).

Beim One-Hot Encoding wird zunächst jedem Element eine eindeutige Zahl (ein Index) zugeordnet. Jedes Element wird dann in einen n-dimensionalen, binären Vektor umgewandelt, wobei *n* für die Größe des Vokabulars steht. Dieser Vektor besteht nur aus Nullen, außer an der Stelle des Indexes des jeweiligen Elements (vgl. ebd., 181).

Embeddings sind ebenfalls Vektoren, die für Elemente eines Textes stehen (meistens werden Embeddings allerdings auf Wörtern aufgebaut). Diese haben im Unterschied zur One-Hot Kodierung eine vordefinierte, meist kleinere Größe und sind nicht binär, besitzen also mehr Werte als nur 0 und 1. Diese Vektoren müssen zunächst erlernt werden, was auf Grundlage gemeinsamer Vorkommnisse der Wörter geschieht. Die Dimensionen des Vektors bilden dabei unterschiedliche semantische Beziehungen zwischen den Wörtern ab.

Dass die Embeddings die Beziehungen zwischen Wörtern gut abbilden können, setzt voraus, dass sie auf möglichst vielen Daten erlernt beziehungsweise trainiert werden. Es gibt daher zwei Möglichkeiten, wie Embeddings eingesetzt werden können. Wenn das Korpus nicht genügend Daten umfasst, können Embeddings zunächst auf einem weitaus größeren, allgemeinen Korpus trainiert und dann eingesetzt werden. Bekannte, vortrainierte Wort Embeddings wurden bereits in Kapitel 4 erwähnt (Word2Vec, GloVe, FastText). Bei ausreichend Daten können die Embeddings von Grund auf trainiert werden. Ein Kompromiss aus beiden Ansätzen ist, vortrainierte Embeddings auf dem eigenen Datensatz weiter zu trainieren. Dies ist zum Beispiel hilfreich, wenn nicht genügend Vokabular durch die vortrainieren Embeddings abgebildet wird.

Die Daten müssen außerdem in Trainings-, Evaluations- und Testset aufgeteilt werden. Das Trainingsset ist die Lerngrundlage des Modells. Das Evaluationsset dient dem Anpassen der Hyperparameter. Nach jeder Trainingsiteration werden meist einige Parameter wie zum Beispiel die Größe der einzelnen Ebenen oder die Größe des Dropouts im Hinblick auf mögliche Verbesserungen angepasst. Verbesserungen oder Verschlechterungen können dann anhand der erreichten Genauigkeit auf dem Evaluationsset abgelesen werden. Nur anhand des Testsets kann dann die tatsächliche Ergebnisqualität des Modells gemessen werden, da die Daten gänzlich ungesehen bleiben. Somit wird Overfitting, welches durch das Anpassen der Hyperparameter durch den Menschen entstehen kann, vermieden. „Overfitting“ bedeutet, dass ein Modell zwar sehr gute Ergebnisse auf „bereits gesehenen“ Daten (dies betrifft das Trainingsset aber auch indirekt das Evaluationsset) erzielt, allerdings auf neuen Daten (dem Testset) deutlich schlechter abschneidet. Das Modell ist damit zu stark angepasst.

Bei der Aufteilung des Korpus ist es wichtig, dass alle Klassen in den Sets gleich häufig vertreten sind, damit alle Klassen gleich gut erlernt werden können. Daher sind für diese Arbeit die Klassen in allen Sets gleichverteilt. Das Trainingsset besteht aus 12.000 Tweets und das Evaluations- sowie Testset bestehen jeweils aus 3.000 Tweets. Die Tweets wurden einmalig zufällig aus den Klassen ausgewählt. Um trotzdem das gesamte Korpus ausschöpfen zu können, könnten bei jedem Durchlauf eines Modells die Tweets auch neu zufällig ausgesucht werden. Um allerdings die Schwankungen hinsichtlich der Ergebnisse so gering wie möglich zu halten wurde dies nicht unternommen. Andere Evaluationsmethoden werden in Kapitel 7 nochmals angesprochen.

Auch wenn die Modellerstellung für die unterschiedlichen Repräsentationen variieren kann bzw. sollte, gibt es Eigenschaften, die gleichbleibend sind. Der Ausgabelayer ist, im Gegensatz zum Eingabelayer, immer derselbe: ein Dense Layer mit Softmax-Aktivierung, welcher die Wahrscheinlichkeitsverteilung über die drei Klassen wiedergibt. Einige weitere Parameter wurden während der Erstellung des ersten Modells festgelegt. Dabei wurde wie folgt vorgegangen: zunächst wurde ein Modell, entsprechend an Kapitel 4 orientiert aber ohne Dropout, aufgebaut und für 50 Epochen trainiert. Anhand der Abweichungen der Trainings- und Validierungsgenauigkeiten („Accuracy“) voneinander wurde dann ein Dropout eingebaut. Dieser wurde für jeden neuen Trainingsdurchlauf erhöht, bis die Genauigkeiten ausgeglichen waren. Anhand des Verlaufs der Genauigkeiten über die Epochen konnte auch abgeschätzt werden, wie viele Epochen in etwa für das Training benötigt werden, bis keine Verbesserung mehr erzielt wird. Der somit ermittelte Dropout (0,5) und die Anzahl an Epochen (15) wurden auf alle anderen Modelle übertragen. Als Lernrate wurde immer RMSprop verwendet. Zur Bewertung der Modelle wird der F1 Score herangezogen.

Ein wesentlicher Fokus dieser Arbeit besteht darin, unterschiedliche Repräsentationen für Text für neuronale Netze und deren Einfluss auf die Ergebnisse zu untersuchen. Die Ansätze in allen folgenden Sektionen dieses Kapitels sind von den vorgestellten Studien aus Kapitel 4 inspiriert. Im Folgenden sollen die einzelnen Schritte sowie einige Varianten der in dieser Arbeit verfolgten Ansätze, gegliedert in Wort-, Buchstaben und N-Gramm-Repräsentation, genauer beschrieben und begründet werden.

Die Erhebung der Ergebnisqualität eines Modells wird anhand des F1 Scores und anhand der Wahrheitsmatrix vorgenommen. Programmiert wurde in der Sprache Python, als Bibliothek

wurde Keras[8] mit Tensorflow[9] Back-End verwendet. Der Programmcode für die Auswertungsmethoden wurde der Bibliothek SciKit-Learn[10] entnommen.

5.1 Vorverarbeitungsschritte und Embeddings

5.1.1 Wortrepräsentation

Die Repräsentation eines Tweets als eine Aneinanderreihung von Wörtern zu verstehen bedeutet als erstes zu bestimmen, wie genau die Tokenisierung vorgenommen wird. Tokenisieren bedeutet die Zerlegung einer Zeichenkette in einzelne Tokens. Genauer gesagt entsteht in diesem Fall eine Liste aus Wörtern. Dies kann sehr einfach ausfallen, indem man Wortgrenzen als jedes Zeichen außer einem Buchstaben definiert. In der Praxis sollten allerdings viel mehr Regeln als diese die Tokenisierung ausmachen (beispielsweise die Behandlung von Apostrophen oder Bindestrichen). Für einen ersten Versuch wurde der Tokenisierer von Keras verwendet. Dieser ist recht simpel aufgebaut: alle Wörter werden in Kleinbuchstaben umgewandelt und alle vorher definierten Zeichen werden als Wortgrenze angesehen.

Um dies besser zu handhaben wurde der TweetTokenizer von NLTK in den Workflow eingebunden. Durch den TweetTokenizer werden auch zunächst alle Zeichen der Tweets in Kleinbuchstaben umgewandelt. Außerdem werden Buchstaben, die mehr als dreimal hintereinander wiederholt werden, zu einer Länge von drei vereinheitlicht. Eine Wiederholung eines Buchstabens kann als eine Betonung angesehen werden; diese soll auch als solche erhalten bleiben. Somit wird beispielsweise das Wort *big* von den Wörtern *biiig* und *biiiiig* abgegrenzt. Letztere sind dann gleichwertig. Es werden zudem alle Zahlen und Zeichen, welche nicht innerhalb eines englischen Ausdrucks vorkommen können, entfernt. Dies gilt auch für das Zeichen #. Somit wurde sich dazu entschieden, Hashtags zu normalen Wörtern umzuformen (*cat* bedeutet dasselbe wie *#cat*). Benutzernamen (gekennzeichnet durch @) werden dahingegen gänzlich entfernt. Diese Entscheidung wurde getroffen, um Vorurteile bezüglich der in Tweets erwähnten Personen nicht mit in das Modell einfließen zu lassen. Denn nur weil beispielsweise ein Benutzer häufig in Hasskommentaren erwähnt wird, soll die

[8] https://github.com/keras-team/keras (Stand 21.05.2019).
[9] https://github.com/tensorflow/tensorflow (Stand 21.05.2019).
[10] https://scikit-learn.org/stable/ (Stand 21.05.2019).

Wahrscheinlichkeit dafür, dass ein Tweet, das diesen Benutzernamen enthält, auch Hassrede ist, nicht ansteigen.

Schließlich galt besondere Aufmerksamkeit der in den Tweets verwendeten Emojis. Diese sind teils als HTML-Entities und teils in Unicode kodiert. Beide Varianten wurden in ihre textuelle Repräsentation (z.B. *face with tears of joy*) umgewandelt und die einzelnen Wörter wiederum der Wörterliste des Tweets angehängt. So können die Bedeutungen der Emojis auch in vortrainierten Wort-Embeddings gefunden werden. Letztlich wurde das Token *i'd* in *i* und *would* aufgelöst; dies liegt daran, dass nur dieses Token nicht in den vortrainierten Embeddings gefunden werden konnte.

Außerdem galt es zu bestimmen, auf welche Länge und auf welches Vokabular die Tweets standardisiert werden sollten. Für die maximale Länge wurden 100 Tokens festgelegt, sodass alle Tweets in ihrer gesamten Länge erfasst werden können. In das Vokabular wurden jene Wörter aufgenommen, welche öfter als 10-mal im gesamten Korpus vorkommen. Wörter, welche nur 1-mal im Korpus vorkommen sind nicht relevant, da sie keinen Wiedererkennungswert besitzen. Durch die Tatsache, dass grundsätzlich in einem Durchlauf nicht von allen Daten Gebrauch gemacht wird, wurde von mindestens zwei Vorkommnissen auf mindestens 11 Vorkommnisse erhöht. So soll sichergestellt sein, dass Wörter aus dem Training auch in den Testdaten erkannt werden können. Alle URLs werden somit auch automatisch aussortiert.

Schließlich wurden noch Wörter, die zwar häufig vorkommen aber jedoch störend sind, aus dem Vokabular ausgeschlossen. Einerseits betraf dies das Token *rt*, welches ein Retweet signalisiert. Laut Vora et. al. (2017) ist dieses Token als Störsignal aufzufassen und trägt nicht zu einer Verbesserung der Klassifikation bei. Andererseits wurden auch die Tokens *yankees* und *charlie* herausgenommen. Diese Tokens sind eindeutig zu einem speziellen Thema zugeordnet (vgl. Kapitel 3). Die Maskierung sollte helfen, dieses Themensignal herauszunehmen, sodass eine Klassifikation nicht diesen beiden Tokens zugrunde liegen kann. Das Wort *bird* wurde zwecks dessen Mehrdeutigkeit nicht maskiert. Die Tweets hätten auch gänzlich aus dem Korpus entfernt werden können. Allerdings wären dann auch bedeutsame Tweets, bei denen das Wort tatsächlich als Beleidigung auftritt, gelöscht worden und dieser Fall von Hassrede wäre nicht mehr abgedeckt. Der gewählte Weg soll einen Kompromiss darstellen. Letztlich wurde die Möglichkeit implementiert, Buchstaben N-Gramme aus den so vorverarbeiteten Wörtern zu generieren.

Die bereits angesprochene, maschinenlesbare Auflösung der Wörter eines Tweets ist in diesem Fall ein 100-dimensionaler Zahlenvektor, bei welchem jede Zahl für genau ein Wort steht. Die Zahlen 0 und 1 stehen dabei jeweils für das Padding (um ein Tweet auf die Länge 100 zu strecken) und für Wörter außerhalb des Vokabulars. Diese Repräsentation allein ist allerdings nicht ausreichend, denn noch sind die einzelnen Wörter nur durch eine einzige Zahl vertreten und sind noch keine Vektoren. Hier wurden vortrainierte FastText Embeddings (Grave et. al. 2018) eingesetzt, da diese bereits ein etablierter Standard sind und das Korpus zu klein ist, um die Embeddings selbst zu trainieren.

FastText Embeddings bieten sich besonders an, da Wörter außerhalb des FastText Vokabulars behandelt werden können. Ein Problem beim Umgang mit Tweets ist, dass viele unterschiedliche Schreibweisen von Wörtern auftreten. Diese sind nicht nur Rechtschreibfehler, sondern auch gewollte Alternativen. Die betroffenen Wörter können deshalb nicht einfach aussortiert oder ersetzt werden. Vortrainierte Embeddings entstehen meist auf Datengrundlagen wie Wikipedia – anders geschriebene Wörter oder auch Slang werden dann in den Embeddings nicht gefunden. Ein großer Vorteil von FastText gegenüber Word2Vec oder auch GloVe ist, dass falls es kein Embedding für ein Wort gibt, trotzdem ein Vektor aus den N-Grammen des Wortes generiert werden kann (Bojanowski et. al. 2017). Jedes Embedding wird dann einem Wort über dessen Index zugeordnet. Im Modell werden daraufhin im ersten Schritt die Embeddings für die Indexe in den Eingabevektoren eingesetzt.

Zusätzlich zu den Embeddings von FastText wurden auch die Embeddings von BERT getestet. BERT ist ein bereits von Google vortrainiertes Sprachmodell, welches für jede Art von Natural Language Processing angepasst und verwendet werden kann (vgl. Devlin et. al. 2018). Die Embeddings mussten dabei erst aus diesem Sprachmodell extrahiert werden. Benutzt wurde nur der letzte Layer von BERT; die Embeddings haben damit eine Größe von 768. Im Unterschied zu Word2Vec, GloVe und FastText sind die Embeddings kontextualisiert. Das bedeutet, dass das gleiche Wort in zwei unterschiedlichen Kontexten auch unterschiedliche Embeddings zugeordnet bekommt.

Die Vorverarbeitung der Tweets kann wie folgt zusammengefasst werden:

- Umwandlung in Kleinbuchstaben,
- Umwandlung der Emojis zu Text,
- Entfernen von Nutzernamen,
- Kürzen von sich wiederholenden Buchstaben in Wörtern,

- Entfernen der meisten Zeichen und allen Zahlen,
- Maskieren von problematischen Wörtern,
- Ausschließen der seltensten Wörter.

Einige in Kapitel 4 aufgeführte Schritte wurden nicht angewandt. Dazu gehört zum Beispiel das Herausfiltern von Stoppwörtern, auch wenn es in dieser Arbeit getestet wurde. Da die Modelle eher schlechter mit entfernten Stoppwörtern performten (vgl. Anhang, Abbildung A22), wurde dieser Schritt nicht beibehalten. Rechtschreibkorrektur sowie auch die Auftrennung von zusammengeschriebenen Wörtern wurde nicht unternommen, da beide Arbeitsschritte indirekt durch die hier verwendeten Embeddings behandelt werden. Dieses Thema wird in Kapitel 7 nochmals ausführlich diskutiert.

5.1.2 Buchstabenrepräsentation

Genauso wie Wörter eine Repräsentationsebene eines Textes sein können, können ebenso Buchstaben an ihre Stelle treten. Für diesen Ansatz wird als ersten Schritt ein Tweet in eine Liste von Zeichen umgeformt. Statt sich auf ein ausgewähltes Alphabet zu beschränken wurden alle Zeichen erfasst. Die Emojis wurden bei dieser Variante nicht umgeformt, da sie als eigene Zeichen gezählt werden sollten. Statt die Wörter *yankees* und *charlie* zu maskieren wurden sie gänzlich aus dem Korpus entfernt. Als maximale Länge für die Tweets wurde 280 festgelegt – dies ist die Konvention von Twitter, die Eingabe von maximal 280 Zeichen pro Tweet. Somit wird auch bei dieser Repräsentation ein Tweet in seiner gesamten Länge erfasst.

Genauso wie es vortrainierte Embeddings für Wörter gibt, existieren solche auch für Buchstaben. Allerdings beinhalten diese meist keine Emojis und werden häufig auch nicht speziell mit Tweets trainiert[11]. Es wurden deswegen 300-dimensionale Embeddings auf einem größeren Twitter Korpus mit FastText trainiert und anschließend auf die eigenen Daten (alle 37.555 Tweets) abgestimmt. Dieses zusätzliche Twitter Korpus beinhaltet drei Millionen Tweets zum Thema „Kundenbetreuung“[12]. Die Embeddings wurden dann für 20 Epochen auf dem „Hassrede“-Korpus weitertrainiert. Zusätzlich wurde die Möglichkeit, die Tweets in One-Hot Kodierung umzuwandeln und in dieser Form an das Modell weiterzureichen, getestet, um dies mit den vortrainierten Embedding zu vergleichen. Aus demselben Grund wurde ebenso die

[11] Getestet wurden die Buchstaben Embeddings von Max Woolf (https://github.com/minimaxir/char-embeddings), welche aufgrund der in ihnen nicht vorhanden Emojis nicht verwendet werden konnten (Stand 21.05.2019).
[12] https://www.kaggle.com/thoughtvector/customer-support-on-twitter/version/10 (Stand 21.05.2019).

Variante, die Embeddings nur auf den eigenen Daten zu trainieren, offengehalten (vgl. Kapitel 5.2.2).

5.2 Modellerstellung

In diesem Abschnitt geht es um die Auswahl der Architektur des neuronalen Netzes für die unterschiedlichen Textrepräsentationen – analog zu den Unterkapiteln von Abschnitt 4.2. Es sei darauf hingewiesen, dass dies nicht im Hauptfokus der Arbeit stand. Die Bestimmung der besten Netzarchitektur kann sicherlich ausführlicher untersucht werden. Für den Zweck der vorliegenden Thesis war es ausreichend, ein gut geeignetes neuronales Netz mittels vorangegangener Forschungsarbeiten zu bestimmen. Ob es nun tatsächlich ideale neuronale Netze sind, ist für die Ergebnisqualität im Sinne dieser Arbeit unerheblich. Es wird daher nicht ausgeschlossen, dass es jeweils geringfügig bessere Modelle für die einzelnen Repräsentationen als jene, die am Ende der einzelnen Sektionen ausgewählt werden, geben könnte.

5.2.1 Wortrepräsentation

Es wurden vier Modelle getestet, welche von den Arbeiten aus Kapitel 4.2.1 inspiriert sind. Dies umfasst ein LSTM Modell, ein CNN-LSTM Modell, ein Bi-LSTM Modell und ein Dense Modell mit drei Layern. Jedes Modell wurde in 20 Durchläufen getestet, um die Varianz in den Ergebnissen gering zu halten. Alle Modelle wurden für 15 Epochen auf einer Batchgröße von 32 trainiert. Außerdem wurde bei allen Modellen ein Dropout von 0,5 eingebaut, um Overfitting auf dem Evaluationsset zu vermeiden. All dies wurde auf dem wortbasierten sowie auch auf dem N-Gramm-basierten Datenset durchgeführt. Es wurden nur Tetragramme getestet. Weitere Parameter der einzelnen Modelle können im Anhang (vgl. A1) eingesehen werden. Die Konfusionsmatrizen für diesen Abschnitt sind ebenfalls im Anhang (vgl. A2) einsehbar. Der Durchschnitt aller 20 F1 Scores je Modell wurde in Tabelle 5.1 festgehalten.

	LSTM	3 x Dense	CNN-LSTM	Bi-LSTM
F1 Score (Wörter)	0,9063	0,8959	0,9010	0,9073
F1 Score (N-Gramme)	0,9134	0,9062	0,9078	0,9127

Tabelle 5.1: Durchschnittliche F1 Scores (gewichtet) auf dem wortbasierten und N-Gramm basierten Testset aus 20 Durchläufen vier getesteter Modelle.

Es zeigt sich, dass LSTM und BiLSTM Modelle im Durchschnitt die besten Performanzen aufweisen. Das Hinzufügen von CNNs scheint dem Modell eher zu schaden. Dies gibt zu erkennen, dass komplexere Modelle (wie zum Beispiel CNN-LSTM aber auch BiLSTM) nicht unbedingt die angebrachteste Lösung für ein Deep Learning Problem sind, da einfachere Modelle ähnlich gute Ergebnisse erzielen können. Auch zeigt sich, dass N-Gramme insgesamt besser für die Klassifikation geeignet sind als Wörter. Da LSTM zwar simpler ist, sich aber als ähnlich effektiv wie BiLSTM herausgestellt hat, ist LSTM-ngram als finales, wortbasiertes Modell gewählt worden.

Da sich LSTM als beste Architektur statistisch erwiesen hat wurden nur in diesem Modell die Embeddings von FastText durch die Embeddings von BERT ersetzt. Da BERT einen eigenen Tokenisierer verwendet, welcher unbekannte Wörter in N-Gramme auftrennt, wurden als Eingabe die vorverarbeiten Wörter und nicht N-Gramme verwendet. Die Daten wurden auch wie in 5.1.1 beschrieben vorverarbeitet, bevor sie an BERT weitergegeben wurden.

Für den ersten Versuch, die BERT Embeddings zu extrahieren, wurden einige Skripte von GitHub[13] verwendet. Die Wortvektoren in der resultierenden Ausgabedatei wurden dann für LSTM-ngram wieder eingelesen. Beim ersten Durchlauf des Modells ist allerdings aufgefallen, dass für jede Epoche nur eine Genauigkeit von 0,33 auf dem Validierungsset erreicht werden konnte. Die Vorhersagen des Modells auf dem Testset zeigen, dass ein Tweet zu 33% jeder Klasse zugeordnet wird. Es wurde deswegen eine weitere Möglichkeit getestet, BERT Embeddings zu extrahieren. Hierzu wurde BERT-as-service[14] verwendet, allerdings ohne eine Ergebnisverbesserung zu erzielen. Mit BERT-as-service wurde dann die durch BERT vorgenommene Tokenisierung (vor allem die Umwandlung in Buchstaben N-Gramme) überprüft, indem nicht nur die Embeddings sondern auch die Tokens zurückgegeben wurden. Es hat sich gezeigt, dass wenn die Tokenisierung durch BERT ausbleibt, vor allem stark negativ konnotierte Wörter nicht erkannt werden. Der Sinn dieser Wörter ergibt sich für BERT demnach nur aus deren N-Grammen. Es wurden deshalb zwei weitere Versuche aufgestellt, in dem von einigen Wortgruppen die Embeddings extrahiert wurden, welche wiederrum mit einer PCA (Pearson 1901) visualisiert wurden (vgl. Anhang A4). Die Versuche zeigen, dass BERT die positiven Wörter den Negativen zuordnet (vgl. Kapitel 7.3). Schließlich wurde statt die „nur“ vortrainierte Version von BERT für die Extraktion zu benutzen auch das auf die eigenen Daten abgestimmte, fine-tuned BERT (siehe Ende des Unterkapitels) verwendet, sodass die

[13] https://github.com/google-research/bert (Stand 13.06.2019).
[14] https://github.com/hanxiao/bert-as-service (Stand 13.06.2019)

fehlenden Wörter und deren Bedeutung eventuell so aufgenommen wurden. Allerdings hat sich auch hier keine Verbesserung der Ergebnisse gezeigt.

In einem letzten Versuch wurde in jenen Workflow Daten aus Wikipedia statt dem Hassrede-Korpus eingebunden, womit eindeutig sinnvolle Ergebnisse erzielt werden konnten. Das Problem liegt somit bei den Hassrede-Daten. Die BERT Embeddings scheinen also für diese Art von Klassifikation unbrauchbar zu sein, da sie den Unterschied zwischen den Klassen nicht abbilden können. Warum genau die BERT Embeddings möglicherweise ungeeignet sind, wird in Kapitel 7.3 nochmals diskutiert.

Zusätzlich wurde BERT selbst verwendet. Statt die Embeddings aus dem Modell zu extrahieren und in das eigene Modell einzubinden wurde BERT auf die eigenen Daten abgestimmt (Fine-tuning). Hierzu wurden ebenfalls Skripte von GitHub[13] verwendet und teilweise angepasst. Ebenso wie für die BERT Embeddings wurden für die Eingabe die vorverarbeiteten Tweets in Wortform vorgezogen. Das daraus entstehende Modell wurde dann ohne weitere Anpassungen für die Vorhersagen auf dem Testset verwendet. Fine-tuned BERT erreicht einen F1 Score von 0,934. Da dieser Wert bereits vergleichsweise deutlich besser ist als jeder der einzelnen F1 Scores der anderen Architekturen (kein vorheriger Wert konnte 0,919 überschreiten) wurde dies nur einmal getestet. Es übertrifft damit alle anderen getesteten Modelle.

5.2.2 Buchstabenrepräsentation

Was die Buchstabenebene angeht, so wurde hauptsächlich ein einziges Modell erstellt, welches allerdings mit unterschiedlichen Eingabelayern ausgestattet wurde. Das Modell besteht aus drei CNN Layern und einer weiteren Dense Layer vor der Ausgabelayer. Es wurde außerdem ebenfalls ein Dropout von 0,5 eingebaut. Drei verschiedene Eingabemethoden wurden getestet: die in 5.1.1 beschriebenen, vortrainierten Embeddings (CNN-Pre-Emb), One-Hot Kodierung (CNN-One-Hot) und nur auf dem Korpus trainierte Embeddings (CNN-Emb), welche in diesem Fall nur innerhalb des Modells erlernt werden. Wie bei den wortbasierten Modellen auch umfasst ein Durchlauf 15 Epochen. Tabelle 5.2 zeigt die Ergebnisse.

	CNN-Pre-Emb	CNN-One-Hot	CNN-Emb
F1 Score	0,7405	0,7397	0,7469

Tabelle 5.2: Durchschnittliche F1 Scores (gewichtet) auf dem buchstabenbasierten Testset aus 20 Durchläufen drei getesteter Modelle.

Die Ergebnisse aller charakterbasierten Modelle ähneln sich auf den ersten Blick. Insgesamt übertreffen die wortbasierten Modelle die buchstabenbasierten Modelle deutlich. Vor allem zeigt sich unerwarteterweise, dass die vortrainierten Embeddings nicht zu besseren Ergebnissen verhelfen können als die selbsttrainierten Embeddings. One-Hot-Kodierung weist die schlechtesten Ergebnisse auf. Die Konfusionsmatrizen der Modelle werden in Abbildung 5.1 dargestellt. Die Konfusionsmatrizen wurden aus denselben 20 Durchläufen der Modelle erstellt.

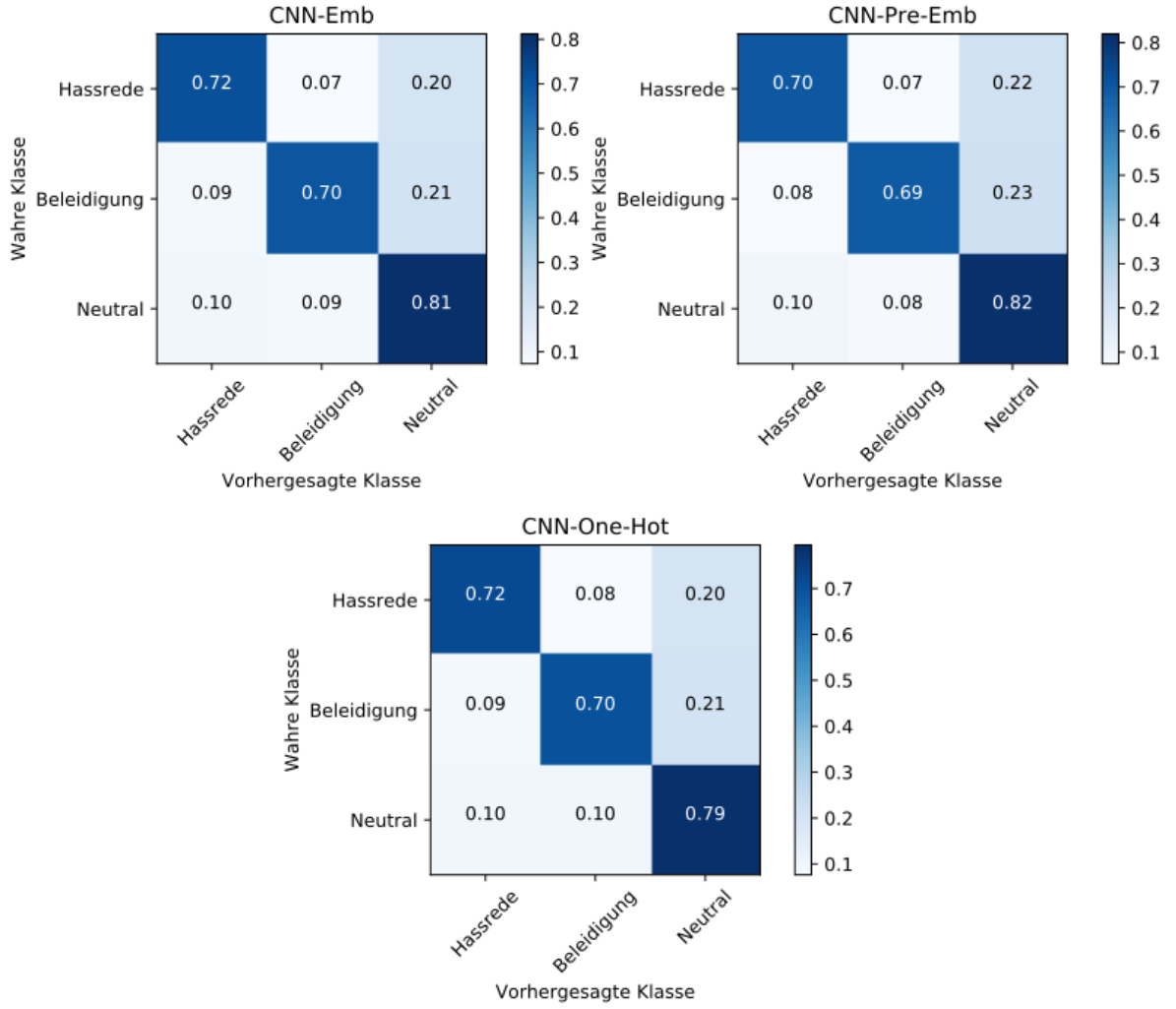

Abbildung 5.1: Gemittelte Konfusionsmatrizen von CNN-Emb (oben links), CNN-Pre-Emb (oben rechts) und CNN-One-Hot (unten) aus jeweils 20 Durchläufen.

Die Fehlerverteilungen ähneln sich ebenfalls. Bei allen drei Varianten sind vor allem Verwechslungen mit der Klasse „Neutral“ die Ursache für die schlechtere Performanz gegenüber den wortbasierten Modellen. Die Klasse „Neutral“ wird demnach von allen Varianten am besten wiedererkannt. Warum die vortrainierten Embeddings möglicherweise nicht zu verbesserten Ergebnissen führen konnten, wird in Kapitel 7 diskutiert. Sich für eines der Modelle zu entscheiden, ist in diesem Fall schwierig, da die Ergebnisse trotz der 20 Durchläufe kleinen Schwankungen unterliegen können. CNN-Emb wurde schließlich

ausgewählt, gerade da die vortrainierten Embeddings in CNN-Pre-Emb interessanterweise eine erhoffte Verbesserung nicht lieferten.

5.2.3 Gemischter Ansatz

Ein letzter Ansatz wurde aus der Vermischung der beiden besten Modellen (LSTM-ngram und CNN-Pre-Emb) konstruiert, um herauszufinden, ob eine Kombination der beiden Repräsentationen zu einer Verbesserung der Performanz führt. Wie in Kapitel 4.3 beschrieben gibt es mehrere Möglichkeiten, verschiedene Repräsentationen gleichzeitig von einem Modell verarbeiten zu können. Um die Modelle zu vereinen, werden zunächst Wörter und Buchstaben separat verarbeitet. Die jeweiligen, einzelnen Modelle verarbeiten die Embeddings dieser. Die Ausgabevektoren beider letzten Layer werden dann konkateniert, nach welchem ein Dropout von 0,5 angewendet wird. Danach folgt der Ausgabelayer. Auch dieses Modell (LSTM-CNN-Mixed) wurde für 20 Durchläufe je 15 Epochen trainiert und die Ergebnisse gemittelt. Der gemittelte F1 Score von LSTM-CNN-Mix liegt bei 0,9100 und zeigt damit ein ähnlich gutes Ergebnis wie LSTM-ngram. Durch die höhere Effizienz von LSTM-ngram allerdings ist dieses allein wortbasierte Modell damit (abgesehen von Fine-tuned BERT) das Leistungsfähigste der untersuchten Modelle.

5.3 Zusammenfassung der Ergebnisse

Da sich LSTM-ngram als bestes Modell erwiesen hat, wurde als letzter Schritt ein Hyperparametertraining mit Hyperas[15] vorgenommen. Hiermit wurden Parameter, die vorerst über die Ergebnisse vorangehendender Arbeiten bzw. durch strukturiertes Ausprobieren bestimmt wurden, noch einmal systematisch optimiert. Dies betrifft den Dropout, die Anzahl an LSTM-Layern (eine oder zwei) sowie die Größe dieser, und die Batchgröße. Aus 70 Durchläufen je 15 Epochen wurde wieder das leistungsfähigste Modell bestimmt. Es hat sich ergeben, dass statt Einem zwei LSTM Layer eingesetzt werden sollten. Der Dropout wird auf 0,44 festgelegt und die Batchgröße auf 32. Der beste F1 Score dieses Modells auf demselben Testest liegt bei 0,917. Bei 20 Durchläufen wird ebenfalls ein durchschnittlicher F1 Score von

[15] https://github.com/maxpumperla/hyperas (Stand 12.04.2019).

0,917 erzielt. Im weiteren Verlauf der Arbeit steht LSTM-ngram für diese angepassten Parameter.

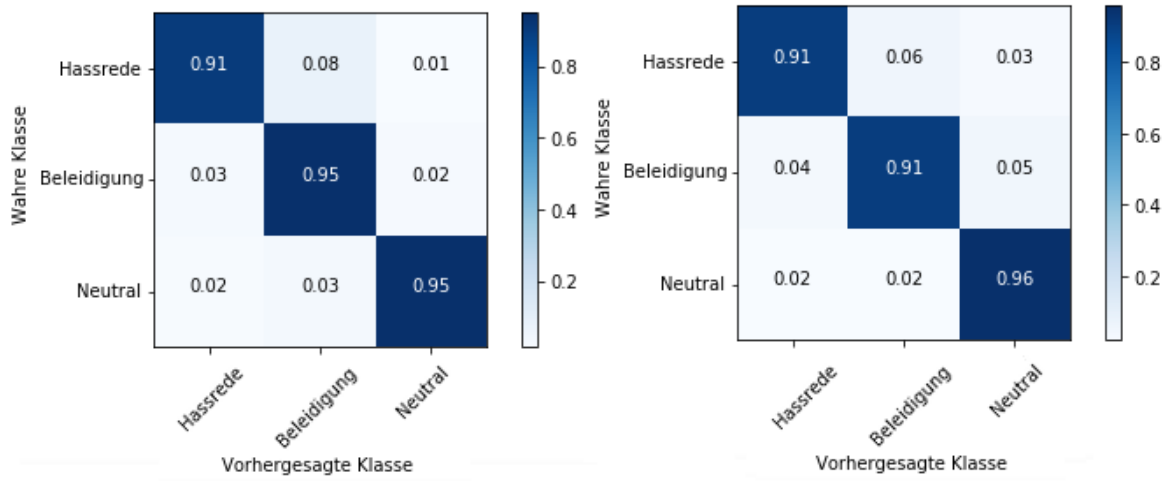

Abbildung 5.2: Konfusionsmatrizen von Fine-tuned BERT (links) und LSTM-ngram (rechts).

Die Ergebnisse von LSTM-ngram konnten die Bisherigen aus vorangegangenen Studien übertreffen. Davidson et. al. (2017) erreichten mit etwa 25.000 Tweets einen F1 Score von 0,91, wobei allerdings nur eine Genauigkeit von 61% bei der Klasse „Hassrede“ erreicht werden konnte. Mit mehr Daten für die unterrepräsentierten Klassen sowie durch insgesamt gleichverteilte Klassen in den Datensets wurde derselbe F1 Score erreicht, allerdings mit deutlich besseren Ergebnissen für die Klasse „Hassrede“ (vgl. Abbildung 5.2). Dies wurde zudem mit insgesamt weniger Daten erreicht (18.000 Tweets). Die Klassen „Hassrede“ und „Beleidigung“ wurden außerdem nicht, wie in den Arbeiten von Zhang et. al. (2018) und Robinson et. al. (2018), zusammengelegt und die Tweets wurden nach Duplikaten durchsucht, sodass der F1 Score nicht unter anderem durch Wiedererkennung gleicher Tweets, so wie in der Arbeit von Gaydhani et. al. (2018), zustande gekommen ist.

Zudem hat sich gezeigt, dass Fine-tuned BERT eine deutlich bessere Performanz als alle anderen getesteten Modelle aufweist. Dies beweist, wie leistungsfähig ein vortrainiertes Sprachmodell für alle NLP Aufgabenstellungen sein kann.

6. Evaluation

In diesem Kapitel sollen die beiden besten Modelle, LSTM-ngram und Fine-tuned BERT genauer ausgewertet und verglichen werden. Da BERT ein Sprachmodell ist und sich damit

prinzipiell von allen anderen in dieser Arbeit erstellten Modellen abhebt, sind deutliche Unterschiede zu Modell LSTM-ngram möglich. Es geht hierbei einerseits um die Identifikation möglicher Schwachstellen der Modelle und andererseits um die Art und Weise, wie die beiden Modelle die Testdaten unterschiedlich klassifizieren könnten. Für beide Modelle wurden zudem die Vorhersagegenauigkeiten für alle Themen und Klassen erfasst, um abschätzen zu können, wie sicher die Modelle bei der Klassifikation sind, um Willkürlichkeit auszuschließen. Eine Vorhersagegenauigkeit beispielsweise für die Klasse Hassrede für das Thema Sexismus von 0,854 bedeutet, dass im Durchschnitt diese bereits klassifizierten Aussagen mit einer Wahrscheinlichkeit von 85,4% der Klasse zugeordnet werden. Willkürlichkeit würde sich nahe einer Vorhersagegenauigkeit von 0,33 ausdrücken. Diese Tabellen können im Anhang (vgl. A3) eingesehen werden. Zunächst soll allerdings erläutert werden, warum es Schwachstellen bei KI Modellen geben kann.

6.1 Bias

Beim Trainieren einer künstlichen Intelligenz mit Deep Learning, oder aber auch anderen maschinellen Lernverfahren, wird die KI häufig Vorurteile erlernen, die den Trainingsdaten entstammen. Dies ist beispielweise bei einem von Amazon entwickelten Werkzeug zur automatischen Identifikation der besten Bewerber aufgrund deren Lebensläufe geschehen (vgl. Dastin, 2018). Das Modell wurde auf Lebensläufen trainiert, die über einen Zeitraum von zehn Jahren bei der Firma eingingen. Da diese hauptsächlich von Männern stammten, hat das Modell Männer bevorzugt. Selbiges passiert, wenn die Information, dass in der Vergangenheit hauptsächlich Männer in die heute typischen, Männer-dominierten Berufe eingestellt wurden, in den Daten enthalten ist (vgl. Cairns 2019). In beiden Fällen bekommt das Modell insgesamt weniger Beispiele für erfolgreiche Bewerbungen von Frauen als von Männern vorgelegt. Auch ein Blogpost von Speer (2017) zeigt, wie sich Vorurteile in künstlicher Intelligenz widerspiegeln können. Bei der in dem Blogpost vorgestellten Sentiment Klassifizierung hat sich zum Beispiel herausgestellt, dass stereotypisch weiße Vornamen positiver bewertet werden als stereotypisch schwarze Vornamen.

Oftmals sind diese, sich in der KI widerspiegelnden Vorurteile verbunden mit Sexismus und Rassismus. Im Falle dieser Arbeit sind beide Konzepte unter anderem gerade das, was identifiziert werden soll. Aber auch hierbei können sich Vorurteile in den Daten auftun. Denn innerhalb der einzelnen Hassrede-Kategorien ist verstärkt nur eine Seite betroffen. Im Falle von

Sexismus sind es hauptsächlich Frauen; im Falle von Rassismus sind es hauptsächlich Minderheiten, denen Hass wiederfährt. Diese Art von Hassrede wird allerdings von dem größten Teil der Gesellschaft bestraft. Im Gegensatz dazu wird beispielsweise Hass gegen Weiße und Hass gegen Männer, wenn vielleicht auch weniger vorhanden, weniger diskutiert und somit auch eher toleriert. Diese Denkweise spiegelt sich beispielsweise in Artikeln wie „The Year in Male Tears. 2015 was the year misandry went mainstream, and that's a good thing." (Summers 2015) wieder. Dasselbe kann beim Thema Rassismus beobachtet werden. Es besteht die Ansicht, dass es weniger schlimm ist Weiße zu stereotypisieren als Schwarze und andere Minderheiten (vgl. The Economist 2018). Es kann zurecht argumentiert werden, dass dieser „umgekehrte Rassismus" weniger Schaden anrichten kann als konventioneller Rassismus. Trotzdem ist der Ausdruck von Hass gegenüber Menschen aufgrund ihrer Hautfarbe nie gerechtfertigt, auch wenn in diesem Fall keine systemische Unterdrückung im Laufe der Geschichte stattfand (vgl. ebd.). Obwohl diese Denkweisen in keiner Definition von Hassrede erwähnt werden, prägen sie doch die Sicht der Gesellschaft auf das Thema Hassrede maßgeblich.

Im Folgenden soll also geprüft werden, ob sich diese Vorurteile anhand der Modelle bestätigen lassen. Dafür wurde der Twitter Stream mit Schlüsselwörtern nach Hassrede gefiltert, welche Misandrie oder Rassismus gegen Weiße enthält. Auch wurden einige Kommentare auf einer Internetseite manuell herausgesucht[16]. Die Beurteilung, ob diese ausgesuchten Aussagen Hassrede sind, wurde lediglich durch die eigene Person getroffen. Sicherlich wäre diskutierbar, ob nicht einige der Tweets eher in die Kategorie Beleidigung fallen. Da es bei der Beurteilung allerdings hauptsächlich um den Unterschied zwischen hasserfüllten und neutralen Aussagen geht, wurde kein weiterer Aufwand betrieben, um möglicherweise geeignetere Aussagen auszuwählen. Um eine größere Datengrundlage zu besitzen wurden die Tweets der beiden Fälle verdoppelt. Für die Klasse Rassismus konnten so 56 und für die Klasse Sexismus 84 Aussagen gesammelt werden.

Zunächst sollen die Ergebnisse von LSTM-ngram betrachtet werden. Von 56 rassistischen Aussagen werden 20 als „Hassrede" und 32 als „Neutral" klassifiziert. Nur 4 fallen unter „Beleidigung". Es zeigt sich damit eine deutliche, überraschende Abgrenzung zwischen den beiden Extremen. Auf den ersten Blick fällt auf, dass die Tweets, welche als Hassrede eingestuft werden, erheblich mehr Schimpfwörter enthalten. Das Wort *f*ck* kommt beispielsweise in den

[16] http://misandry.tripod.com/id6.html (Stand 13.05.2019).

neutralen Aussagen nur zwei Mal vor, während es bei den hasserfüllten Aussagen 20 Vorkommnisse gibt. Das Hashtag *f*ckwhitepeople* scheint hierbei außerdem ein Kriterium für Hassrede zu sein. Ansonsten finden sich keine weiteren, offensichtlichen Zusammenhänge zwischen den Klassen und Aussagen.

Diese Abgrenzung zwischen den Extremen lässt sich auch bei der Klassifikation durch Fine-tuned BERT feststellen. Dabei gelten 22 Aussagen als „Hassrede" und 34 als „Neutral"; keine der Aussagen wird als „Beleidigung" klassifiziert. Auch wenn die Zahlen ähnlich sind gibt es doch Unterschiede zu der Klassifikation von LSTM-ngram. Denn es sind nur etwa 50% der von LSTM-ngram als „Hassrede" klassifizierten Aussagen in der Klassifikation durch BERT enthalten und umgekehrt. Dasselbe gilt also auch für die Klasse „Neutral". Sichtbar sind bei diesem Unterschied vor allem die durch Schimpfwörter gekennzeichneten Aussagen, von welchen nun einige durch BERT als „Neutral" gelten. Umgekehrt lassen sich einige weniger vulgäre, aber dennoch hasserfüllte Aussagen wie beispielsweise *white lives don't matter* bei der Klassifikation durch BERT in der Klasse „Hassrede" finden.

Bei den sexistischen Aussagen zeigt sich eine etwas andere Klassenverteilung. Von insgesamt 84 Aussagen werden 40 als neutral eingestuft, während 38 Aussagen in die Kategorie „Beleidigung" fallen und wiederum 6 als Hassrede gelten. Die Abgrenzung zwischen den Extremen, wie sie bei der Klassifikation der rassistischen Aussagen zu finden ist, gibt es bei diesem Thema damit nicht. Stattdessen sind die Aussagen ähnlich auf die Klassen „Neutral" und „Beleidigung" verteilt; mit einem leichten Überschuss der neutralen Klasse. Qualitativ gesehen erscheint die Klassifikation relativ willkürlich, ähnlich wie bei den rassistischen Aussagen auch. Einige der brutalsten Hassbekundungen finden sich in der neutralen Klasse wieder, so wie zum Beispiel *all men are good for is f*cking, and running over with a truck*. Im Gegensatz dazu sind die Aussagen der Klasse „Hassrede" eher subtil und enthalten zudem weniger vulgäre Ausdrücke.

Die Fine-tuned BERT Klassifikation der sexistischen Aussagen unterscheidet sich stark von der LSTM-ngram Klassifikation. Denn hier werden 80 Aussagen als „neutral" eingestuft; jeweils 2 Aussagen finden sich in den beiden anderen Klassen. Die Tendenz, dass die Klasse „Hassrede" eher wenige Aussagen zugeordnet bekommt, bleibt somit allerdings bestehen.

Insgesamt zeigt sich, dass sich der Bias in der Klassifikation von beiden Modellen widerspiegelt, auch wenn die Klassen „Hassrede" und „Beleidigung" für die Ergebnisse zusammengefasst werden würden. Denn trotzdem werden etwa 50% aller Aussagen (hier für

LSTM-ngram) fälschlicherweise als „Neutral“ klassifiziert; darunter sogar einige der Aggressivsten. Diese Tatsache lässt sich auf unterschiedliche Arten erklären. Oftmals enthält Hassrede Verunglimpfungen, welche sich auf die Opfer beziehen. Für Frauen sowie auch Minderheiten gibt es davon viele. Für Männer und Weiße allerdings existieren weitaus weniger solche Begriffe (zumindest dem englischsprachigen Raum entstammend). Diese Verunglimpfungen stellen vermutlich einen großen Einfluss auf die Klassifikation dar. Es sind wenig Schlüsselwörter denkbar, die alleinstehend einen Angriff explizit auf Weiße oder Männer symbolisieren. Die Abwesenheit einer Verunglimpfung ist allerdings nicht gleichzusetzen mit der Abwesenheit von Hass. Ohne solche Begriffe und stattdessen mit subtileren Angriffen ist es möglicherweise für das Modell schwerer, Hassrede zu identifizieren.

Zusätzlich wurde das Korpus durch das Filtern durch Schlüsselwörter, so auch durch typische Verunglimpfungen, zusammengestellt. Dass weniger Verunglimpfungen für Weiße oder für Männer existieren bedeutet damit auch, dass es schwieriger ist, durch jene Methode solche Instanzen von Hassrede herauszufiltern. Somit haben diese Arten von Hassrede insgesamt weniger Wiedererkennungswert bei der Klassifikation, da vermutlich weniger Trainingsdaten dafür vorhanden sind. Diese Hypothese könnte zusätzlich qualitativ überprüft werden, was allerdings den Rahmen der Arbeit übersteigen würde.

Die herausgesuchten Beispiele, vor allem für Misandrie, entstammen damit hauptsächlich impliziter Hassrede. Dies könnte bedeuten, dass die Klassifikation nicht wegen des Themas, sondern wegen der impliziten Natur der Aussagen so ausgefallen ist. Allerdings wurden im vorliegenden Fall gerade die explizitesten Aussagen als „Neutral“ klassifiziert, was gegen diese Vermutung spricht. An dieser Stelle wäre es auch interessant zu prüfen, ob es unter Umständen einen Zusammenhang zwischen impliziter Hassrede und weniger stark repräsentierten Hassrede-Themen wie Misandrie gibt, gerade weil hier keine Verunglimpfungen existieren. Weitere Untersuchungen zu impliziter Hassrede sind an dieser Stelle notwendig, um die bisherigen Ergebnisse zu vervollständigen. Weiter diskutiert wird jener Ansatz in Kapitel 7.

6.2 Schwarzer Humor

Eine weitere Schwachstelle des Modells könnte schwarzer Humor sein. In der Definition von Hassrede auf Facebook wird Humor, insbesondere Parodien, explizit ausgeschlossen (vgl. Kapitel 2). Hier ist allerdings nicht geklärt, ob gerade schwarzer Humor ein Teil von Hassrede und somit inakzeptabel sein kann oder nicht, was diskutierbar ist. Zugunsten der folgenden

Argumentation und Analyse wird hier schwarzer Humor als „beleidigend“ eingestuft (wie im Englischen – „offensive jokes“) und somit von Hassrede abgegrenzt. In diesem Abschnitt wird also der umgekehrte Fall zu dem vorherigen Abschnitt erörtert: Aussagen, welche nicht Hassrede sind, aber möglicherweise als solche klassifiziert werden.

Unter schwarzen Humor fallen all jene Witze, welche in ihrer Natur in irgendeiner Form grotesk, makaber, pervers, absurd oder ähnliches sind (vgl. O’Neill 1983, 145). Diese Form von Humor reicht weit bis in die Vergangenheit zurück (vgl. ebd., 152) und wird daher einerseits als Teil der menschlichen Natur betrachtet (vgl. ebd., 145). Laut O’Neill (1983) entstammt schwarzer Humor aus der Verbindung zwischen Pessimismus und Optimismus; aus der Vergeblichkeit der eigenen Handlungen, um ein Ideal und die Realität zusammenzuführen. In dem Sinn funktioniert schwarzer Humor als ein verzweifelter Versuch, Kritik am menschlichen Miteinander auszuüben (155). Andererseits kann ebenso argumentiert werden, dass durch schwarzen Humor beispielsweise negative Stereotypen nur weiter verstärkt werden.

Schwarzer Humor funktioniert ähnlich wie Hassrede in dem Sinne, als dass verletzliche Personengruppen (Homosexuelle, Minderheiten etc.) thematisiert werden. Laut Bicknell (2007, 463) ist die Verletzlichkeit einer Personengruppe essenziell, wenn es darum geht, zu entscheiden, ob oder wie stark ein Witz moralisch verwerflich ist, da schwarzer Humor diese Verletzlichkeit versucht auszunutzen. Dies impliziert, dass ein Witz weniger beleidigend ist, wenn beispielsweise Weiße statt Schwarze Thema des Witzes sind. Es lässt sich damit sagen, dass der im vorherigen Abschnitt beschriebene Bias hier besonders zum Ausdruck kommt. Würde sich dieses Gefälle an Verletzlichkeit einer Personengruppe umkehren, so würde sich auch die Verbreitung von schwarzem Humor über diese Gruppe je nachdem ändern (vgl. ebd., 463). Unter diese Definition von schwarzem Humor fallen allerdings auch weitere Personengruppen, welche grundsätzlich nicht wie die zuvor Genannten mit Hassrede verbunden werden. Das sind zum Beispiel Kinder oder Opfer von Katastrophen.

Ähnlich wie im vorherigen Kapitel soll überprüft werden, wie schwarzer Humor von den Modellen klassifiziert wird. Dafür wurden aus einem Reddit-Thread[17] 66 Beispiele für schwarzen Humor herausgesucht. Wie schon erwähnt gelten alle Witze für diese Analyse als „beleidigend“. Durch LSTM-ngram werden 36 als „Neutral“ klassifiziert, außerdem 12 als „Hassrede“ und 18 als „Beleidigung“. Hierbei fällt auf, dass vor allem die Witze, in welchen

[17] https://www.reddit.com/r/AskReddit/comments/hjyzr/hey_reddit_what_is_the_most_jaw_droppingly/ (Stand 19.04.2019).

Gewalt zur Sprache kommt, als Hassrede eingestuft werden. Witze, in denen nur von negativen Stereotypen Gebrauch gemacht wird, werden eher der Klasse „Neutral“ zugeordnet; so wie zum Beispiel *what did the Mexican kid get for Christmas? ... My bike*. Die „neutralen“ Witze enthalten ebenfalls seltener Schlüsselwörter wie beispielsweise *black* oder *jew* und sind insgesamt subtiler. Im Gegensatz dazu kommt in allen Aussagen der Klasse „Beleidigung“ das Wort *black* vor. Die Ergebnisse dieser Klassifikation des Modells sind damit zumindest nachvollziehbar.

Mit der Klassifikation durch Fine-tuned BERT wird die Mehrheit der Witze (59) als „Neutral“ eingestuft. Weitere 5 gelten als „Beleidigung“ und 2 als „Hassrede“. Dabei ist allerdings nicht ersichtlich, warum gerade diese 7 nicht auch als „Neutral“ klassifiziert worden sind. Sie unterscheiden sich weder in der dargestellten Brutalität noch in den verwendeten Ausdrücken von den „neutralen“ Gegenstücken.

Es zeigt sich, dass in diesem Fall hauptsächlich die Klassifikation durch LSTM-ngram nicht ideal verläuft. Die Klassifikation von Fine-tuned BERT entspricht dagegen sogar fast dem Anspruch; die Fehlerwahrscheinlichkeit von etwa 7% einbezogen. In diesem Fall ist Fine-tuned BERT LSTM-ngram deutlich vorzuziehen, da hier augenscheinlich der Humor aufgenommen wird. Insgesamt verläuft die Klassifikation für schwarzen Humor präziser als für Sexismus und Rassismus. Da es einen deutlichen Unterschied zwischen den Ergebnissen der Modelle gibt, ist auszuschließen, dass die fehleranfällige Klassifikation des schwarzen Humors bei LSTM-ngram nur an mangelnden Trainingsdaten oder ungünstig ausgewählten Testdaten liegt.

7. Problemdiskussion und Möglichkeiten zur Verbesserung

7.1 Korpuserstellung

Bei der Durchführung dieses Projekts sind einige Dinge aufgefallen, die die Arbeit noch erweitern beziehungsweise aufwerten könnten, welche allerdings aus zeitlichen Gründen nicht möglich waren oder außerhalb des vorher gesteckten Rahmens liegen. Dies betrifft zunächst die Erstellung des Korpus. Denn ein Modell kann grundsätzlich nur so gut klassifizieren, wie es die Daten, aus denen es lernen soll, erlauben.

In dieser Arbeit wurde ein bereits vorgefertigtes Korpus verwendet, welches durch Teile weiterer Korpora ergänzt werden musste. Die Tweets wurden hauptsächlich durch

Schlagwörter aus Twitter gefiltert. In Kapitel 3 hat sich gezeigt, dass die Auswahl an Schlagwörtern, durch welche die Tweets herausgefiltert werden, einen sehr großen Einfluss auf das Korpus nehmen. Ein Wort wie zum Beispiel *yankees* führt durch dessen Zweideutigkeit zu ungewollt überrepräsentierten Themen in den Daten. Andersherum kann es ebenso sein, dass implizite Hassrede (vgl. Kapitel 2) nicht erfasst wurde, weil das entsprechende Schlagwort nicht zum Filtern benutzt wurde. Möglicherweise gibt es auch Hassrede, welche nicht durch so ein Schlagwort zusammengefasst werden kann und somit durch diese Methode gar nicht erfassbar ist. Angenommen eine Minderheit wird als „Tiere“ oder „Parasiten“ bezeichnet. Beides sind grundsätzlich keine verletzenden Schlüsselwörter, lediglich der verwendete Kontext gibt Aufschluss über die Absicht des Verfassers (vgl. Saleem et. al. 2017).

Der Schlüsselwort-basierte Ansatz ist, wie bereits angedeutet, auch anfällig für Tweets, die zwar erfasst werden, aber nicht zur Kategorie Hassrede gehören. Im Fall von Davidson führte dies nur zu einem relativ großen, neutralen Anteil von Tweets mit dem Schlüsselwort *yankees*. Im Fall von ElSherief wurde mit diesem Problem anders umgegangen. Hier wurden Schlüsselwörter, die aus Erfahrung sehr kontextsensibel sind, einfach nicht zum Filtern verwendet (vgl. ElSherief et. al. 2018). Um mehr Kategorien von Hassrede abdecken zu können, wäre es besser gewesen, diese kontextsensiblen Schlüsselwörter beizubehalten und zusätzlich die gefilterten Tweets intensiver zu kontrollieren.

So wie einige Annotationen durch Crowd-Sourcing entstanden sind, so hätten auch die Tweets zum Beispiel durch Einsendungen gesammelt werden können. Es gilt: umso mehr menschliche Kontrolle über die Auswahl der Daten, desto genauer können die Tweets die Klassen auch abbilden. Natürlich muss dabei der Aufwand abgewogen werden. In diesem Fall wurden einige Mängel am Korpus durch die Vorverarbeitung behoben. Für die Erstellung eines neuen Korpus würde sich anbieten, zunächst die Tweets durch eine besonders große Schlagwortauswahl zu filtern (oder auf Crowd-Sourcing zurückzugreifen) und dann manuell Tweets mit sehr spezifischen Kontexten auszusortieren, in welchen Schlagwörter überdurchschnittlich oft vertreten sind. So könnte ein großer Teil von Hassrede abgebildet werden, ohne dass zufällige Zuordnungen vom Modell erlernt werden. Zusätzlich würden dann vermutlich extremere Hasskommentare in den Daten auftauchen, da durch die gezwungene Auflösung der IDs in die Tweets einige Daten verloren gegangen sind (vgl. Kapitel 3).

Ein weiteres Kriterium ist die Art und Weise, die Tweets zu annotieren. Besonders bei einem sehr sensiblen Thema wie Hassrede ist es wichtig, dass die Tweets nicht nur von einer Person

annotiert werden. Sonst kann es sein, dass sich ein starkes Bias in den Daten niederschlägt. In diesem Fall wurden alle Tweets von mehreren Personen annotiert oder zumindest überprüft. Aber auch bei einer Gruppe von Personen besteht die Gefahr, dass sich eine gewisse Subjektivität abzeichnet. Es hat sich gezeigt, dass Menschen grundsätzlich rassistische Aussagen eher als Hassrede klassifizieren als sexistische Aussagen (vgl. Waseem et. al., 2017). Zusätzlich kommt es auch darauf an, ob es sich bei diesen Personen um Experten handelt. Waseem (2016) hat gezeigt, dass es einen wesentlichen Unterschied macht, ob die Daten von Experten oder Amateuren annotiert werden, da Annotationen von Experten grundsätzlich zu genaueren Klassifikationen führen.

7.2 Vorverarbeitung und Embeddings

Ein weiteres Problemfeld ist grundsätzlich das Auswählen der Vorverarbeitungsschritte. Gerade bei nutzergenerierten Inhalten sollten die Daten normalerweise soweit wie möglich bereinigt und vereinheitlicht werden. Dazu kann zum Beispiel eine Rechtschreibkorrektur eingesetzt werden (vgl. Kapitel 4). Allerdings hat sich bereits in Kapitel 3 gezeigt, dass unterschiedliche Schreibweisen von Wörtern in verschiedene Klassen stärker vertreten sind. Das bedeutet, dass Varianten von Wörtern vom Menschen unterschiedlich wahrgenommen werden können. Hierbei stellt sich die grundsätzliche Frage, ob das Phänomen, dass ein Wort in unterschiedlichen Schreibweisen abweichende Bedeutungen (in diesem Fall unterschiedliche Stufen von Aggressivität) aufweist, auch von einem Modell abgebildet werden sollte. Die Varianten zu vereinheitlichen würde bedeuten, dass die von der Korrektur betroffen Tweets möglicherweise falsch klassifiziert werden würden – es wäre allerdings auch vergleichsweise robust und durchschaubar. Sie nicht zu vereinheitlichen, so wie in dieser Arbeit, bedeutet eine präzisere Klassifikation; es setzt allerdings auch eine variantenreichere Datengrundlage für das Training voraus und das Modell ist insgesamt anfälliger für Ausreißer.

Angenommen dieses Modell (welches verschiedene Varianten nicht vereinheitlicht) würde als Filter für eine Internetseite eingesetzt werden und die Benutzer würden so einen Filter umgehen wollen. Dann kann es sein, dass nur weil eine rassistische Beleidigung absichtlich deutlich anders geschrieben worden ist (und diese in den Trainingsdaten nicht vorhanden war) das Modell diese nicht mehr als Feature für die Klasse „Hassrede“ erkennen würde, selbst wenn es für einen Menschen klar wäre (vgl. Warner und Hirschberg 2012, 21). Darauf müsste das Modell nachträglich angepasst werden. Beide Möglichkeiten haben somit ihre Vor- und

Nachteile. Allerdings sollte, wenn möglich, die zweite Variante vorgezogen werden, da sie die Wirklichkeit besser widerspiegelt – auch wenn die Pflege jenes Modells einen größeren Aufwand erfordern würde Im Falle dieser Arbeit kann gesagt werden, dass die FastText Embeddings einen Teil dieses Aufwands übernehmen. Sie können den Text in Wortform am besten kodieren, da unterschiedliche Varianten von Wörtern zwar vergleichsweise ähnliche, aber dennoch unterschiedliche Vektoren zugeordnet bekommen.

Ein weiterer, diskutierbarer Punkt, welcher den Vorverarbeitungsschritten zugehörig ist, ist der Ausschluss von Zahlen und Zeichen aus dem Alphabet beziehungsweise das Entfernen dieser aus den Wörtern. Dies wurde vorgenommen, damit die Embeddings besser den Wörtern zugeordnet werden können, da bestimmte Zeichen und Zahlen grundsätzlich nicht zum Sinn eines Textes beitragen (so wie zum Beispiel *+#cat#+*). Bei nutzergenerierten Inhalten kann, wie bereits angesprochen, eben nicht angenommen werden, dass die Texte immer den Standards angepasst sind. Da der Ausschluss der Zeichen automatisch vorgenommen wurde, kann es sein, dass Wörter, in denen zum Beispiel eine Zahl einen Buchstaben innerhalb des Worts ersetzt hat, der Sinn des Worts dann nicht mehr durch die Embeddings korrekt abgebildet werden kann. Dies könnte für Zeichenketten wie beispielsweise *a$$ho1e* gelten. Auf der einen Seite kann das Löschen der überflüssigen Zeichen also eine einfache Bereinigung des Textes darstellen, aber in manchen Fällen auch Probleme bereiten, wenn diese Zeichen andere ersetzen. Es kann gesagt werden, dass das Buchstaben-basierte Modell auf solche Fälle vermutlich besser angepasst ist. Dieser Hypothese bedarf es damit weiterer Analyse.

Ein ähnliches Problem betrifft das Entfernen der URLs sowie der Benutzernamen. Wie auch schon in Arbeiten aus Kapitel 4 demonstriert sind die URLs durch eliminieren der am wenigsten vorkommenden Wörter entfernt worden, da sie als Störsignal aufgefasst werden. Ähnlich wie im vorherigen Abschnitt beschrieben ist allerdings vorstellbar, dass zum Beispiel rassistische Ausdrücke durch die URL-Schreibweise getarnt und somit nicht mehr als Hassrede erkannt werden. Deswegen ist denkbar, URLs in der konkatenierten Form in den Tweets beizubehalten. Dasselbe ist auch für die Benutzernamen denkbar. Benutzernamen wurden ebenfalls trotzdem entfernt, da die Erkennung von Hassrede in Benutzernamen vermutlich eher ein eigenes Problem ist.

Der Grund für das Entfernen der Benutzernamen bestand darin, dass diese nicht versehentlich vom Modell als eine Eigenschaft einer Klasse verstanden werden. Dies gilt natürlich auch für alle anderen Namen von Personen, die in Tweets vorkommen. Besser wäre es also, alle

Vornamen und Nachnamen im Rahmen der Vorverarbeitung aus den Tweets zu entfernen, was allerdings einen größeren manuellen Aufwand erfordern würde.

In dieser Arbeit hat sich gezeigt, dass die vortrainierten Embeddings für Buchstaben, obwohl sie auf einem Korpus, welches drei Millionen Tweets umfasst, vortrainiert wurden, nicht die Embeddings übertreffen, die auf nur 12.000 Tweets trainiert wurden. Dies kann daran liegen, dass das Korpus bereits zu speziell auf ein Thema zugeschnitten ist. Es ist möglich, dass Buchstabenembeddings, welche auf einem großen, allgemeinen Twitter Korpus vortrainiert werden, besser geeignet sind.

Zusätzlich ist erneut anzumerken, dass die Embeddings, die aus BERT extrahiert wurden, nicht tauglich für diese Arbeit waren. Hierbei ist aufgefallen, dass vor allem anstößige Wörter beispielsweise *r*tard* oder *f*ggot* überhaupt nicht im Vokabular von BERT enthalten sind. Ohne die Tokenisierung durch BERT werden diese nicht erkannt. Nur durch die Aufteilung der Begriffe in N-Gramme ist es BERT möglich, sie zu verarbeiten. Somit kann es sein, dass die Aussagekraft dieser Begriffe zum Teil bereits verloren geht.

Weiterhin kann es auch an der Abhängigkeit der Embeddings vom Kontext liegen, warum die Klassifizierung fehlschlägt. Um die von BERT generierten Embeddings besser zu verstehen, wurden die Embeddings einiger Wörter durch eine PCA visualisiert (vgl. Anhang A4). Mehrere, normalerweise voneinander getrennte Wortgruppen wurden dafür herangezogen: Zahlen, Monate, Farben, Wörter bezogen auf die Landschaft sowie positive und negative Ausdrücke, die auch ohne die Tokenisierung durch BERT erkannt werden. Für eine erste Visualisierung wurden alle Wörter in einen Kontext gebracht. Eines dieser Wörter hat somit als Kontext alle verbleibenden Wörter. Dabei hat sich gezeigt, dass nur die Monate klar von allen anderen Wörtern unterschieden werden konnten. Als nächstes wurde jedem der Wörter ein kurzer, sinnvoller Kontext gegeben. Wenn nun die Embeddings dieser Wörter extrahiert wurden, so treten die Gruppierungen hervor. Die Gruppen der positiven und negativen Wörter allerdings sind weiterhin relativ dicht beieinander und überlappen. Dies kann ebenso ein Indikator dafür sein, warum die Klassifikation fehlschlägt. Da die Kontexte von positiven und negativen Wörtern grundsätzlich sehr ähnlich sein können und die BERT Embeddings sich stark am Kontext orientieren ist es möglich, dass die Unterscheidung von positiven und negativen Aussagen nicht stattfinden kann. Die Klassifikation des Satzes *what a lucky … you are* ist zum Beispiel gänzlich abhängig von dem hier ausgelassenen Wort.

Gerade da die BERT Embeddings sich für dieses Projekt nicht als hilfreich erwiesen haben, wäre es sinnvoll, die Embeddings, die aus dem Sprachmodell ELMO (Peters et. al. 2018) gewonnen werden können, ebenfalls zu testen. ELMO Embeddings sind wie BERT Embeddings auch kontextualisiert. Die Ergebnisse mit ELMO Embeddings könnten dann zum Vergleich herangezogen werden und vielleicht auch bestätigen, ob das Problem bei kontextualisierten Embeddings im Allgemeinen liegt.

7.3 Mögliche Erweiterungen des Projekts

Für die Repräsentation der Tweets für das Modell wurden Wörter und Buchstaben beziehungsweise auch N-Gramme gewählt. Eine weitere grundsätzliche Möglichkeit, einen Text für eine Maschine abzubilden, ist die Auftrennung eines Texts in Sätze, welche wiederum durch Satzembeddings (zum Beispiel mit InferSent[18]) repräsentiert werden könnten. So ist es auch möglich, die Bedeutung aller Wörter in einem Satz zu erfassen und dementsprechend die Wörter zu gewichten. Für Tweets ist die Aufteilung in Sätze allerdings etwas schwieriger zu definieren, denn oftmals besteht ein Tweet auch nur genau aus einem Statement. Falls nicht, ist es trotzdem oftmals schwer, Satzgrenzen zu erkennen, weil häufig die Satzzeichen fehlen. Trotzdem wäre dieser Ansatz sicherlich interessant für einen Vergleich mit den bisherigen Ergebnissen.

Auf Twitter wird nicht nur Text, sondern auch Bilder verbreitet. Wie auch in den Richtlinien von Twitter gegen Hassrede festgehalten gibt es Hass schürende Bilder, welche auf Twitter nicht geduldet werden. In dieser Arbeit wurden nur Tweets in Form von Text verarbeitet. Die Arbeit könnte also insofern erweitert werden, als dass zur Identifikation von Hassrede in einem Tweet sowohl der Text als auch ein möglicherweise von Nutzer beigefügtes Bild verarbeitet wird. Dazu könnte zunächst ein neues Korpus angelegt werden, in welchem ebenfalls Tweets mit oder ohne Bilder aller Klassen ausreichend vertreten sind. So könnte das neue Modell die Tweets sowie die Bilder separat verarbeiten, um einen Tweet zu klassifizieren. Eine weitere, einfachere Lösung wäre allerdings, einen Trainingsdatensatz zu erstellen, in welchem nur bereits klassifizierte Bilder ohne Text enthalten sind. So kann ein zusätzliches Modell trainiert werden, welches Hass in Bildern identifizieren kann. Durch Zusammenführen der Ergebnisse

[18] https://github.com/facebookresearch/InferSent (Stand 05.04.2019).

beider Modelle könnten dann Tweets, welche nur aus Text, nur aus einem Bild oder eben aus beidem bestehen, klassifiziert werden.

Des Weiteren wäre es interessant, das hier erstellte Modell auf Daten zu testen, welche nicht Twitter entstammen. Das könnten ebenfalls von Nutzern erstellte Inhalte sein, wie z.B. Kommentare auf Facebook oder auch Suchanfragen auf Google. Aber auch längere Texte sind denkbar. Beispielsweise könnten dies Ausschnitte aus Wahlprogrammen britischer oder amerikanischer Parteien (da das Modell bislang nur englischsprachige Texte verarbeiten kann) sein. Eine weitere Möglichkeit wären Ausschnitte aus den Reden von Politikern. Natürlich müssten diese Textteile zunächst manuell klassifiziert werden, um die Performanz des Modells auf diesen andersartigen Daten zu testen. Andersherum könnte die Leistungsfähigkeit des Modells als gegeben genommen werden, um herauszufinden, wie groß der Anteil an Beleidigungen und ob möglicherweise auch Hassrede in Textstücken der englischsprachigen Politik zu finden ist.

Der Fokus dieser Arbeit lag außerdem auf englischen Texten. Das Projekt könnte also ebenfalls hinsichtlich weiterer Sprachen ergänzt werden. Hierzu wäre es möglich, erst durch einen separaten Workflow anderssprachige Tweets maschinell zu übersetzen, um dann dieselben Modelle auf den übersetzten Daten zu testen. Somit könnte die Übertragbarkeit der Modelle auf andere Sprachen evaluiert werden. Für BERT hieße dies, dann statt dem englischsprachigen Modell das mehrsprachige Modell heranzuziehen.

Weiterhin hat sich in Kapitel 2 schon gezeigt, dass in unterschiedlichen Ländern Hassrede ganz unterschiedlich aufgefasst werden kann. Es wäre interessant zu testen, ob bzw. wie gut mit Deep Learning Modellen dieser Unterschied dargestellt werden kann. Denkbar ist zum Beispiel, einen Datensatz von Experten unterschiedlicher Länder annotieren zu lassen. Um die Unterschiede vielleicht besser sichtbar zu machen, könnten statt drei auch zehn Klassen verwendet werden, um die Skala von „Neutral“ über „Beleidigung“ bis „Hassrede“ präziser darstellen zu können. Auf den Annotationen könnte dann je Land ein Modell trainiert werden. Es würde dann mit den Annotationen der jeweils verbliebenen Länder getestet werden. Die Unterschiede zwischen den Ländern würden sich dann an einem Abfall der Performanz auf den Testdaten zeigen lassen. Eine Voraussetzung für dieses Experiment wäre allerdings, dass möglichst viele unterschiedliche Themen von Hassrede in dem Datensatz repräsentiert werden. Eine weitere Ergänzung in der Art wäre die Anwendung des Modells auf historische Daten,

denn genauso wie bei den geographischen Unterschieden auch lassen sich auch zeitliche Unterschiede bei der Definition von Hassrede finden.

Ein weiterer Punkt, welcher bei diesem Projekt nicht ausgearbeitet worden ist, sind die unterschiedlichen Varianten, Modelle zu bewerten. Hier wurde sich für eine relativ einfache Variante entschieden. Trainings- sowie Testset wurden einmalig festgelegt (Holdout) und die Modelle wurden mehrmals auf denselben Datensets trainiert sowie getestet. Der durchschnittliche F1 Score aller Durchläufe wurde als Endergebnis festgehalten. Statt die Datensets einmalig festzulegen, könnten die Daten je Durchlauf auch variieren. Die Daten hätten in gleich große Sets aufgeteilt werden können, wobei je Durchlauf alle außer einem Set zum Training und das verbliebene Set zum Testen verwendet hätte werden können (Kreuzvalidierung). So würden in allen Epochen alle Daten einmalig zum Testen verwendet werden. Der Vorteil liegt damit darin, dass das Modell insgesamt mehr Daten zum Beurteilen zur Verfügung gestellt bekommt, womit die endgültige Performanz repräsentativer für den gesamten Datensatz ist. Für dieses Projekt gab es allerdings keine Studie, welche genau dieselben Daten in dieser Kombination verwendet hat und die zum Vergleich hätte herangezogen werden können. Es wurde sich daher einfacherweise für die Holdout-Variante entschieden, da diese in diesem Fall ausreichend war. Hätte es bereits eine ähnliche Studie gegeben, welche die Kreuzvalidierung zur Bewertung auf genau diesen Daten verwendet hätte, so hätte auch für diese Arbeit diese Variante zum besseren Vergleich herangezogen werden müssen.

Ebenfalls wäre es von großem Nutzen herauszufinden, welche Eigenschaften eines Tweets tatsächlich für dessen Klassifikation verantwortlich sind. Möglicherweise gibt es bestimmte Wörter, Wortkombinationen oder Satzstellungen, die die Zuweisung zu einer Klasse durch das Modell begünstigen. Für die Klassifikation von Bildern zum Beispiel kann eine Visualisierung eines bereits klassifizierten Bildes als eine Heatmap helfen zu verstehen, welche Stellen auf dem Bild vom Modell besonders fokussiert wurden. Ähnlich könnte man auch für Text vorgehen. Ein Balkendiagramm ist denkbar, auf welchem die Features (Wörter, Buchstaben, N-Gramme etc.) auf der x-Achse und die Relevanz des Features auf der y-Achse aufgetragen sind.

Dem anknüpfend könnte noch eine weitere Ergänzung zu Kapitel 6 gemacht werden. Bereits angesprochen wurde das Problem, dass Hassrede durch eine falsche oder alternative Schreibweise eventuell „versteckt“ werden könnte, sodass die Aussage zwar vom Menschen,

aber nicht mehr von der Maschine als Hassrede erkannt wird (beispielsweise *grass the grews*). Zudem gibt es insgesamt implizite Hassrede, welche in ihrer Natur nur unterschwellig hasserfüllt ist. Durch die Ergebnisse aus Kapitel 6 liegt die Vermutung nahe, dass Hassrede, in welcher keine Verunglimpfungen oder auch keine Gewalt vorkommen, bereits seltener auch als Hassrede identifiziert werden kann und somit zur Kategorie „implizite Hassrede“ gehört. Aber ebenso das Level an Eloquenz oder „Rationalität“ könnte Einfluss auf die Klassifikation nehmen, was bislang nicht getestet wurde. Insofern wäre es ebenfalls hilfreich, die für die Klassifizierung ausschlaggebenden Eigenschaften konkret herauszufinden. Denn so könnten vielleicht die in Kapitel 6 aufgekommenen, augenscheinlich willkürlichen Klassifikationen einiger Aussagen erklärt werden, sowie damit auch getestet werden, ob speziell Gewaltdarstellungen, Verunglimpfungen oder Eloquenz tatsächlich vom Modell als Eigenschaften wahrgenommen werden oder nicht.

Demnach besteht ein weiterführender Ansatz darin, weitere Testdatensätze auf Grundlage dieser Eigenschaften zu erstellen, um somit feststellen zu können, inwieweit diese Eigenschaften Einfluss auf die Klassifikation haben. Je nachdem müssten vielleicht die für Kapitel 6 ausgewählten Tweets, insbesondere was die Kategorie Misandrie angeht, angepasst und die Evaluation nochmals durchgeführt werden. Aber auch bei neutralen Aussagen könnte es sein, dass diese durch ein angesprochenes Thema, welches typisch für Hassrede ist, fälschlicherweise als Hassrede vom Modell klassifiziert werden. Dies könnte ebenfalls anhand von Tweets getestet werden, welche implizite Hassrede aber zudem auch implizite neutrale Aussagen abdecken. Jene könnten beispielsweise Diskussionen sein, in welchen Verunglimpfungen lediglich als Beispiele aufkommen und somit in einem neutralen Kontext stehen. Auch sarkastische Aussagen können zu impliziter Hassrede gezählt werden.

Vorhergehende Studien haben bereits gezeigt, dass es sinnvoll sein kann, statt Hassrede nur im Allgemeinen vorherzusagen auch unterschiedliche Arten von Hassrede für die Klassifikation einzubeziehen. Dies betrifft die Unterscheidung zwischen direkter und indirekter Hassrede sowie die Unterscheidung zwischen impliziter und expliziter Hassrede (vgl. Waseem et. al., 2017). Denn es ist bereits offenkundig, wie divers Hassrede sein kann. Unterschiedliche Arten von Diskriminierung werden bislang zu Hassrede zusammengefasst, wodurch die Modelle sich zwangsläufig der Varianz innerhalb der Klasse anpassen müssen. Feiner gegliederte Klassen könnten es dem Modell ermöglichen, spezifischere Arten von Hass zu identifizieren. Das Modell kann so möglicherweise insgesamt besser für einige Arten performen. Dies ist zudem

eine weitere Möglichkeit, zu bestimmen, inwieweit implizite und explizite Hassrede unterschiedlich gut erkannt werden.

Schließlich gibt es noch eine weitere Verbesserung, die vorgenommen werden könnte. Diese bestünde aus der Aufnahme von Meta-Informationen der Tweets zugunsten deren Klassifikation, so wie beispielsweise dem Geschlecht des Schreibers sowie dem Geschlecht der möglichen Zielperson. Ob eine Aussage Hassrede ist oder nicht, kann, wie schon erläutert, gänzlich vom Kontext abhängen. Dazu gehört zum Beispiel, ob die Person, welche eine Verunglimpfung benutzt, selbst der Gruppe angehört, zu welcher auch die Verunglimpfung gehörig ist. Eigenschaften über den Nutzer können für solche Fälle Aufschluss geben. Aber auch Informationen über eine Zielperson (eventuell über @ gekennzeichnet oder namentlich angesprochen) können essenziell sein, nicht nur was eine präzisere Klassifikation angeht, sondern auch um weitere Erkenntnisse über die Verbreitung von Hassrede zu erlangen. An die Daten zu gelangen, ist allerdings nicht zuletzt aufgrund der Datenschutzbestimmungen von Twitter unmöglich und bleibt somit für Twitter-externe Analysen eine Vision.

Die Möglichkeiten, Hassrede mit Deep Learning zu klassifizieren, sind also lange nicht ausgeschöpft. Die verschiedenen, weiterführenden Ansätze sowie auch die durch diese Arbeit aufgedeckten Kritikpunkte am derzeitigen Stand der Forschung zeigen, wie viel mehr Aufwand in dieses Thema investiert werden kann. Die wichtigsten Punkte sind:

- Die Erstellung eines neuen Hassrede Korpus, für welches nicht nur durch Schlüsselwörter nach Daten gefiltert wurde.
- Die Verbesserung von Buchstaben Embeddings, speziell im Hinblick auf Tweets.
- Die weitere Untersuchung der kontextualisierten BERT Embeddings im Hinblick auf das Thema Hassrede.
- Die Übertragung des Projekts auf weitere Sprachen/weitere Kulturen, vor allem was Wort-basierte Ansätze betrifft.

8. Fazit

Diese Arbeit hat gezeigt, dass es keine allgemeingültige Lösung für ein Deep Learning Problem gibt. Besonders für Textklassifikation zeigt sich, dass den Daten und des Themas zufolge die Vorverarbeitungsschritte sorgfältig bestimmt werden müssen. Obwohl die in Kapitel 4 aufgeführten Arbeiten gezeigt haben, dass die eine oder andere Vorgehensweise besser für die

jeweiligen Daten oder das jeweilige Thema geeignet ist, können diese Lösungswege nicht unreflektiert für das eigene Projekt übernommen werden. Insbesondere hat sich herausgestellt, dass der wortbasierte Ansatz sinnvoller ist als der buchstabenbasierte Ansatz. Weiterhin haben sich zwei klare Trends für den wortbasierten Ansatz bemerkbar gemacht. Zum einen sind LSTM Modelle im Allgemeinen gegenüber anderen Modellarchitekturen eher geeignet. Zum anderen haben sich N-Gramme als die leistungsfähigere Alternative gegenüber der einfachen Wortrepräsentation herausgestellt.

Auch die detaillierteren Vorverarbeitungsschritte sind bei einem Thema wie der Identifikation von Hassrede sorgfältig zu überdenken. Bei dieser Arbeit hat sich herausgestellt, dass das Entfernen der Stoppwörter die Ergebnisse nicht verbessert und somit nicht notwendig ist. Auch die Rechtschreibkorrektur kann, so wie in Kapitel 7 diskutiert, den Ergebnissen schaden. Weiterhin hat das Entfernen von Benutzernamen und Personennamen besondere Priorität.

Fine-tuned BERT hat sich in dieser Arbeit insgesamt als die beste Strategie erwiesen, nicht nur was die insgesamt beste Ergebnisqualität angeht, sondern auch aufgrund der höheren Angepasstheit auf problematische Daten, in diesem Fall schwarzen Humor. Die Tatsache, dass BERT nicht nur für Klassifikation, sondern auch für viele weitere NLP Aufgaben wie beispielsweise Part-of-Speech Tagging, Question Answering oder Named Entity Recognition geeignet ist, macht es zu dem vermutlich mächtigsten Werkzeug im Bereich des NLP zurzeit.

Das Thema Hassrede im Internet wird auch für viele kommende Jahre bedeutend bleiben. Weiteres Verständnis von Hassrede, deren Vorkommen und Struktur ist notwendig, um Lösungen für ihre Eindämmung zu finden. Dazu zählt auch eine einheitliche Definition. Das beste, in dieser Arbeit verwendete Modell (Fine-tuned BERT) erzielt einen F1 Score von etwa 0,93, was zunächst beeindruckend erscheint. Dennoch hat sich gezeigt, dass diese Zahl stark von den für das Training verwendeten Daten und besonders deren Annotation abhängt. Denn es kann sein, dass die Daten nur einen Teilbereich von Hassrede wirklich abbilden. Dies ist abhängig von der verwendeten Auswahlmethode für die Daten sowie auch der darunterliegenden Definition für Hassrede. Es wurden Schlüsselwörter zum Filtern von Hassrede benutzt, wobei viel „versteckter“ Hass womöglich nicht in die Daten aufgenommen werden konnte. Als versteckter Hass gilt zunächst Hassrede ohne typische Schlüsselwörter (wie Verunglimpfungen und Gewaltdarstellungen), aber auch weniger verbreitete Arten von Hassrede, welche zum Beispiel Weiße oder Männer zum Ziel hat.

Selbst wenn jeder „versteckter“ Hass durch die Datengrundlage perfekt abgedeckt werden würde, stellt sich immer noch die Frage, welche Art von Diskriminierungen als Hassrede gelten. Dass Angriffe gegen ein Geschlecht als Hassrede zählen, während Angriffe gegen Anwälte keine Hassrede sind, ist unter Umständen nachvollziehbar (vgl. Bicknell 2007, 462). Was ist aber mit Angriffen gegen eine Kaste oder gegen eine Altersgruppe, so wie es in den Definitionen durch Facebook und Twitter bereits unterschieden wird? Was ist mit Grenzfällen von „typischen“ Arten von Hassrede, wenn beispielsweise der Kontext eines Tweets, in dem eine Verunglimpfung verwendet wird, hergeben könnte, dass es sich gar nicht um Hassrede handelt? Bislang wird also versucht, etwas maschinell zu klassifizieren, wovon noch keine klare Definition, sondern eher nur eine grobe Idee vom Menschen existiert. Das hier verwendete Korpus ist lediglich durch die Entscheidungen einer kleinen Gruppe von Menschen im Rahmen der verschiedenen Definitionen von Hassrede beeinflusst worden.

Der größte Fortschritt bei dieser maschinellen Klassifikation würde also durch eine Verbesserung der Daten erzielt werden. Das bedeutet hauptsächlich:

- die Festlegung einer klaren Definition von Hassrede und anhand dieser
- eine bessere Auswahlmethode für die Trainingsdaten (über Schlüsselwörter hinausgehend) zu wählen, sodass vor allem auch impliziter Hass abgedeckt wird, sowie
- möglichst sorgfältig die Daten in Trainings-, Evaluations- und Testset aufzuteilen, sodass die unterschiedlichen Arten von Hassrede gleich stark in den Datensets vertreten sind.

Dies würde nicht nur enorm viel Zeit und Aufwand verlangen. Es würde ebenso bedeuten, einen Schritt zurückzugehen und einige grundlegende Erkenntnisse zum Thema Hassrede klar abzustecken. Ist dies nicht möglich, so wird ein Modell auch nur einen Teilbereich davon, was Hassrede eigentlich bedeutet, widerspiegeln. Aber selbst bei einem idealen Modell, welches Hassrede perfekt identifizieren kann, stellt sich vielleicht eine weitere Frage: sollte dieses Modell eingesetzt werden, um Hassrede im Internet automatisch zu unterdrücken?

Die Antwort lautet nein, insbesondere nicht ausgehend von den bislang geschaffenen Modellen, welche auf keinem allgemein gültigen Konzept von Hassrede beruhen. Stattdessen sollten andere Lösungsansätze gewählt werden, um Hassrede entgegenzuwirken. Prof. Nadine Strossen beispielsweise argumentiert: „Egal wie gut gemeint, in der Praxis ist das Verbot von ‚Hate speech‘ bestenfalls wirkungslos und im schlimmsten Fall kontraproduktiv. Wir müssen die Art der Aussagen, mit der wir nicht einverstanden sind, mit anderen Aussagen kontern“

(Fox 2018). Dasselbe kann auch für Fake News gesagt werden: „Wir wollen kein Wahrheitsministerium. Lügen ist nicht illegal“ (Věra Jourová in Fox 2018). Gerade da diese Lösung so unausgereift ist, sind mögliche gravierende Folgen so einer Umsetzung hinsichtlich der Meinungsfreiheit schwer abschätzbar.

Zusätzlich sollte beachtet werden, dass Deep Learning keinesfalls menschliches Denken widerspiegelt. Bei Deep Learning werden die Eigenschaften der Klassen ebenfalls durch die Maschine erlernt. Im Laufe der Arbeit wurden einige Vermutungen angestellt, welche Eigenschaften eines Tweets für dessen Klassifikation verantwortlich sein könnten. In Kapitel 6 hat sich auch herausgestellt, dass manche Tweets, die durch das menschliche Verständnis eindeutig als Hassrede klassifiziert werden können, von der Maschine gänzlich anders eingestuft worden sind. Wie schon in Kapitel 7 erwähnt wäre es sinnvoll, die Eigenschaften sichtbar zu machen, welche die Maschine für die Klassifikation benutzt. Denn die Charakteristika, die die Maschine erkennt, sind meistens für den Menschen völlig unverständlich. Denn bei Deep Learning wird lediglich ein Input über kontinuierliche, geometrische Transformationen einem Output zugeordnet. Dies kann bereits zu großen Erfolgen führen, allerdings ist eine derartige Zuordnung weit von einer KI auf einem menschlichen Level entfernt (vgl. Chollet 2018, 325ff.). Die Maschine „versteht“ in dem Sinn die Aufgabe nicht, die sie lernt. Wenn es also darum geht, zu entscheiden, solche Modelle einzusetzen, sollte einem diese Tatsache bewusst sein. Dass Deep Learning bei weitem menschliches Denken nicht ersetzen kann, spricht umso mehr dafür, solche Modelle für ein moralisch aufgeladenes und feinfühliges Thema wie dem Entgegenwirken von Hassrede nicht einzusetzen.

Dagegen bestünde allerdings die Möglichkeit, die Modelle zur Unterstützung bei der Moderation von Nutzerinhalten einzusetzen. Bislang sind Moderatoren von Hinweisen von Nutzern abhängig, um auf aggressive Inhalte aufmerksam zu werden. So könnten die vom Modell als Hassrede klassifizierten (bzw. zu löschenden) Inhalte ebenso erst einem Moderator vorgelegt werden, sodass eine Überprüfung durch den Menschen trotzdem noch stattfindet und nicht ersetzt wird. Auch die Sicherheit, mit welcher eine Aussage vom Modell als Hassrede klassifiziert wird, kann hierbei eine Rolle spielen. So könnten auch nur jene Aussagen den Moderatoren vorgelegt werden, bei welchen die Vorhersagesicherheit einen Schwellenwert unterschreitet.

Weitere Regulierung durch die Regierung, zusätzlich zum NetzDG, ist aus demselben Grund ebenso kritisch anzusehen. Wie bereits in Kapitel 2 angesprochen, reicht die ungenaue Definition von Hassrede, wie sie bislang besteht, nicht aus, um klare Limitierungen zu verordnen. Bislang benötigen viele Aussagen eine sorgfältige Überprüfung, ob sie Hassrede zuzuordnen sind oder nicht. Das bedeutet allerdings nicht, dass nicht grundsätzlich gegen Hassrede vorgegangen werden sollte. Es gibt zum Beispiel bereits einige Kampagnen mit dem Ziel, auf das Problem Hassrede aufmerksam zu machen, so wie die Bewegung „No hate speech Movement“[19] des Europarates. Vielleicht ist auch eine genaue Regulierung denkbar, welche nur einen klar abgrenzbaren Teilbereich von Hassrede unterbindet, wenn beispielsweise die juristische Definition von Hassrede in Deutschland dahingehend ergänzt wird.

Hassrede in Deutschland wird bislang aktiv nur durch die Webseitenbetreiber moderiert, da diese durch das NetzDG dazu aufgerufen werden. Das Löschen von Hassredekommentaren allein stellt allerdings „keine wirksame und vor allem nachhaltige Strategie“ dar (Meßmer und Krause 2018, 10). Aber selbst der rechtliche Weg ist oft nicht erfolgreich, selbst wenn es um justiziable Hassrede geht. Hier müssen „Verfahren regelmäßig eingestellt werden [...], weil Beklagte nicht ermittelt werden können“ (ebd., 16). Denn für Facebook beispielsweise gilt, dass aufgrund der Rechtslage in den US personenbezogene Daten nicht herausgegeben werden können (vgl. ebd., 16).

Ein alternativer, zusätzlicher Lösungsansatz kann beispielsweise beinhalten, das Internet auch als Chance zu verstehen. So wie sich die Radikalisierung über das Internet schneller ausbreiten konnte, so kann vielleicht auch eine geleitete Deradikalisierung über das Internet stattfinden (vgl. Neumann 2018, 6). Dabei gilt „Counter Speech“ als zentraler Lösungsvorschlag. Statt Hassrede zu ignorieren, ironisieren oder deren Absender zu beschimpfen, sollte ihr mit sachlichen Gegenargumenten begegnet werden. Dies hat zum Ziel, die Diskussionskultur als Ganzes zu verändern und zu stärken. So gibt es beispielsweise die Facebook-Gruppe „#ichbinhier“[20] (vgl. Meßmer und Krause 2018, 11).

Es kann argumentiert werden, dass Hassrede (online sowie auch offline) grundsätzlich nur ein Symptom von Radikalisierung und Extremismus ist, und das Vorgehen gegen Hassrede gänzlich unwirksam ist, um Hass als Ganzes zu bekämpfen. Allerdings können Personen, welche bewusst Hass schüren, andere in ihrem Sinne beeinflussen und dies zum Erreichen ihrer

[19] https://www.coe.int/en/web/no-hate-campaign (Stand 13.05.2019).
[20] http://www.ichbinhier.eu/ich-bin-hier (Stand 17.05.2019).

persönlichen Ziele einsetzen. Hassrede ist damit eher ein Werkzeug und kann ebenso ein Hassverstärker sein. Somit gleicht die Beziehung zwischen Hass und Hassrede vielmehr einem Teufelskreis. Gegen Hassrede, genauso wie auch gegen die hassverstärkende Wirkung von Hassrede vorzugehen, kann also helfen, den Teufelskreis zu durchbrechen und Ungleichheiten sowie auch Manipulation durch Hass in der Gesellschaft zu bekämpfen. So wie beispielsweise Aufklärung zur Vermeidung, sodass Menschen sich nicht weiter durch Hassrede instrumentalisieren lassen.

Zuvor wurde erläutert, dass Deep Learning bzw. künstliche Intelligenz im allgemeinen menschliches Denken nicht widerspiegelt. Allerdings ist genau diese Tatsache auch besonders nützlich: es werden Beziehungen zwischen Daten und Klassen hergestellt, die für einen Menschen unsichtbar oder nicht unbedingt offensichtlich sind. Dem Menschen wird durch Deep Learning eine gänzlich andere Sichtweise auf Problemstellungen geboten, was in der Vergangenheit bereits revolutionäre Ergebnisse produzieren konnte. Der Hype um Deep Learning und KI wird daher weiterhin bestehen bleiben. Zukünftig wird KI in mehr und mehr Arbeitsbereichen eingesetzt werden, voraussichtlich auch, um unternehmerische Entscheidungen innerhalb weniger technisch-basierten Betrieben zu leiten. Auch wird KI in weiteren, alltäglichen Aufgaben Anwendung finden (vgl. Bhattacharya 2018) – zusätzlich etwa zu dem weitverbreiteten Empfehlungssystem. Auf der einen Seite kann die Zukunft von KI damit viele Vorteile, so wie wirtschaftliches Wachstum oder auch menschliche Zufriedenheit und Sicherheit, versprechen (vgl. Wang und Siau 2018, 1).

Auf der anderen Seite stellt KI auch ein Risiko dar. Dazu gehört das oftmals angenommene, dystopische Weltbild, dass Menschen zukünftig lediglich den Entscheidungen der Maschinen vertrauen und somit auch in sämtlichen Arbeitsplätzen ersetzt werden (vgl. Makridakis 2017, 9), so wie es mit selbstfahrenden Autos keine Taxi- oder Busfahrer mehr geben wird. Allerdings bedeutet dies auch, dass an anderen Stellen neue Arbeitsplätze geschaffen werden; so wie beispielsweise Supportmitarbeiter für Autofahrsoftware. Weiterhin stellen sich grundsätzlich einige ethische und moralische Herausforderungen. Am einfachsten lässt sich dies anhand selbstfahrender Autos darstellen. In einer Unfallsituation, in der nur die zwei Szenarien, dass entweder Person X oder Person Y umkommt, existieren, kann und wird die KI eine Entscheidung treffen, die für einen Menschen nicht möglich wäre (vgl. Gordon 2018). Es ist fraglich, ob die KI in diesem Fall ohne eine moralische Grundlage überhaupt handeln darf. Ein Verständnis von Moral an eine KI zu vermitteln ist somit eine der wichtigsten und gleichzeitig komplexesten Anforderungen. Ebenso stellt sich die Frage, wer bei einem Fehler der KI

verantwortlich ist, gerade da es oftmals unmöglich auszumachen ist, warum ein Fehler passiert (vgl. Wang und Siau 2018, 3), so wie auch bei den Klassifikationsfehlern aus Kapitel 6.

Im Juni 2019 wurden die Richtlinien gegen Hassrede auf der Online-Videoplattform YouTube verschärft. Statt wie zuvor Videos mit extremistischen und hassschürenden Inhalten schwerer zugänglich zu machen (vor allem durch Ausschluss aus dem Empfehlungssystem und Deaktivieren der Kommentarfunktion) werden diese jetzt gesperrt und die erschwerte Zugänglichkeit auf Grenzfälle von Hassrede und Falschinformation ausgeweitet[21]. YouTube wurde vor allem von Werbekunden unter Druck gesetzt und hat jetzt deshalb als Reaktion zu einer drastischeren Maßnahme gegriffen (Kuhn 2019). Dies hatte allerdings auch fälschliche Sperrungen von Inhalten zufolge; darunter fallen Videos zur geschichtlichen Aufklärung und Dokumentationen. Weitere Fälle von Sperrungen durch YouTube werden hitzig diskutiert, dazu zählen einige kontroverse Uploader (beispielsweise Steven Crowder) aus dem rechten politischen Spektrum. So wird die Debatte zwischen Meinungsfreiheit und Hassrede auch hier wieder angeschürt (Goggin 2019). Es zeigt sich damit weiterhin, dass Webseiten durch externen Druck impulsiv zu einer kurzfristigen Lösung, was Hassrede betrifft, greifen. In Zukunft sollte die Forschung zum Thema Hassrede im Netz aufholen, sodass an einer längerfristigen, weniger radikalen Lösung – statt der einfachen Unterdrückung durch einen Algorithmus – gearbeitet werden kann.

[21] https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html (Stand 12.06.2019).

9. Quellen- und Literaturverzeichnis

Ahluwalia, Resham, Himani Soni, Edward Callow, Anderson Nascimento und Martine de Cock. „Detecting Hate Speech Against Women in English Tweets." In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018), Turin, Italy. CEUR. org.* 2018.

Anderson, Luvell und Ernie Lepore. „What did you call me? Slurs as prohibited words setting things up." *Analytic Philosophy* 54, Nr. 3 (2013): 350–363.

Aroyehun, Segun T. und Alexander Gelbukh. „Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling." In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 90–7. 2018.

Bakircioglu, Onder. „Freedom of expression and hate speech." *Tulsa J. Comp. & Int'l L.* 16 (2008): 1–49.

Baldauf, Johannes, Julia Ebner und Jakob Guhl, Hrsg. *Hassrede und Radikalisierung im Netz.* ISD, 2018. Zuletzt geprüft am 21.05.2019. https://www.isdglobal.org/wp-content/uploads/2018/09/ISD-NetzDG-Report-German-FINAL-26.9.18.pdf.

Bartlett, Jamie, Jeremy Reffin, Noelle Rumball und Sarah Williamson. „Anti-social media." Zuletzt geprüft am 21.05.2019. http://cilvektiesibas.org.lv/site/record/docs/2014/03/19/DEMOS_Anti-social_Media.pdf.

Bhattacharya, Santanu. „AI Predictions for 2019." Towards Data Science. Zuletzt geprüft am 05.06.2019. https://towardsdatascience.com/ai-predictions-for-2019-610b8de56aad.

Bianchi, Claudia. „Slurs and appropriation: An echoic account." *Journal of Pragmatics* 66 (2014): 35–44. doi:10.1016/j.pragma.2014.02.009.

Bicknell, Jeanette. „What is offensive about offensive jokes?" *Philosophy Today* 51, Nr. 4 (2007): 458–465.

Bojanowski, Piotr, Edouard Grave, Armand Joulin und Tomas Mikolov. „Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5 (2017): 135–146.

Brown, Alexander. „What is hate speech? Part 1: The Myth of Hate." *Law and Philosophy* 36, Nr. 4 (2017): 419–468. doi:10.1007/s10982-017-9297-1.

Brown, Alexander. „What is so special about online (as compared to offline) hate speech?" *Ethnicities* 18, Nr. 3 (2018): 297–326.

Cairns, Ann. „Why AI is failing the next generation of women." World Economic Forum. Zuletzt geprüft am 05.06.2019. https://www.weforum.org/agenda/2019/01/ai-artificial-intelligence-failing-next-generation-women-bias/.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau und Yoshua Bengio. „On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259*, 2014.

Chollet, François. *Deep learning with Python.* Shelter Island, NY: Manning, 2018.

Cortes, Corinna und Vladimir Vapnik. „Support-vector networks." *Machine learning* 20, Nr. 3 (1995): 273–297.

Couto, Javier. „The major advancements in Deep Learning in 2018.," Tryolabs. Zuletzt geprüft am 21.05.2019. https://tryolabs.com/blog/2018/12/19/major-advancements-deep-learning-2018/.

Dastin, Jeffrey. „Amazon scraps secret AI recruiting tool that showed bias against women," Reuters. Zuletzt geprüft am 05.06.2019. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Davidson, Thomas, Dana Warmsley, Michael Macy und Ingmar Weber. „Automated Hate Speech Detection and the Problem of Offensive Language." In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 512–5. ICWSM '17. 2017.

Deutscher Bundestag. „Regulierung von Hate Speech und Fake in sozialen Netzwerken durch EU-Mitgliedstaaten." Zuletzt geprüft am 01.07.2019. https://www.bundestag.de/blob/566942/a5eb997872bbe5dbca3f47112eb04c46/wd-10-032-18-pdf-data.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. „Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

Dos Santos, Cicero und Maira Gatti. „Deep convolutional neural networks for sentiment analysis of short texts." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 69–78. 2014.

ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Y. Wang und Elizabeth Belding. „Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media." In *Proceedings of the 12th International AAAI Conference on Web and Social Media.* ICWSM '18. 2018.

Founta, Antigoni-Maria, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali und Ilias Leontiadis. „A unified deep learning architecture for abuse detection." *arXiv preprint arXiv:1802.00385*, 2018.

Fox, Benjamin. „Gesetze gegen „Hassrede" vs. Meinungsfreiheit.," Euractiv. Zuletzt geprüft am 05.06.2019. https://www.euractiv.de/section/digitale-agenda/news/gesetze-gegen-hassrede-vs-meinungsfreiheit/.

Gagliardone, Iginio, Danit Gal, Thiago Alves und Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

Gambäck, Björn und Utpal K. Sikdar. „Using convolutional neural networks to classify hate-speech." In *Proceedings of the first workshop on abusive language online*, 85–90. 2017.

Gaydhani, Aditya, Vikrant Doma, Shrikant Kendre und Laxmi Bhagwat. „Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach." *arXiv preprint arXiv:1809.08651*, 2018.

Goggin, Benjamin. „YouTube's week from hell: How the debate over free speech online exploded after a conservative star with millions of subscribers was accused of homophobic harassment." Zuletzt geprüft am 13.06.2019. https://www.businessinsider.de/steven-crowder-youtube-speech-carlos-maza-explained-youtube-2019-6?op=1.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin und Tomas Mikolov. „Learning Word Vectors for 157 Languages." In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).* 2018.

Graves, Alex, Santiago Fernández und Jürgen Schmidhuber. „Bidirectional LSTM networks for improved phoneme classification and recognition.“ In *International Conference on Artificial Neural Networks*, 799–804. 2005.

Hochreiter, Sepp und Jürgen Schmidhuber. „Long short-term memory.“ *Neural computation* 9, Nr. 8 (1997): 1735–1780.

Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero und Larry Heck. „Learning Deep Structured Semantic Models for Web Search using Clickthrough Data.“ In. ACM International Conference on Information and Knowledge Management (CIKM), 2013. https://www.microsoft.com/en-us/research/publication/learning-deep-structured-semantic-models-for-web-search-using-clickthrough-data/.

Kim, Yoon. „Convolutional neural networks for sentence classification.“ *arXiv preprint arXiv:1408.5882*, 2014.

Kuhn, Johannes. „Youtube verschärft Upload-Regeln.“ Süddeutsche Zeitung. Zuletzt geprüft am 13.06.2019. https://www.sueddeutsche.de/digital/youtube-upload-regeln-problemvideos-1.4476859.

LeCun, Yann, Léon Bottou, Yoshua Bengio, Patrick Haffner und others. „Gradient-based learning applied to document recognition.“ *Proceedings of the IEEE* 86, Nr. 11 (1998): 2278–2324.

Makridakis, Spyros. „The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms.“ *Futures* 90 (2017): 46–60.

Mehdad, Yashar und Joel Tetreault. „Do characters abuse more than words?“ In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 299–303. 2016.

Meßmer, Anna-Katharina und Laura-Kristine Krause. „Wie umgehen mit Hate Speech?“. Zuletzt geprüft am 21.05.2019. http://www.progressives-zentrum.org/wp-content/uploads/2018/11/WIE-UMGEHEN-MIT-HATE-SPEECH_Anna-Katharina-Me%C3%9Fmer_Laura-Kristine-Krause_Das-Progressive-Zentrum.pdf.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado und Jeffrey A. Dean. *Computing numeric representations of words in a high-dimensional space*. Google Patents, 2015.

Mondal, Mainack, Leandro A. Silva und Fabr\'\icio Benevenuto. „A measurement study of hate speech in social media.“ In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94. 2017.

Neumann, Peter. „Vorwort.“ In *Hassrede und Radikalisierung im Netz.* Hrsg. von Johannes Baldauf, Julia Ebner und Jakob Guhl, 5–6. ISD, 2018.

Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad und Yi Chang. „Abusive language detection in online user content.“ In *Proceedings of the 25th international conference on world wide web*, 145–53. 2016.

O'Neill, Patrick. „The comedy of entropy: the contexts of black humour.“ *Canadian Review of Comparative Literature/Revue canadienne de littérature comparée* 10, Nr. 2 (1983): 145–166.

Park, Ji H. und Pascale Fung. „One-step and two-step classification for abusive language detection on twitter.“ *arXiv preprint arXiv:1706.01206*, 2017.

Pavlopoulos, John, Prodromos Malakasiotis und Ion Androutsopoulos. „Deep learning for user comment moderation.“ *arXiv preprint arXiv:1705.09993*, 2017.

Pearson, Karl. „On lines and planes of closest fit to systems of points in space.“ *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, Nr. 11 (1901): 559–572. doi:10.1080/14786440109462720.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel und M. Blondel et al. „Scikit-learn: Machine Learning in Python.“ *Journal of Machine Learning Research* 12 (2011): 2825–2830.

Pennington, Jeffrey, Richard Socher und Christopher Manning. „Glove: Global vectors for word representation.“ In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–43. 2014.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee und Luke Zettlemoyer. „Deep contextualized word representations.“ In *Proc. of NAACL*. 2018.

Post, Robert C. „Racist Speech, Democracy, and the First Amendment.“ *Faculty Scholarship Series* 267, Nr. 32 (1991): 267–327.

Rafael, Simone und Alexander Ritzmann. „Hintergrund: „Das ABC des Problemkomplexes Hassrede, Extremismus und NetzDG“.“ In *Hassrede und Radikalisierung im Netz*. Hrsg. von Johannes Baldauf, Julia Ebner und Jakob Guhl, 11–9. ISD, 2018.

Raiyani, Kashyap, Teresa Gonçalves, Paulo Quaresma und Vitor B. Nogueira. „Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter.“ 2018.

Risch, Julian und Ralf Krestel. „Aggression identification using deep learning and data augmentation.“ In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 150–8. 2018.

Robinson, David, Ziqi Zhang und Jonathan Tepper. „Hate speech detection on twitter: Feature engineering vs feature selection.“ In *European Semantic Web Conference*, 46–9. 2018.

Saleem, Haji M., Kelly P. Dillon, Susan Benesch und Derek Ruths. „A web of hate: Tackling hateful speech in online social spaces.“ *arXiv preprint arXiv:1709.10159*, 2017.

Serra, Joan, Ilias Leontiadis, Dimitris Spathis, J. Blackburn, G. Stringhini und Athena Vakali. „Class-based prediction errors to detect hate speech with out-of-vocabulary words.“ In *Abusive Language Workshop*. Bd. 1. 2017.

Shaw, LaShel. „Hate Speech in Cyberspace: bitterness without boundaries.“ *Notre Dame JL Ethics & Pub. Pol'y* 25 (2012): 279–304.

Siau, Keng und Weiyu Wang. „Ethical and Moral Issues with AI.“ Zuletzt geprüft am 21.05.2019. https://www.researchgate.net/publication/325934375_Ethical_and_Moral_Issues_with_AI.

Speer, Robyn. „How to make a racist AI without really trying.“ Zuletzt geprüft am 05.06.2019. http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever und Ruslan Salakhutdinov. „Dropout: a simple way to prevent neural networks from overfitting.“ *The Journal of Machine Learning Research* 15, Nr. 1 (2014): 1929–1958.

Strother H. Walker und David B. Duncan. „Estimation of the Probability of an Event as a Function of Several Independent Variables.“ *Biometrika* 54, 1/2 (1967): 167–179. http://www.jstor.org/stable/2333860.

Summers, Chelsea G. „The Year in Male Tears: 2015 was the year misandry went mainstream, and that's a good thing.," Vice. Zuletzt geprüft am 05.06.2019. https://www.vice.com/en_ca/article/4wbp9j/the-year-in-male-tears.

The Economist. „Can white people experience racism?". Zuletzt geprüft am 05.06.2019. https://www.economist.com/open-future/2018/09/18/can-white-people-experience-racism.

Vijayaraghavan, Prashanth, Ivan Sysoev, Soroush Vosoughi und Deb Roy. „Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns." *arXiv preprint arXiv:1606.05694*, 2016.

Vora, Parth, Mansi Khara und Kavita Kelkar. „Classification of Tweets based on Emotions using Word Embedding and Random Forest Classifiers." *International Journal of Computer Applications (0975 – 8887)* 178, Nr. 3 (2017). Zuletzt geprüft am 14.01.2019. https://pdfs.semanticscholar.org/f777/c30d5682c4b43ffe1120b731adbeeecd98a7.pdf.

Warner, William und Julia Hirschberg. „Detecting hate speech on the world wide web." In *Proceedings of the Second Workshop on Language in Social Media*, 19–26. 2012.

Warner, William und Julia Hirschberg. „Detecting hate speech on the world wide web." In *Proceedings of the second workshop on language in social media*, 19–26. 2012.

Waseem, Zeerak. „Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter." In *Proceedings of the first workshop on NLP and computational social science*, 138–42. 2016.

Waseem, Zeerak, Thomas Davidson, Dana Warmsley und Ingmar Weber. „Understanding abuse: A typology of abusive language detection subtasks." *arXiv preprint arXiv:1705.09899*, 2017.

Waseem, Zeerak und Dirk Hovy. „Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." In *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: Association for Computational Linguistics, 2016. http://www.aclweb.org/anthology/N16-2013.

Yenala, Harish, Ashish Jhanwar, Manoj K. Chinnakotla und Jay Goyal. „Deep learning for detecting inappropriate content in text." *International Journal of Data Science and Analytics* 6, Nr. 4 (2018): 273–286.

Yong, Caleb. „Does Freedom of Speech Include Hate Speech?" *Res Publica* 17, Nr. 4 (2011): 385. doi:10.1007/s11158-011-9158-y.

Zhang, Shiwei, Xiuzhen Zhang und Jeffrey Chan. „Language-independent Twitter Classification using Character-based Convolutional Networks." In *International Conference on Advanced Data Mining and Applications*, 413–25. 2017.

Zhang, Xiang und Yann LeCun. „Text understanding from scratch." *arXiv preprint arXiv:1502.01710*, 2015.

Zhang, Xiang, Junbo Zhao und Yann LeCun. „Character-level convolutional networks for text classification." In *Advances in neural information processing systems*, 649–57. 2015.

Zhang, Ziqi, David Robinson und Jonathan Tepper. „Detecting hate speech on Twitter using a convolution-GRU based deep neural network." In *European Semantic Web Conference*, 745–60. 2018.

A Anhang

A1 Parameter zur Modellerstellung bezüglich Kapitel 5

BERT

Modell	BERT-Base, Uncased
Max. Sequenzlänge	128
Epochen (Fine-tuning)	2
Lernrate	2e-5

LSTM

Größe	5
Dropout	0,5
Epochen	15

CNN-LSTM

Größe (CNN)	32
Schrittgröße	5
Aktivierung	ReLu
MaxPooling	3
Größe (LSTM)	5
Dropout	0,5
Epochen	15

3 x Dense

Größe	16, 8, 5
Aktivierung	ReLu
Dropout	0,5
Epochen	15

BiLSTM

Größe	5
Dropout	0,5
Epochen	15
Vereinigung	Add

Character CNN

Größe	32
Schrittgröße	5
MaxPooling	3
Aktivierung	ReLu
Dropout	0,5
Epochen	15

A2 Konfusionsmatrizen bezüglich Kapitel 5.2.1

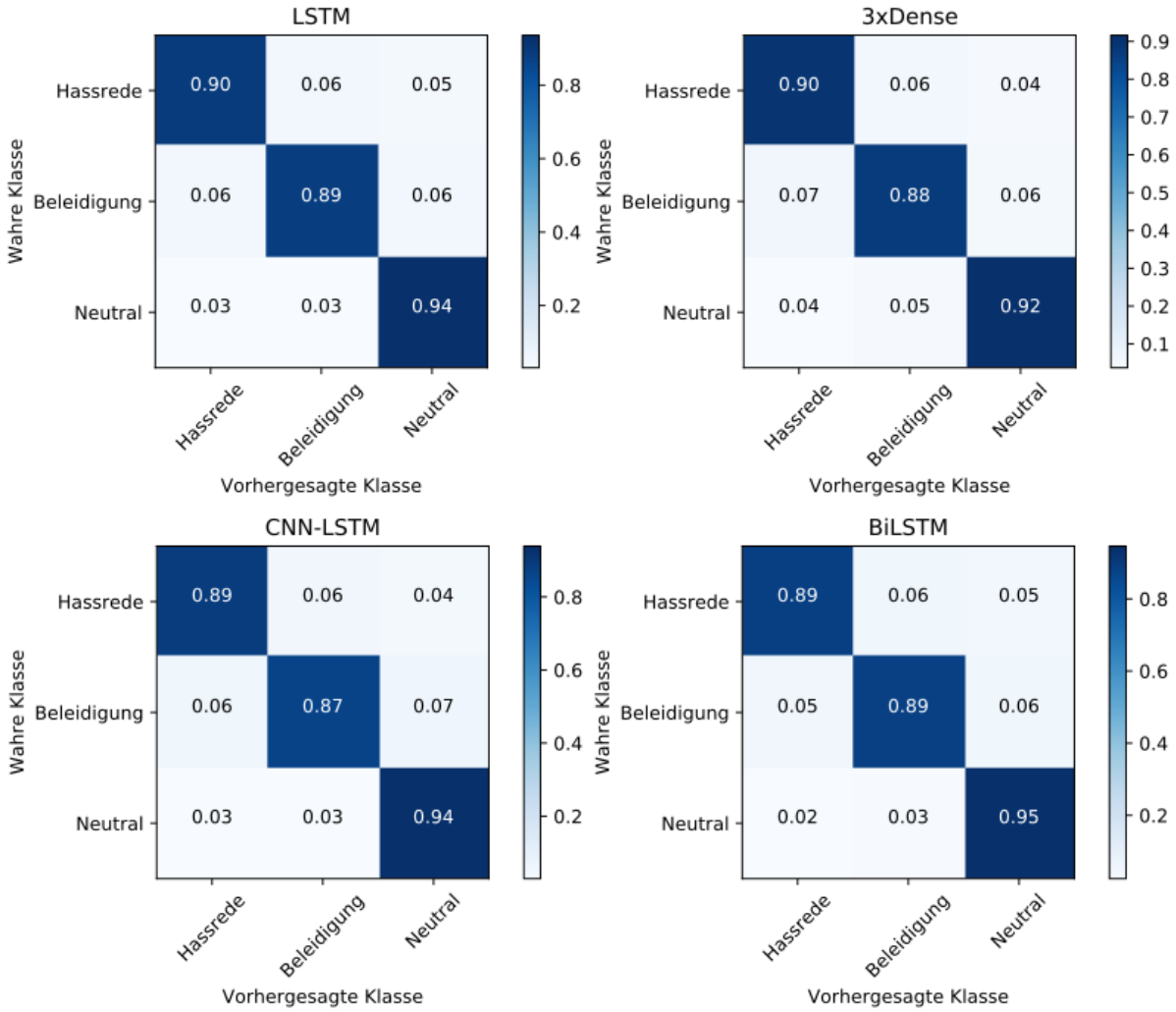

Abbildung A21: Gemittelte Konfusionsmatrizen der wortbasierten Modelle LSTM (oben links), 3xDense (oben rechts), CNN-LSTM (unten links) und BiLSTM (unten rechts) aus jeweils 20 Durchläufen.

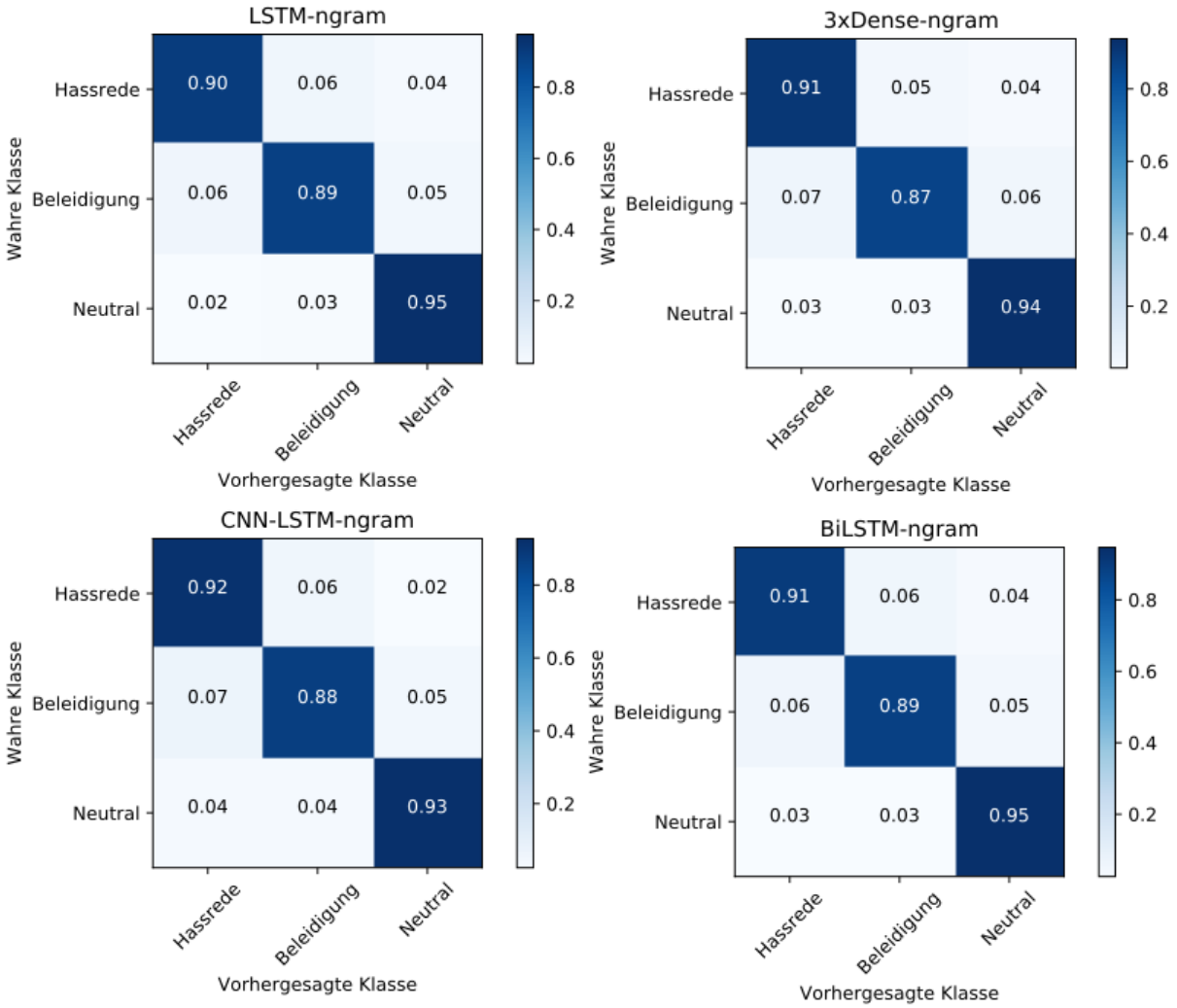

Abbildung A22: Gemittelte Konfusionsmatrizen der N-Gramm-basierten Modelle LSTM-ngram (oben links), 3xDense-ngram (oben rechts), CNN-LSTM-ngram (unten links) und BiLSTM-ngram (unten rechts) aus jeweils 20 Durchläufen.

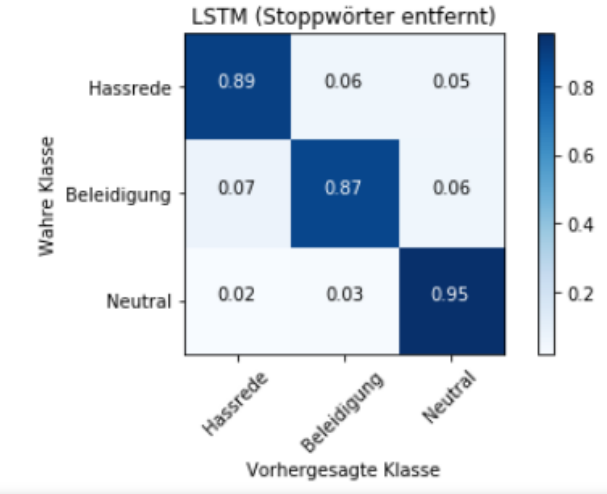

Abbildung A23: Konfusionsmatrix bei Entfernen der Stoppwörter am Beispiel eines Durchlaufs des wortbasierten LSTM-Modells.

A3 Vorhersagesicherheiten bezüglich Kapitel 6

	Hassrede	Beleidigung	Neutral
Schwarzer Humor	0,936	0,915	0,985
Rassismus	0,999	0,902	0,921
Sexismus	0,854	0,853	0,958

Tabelle A31: Vorhersagesicherheiten des Modells „LSTM-ngram“ für alle Klassen je Thema.

	Hassrede	Beleidigung	Neutral
Schwarzer Humor	0,959	0,874	0,939
Rassismus	0,784	-	0,810
Sexismus	0,760	0,788	0,987

Tabelle A32: Vorhersagesicherheiten des Modells „Fine-tuned BERT“ für alle Klassen je Thema.

A4 Hauptkomponentenanalyse der BERT Embeddings

Abbildung A41: PCA der BERT Embeddings einiger, im selben Kontext stehender Wörter.

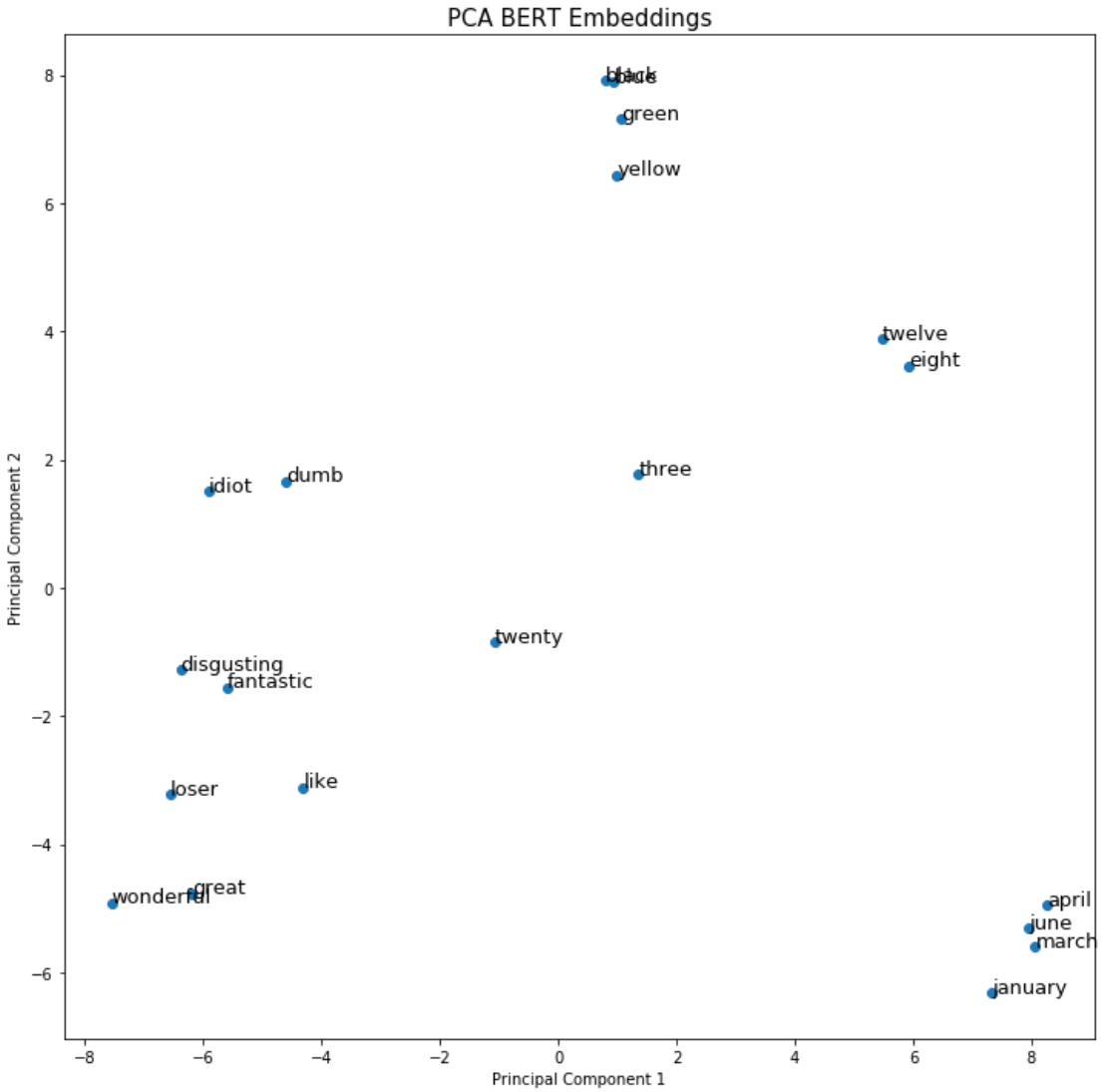

Abbildung A42: PCA der BERT Embeddings einiger, in kurzen, sinnvollen Kontexten stehender Wörter.

Eigenständigkeitserklärung

Ich versichere, dass ich die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sämtliche wörtlichen oder sinngemäßen Übernahmen und Zitate sind kenntlich gemacht und nachgewiesen.
Ferner versichere ich, dass das Thema dieser Arbeit nicht identisch ist mit dem Thema einer von mir bereits für eine andere Prüfung eingereichten Arbeit.

Ich erkläre weiterhin, dass ich die Arbeit nicht bereits an einer anderen Hochschule als Prüfungsleistung eingereicht habe.

_____

Datum, Unterschrift