Hate Speech behandeln: Diagnosewerkzeuge aus der Computerlinguistik

Johannes Schäfer

1 Einführung

Hate Speech ist beim Konsum von sozialen Medien mittlerweile allgegenwärtig. Die hohe Geschwindigkeit, mit der Millionen Nutzer:innen neue Nachrichten produzieren, hat zur Folge, dass diese nicht manuell gesichtet werden können, bevor sie öffentlich zugänglich als Einträge für alle sichtbar werden (so berichtet @raffi, 2013 zum Beispiel von 500 Mio. eingehenden *Tweets* pro Tag). Dadurch ist es leicht möglich, dass problematische Inhalte, insbesondere hier auch Hassbotschaften, scheinbar ungehindert verbreitet werden können. Dies hat auch seinen Ursprung im generellen Aufbau der Kommunikation online. Hier gibt es für verschiedene Bedürfnisse von Nutzer:innen vielfältige Angebote, die im Rahmen der technischen Möglichkeiten diverse Kommunikationskanäle öffnen.

Das Problem Hate Speech lässt sich zusätzlich besser verstehen, wenn man den Unterschied in der Kommunikation online im Vergleich zum Erstellen von Äußerungen offline betrachtet. So könnte man zum Beispiel das Verfassen von Beiträgen in einem sozialen Netzwerk dem Schreiben von Leserbriefen in einem Nachrichtenmagazin gegenüberstellen. Hier hat die Kommunikation online zunächst nur die Voraussetzung, dass, sofern bei Nutzer:innen die technische Ausrüstung vorhanden ist, ein Nutzerkonto erstellt wurde, welches je nach Plattform jedoch auch teils mit unüberprüften Falschangaben versehen werden kann oder gar von vornherein weitestgehend anonymisiert gehalten ist. Das Anschreiben eines Nachrichtenmagazins scheint erheblich aufwändiger und persönlicher, wobei typischerweise vor

J. Schäfer (✉)
Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim, Hildesheim, Deutschland
E-mail: johannes.schaefer@uni-hildesheim.de

S. Jaki und S. Steiger (Hrsg.), *Digitale Hate Speech*,
https://doi.org/10.1007/978-3-662-65964-9_5

dem Erscheinen eines Leserbriefs noch eine Überprüfung des Inhalts stattfindet. Bei sozialen Medien hingegen gibt es meist keine Redaktion, die Nachrichten sichtet, bevor sie veröffentlicht werden. Hier bleibt es üblicherweise anderen Nutzer:innen überlassen, unpassende Inhalte zu melden, wodurch zunächst nur die Betreiber der Plattform auf diese aufmerksam gemacht werden.

Fortschritte in der Kommunikation online scheinen daher gesellschaftlichen Regeln des Miteinanders zu widersprechen. Beschäftigt man sich gezielt mit dem Phänomen Hate Speech, möchte man meinen, sich in einer Pandemie zu befinden, in der Hass wie ein Virus verbreitet wird, um eigene Agenden zu verfolgen. Soziale Medien machen es dem Populismus leicht, schnell viele Menschen ohne großen Aufwand zu erreichen, wie Engesser et al. (2017) zeigen. Nichts eint dabei so sehr wie die gemeinsame Ablehnung, bis hin zum Hass, gewisser Denkweisen, Personen(-gruppen) oder Dinge.

Eine Regulierung des Inhalts, dargestellt auf öffentlichen Webseiten, ist unabdingbar. So sind auch die Betreiber von Plattformen der sozialen Medien in der Verantwortung, da sie den Inhalt, der von Nutzer:innen erstellt wird, für jedermann zugänglich machen; diese Verantwortlichkeit besteht auch nach der Erstveröffentlichung fort (Gillespie, 2018). Eine weiterführende Diskussion der Regulierungsproblematik geben Schünemann und Steiger in diesem Band. Es gilt jedenfalls, geeignete Methoden einzusetzen, um den richtigen Grad einer möglichst unbeschränkten Kommunikation mit all den technischen Möglichkeiten des Internets zu ermöglichen, gleichzeitig jedoch mit den gesellschaftlichen – zum Teil auch gesetzlichen – Vorschriften und Regeln konform zu bleiben.

An dieser Stelle kommt die Computerlinguistik ins Spiel, welche Methoden zur vorrangig automatischen Verarbeitung von sprachlichem Material erforscht. Das Phänomen Hate Speech findet sich in Nachrichten in sozialen Medien, welche man als sprachliche Äußerungen beschreiben kann, meist geäußert in Form von Text, jedoch manchmal auch multimodal (wie es beispielsweise Kiela et al., 2020 oder De Smedt & Jaki, 2018 untersuchen). Hier ist nun speziell die Analyse und Erkennung von relevanten Nachrichten gefragt, in Anwendungen für Betreiber, Autor:innen und Konsument:innen von sozialen Medien.

Am wichtigsten sind wohl die Anwendungen für Konsument:innen: aus gesellschaftspolitischer Sicht zum Schutz der Bevölkerung vor illegalem Inhalt und zur Verhinderung von dessen Weiterverbreitung. Hierbei geht es also um eine Regulierung des Konversationsstils und -inhalts. In manchen Fällen mag es auch ausreichen, Leser:innen eine Warnung anzuzeigen, zum Beispiel bei Behauptungen mit mangelnden Fakten, also möglichen Falschnachrichten. Forschungen dazu präsentieren beispielsweise Shu et al. (2017) und Hardalov et al. (2016).

Anwendungen für Betreiber betreffen primär das Blockieren von inakzeptablem Inhalt, worunter auch Hate Speech fällt. Dies ist motiviert durch gesetzliche Vorschriften, kann aber auch nach eigenen Regeln der Betreiber erfolgen. Ein Beispiel für so eine plattformspezifische Content-Regulierung sind Produktbewertungen auf *Amazon,* bei denen es nicht erlaubt ist, die Qualität der Lieferung miteinzubeziehen. Versuchen Nutzer:innen trotzdem in seiner Rezension darüber zu schreiben, kann es dazu kommen, dass eine Warnmeldung erscheint, dass der Kommentar nicht akzeptiert werden kann. Dahinter steckt ein automatisches (computerlinguistisches) System, welches den Inhalt von Nachrichten bezüglich der Plattformrichtlinien analysiert.

Das obige Beispiel kann auch als eine Anwendung für Autor:innen verstanden werden, wobei diese Kategorie in der Praxis selten integriert ist. Eine Anwendung für (die Interaktion mit) Autor:innen würde über eine Rechtschreibkorrektur, welche mittlerweile weit verbreitet angeboten wird, hinausgehen und gezielte automatische Analysen auch zur Bedeutung des Inhalts von Nachrichten einschließen. Bei Grenzfällen zu Hate Speech wären Warnungen denkbar, dass Nachrichten zum Beispiel falsch verstanden werden könnten, doppeldeutige oder eventuell in einem Kontext verletzende Begriffe beinhalten könnten. Damit könnten Nutzer:innen in der Rolle von Autor:innen auf ein mögliches Fehlverhalten hingewiesen werden. Bei Fällen von Hate Speech stellt sich allerdings die Frage, ob deren Autor:innen sich dessen nicht sowieso meist bewusst sind, was die Wirkung eines Warnsystems in Frage stellt.

Im Fokus, sowohl bei der Nachfrage nach Anwendungen als auch in der computerlinguistischen Forschung zu Hate Speech, sind sicherlich Methoden zu deren automatischer Erkennung. Einen Überblick zu dieser Problematik geben Schmidt und Wiegand (2017). In diesem Anwendungsfall geht es für ein System darum, für jede gegebene Nachricht zu entscheiden, ob sie als Hate Speech kategorisiert werden könnte, oder ob sie komplett aus unkritischem Inhalt besteht. Dabei besteht eine große Schwierigkeit darin, präzise vorab zu definieren, wie diese Entscheidung zu treffen ist. In diesem Artikel möchte ich dieses Definitionsproblem allerdings nicht tiefer diskutieren, da dies nicht eine speziell computerlinguistische Aufgabe ist, sondern interdisziplinär beantwortet werden muss. Ich möchte jedoch auf einige geeignete computerlinguistische Methoden hinweisen, die bei der Erforschung dieses Problems Anwendung finden. In Abb. 1 ist eine thematische Unterteilung der Erkennung von Hate Speech in der Computerlinguistik dargestellt.

Das übliche Vorgehen gestaltet sich so, dass im Rahmen von Projekten eine ungefähre Definition von Hate Speech vorgegeben wird, gegebenenfalls erklärt mit Hilfe von ein paar wenigen Beispielen. Im nächsten Schritt ist es die Aufgabe von mehreren Annotator:innen, eine größere Menge von empirisch gesammelten Daten

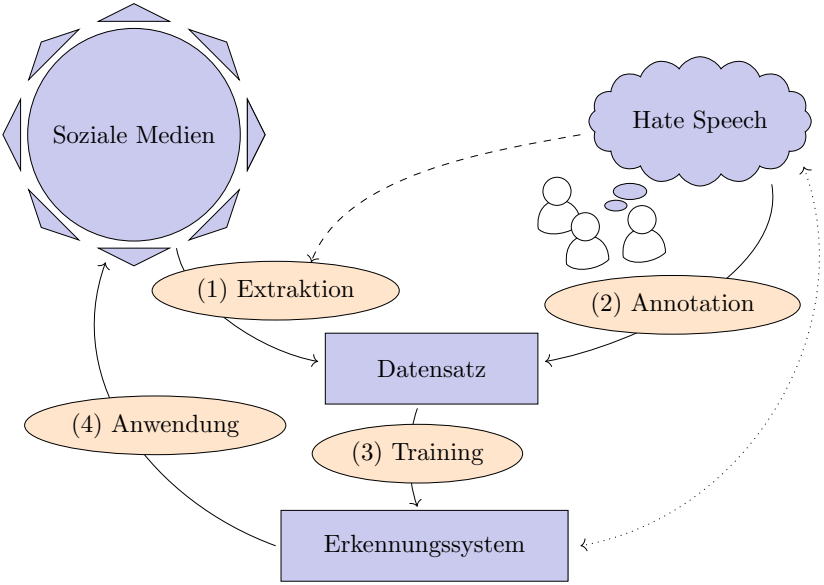

Abb. 1 Computerlinguistische Ressourcen und Prozesse bei der Erkennung von Hate Speech. Als Ressourcen sind hier dargestellt: die sozialen Medien links oben, das Wissen oder Vorstellungen über das Phänomen Hate Speech rechts oben, ein maschinelles Erkennungssystem unten und ein Datensatz mit Textbeispielen im Zentrum. Zu deren Verbindung sind vier Prozesse gekennzeichnet: (1) Die Extraktion von Hate-Speech-Beispieldaten aus den sozialen Medien, bei der die menschliche Vorstellung vom Phänomen auch eine Rolle spielt, wie ich weiterführend in Abschn. 2 diskutiere. (2) Die manuelle Annotation von Daten, bei der das Wissen über das Phänomen dem Datensatz explizit angereichert wird (auch dazu mehr in Abschn. 2). (3) Das Training des Erkennungssystems mit einem annotierten Datensatz. Hierbei ist es das Ziel, das System auf Aufgaben vorzubereiten, die es bei (4), der Anwendung auf neue Daten, ausführen soll. Dazu betrachte ich in Abschn. 3 mögliche Aufgaben im Detail und gebe in Abschn. 4 einen Überblick zu grundlegenden Methoden der Erkennung. Im Idealfall erhält man ein System, das die menschliche Vorstellung des Phänomens nachbildet, was in der Grafik mit einer gepunkteten Verbindung angedeutet ist

anhand der Definition zu annotieren. Das bedeutet, für jede einzelne Nachricht ist zu entscheiden, in welche Kategorie sie nach der Definition einzuordnen wäre. Schwierigkeiten bei der unvoreingenommenen Sammlung von Daten, die möglichst alle verschiedenartigen Vorkommen von Hate Speech gleichmäßig abdecken sollte, beleuchte ich detaillierter in Abschn. 2. Zum Erstellen solcher Ressourcen berichtet auch Mandl in diesem Band. Die teils auf subjektiven Interpretationen basierenden

Entscheidungen der Einzelpersonen können dadurch relativiert werden, dass die Markierungen mehrerer Annotator:innen vereint werden. Zusätzlich handelt es sich im Idealfall um eine so große Menge an Daten, dass einzelne Falschentscheidungen nicht schwer ins Gewicht fallen. Bei diesem Prozess werden diverse, computergestützte Werkzeuge zur manuellen Annotation von Texten eingesetzt, die den Prozess beschleunigen und vereinfachen; zum Beispiel gibt es hierfür die Werkzeuge Webanno (de Castilho et al., 2016) oder BRAT (Stenetorp et al., 2012). Als Ergebnis entsteht ein Goldstandard von manuell annotierten Texten, wobei eine bestmögliche menschliche Einschätzung angestrebt wird, welche nur zu einem geringen Grad fehlerhaft und inkonsistent sein soll. Automatische Systeme werden nun üblicherweise mit solchen Daten optimiert und auch evaluiert. Dabei wird eine der menschlichen Einschätzung entsprechende Leistung als obere Grenze für die Erkennungsrate von Systemen angenommen.

Im folgenden Abschnitt (Abschn. 2) möchte ich zunächst aus eigener Forschung zur empirischen Datensammlung für das Phänomen Hate Speech berichten und dabei auf weiter bestehende Probleme hinweisen. Für den darauf folgenden Abschnitt sehe ich es als gegeben an, dass von Menschen entschieden werden kann, welche Nachrichten als Hate Speech gelten und welche nicht. Aufgaben für computerlinguistische Werkzeuge stellen sich nun anhand dieser Eingabe. In Abschn. 3 diskutiere ich diverse Arten von Aufgaben für die Computerlinguistik zur Analyse von Hate Speech, die in der Forschung unterschiedlich prominent untersucht werden. Daran anschließend gebe ich in Abschn. 4 einen Überblick über grundlegende Methoden zur automatischen Erkennung von Hate Speech. Ich schließe in Abschn. 5 mit einem Ausblick, in dem ich zusammenfasse und diskutiere, welche Fragen für die zukünftige Forschung offen bleiben.

2 Empirische Datensammlung

Das Ziel einer Sammlung von Daten für das Phänomen Hate Speech formuliere ich hier so, dass es gilt, eine möglichst große und zum Phänomen passende Menge an Beispielen in einem Datensatz zu vereinen. Dabei bietet sich eine empirische Methodik an, also eine Sammlung von real existierenden, geäußerten Beiträgen von verschiedenen Autor:innen, um ein möglichst realitätsnahes Abbild des Phänomens zu erzielen. Eine Vorgabe von Beispielen aus sich heraus im Stil von „Armchair"-Linguist:innen würde für das komplexe Phänomen Hate Speech sicherlich nicht genügen, da es hier darum gehen soll, die gesamte Varietät des Phänomens abzudecken, das von verschiedenen Nutzer:innen in öffentlich zugänglichen sozialen Medien verbreitet wird. In der Forschung werden auch synthetisch erstellte Daten-

sätze verwendet, zum Beispiel zur Evaluierung von Systemen, wie der Datensatz HatemojiCheck von Kirk et al. (2021). Solche sind jedoch für eine grundlegende Phänomenerfassung ungeeignet. Angestrebt werden sollte hier also eine möglichst unvoreingenommene Sammlung von Beispielen, bei der unter anderem auch auf die Varietät im Bezug auf die Autor:innen von Nachrichten, thematische Bezüge und anderen sprachlichen Realisierungen des Phänomens geachtet werden sollte. Als Resultat soll ein in seiner Struktur konsistentes Sprachkorpus entstehen, welches automatisch weiterverarbeitet werden kann für aufbauende Analysen.

Den Zielen gegenüber stehen diverse Möglichkeiten, Grenzen beziehungsweise Einschränkungen, die die Online-Welt mit sich bringt. So muss bei der Wahl der Plattformen, die als Quellen für eine Datensammlung dienen sollen, miteinbezogen werden, inwiefern diese eine automatische Sammlung ermöglicht. Diese wird oft durch Filtermöglichkeiten unterstützt, welche beispielsweise eine Suche nach bestimmten Begriffen zulässt. Eine zufällige Sammlung von Beiträgen ist für das Phänomen Hate Speech unzureichend, da dieses an der Gesamtmenge von erstellten Beiträgen gemessen relativ selten ist und damit einen unrealistisch hohen manuellen Aufwand zur Beispielsuche erfordern würde. Daher muss abgewogen werden, inwiefern die automatische Suche durch die Wahl bestimmter Filter eingeschränkt werden kann. Einerseits sollte durch einen auf das Phänomen passenden Filter der Nachbearbeitungsaufwand reduziert werden, andererseits trotzdem eine möglichst hohe Diversität des Phänomens durchgelassen werden. Bei der Datensammlung müssen außerdem Rechte von Autor:innen betrachtet werden, im Bezug auf den Schutz der Daten von Nutzer:innen, welche als Metadaten im Korpus gespeichert werden können, und auch auf den Kopierschutz der Beiträge an sich. Dies spielt in der Forschung insbesondere für die freie Weiterveröffentlichung der Daten eine Rolle, denn hier soll ein Ziel sein, das Wissen, welches bei der Datensammlung und einer eventuell anschließenden Annotation in das Korpus fließt, anderen Forscher:innen weiterzugeben. Auch entscheidend für die Auswahl von Quellen ist der Aufwand für eine manuelle Weiterverarbeitung oder Annotation, weshalb meist der Fokus auf Kurznachrichten liegt.

Eine manuelle Annotation stellt den üblicherweise durchgeführten, ersten systematischen Schritt zur Analyse von empirisch gesammelten Daten dar. Dieser Schritt der manuellen Durchsicht ist hier meist notwendig, da, je nach Methode der Sammlung, nur ein ungewisser Teil der gesammelten Daten wirklich dem Phänomen zuzuschreiben ist. Auf jeden Fall gilt es, dies festzustellen, wobei der übrige Teil der Daten als Gegenbeispiele weiterverwendet werden kann. Für die manuelle Annotation von Äußerungen bezüglich Hate Speech stellt sich nun zunächst die Frage, was genau in den Daten annotiert werden soll. Hierzu ist eine Phänomendefinition nötig, die entweder vage sein kann, um die eigenen Vorstellungen der Annotator:innen zu

nutzen, oder möglichst präzise vorgegeben wird, um eine einheitliche Annotation zu erzielen. Ein Beispiel solcher Annotationsrichtlinien präsentieren Ruppenhofer et al. (2018). In Annotationsexperimenten, bei denen mehrere Annotator:innen die selben Äußerungen annotieren, kann zudem eine Auswertung der Annotation erfolgen. Dadurch können Aussagen getroffen werden über die Qualität der Annotation, aber auch bezüglich der Schwierigkeit der Aufgabe für menschliche Entscheider, woraus man oft eine obere Grenze für automatische Systeme ableitet.

Auf der Basis dieser Vorüberlegungen berichte ich nun beispielhaft aus einer Forschungsarbeit (Schäfer & Boguslu, 2021), in der wir den Prozess einer Datensammlung und Annotation vollzogen haben. In unserer Arbeit war eine wichtige Fragestellung, unter welchen Gesichtspunkten eine Hate-Speech-Nachricht illegal ist. Nachrichten, die gesetzeswidrig sind, müssen in Anwendungen gesondert gehandhabt werden. Sie werden in manchen Fällen nicht einfach nur gelöscht, sondern sollten gemeldet werden. Daher sollte auch hierfür eine Klassifikation angestrebt werden. Dabei haben wir die Unterscheidung von Hate Speech in legal vs. illegal zuzüglich zur Abgrenzung von sonstigen Nachrichten so realisiert, dass wir folgende drei Kategorien definiert haben:

1. (illegale) Hate Speech,
2. (legale) Offensive Language,
3. sonstige, neutrale Nachrichten.

Generell waren wir besonders an einer möglichst exakten Phänomendefinition interessiert, um dafür fundierte Merkmale entwickeln zu können, mit denen automatische Erkennungssysteme bei der Auswertung spezifischer überprüft werden können. Daher verwendeten wir Gesetzestexte und Gerichtsurteile (aus Deutschland) als Basis. Diese sollten als Quellen für Beispiele und damit die Definition des Phänomens Hate Speech dienen. Die Online-Suchportale von Gerichten zeigten sich jedoch als nicht ergiebig genug, um einen größeren Datensatz zum Thema illegale Hate Speech in den sozialen Medien zu gewinnen. So finden sich zwar passende Gerichtsprotokolle, allerdings erwähnen diese nur in seltenen Fällen spezielle Nachrichtenbeiträge oder stützen sich gar auf einzelne davon in der Urteilsbegründung. Dadurch lassen sich keine automatisch verarbeitbaren Beispiele des Phänomens sammeln. Die im Kontext interpretierten und bewerteten Passagen in Gerichtsurteilen passen nicht auf die reduzierten Daten, die später für Erkennungssysteme verarbeitet werden (welcher Struktur diese Daten sind, betrachte ich genauer in Abschn. 3). Um trotzdem der Grundmotivation zu folgen, verwendeten wir in der Forschungsarbeit die Gesetzestexte und Gerichtsurteile als Basis für unsere Annotationsrichtlinien und spezifizierten dadurch Unterkategorien von illegaler Hate

Speech. In einem zweiten Schritt wendeten wir diese Richtlinien auf neue, empirisch gesammelte Daten an. Hierbei führten wir eine Datensammlung auf Twitter durch, wobei wir nur deutsche Beiträge gesucht haben. Durch die Verwendung von Suchbegriffen fanden wir Beispiele des Phänomens Hate Speech. Um dabei eine möglichst unvoreingenommene Suche durchzuführen, wählten wir vorrangig Suchbegriffe, die nicht ausschließlich im Zusammenhang mit Hate Speech stehen. Zum Beispiel verwendeten wir keine Schimpfworte, sondern wählten Begriffe, die wir sowohl in Kontexten von Hate Speech, als auch in neutralen Kontexten vermuteten. Beispiele hierfür wären „Schwein“ oder „Polizei“, welche jeweils verschieden verwendet werden können, also nicht immer zu einer Interpretation als Hate Speech führen. Als Resultat produzierten wir einen Datensatz, der in verschiedenen Annotationsexperimenten ausgewertet wurde. Hierbei hat sich gezeigt, dass speziell mit dem Phänomen vertraute Annotator:innen eine höhere Übereinstimmung in den Bewertungen haben.

Bei der empirischen Datensammlung ist es wichtig, sich darüber im Klaren zu sein, inwiefern es Einschränkungen für das Resultat im Vergleich zur Realität gibt. Der gewonnene Datensatz ist immer nur ein Ausschnitt aus einer Sprechpraxis, jedoch sollte versucht werden, eine möglichst repräsentative Abbildung der Realität zu erreichen. Ein Nebeneffekt der empirischen Datensammlung sind Verzerrungen (Bias) der Daten verschiedenster Art. Wird der Datensatz nun zum Trainieren eines automatischen Systems verwendet, lernt dieses diese Verzerrung mit. Einen Bias bezüglich des Phänomens Hate Speech gibt es bei einer Suche, wie oben beschrieben, zunächst aufgrund der Wahl der Suchbegriffe: Im Ergebnis erscheinen nur Beiträge, die diese beinhalten. Daher sollte darauf geachtet werden, dass deren Vorkommen nicht übermäßig mit den Kategorien der Klassifikation korrelieren. Außerdem kann es zu Verzerrungen zu gewissen Themen (Topic-Bias) oder Identitätsbegriffen (Identity-Term-Bias) kommen, welche vom Suchzeitraum, von der Auswahl der Quellen/Plattform oder erneut von der Wahl der Suchbegriffe begünstigt sein können. Eine Diskussion dieser Problematik liefern zum Beispiel Davidson et al. (2019), indem sie hier speziell den Bias von Hate-Speech-Daten bezüglich der Nennung bestimmter Rassenbezeichnungen untersuchen. Abschließend lässt sich zur Datensammlung sagen, dass auch hier weitere Lösungen zu finden sind, um das Phänomen Hate Speech genauer beschreiben zu können. Trotz alledem lässt sich mit der dargestellten Methodik umfangreiches Material zum Phänomen Hate Speech sammeln, welches wir im folgenden Abschnitt als Grundlage nutzen möchten, um Aufgaben zur weiteren Analyse für die Computerlinguistik herauszuarbeiten.

3 Aufgaben für die Computerlinguistik

Generell untersucht die Computerlinguistik Methoden zur computergestützten Verarbeitung von linguistischem Material. Äußerungen von Hate Speech gelten im Allgemeinen als sprachliche Äußerungen, da sie vorrangig in Form von Text ausgedrückt werden. Es gibt jedoch auch multimodale Äußerungen, bei denen nur durch eine Kombination von Bild und Text eine Nachricht als Hassbotschaft interpretierbar wird (siehe auch Jaki in diesem Band). Das Medium, also ob es sich um eine schriftliche oder (audio-)visuelle Äußerung handelt, ist für die vorliegende Diskussion sekundär. Ich möchte hier auf sprachliche Äußerungen in schriftlicher Form fokussieren, da dies die wohl häufigste vorkommende Form von Hate Speech ausmacht.

Die größten Sammlungen computerlinguistischer Forschungen zum Thema Hate Speech finden statt in Form von Shared Tasks (Bosco et al., 2018; Wiegand et al., 2018; Zampieri et al., 2019; Basile et al., 2019; Struß et al., 2019; Mandl et al., 2020). Diese werden in einem Modus abgehalten, bei dem von einem Team von Organisator:innen diverse Gruppen von Forscher:innen eingeladen werden, um gemeinsam zu einem bestimmten Thema Beiträge einzubringen. Bei bisher veranstalteten Shared Tasks zu Hate Speech hat sich bewährt, dass ein vorgegebener, annotierter Datensatz für gemeinsame Untersuchungen bereitgestellt wird. Als Grundaufgabe steht üblicherweise die Erkennung von Hate Speech in Kurznachrichten im Vordergrund. Hier kann man die Erkennung von Hate Speech formal definieren als die binäre Klassifikation von Kurznachrichten bezüglich dessen, ob die jeweilige, aus einem Kontext gegriffene Nachricht eine Teiläußerung beinhaltet, die als Hate Speech interpretiert werden kann. Eine Formulierung als Regressionsproblem, das heißt das Messen der Stärke des Hasses in einer Nachricht, ist möglich, wie zum Beispiel in einer Arbeit von Ross et al. (2016), allerdings selten.

Wie schon oben angedeutet, werden bei der Aufgabenstellung aus vielen Gründen häufig mehrere Vereinfachungen getroffen. So werden Beiträge separat analysiert, also ohne den Kontext auf der Plattform, in dem sie geschrieben und später auch dargestellt werden, was das Problem vereinfacht. Aufgrund von Datenschutzrechten ist es außerdem oft nicht möglich, die Nachrichten mit den gesamten Metadaten zu speichern, weiterzuverbreiten und für die Forschungen zu nutzen. Daher ist es oft notwendig, dass Anonymisierungsverfahren angewandt werden, bei denen zum Beispiel Namen von Nutzer:innen und andere personenbezogene Daten gelöscht werden; eine Diskussion dieser Problematik wird beispielsweise von Townsend und Wallace (2018) präsentiert. Lösungsansätze, die mit solchen reduzierten Datensätzen arbeiten, trainiert und evaluiert werden, können daher nur diese verminderte Repräsentation der Realität als Basis nutzen.

Weitere Aufgabenstellungen untersuchen oftmals Unterkategorien von Hate Speech oder verwandte Kategorien wie Obszönitäten, abhängig von der jeweilig gewählten Definition von Hate Speech, wie zum Beispiel in der Arbeit von Ruppenhofer et al. (2018). Diese Aufgabe bezeichnet man in der Computerlinguistik als Klassifizierungsproblem mit mehreren Klassen. Hierbei ist die Aufgabe, für jede gegebene Nachricht eine Kategorie aus einer Menge auszuwählen, welche vom Werkzeug passend zu annotieren ist. Eine Annotation von mehreren Kategorien gleichzeitig für eine Nachricht ist hingegen äußerst selten, da dies die Aufgabe ungemein verkompliziert. Dies wäre allerdings wohl realitätsnaher, wenn man betrachtet, dass zum Beispiel ein Beitrag mehrere Teiläußerungen und auch ganze Sätze beinhalten kann, von denen nur ein Teil als Hate Speech gelten könnte. Demnach findet man auch Äußerungen, von denen verschiedene Teiläußerungen unterschiedlichen Hate-Speech-Kategorien zuzuordnen wären. In solchen Fällen müsste die Gesamtäußerung dann mit mehreren Kategorien annotiert werden. Grundsätzlich wird in der Forschung jedoch primär die Frage gestellt, ob der Beitrag auf einer Plattform der sozialen Medien als Ganzes als Hate Speech einzustufen wäre. Ich möchte hier explizit darauf hinweisen, dass die üblicherweise untersuchte Erkennung von Hate Speech daher also nicht als Satzklassifikation bezeichnet werden kann, was eine häufige computerlinguistische Aufgabe ist, sondern allgemeiner als eine Klassifikation von Kurznachrichten.

Eine computerlinguistische Analyse von Hate-Speech-Nachrichten könnte im Detail wie folgt durchgeführt werden: Im ersten Schritt untersucht man die Wortebene und damit die Bedeutung der einzelnen Wörter, deren Semantik. Hierbei wäre es zum Beispiel möglich, durch die Erfassung in Lexika beleidigende Wörter oder allgemeiner Wörter mit einem negativen Sentiment zu identifizieren. Auch andere Charakteristika von Hate Speech lassen sich damit abdecken, wie zum Beispiel das Target (Ziel) der Äußerung. Viele Formen von Hate Speech, wie Beleidigungen oder Äußerungen der Volksverhetzung, sind zielgerichtet, das heißt sie werden geäußert, um eine spezielle Person oder Gruppe anzugreifen. Diese wird oft explizit genannt und kann daher auf Wortebene erkannt werden, wie beispielsweise ElSherief et al. (2018) zeigen.

Wortkomponenten können mit morphologischen Analysewerkzeugen untersucht werden, wie zum Beispiel mit dem Werkzeug SMOR (Schmid et al., 2004). Hiermit können zum Beispiel Komposita zerlegt werden, wozu zum Beispiel Cap (2014) Methoden präsentiert. Damit können Hate-Speech-Nachrichten, in denen nur Teile von Wörtern beleidigende Ausdrücke sind, speziell untersucht werden. Ein Rückfall auf die Zerteilung von Wörtern durch n-Gramme von Buchstaben, also Sequenzen von aufeinanderfolgenden Buchstaben der Länge n, ist eine weitere, allerdings linguistisch weniger motivierte Möglichkeit. Auch ist es manchmal sinnvoll, Wörter

auf ihre Grundformen zurückzuführen, um eine erhöhte Verallgemeinerungsfähigkeit in der weiteren Analyse zu ermöglichen. Hierfür bietet die Computerlinguistik Lemmatisierungswerkzeuge an, wie zum Beispiel die Systeme von Müller et al. (2015) und Bergmanis und Goldwater (2018).

Üblicherweise bedient sich die semantische Analyse von Wörtern sogenannter Word Embeddings, die numerische Bedeutungsrepräsentationen sind. Eine häufig genutzte Methode zu deren Training ist Word2Vec (Mikolov et al., 2013). Diese Embeddings können auf großen Datenmengen vortrainiert werden. Nach dem Prinzip der distributionellen Semantik (Harris, 1954; Firth, 1957) vereinen sie dadurch präzise in sich diverse semantische Eigenschaften von Wörtern. Daher sind Word Embeddings gut geeignet für eine automatische Weiterverarbeitung.

Um nun weiter die Bedeutungen einzelner Wörter in einer Äußerung zu kombinieren, erfolgt die Analyse im nächsten Schritt auf Satzebene. Hierbei wird die Anordnung der Wörter, also die Struktur im Satz, die Syntax, betrachtet. Diese Betrachtung erlaubt es zum Beispiel oft auch, mehrdeutige Wörter zu disambiguieren, weil Wörter aus dem unmittelbaren Kontext Hinweise auf die gemeinte Bedeutung liefern. Mittels automatischer syntaktischer Analysewerkzeuge, sogenannter Parser (beispielsweise Bohnet, 2010), ist es außerdem möglich, die Bedeutung von Wortkombinationen in Hate-Speech-Nachrichten zu analysieren. Zum Beispiel ergibt sich, wenn mehrere Personennamen in einer Beleidigung vorkommen, erst aus der Satzstruktur, welche Person beleidigt werden soll.

Der logische nächste Schritt, nachdem eine semantische und syntaktische Analyse auf Wort- und Satzebene durchgeführt wurde, wäre eine pragmatische Diskursanalyse, bei der die Bedeutung im Kontext interpretiert wird, wie es beispielsweise von Assimakopoulos et al. (2017) präsentiert wird. Viele Forschungen zu Hate Speech jedoch wagen diesen Schritt noch nicht, da er die Aufgabe deutlich komplexer macht. Außerdem sind auch oft, wie oben bereits erwähnt, Zusatzinformationen aus dem Kontext nicht abrufbar, wie Metadaten oder andere Nachrichten aus dem direkten Diskurs auf der Plattform.

Ich möchte an dieser Stelle darauf hinweisen, dass die so beschriebene Methodik nach der Analyse auf Satzebene nicht immer komplette Nachrichten, die bezüglich Hate Speech zu analysieren sind, abdecken würde, da diese aus mehreren Sätzen, oder zusätzlichen Charakteristika von sozialen Medien, wie Emojis, bestehen können. Um diese Diskrepanz zu umgehen, wird der Schritt häufig vereinfacht. So betrachtet man oft einen Beitrag als ein einziges, komplettes Element, also eine Äußerung, die nicht unterteilt wird, und lässt die syntaktische Detailanalyse außen vor. Auch ist die Aufgabe meist so formuliert, dass Systeme nur erkennen sollen, ob ein Beitrag Hate Speech enthält und nicht, welcher Teil des Beitrags genau Hate Speech ausmacht. Auch wird oft angenommen, dass die zu analysierenden Beiträge

Kurznachrichten sind und damit so kurz, dass sich viele Analysemethoden für Sätze direkt auf sie übertragen lassen.

4 Methoden zur Erkennung

In diesem Abschnitt diskutiere ich nun grundlegende Methoden für die Erkennung von Hate Speech. Die Aufgabe für diese Methoden habe ich oben als binäre Kurztextklassifikation bezeichnet, mit den zwei disjunkten Klassen, die es zu unterscheiden gilt: Hate Speech und zulässiger/neutraler Inhalt. Ich nehme diese Kategorisierung nun als gegeben an, in Form von Richtlinien und von manuell vorannotierten Daten. Lösungsansätze werde ich im Folgenden mit der groben Einteilung in drei Kategorien diskutieren: Lexikonbasierte Erkennung, Methoden auf der Basis von erklärbaren maschinellen Lernsystemen und Methoden auf der Basis von neuronalen Netzwerken.

4.1 Lexikonbasierte Erkennungsansätze

Lexikonbasierte Erkennungsansätze verwenden ein vorab erstelltes Lexikon, welches Wörter und Wortverbindungen enthält, die als Merkmale für die Erkennung von Hate Speech hilfreich sind. Beispielanwendungen solcher Methoden werden von Spertus (1997), Gitari et al. (2015) und Del Vigna et al. (2017) gezeigt. Verwendete Lexika müssen dabei nicht nur aus Begriffen bestehen, die eindeutig über die Kategorisierung einer Äußerung entscheiden, zum Beispiel eindeutig als Beleidigung zu verstehende Begriffe. Genauso können auch Lexika für andere Charakteristika von Hate Speech genutzt werden, zum Beispiel Sammlungen von typischen Targets von Hate Speech können zusätzliche Merkmale zur Identifikation von potentiell relevanten Äußerungen liefern. Auch können Lexika anhand von gegebenen Trainingsdaten trainiert werden, wie von Razavi et al. (2010) gezeigt. Hierbei können zum Beispiel Gewichte maschinell gelernt werden, also einzelne Einträge der Lexika als in höherem bzw. geringerem Maß ausschlaggebend für die Klassifikation markiert werden. Schließlich gibt es beim Einsatz von Lexika auch viele Wahlmöglichkeiten, wie diese verwendet werden. Eine einfache Klassifikation beim Vorkommen eines einzelnen Begriffs aus dem Lexikon scheint zu trivial, um das Problem zu lösen, auch deswegen, weil eine solche Methodik leicht zu umgehen wäre. Vielmehr erscheint es sinnvoller, Regeln aufzustellen für das kombinierte Auftreten mehrerer Begriffe oder mehrerer unterschiedlicher Merkmalskategorien, die sich aus Lexika ableiten lassen, um verschiedenartige Äußerungen von Hate Speech erkennen zu können.

Wenngleich Entscheidungen von lexikonbasierten Systemen immer direkt nachvollziehbar sind, haben sie auch diverse Nachteile. Ein Problem ist, dass das Erstellen von guten Lexika einen extrem hohen manuellen Arbeitsaufwand fordert und das Ergebnis bei der Unendlichkeit der Sprache auch nie vollständig sein kann. So ist es immer nötig, Lexika regelmäßig zu erweitern und zu pflegen, da sich der Sprachgebrauch und damit auch Hate Speech im ständigen Wandel befindet. Ein reichhaltiges Lexikon präsentieren beispielsweise De Smedt et al. (2020). Allein nur durch regelbasierte Systeme mit Lexika ist die Anpassbarkeit an ein komplexes Phänomen, wozu eine umfassende Definition von Hate Speech unbestreitbar zählt, schwer oder gar unmöglich zufriedenstellend zu realisieren. Es scheint unmöglich, alle Regeln für das Phänomen Hate Speech aufzustellen, wenn selbst Menschen große Schwierigkeiten haben, eine eindeutige Definition für die gesamte Reichweite des Begriffs aufzustellen.

4.2 Erkennungsmethoden auf der Basis von erklärbaren maschinellen Lernsystemen

Maschinelle Lernsysteme hingegen verwenden üblicherweise Methoden des überwachten Lernens, die ihre Entscheidungen für bisher noch nicht gesehene Daten hauptsächlich anhand von Ähnlichkeiten von Merkmalen im Vergleich zu Trainingsdaten treffen. Sie werden als erklärbar bezeichnet, wenn der Pfad der Entscheidungsfindung vom System ausgegeben werden kann und dieser für Menschen relativ einfach nachvollziehbar dargestellt werden kann, wie von Doran et al. (2018) beschrieben. Dies wäre zum Beispiel der Fall, wenn ein System die Regeln, nach denen es seine Entscheidung getroffen hat, ausgeben würde, also zum Beispiel die speziell gefundenen Merkmale in einer Äußerung und deren Gewichte. Auch lexikonbasierte Systeme könnte man dazu zählen, wenn zum Beispiel, wie oben erwähnt, Gewichte für Einträge aus den Trainingsdaten gelernt wurden.

Ich möchte zu dieser Kategorie hier jedoch hauptsächlich Systeme zählen, bei denen das maschinelle Lernen im Fokus steht, auch wenn nicht notwendigerweise Wortlisten oder Lexika verwendet werden. Hierbei gestaltet sich der Prozess üblicherweise zweistufig. Zunächst wird definiert, wie aus einer Eingabe (hier eine Kurznachricht) Merkmale extrahiert werden. Dabei können auch anfänglich eventuell abstrakt erscheinende Merkmale, wie zum Beispiel die Anzahl der Wörter in der Nachricht oder die Anzahl der verwendeten Satzzeichen, im Gesamtsystem einen gewinnbringenden Effekt haben. In der Forschung wurden außerdem auch computerlinguistische Merkmale aus semantischen und syntaktischen Analysen der Eingabe sowie die Verwendung von Wortlisten oder Lexika erprobt. Nach

der Merkmalsdefinition wird in der zweiten Stufe eine Lernmethode angewandt, bei der Gewichte und Kombinationen der Merkmale anhand von Trainingsdaten gelernt werden, wobei der Entscheidungsprozess dieser Systeme meist noch nachvollziehbar bleibt. Solche Systeme implementieren Algorithmen auf der Basis von beispielsweise Entscheidungsbäumen oder Support Vector Machines. Als Ergebnis erhält man ein trainiertes Modell, mit dem das System eine Vorhersage für ungesehene Daten anhand nachvollziehbarer Extraktion von Merkmalen, deren Gewichtung und deren Kombination treffen kann. Diverse solcher Systeme wurden zum Beispiel von Alfina et al. (2017), MacAvaney et al. (2019) und Rother und Rettberg (2019) angewandt. Ein Nachteil dieser Methoden ist, ähnlich wie bei lexikonbasierten Ansätzen, dass es sehr aufwändig sein kann, sinnvolle Merkmale für komplexe Probleme vorzugeben. Hierbei ist eine hohe Quantität der Merkmale entscheidend, um möglichst die gesamte Varietät des Phänomens zu erfassen. Zusätzlich gilt es aber auch, Merkmale mit einer hohen Qualität bezüglich deren Nutzen zur Erkennung des Phänomens zu finden, da es dem System sonst nicht gelingt, damit vernünftig zu lernen. Die angesprochenen Methoden haben allerdings den Vorteil, dass nicht geeignete Merkmale auch im Trainingsprozess identifiziert werden können. Außerdem ist die Anpassbarkeit an Veränderungen im Phänomen üblicherweise höher als bei rein lexikonbasierten Ansätzen.

4.3 Erkennungsmethoden auf der Basis von neuronalen Netzwerken

Die in den letzten Jahren populär genutzten Methoden auf der Basis von neuronalen Netzwerken unterscheiden sich von den bereits diskutierten Ansätzen insofern, dass ihre Entscheidungsprozesse nicht trivial nachvollziehbar sind. Sie bedienen sich beim maschinellen Lernprozess vielmehr einer äußerst komplexen und mehrfachen Kombination von Merkmalen. Üblicherweise wird hierbei so vorgegangen, dass Eingabedaten in einem möglichst unverarbeiteten Zustand dem Netzwerk zur Verfügung gestellt werden. Durch die manuelle Definition von Merkmalen, wie sie bei den obigen Systemen angewandt wird, wird eine Vorauswahl getroffen, die Teilinformationen herausfiltert. Dies wird nur angewandt, wenn es nötig ist, die Informationsmenge der Eingabe zu reduzieren. Neuronale Netzwerke hingegen sind dafür konzipiert, mit einer großen Menge an Eingabewerten zu arbeiten und wählen daher einen anderen Weg, indem sie direkt mit der unverarbeiteten Eingabe arbeiten. Sie haben den großen Vorteil, dass sie von der Architektur her so aufgebaut sind, dass sie nicht nur die Gewichtung und Kombinationen von Merkmalen selbst lernen können, sondern auch die Merkmalsextraktion. Außerdem sind die

Kombinationsmöglichkeiten von Merkmalen ungemein komplex. Daher ist es für einen Menschen nicht möglich, den Entscheidungsprozess nachzuvollziehen. Das Training des Netzwerks erfolgt im maschinellen Lernprozess, allerdings nur auf der Basis der gegebenen Daten. Das bedeutet, dass das Netzwerk sein Wissen über das Phänomen nur aus den Trainingsdaten zieht und dadurch seine Entscheidungen begründet werden können. Neuronale Netzwerke sind jedoch bekannt dafür, überaus hungrig nach Trainingsdaten zu sein, da sie mehrere Millionen von Gewichten lernen müssen. Im Bezug auf die Leistungsfähigkeit haben sich in den letzten Jahren neuronale Netzwerke in mehreren Anwendungen der Computerlinguistik als führend erwiesen. Richtungsweisend dabei ist die Methode im System BERT (Devlin et al., 2019), welches auf einem Transformer-Modell (Vaswani et al., 2017) basiert.

In Systemen werden zur Erkennung von Hate Speech mit neuronalen Netzwerken diverse Strukturen verwendet, die meist auf den Word Embeddings der Eingabe aufbauen. So rechnet ein Multilayer Perceptron (MLP) direkt mit allen beliebigen gewichteten Kombinationen der Eingabemerkmale. Bei einem Convolutional Neural Network (CNN) hingegen werden schrittweise aufeinanderfolgende Gruppen von Eingabemerkmalen ausgewählt und gewichtet kombiniert. Dies wird meist so konfiguriert, dass damit immer kurze Wortsequenzen (n-Gramme von Worten) betrachtet werden, was gut auf die Aufgabe der Erkennung von Hate Speech zu passen scheint, da es hierbei ja oft der Fall ist, dass nur eine Teilsequenz einer Äußerung Ausdruck von Hass ist. Bei einem Recurrent Neural Network (RNN) wird hingegen die Eingabe als eine komplette Sequenz verarbeitet, was längere Strukturen direkter in Kombination analysiert. Verschiedene neuronale Architekturen wurden zum Beispiel angewandt von Gröndahl et al. (2018), Founta et al. (2019) und Schäfer (2018). Richtungsweisende Forschungen der letzten Jahre haben darauf fokussiert, die Enkodierungsmethode auf Basis der Embeddings der Eingabe zu verbessern. So ist es mit dem Attention Mechanism (Bahdanau et al., 2015) möglich, für jedes Wort eine Gewichtung zu lernen, die aussagt, wie sehr es in der Bedeutung für eine Anwendung von seinen Kontextworten beeinflusst wird. Dies ist ein wichtiger Baustein von hochperformanten Transformer-Modellen. Mit diesen ist es möglich, kontextabhängige Word Embeddings vorzutrainieren (Devlin et al., Devlin et al. 2019), welche in vielen Anwendungen, wie auch bei der Erkennung von Hate Speech, mit die besten Ergebnisse bei statistischen Evaluierungen liefern. Anwendungen solcher Methoden auf die Hate-Speech-Erkennung präsentieren beispielsweise Risch et al. (2019), Paraschiv und Cercel (2019), Liu et al. (2019) und Wiedemann et al. (2020).

5 Ausblick

Methoden zur computerlinguistischen Behandlung von Hate Speech zeichnen sich meist dadurch aus, dass sie auf Trainingsdaten basieren und versuchen, daraus wiederkehrende Merkmalskombinationen zu lernen. Selbst neuronale Netzwerke basieren grundständig auf manuell annotierten Daten und implementieren keine unkontrollierte künstliche oder irgendwie kreative Intelligenz. Allerdings möchte ich damit schließen, dass neuste Systeme dennoch intelligent dabei vorgehen, durch Auswahl, Gewichtung und Kombination von Merkmalen zu lernen, womit man gegen Hate Speech vorgehen kann. Automatische Werkzeuge können dazu einen erheblichen Beitrag leisten, ersetzen jedoch die manuelle Moderation von Inhalten nicht vollständig.

Zusammenfassend lässt sich sagen, dass bei der Erkennung von Hate Speech in der Forschung oftmals detaillierte Analyseergebnisse computerlinguistischer Verfahren bislang nur begrenzt genutzt werden. Gerade die Extraktion und präzise Analyse von Teiläußerungen von als Hate Speech zu klassifizierenden Nachrichten durch computerlinguistische Werkzeuge scheint untererforscht. Vielversprechende Ergebnisse könnten sich jedoch in der Kombination diverser Methoden zeigen.

Literatur

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 international conference on advanced computer science and information systems (ICACSIS)* (S. 233–238). IEEE. https://doi.org/10.1109/ICACSIS.2017.8355039.

Assimakopoulos, S., Baider, F. H., & Millar, S. (2017). *Online hate speech in the European Union: A discourse-analytic perspective*. Springer Nature. https://doi.org/10.1007/978-3-319-72604-5.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Hrsg.), *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings*. https://doi.org/10.48550/arXiv.1409.0473.

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *13th international workshop on semantic evaluation* (S. 54–63). Association for Computational Linguistics.https://doi.org/10.18653/v1/S19-2007.

Bergmanis, T., & Goldwater, S. (2018). Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (S. 1391–1400). https://doi.org/10.18653/v1/n18-1126.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)* (S. 89–97). https://aclanthology.org/C10-1011.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. In *EVALITA 2018-sixth evaluation campaign of natural language processing and speech tools for Italian* (Bd. 2263, S. 1–9). CEUR. http://ceur-ws.org/Vol-2263/paper010.pdf.

Cap, F. (2014). *Morphological processing of compounds for statistical machine translation*. Dissertation, Universität Stuttgart. https://doi.org/10.18419/opus-3474.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the third workshop on abusive language online, Florenz, Italien* (S. 25–35). https://doi.org/10.18653/v1/W19-3504.

de Castilho, R. E., Mujdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH)* (S. 76–84). https://aclanthology.org/W16-4011/.

De Smedt, T., & Jaki, S. (2018). The Polly corpus: Online political debate in Germany. In *Proceedings of the 6th conference on computer-mediated communication (CMC) and social media corpora (CMC-corpora 2018)* (S. 33–36). https://doc.anet.be/docman/docman.phtml?file=.irua.de0576.153416.pdf#page=39.

De Smedt, T., Voué, P., Jaki, S., Röttcher, M., & De Pauw, G. (2020). Profanity & offensive words (POW): Multilingual fine-grained lexicons for hate speech. In *Textgain technical reports*. ISSN 2684-4842. https://www.textgain.com/portfolio/profanity-offensive-words/.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)* (S. 86–95). http://ceur-ws.org/Vol-1816/paper-09.pdf.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (S. 4171–4186). https://doi.org/10.18653/v1/N19-1423.

Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable ai really mean? A new conceptualization of perspectives. In *Proceedings of the first international workshop on comprehensibility and explanation in AI and ML 2017*. http://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the twelfth international AAAI conference on web and social media (ICWSM 2018)* (S. 42–51). https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17910.

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2017). Populism and social media: How politicians spread a fragmented ideology. *Information, Communication & Society, 20*(8), 1109–1126. https://doi.org/10.1080/1369118X.2016.1207697

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis,* 1–32.

Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science* (S. 105–114). https://doi.org/10.1145/3292522.3326028.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://yalebooks.yale.edu/book/9780300261431/custodians-of-the-internet/.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is „love": Evading hate-speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security (AISec'18)* (S. 2–12). https://doi.org/10.1145/3270101.3270103.

Hardalov, M., Koychev, I., & Nakov, P. (2016). In search of credible news. In *Artificial intelligence: Methodology, systems, and applications (AIMSA 2016)* (S. 172–180). Springer International Publishing. https://doi.org/10.1007/978-3-319-44748-3_17.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in neural information processing systems 33 (NeurIPS 2020)* (S. 2611–2624). https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html.

Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., & Hale, S. A. (2021). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. arXiv preprint arXiv:2108.05921. https://doi.org/10.48550/arXiv.2108.05921.

Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation, Minneapolis, Minnesota, USA* (S. 87–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2011.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS One, 14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Mandl, T., Modha, S., Shahi, G. K., Jaiswal, A. K., Nandini, D., Patel, D., Majumder, P., & Schäfer, J. (2020). Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European languages. In *Working notes of FIRE 2020 – Forum for information retrieval evaluation, Hyderabad, India* (S. 87–111). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2826/T2-1.pdf.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems 26 (NIPS 2013)* (S. 3111–3119). https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (S. 2268–2274). https://doi.org/10.18653/v1/D15-1272.

Paraschiv, A., & Cercel, D. C. (2019.) UPB at GermEval-2019 task 2: BERT-based offensive language classification of German Tweets. In *Proceedings of the 15th conference on*

natural language processing (KONVENS 2019). https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/Germeval_Task_2_2019_paper_9.UPB.pdf.

@raffi. (2013). New Tweets per second record, and how! https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html.

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In *Advances in artificial intelligence, 23rd Canadian conference on artificial intelligence (Canadian AI 2010), Berlin, Heidelberg, Deutschland* (S. 16–27). Springer. https://doi.org/10.1007/978-3-642-13059-5_5.

Risch, J., Stoll, A., Ziegele, M., & Krestel, R. (2019). hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model. In *Proceedings of the 15th conference on natural language processing (KONVENS 2019)*. https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/Germeval_Task_2_2019_paper_10.HPIDEDIS.pdf.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Hrsg.), *Bochumer Linguistische Arbeitsberichte 17, NLP4CMC III: 3rd workshop on natural language processing for computer mediated communication* (S. 6–9). http://dx.doi.org/10.17185/duepublico/42132.

Rother, K., & Rettberg, A. (2019). German Hatespeech classification with Naive Bayes and Logistic Regression-hshl at GermEval 2019-Task 2. In *Proceedings of the 15th conference on natural language processing (KONVENS 2019)*. https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/Germeval_Task_2_2019_paper_2.HSHL.pdf.

Ruppenhofer, J., Siegel, M., & Wiegand, M. (2018). Guidelines for IGGSA shared task on the identification of offensive language. http://www.melaniesiegel.de/publications/2018_GermEval_Guidelines.pdf.

Schäfer, J. (2018). HIIwiStJS at GermEval-2018: Integrating linguistic features in a neural network for the identification of offensive language in microposts. In *Proceedings of GermEval 2018, 14th conference on natural language processing (KONVENS 2018)* (S. 104–112). https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/GermEval2018_Proceedings.pdf#page=110.

Schäfer, J., & Boguslu, K. (2021). Towards annotating illegal hate speech: A computational linguistic approach. In *Detect Then Act (DTCT) technical report 3*. ISSN 2736-6391. https://dtct.eu/wp-content/uploads/2021/10/DTCT-TR3-CL.pdf.

Schmid, H., Fitschen, A., & Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of the fourth international conference on language resources and evaluation (LREC 2004)* (S. 1263–1266). http://www.lrec-conf.org/proceedings/lrec2004/summaries/468.htm.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media (SocialNLP@EACL 2017), Valencia, Spanien* (S. 1–10). https://doi.org/10.18653/v1/w17-1101.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22–36. https://doi.org/10.1145/3137597.3137600.

Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Proceedings of the four-teenth national conference on artificial intelligence and ninth conference on innovative applications of artificial intelligence (AAAI'97/IAAI'97)* (S. 1058–1065). http://www.aaai.org/Library/IAAI/1997/iaai97-209.php.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the demonstrations at the 13th conference of the European chapter of the association for computational linguistics, Avignon, Frankreich* (S. 102–107). https://aclanthology.org/E12-2021/.

Struß, J., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th conference on natural language processing (KONVENS 2019), Erlangen, Deutschland* (S. 354–365). German Society for Computational Linguistics & Language Technology. https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/GermEvalSharedTask2019Iggsa.pdf.

Townsend, L., & Wallace, C. (2018). The ethics of using social media data in research: A new framework. In K. Woodfield (Hrsg.),*The ethics of online research, advances in research ethics and integrity* (Bd. 2,S. 189–207). Emerald Publishing Limited. https://doi.org/10.1108/S2398-601820180000002008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems 30 (NIPS 2017)*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wiedemann, G., Yimam, S. M., & Biemann, C. (2020). UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the fourteenth workshop on semantic evaluation (SemEval@COLING 2020)* (S. 1638–1644). https://doi.org/10.18653/v1/2020.semeval-1.213.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th conference on natural language processing KONVENS 2018), Wien, Österreich* (S. 1–10). Österreichische Akademie der Wissenschaften. https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/GermEval2018_Proceedings.pdf#page=7.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th international workshop on semantic evaluation (SemEval@NAACL-HLT 2019)* (S. 75–86). https://doi.org/10.18653/v1/s19-2010.