Automatisierte Analyse Radikaler Inhalte im Internet

Inna Vogel[1], Roey Regev[1], Martin Steinebach[1]

Abstract: Rassismus, Antisemitismus, Sexismus und andere Diskriminierungs- und Radikalisierungsformen zeigen sich auf unterschiedliche Arten im Internet. Es kann als Satire verpackt sein oder als menschenverachtende Parolen. Sogenannte Hassrede ist für die Kommunikationskultur ein Problem, dem die betroffenen Personen oder Personengruppen ausgesetzt sind. Zwar gibt es den Volksverhetzungsparagraphen (§ 130 StGB), Hassrede liegt allerdings nicht selten außerhalb des justiziablen Bereichs. Dennoch sind hasserfüllte Aussagen problematisch, da sie mit falschen Fakten Gruppierungen radikalisieren und Betroffene in ihrer Würde verletzen. 2017 stellte die Bundesregierung das Netzwerkdurchsetzungsgesetz vor, welches die sozialen Netzwerke dazu zwingt, Hassrede konsequent zu entfernen. Ohne eine automatisierte Erkennung ist dieses aber nur schwer möglich. In unserer Arbeit stellen wir einen Ansatz vor, wie solche Inhalte mithilfe des maschinellen Lernens erkannt werden können. Hierfür werden zunächst die Begriffe Radikalisierung und Hate Speech sprachlich eingeordnet. In diesem Zusammenhang wird darauf eingegangen wie Textdaten bereinigt und strukturiert werden. Anschließend wird der k-Nearest-Neighbor-Algorithmus eingesetzt, um Hate Speech in Tweets zu erkennen und zu klassifizieren. Mit unserem Vorgehen konnten wir einen Genauigkeitswert von 0,82 (Accuracy) erreichen - dieser zeigt die Effektivität des KNN-Klassifikationsansatzes.

Keywords: Hassrede, Hate Speech, Soziale Netzwerke, NLP, KNN-Algorithmus, Twitter

1 Einleitung

Die Debattenkultur in sozialen Netzwerken ist nicht selten beleidigend, verletzend, aggressiv oder aber auch hasserfüllt und bedrohlich. Die Sprache radikalisiert sich immer mehr und der verbale Kulturkampf eskaliert seit Jahren zunehmend. Die Feindbilder sind altbekannt: Juden, Linke, Schwarze, Muslime, Homosexuelle, Feministinnen oder Flüchtlinge. Doch Sprache ist nicht nur ein Narrativ. Sie deutet Gegebenheiten und bereitet auf das Handeln vor. Die Diskussionen im Netz dienen nicht nur als Ort des Austausches, sondern auch als Ort konkreter Verabredungen und Planungen von Aktionen, wie aus den folgenden Beispielen abgeleitet werden kann.

- *„So kennt man die Musels. Ist halt ein hinterhältiges Pack. Spielen wir Bimbos versenken ...“*
- *„Diese drecks Bild, wieso steht das nicht auf der 1 Seite-Mauelkorb von Merkel-Leute kauft keine Bild mehr, die veraschen uns so und so nur. Afrikaner...Nigger-schlagt ihn die Köpfe ab, widerliches Pack [...]“*
- *„Lasst uns diese abartigen Asylis entfernen, die haben hier nichts zu suchen“* (Quelle: Twitter)[2]

[1] Fraunhofer-Institut für sichere Informationstechnologie SIT, Rheinstr. 75, 64295 Darmstadt, Germany, {inna.vogel,roey.regev,martin.steinebach}@sit.fraunhofer.de

[2] Die Beispiele in dieser Arbeit illustrieren die Schwere des Problems der Hassrede. Sie stammen aus sozialen Netzwerken und spiegeln in keinster Weise die Meinung der Autoren wider.

doi:10.18420/inf2019_27

„Hassrede“ ist kein juristisch abgegrenzter Begriff, da die unzulässige Meinungsäußerung nicht aus dem jeweiligen Kontext gelöst werden kann. Der Kontext ist meist von der nationalstaatlichen Ordnung geprägt. Als juristischer Ausgangspunkt dient in Deutschland die freie- sowie die unzulässige Meinungsäußerung. Das Grundgesetz stellt in Artikel 5 Abs. 1 Satz 1 fest, dass jeder Mensch das Recht hat, „seine Meinung in Wort, Schrift und Bild frei zu äußern“. Die freie Meinungsäußerung endet jedoch im Artikel 5 Absatz zwei, wenn die persönliche Ehre verletzt wird. Verboten werden kann beispielsweise die Schmähkritik, die stets auf den Kontext ankommt und deshalb eine stets auf den Einzelfall zu prüfende Frage bleibt. Bei Volksverhetzung im Internet droht eine Freiheitsstrafe von bis zu fünf Jahren[3]. Jenseits der schwierigen Rechtslage ist es zudem aufgrund der schieren Masse problematischer Beiträge und Verbreitungsgeschwindigkeit eine Herausforderung Rassismus, Fremdenfeindlichkeit, Antisemitismus oder andere Formen von Intoleranz in ihren Facetten und Dimensionen in den Kommentarspalten des Internets zu identifizieren und einzudämmen.

Anfang 2017 trat das Netzwerkdurchsetzungsgesetz[4] in Kraft. Soziale Netzwerke, darunter Twitter, Facebook und YouTube sind seitdem verpflichtet, „rechtswidrige Inhalte“ innerhalb von 24 Stunden nach Eingang einer Beschwerde zu entfernen oder zu sperren, sonst drohen Bußgelder. Da die Unternehmen hohe Millionen-Strafen fürchten, werden potentiell auch nicht strafbare oder nicht rechtswidrige Inhalte gelöscht[5], was die Meinungsfreiheit der Nutzer einschränkt.

Eine geeignete Lösung hierfür wäre eine Automatisierung oder zumindest eine signifikante semi-automatische Unterstützung, um erfolgreiche Strategien zur Bekämpfung zu entwickeln. In unserer Arbeit stellen wir einen Ansatz vor, wie mithilfe von maschinellen Lernverfahren problematische Beiträge in sozialen Netzwerken automatisch erkannt werden können. Unsere Arbeit ist wie folgt strukturiert: Zunächst werden die Begriffe „Radikalismus“ und „Hassrede“ definiert und voneinander abgegrenzt. In Kapitel 4 stellen wir den verwendeten Textkorpus vor und gehen darauf ein, wie die Twittertexte im Rahmen einer Vorverarbeitung strukturiert und standardisiert werden. Um Hasskommentare automatisch zu klassifizieren, wurde der KNN-Klassifikator (KNN; engl. „k-Nearest-Neighbor“) verwendet. Im Rahmnen unserer Analyse wurden zwei unterschiedliche Frameworks verwendet, deren Unterschied maßgeblich in der Art der Datenrepräsentation liegt. Die vorliegende Arbeit schließt mit einer Zusammenfassung der Ergebnisse und einem Resümee ab.

2 Stand der Forschung

Eine zentrale Herausforderung für das automatische Erkennen von Hasskommentaren in sozialen Medien ist die Trennung zwischen Hate Speech und „nur“ beleidigender Sprache. Davidson et al. [Da17a] haben mithilfe eines Lexikons für beleidigende Sprache Tweets gesammelt und diese anschließend anhand der drei folgenden Kategorien händisch klassifiziert: Hassrede, beleidigende Sprache und neutrale Tweets. Anhand dieses Diktionärs wurde ein Multiklassen-Klassifikator trainiert, um (neue) Tweets anhand dieser drei verschiedenen Kategorien einzuordnen. Eine Analyse der Vorhersagen und Fehler zeigt, dass rassistische und homophobe Tweets eher als Hassrede klassifiziert werden. Sexistische Tweets dagegen werden bevorzugt als beleidigend eingestuft.

Burnap und Williams [BW14] sammelten in den ersten zwei Wochen nach dem Mord an Lee Rigby rund 450.000 Tweets, welche im Zusammenhang mit dem Mord gepostet wurden. Der 25-jährige britische Soldat wurde im Jahr 2013 auf offener Straße in London von zwei mutmaßlichen Islamisten

[3] Strafgesetzbuch (StGB) § 130 Volksverhetzung: https://www.gesetze-im-internet.de/stgb/__130.html

[4] Netzwerkdurchsetzungsgesetz - NetzDG vom 1 September 2017 (BGBl. I p. 3352)

[5] „Soziale Netzwerke löschen tausende Beiträge“: https://www.mdr.de/nachrichten/politik/inland/bilanz-netzdg-twitter-facebook-daten-beschwerden-100.html (29.04.2019)

ermordet. Die Autoren haben unterschiedliche maschinelle Lernverfahren trainiert, um Hasstweets zu klassifizieren, welche sich gegen eine bestimmte Rasse, Ethnie oder Religion richten. Mit ihrem Ansatz konnten sie einen Genauigkeitswert von 0,95 ($F_1 - Score$) erreichen. Der Klassifikator soll Politiker und Entscheidungsträger dabei unterstützen die öffentliche Reaktion auf großräumige emotionale Ereignisse besser einschätzen zu können.

Del Vigna et al. [De17] verwendeten für ihren Ansatz öffentliche Kommentare von italienischen öffentlichen Facebook-Profilen. Nachdem die Hasskommentare von bis zu fünf verschiedenen Annotatoren binär klassifiziert wurden (Hate Speech und neutrale Posts), wurden morpho-syntaktische Merkmale, Sentiment-Polarität und Word Embeddings als Features verwendet, um Hate Speech in Facebook-Kommentaren zu erkennen. Das Trainieren einer Support Vector Machine (SVM) und eines Recurrent Neural Networks (Long Short Term Memory - LSTM) haben die Effektivität der beiden untersuchten Klassifikationsansätze gezeigt.

3 Terminologie

Hassrede (engl. „Hate Speech“)[6] und Radikalisierung sind nicht nur sprachwissenschaftliche, sondern auch politische Phänomene mit Bezügen zu juristischen Tatbeständen. Um radikale Inhalte automatisch zu erkennen, bedarf es eines Algorithmus, der in der Lage ist, Inhalte zu klassifizieren, welche zum Hass gegen Teile der Bevölkerung aufstacheln, zu Gewaltmaßnahmen gegen sie auffordern oder aber die Menschenwürde anderer dadurch angreifen, dass Teile der Bevölkerung beschimpft, böswillig verächtlich gemacht oder verleumdet werden (StGB, §130(1)). Und obwohl Begriffe wie „Radikalismus“, „Extremismus“ und „Hassrede“ im Alltag sowie in den Medien allgegenwärtig sind, ist keineswegs klar, was diese im Einzelnen genau bedeuten. Nicht selten erfolgt ihre Verwendung synonym, da sich ihre Abgrenzung selbst für Wissenschaftler schwierig gestaltet. Um ein gemeinsames Verständnis für die Konzepte zu schaffen, werden im Folgenden die Begriffe kurz definiert und zueinander ins Verhältnis gesetzt.

3.1 Radikalisismus

Radikalismus ist im sozialen und politischen Rahmen eine Geisteshaltung, die eine Änderung von etwas Bestehendem anstrebt [Ne13]. Dies kann beispielsweise das bestehende soziale oder politische System sein. McCauley und Moskalenko [MM08] beschreiben Radikalismus als die Veränderung von Überzeugungen, Gefühlen und Verhaltensmustern. Gruppenkonflikte werden im Zuge dessen zunehmend gerechtfertigt und Opfer zur Verteidigung der Gruppe gefordert. Allerdings muss die Radikalisierung nicht zwingend durch Gewalt gekennzeichnet sein, sondern kann sich in gewaltfreien Widerstands- und Auseinandersetzungsformen manifestieren (z.B. Proteste oder Boykott- und Streikaktionen). Gemäß Wiktorowiczs [Wi05] ist die Radikalisierung ein gradueller und sukzessiver Prozess. Die Person oder Personengruppen passen sich ständig an Normen, Ideologien und Sitten an, die von dem normativen Status quo abweichen. Im Laufe des Radikalisierungsprozesses kommt es zu passiven sowie aktiven Interaktionen eines Individuums in extremistischen Milieus. Dabei werden Handlungen und Ideen befürwortet, die den gängigen Werten der Gesellschaft entgegenstehen [Ne13]. Die Ideologie ist anti-demokratisch, lehnt das bestehende System ab und hat das Ziel dieses mit etwas Neuem zu ersetzen.

Untersuchungen des Bundeskriminalamtes [Bu15] zur Folge steigt die Bedeutung der Radikalisierung in sozialen Medien stetig an, da das Internet extremistischen Gruppen eine (große) Bandbreite an

[6] Die Begriffe Hate Speech, oder deutsch Hassrede, werden nachfolgend synonym verwendet.

Plattformen für den Informationsaustausch und Meinungsäußerung bietet. Zudem erleichtern diese auch die Verbreitung von (Online) Hate Speech und extremistischer Propaganda.

3.2 Hate Speech

Im europäischen Zusammenhang wird Hate Speech zusammengefasst als:

> *„Jegliche Ausdrucksformen, welche Rassenhass, Fremdenfeindlichkeit, Antisemitismus oder andere Formen von Hass, die auf Intoleranz gründen, propagieren, dazu anstiften, sie fördern oder rechtfertigen, einschließlich der Intoleranz, die sich in Form eines aggressiven Nationalismus und Ethnozentrismus, einer Diskriminierung und Feindseligkeit gegenüber Minderheiten und Einwanderern [...] ausdrückt“* (Ministerkomitee des Europarats, Empfehlung R (97) 20, 30.10.1997.

Hate Speech kann sich gegen Hautfarbe, Nationalität, Herkunft, Religion, Geschlecht, sexuelle Orientierung, sozialen Status, Gesundheit, Aussehen, oder eine Kombination davon richten. Die Liste ist keineswegs vollständig, da im Prinzip jede Eigenschaft eines Individuums zum Gegenstand von Hass werden kann [Me13]. Hassäußerungen können unterschiedliche Formen annehmen, sodass es selbst für Menschen nicht immer einfach ist, diese zu entdecken. Sie können direkt oder indirekt geäußert werden. Eine direkte Abwertung von Einwanderern wären beispielsweise folgende Formulierungen: *„Dreckspack entsorgen“*, *„RAUS mit dem PACK“* (Quelle: Twitter). Eine indirekte Herabsetzung wäre beispielsweise: *„Morde, Vergewaltigungen, Messerstechereien...Das ist multi Kulti“* (Quelle: Twitter). Das Gefährliche an Hate Speech ist nicht nur die Verbreitung von verbalen Hassaussagen, sondern oftmals auch die Motivation zum gelebten Gewaltexzess. Hassrede kann zu Übergriffen und Ermordungen an Menschen aufgrund ihrer Hautfarbe, ihrer Religion, ihrer Geschlechtsidentität oder ihrer Sexualität anstiften.

4 Erstellung und Vorverarbeitung der Textdaten

Soziale Netzwerke bieten extremistischen Gruppierungen eine große Auswahl an Plattformen für die Verbreitung von Hassrede. Nicht jeder, der eine Ideologie oder Idee im Netz teilt, ist bereit den Weg der Radikalisierung bis hin zur Gewaltanwendung zu gehen. Allerdings verleitet die vermeintliche Anonymität des Netzes häufig zu sprachlicher Verrohung und zum Verzicht auf Respekt gegenüber Mitmenschen. Sprache wird dabei benutzt, um Ideen und Ideologien zu teilen, bis hin zur Aufforderung von Gewaltanwendung, beispielsweise gegen Einwanderer, Andersgläubige oder andere Minderheiten. Die folgenden Auszüge aus sozialen Medien sollen beispielhaft zeigen wie die Sprache dazu verwendet wird, um Menschen herabzusetzen oder zu verunglimpfen. Das erste Beispiel macht die gefühlte Ungerechtigkeit eines Individuums deutlich sowie den steigenden Rassismus und Ablehnung gegen die Regierung. Das zweite Beispiel offenbart die Vertretung von islamistischer Ideologie. Es wird auf den Gottesstaat referiert, in dem terroristische Gewalt ein Mittel gegen „Ungläubige“ und sogenannte korrupte Regime ist.

> „Politiker bekommen Personenschutz und wer schützt unsere Bevölkerung vor diesen Verbrechern und Abschaum? Das ja den GROSSEN nichts passiert. Haben Angst um ihr bisschen Leben. Was ist mit der Angst die diese alte Frau jetzt hat? Die sollte Schmerzensgeld von der Bundesregierung einklagen die dieses Unrat in unser Land holt.....“ (Quelle: Facebook)

> „DRECKS UNGLÄUBIGE WESTLERABSCHAUM ZIONISTENSCHWEINE PACKT EURE MÄRCHENIDEOLOGIEN UND EUER GRUNDGESETZ HURENSOHN EIN UND VERPISST EUCH BEVOR IHR DER ENDLÖSUNG ZUGEFÜHRT WERDET ; ISIS HÖRT MIT UND DAS SCHARFE SCHWERT REICH FÜR EUCH ZIONISTEN ALLE AUS!!!!! VERPISST EUCH SELBST BEVOR IHR WIE ISRAEL UND DIE RUSSEN SERBEN AUSGERTOTTET WERDET!!!“ (Quelle: YouTube)

Vom „Effekt der sprachlichen Identifikation“ [Di80] wird in der Literatur gesprochen, wenn die Sprache dazu verwendet wird, um sich mit einer Gruppe zu identifizieren [Di80]. Der Effekt besagt, dass innerhalb einer Gruppe dieselbe Sprache gesprochen wird. Dadurch ist es möglich, Hassrede und sich aufbauende Radikalisierung mithilfe von Werkzeugen der linguistischen IT-Forensik und des maschinellen Lernens zu erkennen, bevor es zur Gewaltausübung kommt.

4.1 Korpus

Um Hate Speech automatisch zu identifizieren, haben wir einen englischen Textkorpus mit 25.296 manuell klassifizierten Twitternachrichten[7] [Da17b] verwendet. Twittertexte zeichnen sich dadurch aus, dass jede Meldung eine Maximallänge von 280 Zeichen hat. Die Texte im Korpus wurden von sechs Annotatoren händisch 3 Klassen zugeordnet.

- Klasse 0 = Hate Speech 18.892 Nachrichten
- Klasse 1 = Beleidigende Sprache 1.694 Nachrichten
- Klasse 2 = Neutrale Tweets 1.200 Nachrichten

Der Datenkorpus zeigt, dass die Klassenverteilung ungleich bzw. „unbalanciert“ ist. Es sind deutlich mehr Hate Speech-Tweets im Korpus vorhanden, als neutrale oder beleidigende Nachrichten. Es existieren unterschiedliche Methoden, um dem Problem des Ungleichgewichts im Datensatz entgegenzuwirken. Sampling-basierte Methoden sind in der Datenanalyse Techniken, mit denen die Verteilung der Daten in den Klassen angepasst werden kann. Hierbei unterscheidet man zwischen Undersampling, Oversampling und Hybrid-Verfahren. Hybridverfahren wenden eine Mischform zwischen Under- und Oversampling an, um die Häufigkeit der Daten innerhalb der Klassen anzugleichen.

Beim Undersampling werden Elemente aus der größten Klasse eliminiert, um die Häufigkeitsverteilung der Daten pro Klasse auszugleichen. Das hat allerdings zur Folge, dass weniger Daten der jeweiligen Klasse zum Trainieren zur Verfügung stehen. Beim Oversampling wird die kleinere Klasse synthetisch vervielfältigt, indem beispielsweise zufällig ausgewählte Textdokumente vervielfacht werden. Der Vorteil des Oversamplings besteht darin, dass keine Informationen aus dem Trainingsset verloren gehen, da die Daten sowohl aus der Minderheits- als auch aus der Mehrheitsklasse erhalten bleiben. Der Nachteil ist jedoch, dass der Korpus nicht mit neuen Daten angereichert wird, sondern dieselben Daten lediglich reproduziert werden und dadurch die Gefahr des Overfittings erhöht wird.

Da im verwendeten Korpus Hate Speech über- und neutrale Tweets unterrepräsentiert sind, haben wir uns zunächst für das Oversampling entschieden, um die gleiche Häufigkeitsverteilung der Daten in den Klassen zu erzielen. Statt allerdings synthetisch den Korpus zu erweitern, haben wir die Klasse mit neutralen Tweets erweitert, indem wir den Korpus mit Twittertexten angereichert haben,

[7] Hate Speech Korpus: `https://github.com/t-davidson/hate-speech-and-offensive-language`

die nicht einer einzigen speziellen Domäne zugeordnet werden können. Hierfür wurde der Korpus für Sentimentanalyse von Go et al. [GBH09] verwendet. Der Korpus umfasst rund 1,6 Mio. Tweets sieben unterschiedlicher Domänen (z.B. „Company“, „Event“, „Location“, „Movie“, „Person“ etc.). Dieser Korpus wurde ausgewählt, da die Autoren darauf hinweisen, dass viele Tweets keine Stimmung beinhalten. Vor der Anwendung der Tweets haben wir eine manuelle Überprüfung durchgeführt, um sicherzugehen, dass die Daten nicht aufgrund eines bestimmten Vokabulars oder Topics verzerrt sind. Die Tweets wurden nach dem Zufallsprinzip aus dem Datensatz extrahiert. Es wurde zudem darauf geachtet, dass die Nachrichten in etwa die gleiche Länge und dieselben Textspezifika aufweisen. Twitternachrichten sind speziell, da sie maximal 280 Zeichen lang sind. Zudem werden Hashtags (#), User-Mentions (z.B. @John) sowie Emojis und Emoticons verwendet. Meistens wenden User Umgangssprache bzw. Alltagssprache beim Schreiben an. Der Unterschied zur Fachsprache besteht darin, dass die Alltagssprache, die im täglichen Umgang benutzt wird, keinem bestimmten Soziolekt entspricht wie etwa die Fachsprache (z.B. Wissenschafts- oder Medizinsprache).

Die Domänenunabhägigkeit der Gegenklasse ist wichtig, da sonst der Algorithmus Muster eines bestimmten Themas oder Schreibstils klassifizieren würde[8]. In der folgenden Tabelle sind Beispiele der klassifizierten Hasstweets und neutralen Tweets aus dem verwendeten englischen Datenset aufgeführt:

Hate Speech Tweets	**Neutrale Tweets**
Why people think gay marriage is okay is beyond me. Sorry I don't want my future son seeing 2 fags walking down the street holding hands	Peel up peel up bring it back up rewind back where I'm from they move Shaq from the line„“ ooooow who tf said that trash!!?
#AZmonsoon lot of rain, too bad it wasn't enough to wash away the teabagger racist white trash in the state. #Tcot #teaparty #azflooding	10 birds your grandkids may never see, thanks to climate change http://t.co/XqmXHkAsWt #Climate
#Dutch people who live outside of #NewYorkCity are all white trash.	@SportsCenter: Eli Manning just threw his NFL-leading 27th interception of the season. Lmao trash

Tab. 1: Beispiele für Hasskommentare und neutrale Tweets (Quelle: Twitter)

Um den Datenkorpus zu balancieren, wurde die neutrale Klasse mit zusätzlichen rund 17.996 Tweets erweitert. Je Klasse wurden folglich insgesamt rund 19.200 Tweets verwendet[9]. Die Klasse mit beleidigenden Tweets wurde nach einigen Testversuchen verworfen, da nicht genügend Trainingsmaterial zur Verfügung steht. Allerdings haben wir Tweets, die von mindestens zwei Annotatoren als Hate Speech gelabelt wurden (in der Gesamtbewertung aber als beleidigende Sprache gelabelt wurden), ebenfalls in unserem Korpus aufgenommen. Es wurden folglich die Klassen Hassrede und neutrale Tweets betrachtet und analysiert.

[8] Wenn als Gegenklasse beispielsweise Reviews zu Fotokameras verwendet werden würden, dann könnte der Algorithmus auf der lexikalischen Ebene Unterschiede finden, z.B. zwischen Kamerafeatures und Beleidigungen

[9] Insgesamt weist unser Korpus 19.196 Hate Speech-Kommentare, sowie dieselbe Anzahl an neutrale Tweets, auf

4.2 Datenbereinigung

Um Texte zu strukturieren und zu standardisieren, müssen die Daten einem Preprocessing, d.h. einer Vorverarbeitung, unterzogen werden. Folgende Schritte wurden unternommen:

- Entfernung aller Tweets, die aus weniger als drei Tokens bestehen sowie aller Markup-Tags z.B. <*City*>London<*/City*>
- Entfernung aller User-Mentions (z.B. @John), da diese nicht bedeutungstragend sind
- Kleinschreibung aller Wörter, um das Vokabular zu normalisieren
- Entfernung von Stoppwörtern wie z.B. Artikel („the“, „that“, „a“, „an“), Konjunktionen („and“, „or“, „but“, „because“ etc.) und häufig gebrauchte Präpositionen (z.B. „at“, „in“, „on“ etc.) sowie die Negation „not“. Häufig tragen diese Wortgruppen weniger zum Inhalt bei als die Wortklassen Nomen, Verben, Adjektive und Adverbien
- Entfernung von Sonderzeichen (z.B. <> ()[]{} ∗ +/})
- Negative Smileys und Emojis werden durch CLDR (Common Locale Data Repository) ersetzt, d.h. durch Kurzzeichennamen oder Schlüsselwörter, um diese als Feature nutzen zu können. Die restlichen Smileys werden entfernt. Die Annahme ist, dass negative Smileys und Emojis verwendet werden, um Hate Speech zu untermauern bzw. zu verstärken (z.B. „*Why the fuck do niggers act so different when girls are around?:-S*“)
- Normalisierung der Interpunktion und Buchstabenzeichen. Um dem Gesagten mehr Ausdruck zu verleihen, verwenden Social Media User entweder Großbuchstaben oder eine Aneinanderreihung von gleichen Buchstaben. Damit der Computer erkennt, dass „heeeeeeey“ und „hey“ ein und dasselbe Wort ist, wurden Zeichen, die sich mehr als zweimal wiederholen, zu einem Zeichen reduziert. So wird „heeeeeeey“ zu „hey“ und Wörter wie „arriving“ bleiben unverändert. Angemerkt an dieser Stelle sei, dass dieses triviale Vorgehen zu Fehlern führt. Wird das Wort „Hello“ als „Hellllloooo“ geschrieben, wird es fälschlicherweise zu „Helo“ reduziert. Hier kann mit Rechtschreibprüfungsprogrammen entgegengewirkt werden
- Entfernung von URLs. Es gibt Fälle, bei denen die URL bedeutungstragend ist wie z.B. *„you whore belong to www.pornhub.com“*, aber da diese Fälle selten sind, werden alle URLs gelöscht (z.B. *„www.pornhub.com accepts now Verge, amazing!“*)
- Im letzten Vorverarbeitungsschritt erfolgt die Tokenisierung der Texte

4.3 Hashtags

Hashtags (z.B. *#EarlyChristmas* oder *#FreeMoney*) wurden gesondert behandelt, da diese erheblich zum Inhalt beitragen können. Einerseits sind sie bedeutungstragend, andererseits werden teilweise mehrere Wörter zu einem zusammengefasst, was die maschinelle Verarbeitung erschwert. Zunächst haben wir das Hashtag-Zeichen entfernt und geprüft, ob einem Kleinbuchstaben ein Großbuchstabe folgt. Wenn ja, wurden die Wörter an dieser Stelle voneinander getrennt (z.B. wurde aus *”#FreeMoney”* - *”Free”* und *”Money”*). Anschließend wurde geprüft, ob das Wort in einem Wörterbuch vorhanden ist, wenn ja, wurde es behalten (z.B. *#toys, #shots, #Pisces*). Wenn das Wort nicht in einem Wörterbuch zu finden war, wurde dieses in Fragmente aus drei Buchstabenfolgen, Trigramme genannt, zerlegt und mit den Trigrammen einer Liste mit beleidigenden Wörtern verglichen. Um die Ähnlichkeit der Wörter zu vergleichen, wurde der Jaccard-Koeffizient (J) verwendet. Der Koeffizient nimmt die Mengen A

und B, in unserem Fall die Trigramme, und berechnet daraus den Quotienten der Schnittmenge und deren Vereinigungsmenge ($J(A,B) = \frac{|A \cap B|}{|A \cup B|}$). Der errechnete Wert ist stets zwischen Null und Eins. Je höher der Wert, desto ähnlicher sind sich die Mengen. Die Trigramme werden folglich miteinander verglichen und deren Schnittmenge ermittelt. Hierbei haben wir einen Schwellwert von 0,5 festgelegt. Wurde dieser Schwellwert überschritten, wurde dieses Wort ohne Hashtag im Text behalten. Die restlichen Hashtagwörter wurden entfernt. Der Schwellwert wurde mithilfe eines Sensitivitätstests errechnet. Hierfür wurden 5.000 zufällig gewählte Wörter analysiert. Rund 70% der Wörter konnten durch dieses einfache Verfahren dem richtigen Ursprungswort zugeordnet werden (Beispiel: *"ucunt" - "cunt", "niggas" - "nigger", "abitch" - "bitch"*).

5 Klassifizierung von Hassrede in englischen Twitternachrichten

Um Hate Speech automatisch zu klassifizieren, wurde der Nächste-Nachbarn-Klassifikator (KNN; engl. „k-Nearest-Neighbor“) ausgewählt. Bei diesem Klassifikationsverfahren wird die Klassenzugehörigkeit unter Berücksichtigung der ausgewählten k-nächsten-Nachbarn vorgenommen. Die Dateninstanz wird entweder der Klasse zugeordnet, welche am häufigsten unter den k-Nächsten-Nachbarn vertreten ist oder die Klassenzugehörigkeit erfolgt gemäß des nächsten Nachbarns, gemessen an der geringsten Distanz (k=1). Trotz der relativ simplen Funktionsweise gehört der Klassifikator zu den erfolgreichsten maschinellen Lernverfahren [Ko15].

Die Funktionsweise des KNN-Algorithmus ist in Abbildung 1 beschrieben. Es wird ein binäres Klassifikationsproblem mit den Klassen A und B betrachtet. Werden k=3 nächste Nachbarn in Betracht gezogen, erfolgt die Zuordnung des unbekannten Objekts zur Klasse B. Werden k=6 nächste Nachbarn ausgewählt, ähnelt das Objekt laut dem Algorithmus der Klasse A. Hier offenbart sich ein Nachteil des Verfahrens. Wird ein zu „kleines“ k ausgewählt (z.B. k=2), weist der Algorithmus eine hohe Sensitivität gegenüber Ausreißern auf. Wird k zu hoch angesetzt (z.B. k=30), werden zu viele Objekte der Gegenklasse in der Entscheidungsmenge vorkommen und vom Klassifikator „favorisiert“.

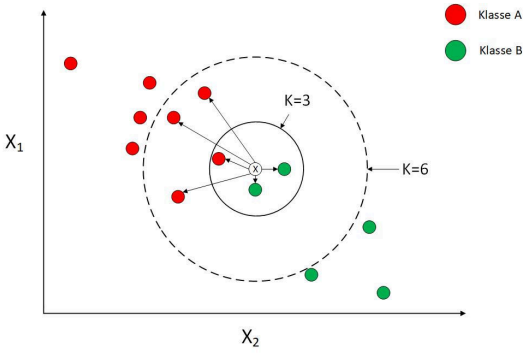

Abb. 1: KNN-Klassifikator mit k=3 und k=6 Nachbarn

5.1 NuPIC KNN-Framework

Der KNN-Klassifikator wurde auf zwei unterschiedliche Weisen verwendet. Das erste Framework wurde mittels „NuPIC“ (Numenta Platform for Intelligent Computing) integriert, einer Pythonbibliothek für den HTM-Lernalgorithmus (HTM - engl. „Hierarchical Temporal Memory“). HTM entstammt aus dem Forschungsgebiet der Neurowissenschaft und wurde von Jeff Hawkins und George Dileep [Ha19] entwickelt. HTM ist eine detaillierte Berechnungstheorie, welche versucht, die strukturellen und algorithmischen Eigenschaften des Neokortex in einem Bayes'schen Netz abzubilden. Das Herzstück von HTM sind zeitbasierte, kontinuierliche Lernalgorithmen, die räumliche und zeitliche Muster speichern und abrufen. NuPIC bietet neben dem HTM-Lernalgorithmus unterschiedliche Klassifikatoren an[10], von denen wir den „KNN Classifier“ verwendet haben.

Als Eingabe erfordert NuPICs KNN eine feste Länge des Eingabevektors, weshalb alle Tweets auf dieselbe Zeichenlänge aufgefüllt wurden. Als Eingabevektor wurde die Maximallänge der Tweets von 280 Zeichen gewählt. Die Eingabedaten müssen bei NuPIC spärlich repräsentiert sein (SDR; engl. „Sparse Distributed Representation“). Für das Encoding wurden die Buchstabenzeichen folglich binär in One-Hot-Vektoren umgewandelt (z.B. [00101010] = [2,4]). Der „SparsePassThroughEncoder“ wurde verwendet, um die One-Hot-Vektoren an den KNN-Klassifikator zu übergeben.

Während der Trainingsphase wird eine Menge der zur Klassifikation ausgewählten k-nächsten Nachbarn betrachtet. Je ähnlicher die Dokumente, desto näher zueinander befinden sich diese im n-dimensionalen Vektorraum, wobei n die Dimension des Eingabevektors ist. Je unterschiedlicher die Dokumente, desto weiter sind diese im Vektorraum voneinander entfernt. In der Testphase wurde abschließend die euklidische Distanz genutzt, um zu eruieren, wie präzise das Verfahren klassifiziert. Andere Distanzmaße wie bspw. die Manhatten-Distanz sind ebenso denkbar. Die Testergebnisse haben eine Accuracy von 0,71 erzielt. Das Entfernen der Stoppwörter resultiert in einer marginalen Verschlechterung.

Der Genauigkeitswert kann weiter gesteigert werden, indem die Wörter im Korpus beispielsweise auf ihre Grundform reduziert werden. Dieses sogenannte „Stemming“ (Grundformenreduktion) ist ein Verfahren, mit dem verschiedene morphologische Varianten eines Wortes auf ihren gemeinsamen Wortstamm (stem) zurückgeführt werden, z.B. „essen“ - „isst“ - „gegessen“ = „ess“. Dieser Ansatz reduziert das Vokabular deutlich und führt bei unterschiedlichen Verfahren zu einer besseren Klassifizierungsperformance.

5.2 Scikit-learn KNN-Framework

Das zweite KNN-Framework wurde mittels „scikit-learn“ integriert. Der Unterschied zum ersten Framework ist hauptsächlich in der Datenrepräsentation begründet. Um die Twitternachrichten zu vektorisieren, wurde das Tf-idf-Maß (engl. „Term Frequency / Inverse Document Frequency“) verwendet. Das Maß wird dazu verwendet, um die Relevanz eines Terms im Dokument im Vergleich zur Dokumentenkollektion zu beurteilen. Zunächst wird die Termfrequenz berechnet, d.h. wie häufig ein Wort im Dokument verwendet wird. Anschließend wird ermittelt in wie vielen Dokumenten eine Kollektion dieses Wortes vorkommt. Die Relevanzhypothese besagt, dass Schlüsselbegriffe, die in einem Dokument (themenbezogen) relativ häufig vorkommen, in der Gesamtheit der Dokumente aber relativ selten, ein guter Indikator für den Inhalt eines Dokumentes sind. Artikel und Konjunktionen

[10] NuPIC API Documentation: `http://nupic.docs.numenta.org/1.0.0/index.html`

kommen dagegen in allen Dokumenten etwa gleich oft vor und sind deshalb weniger relevant für die Bestimmung eines Topics. Diese sind beispielsweise relevant für die Stilerkennung eines Textes.

$$TF(i,j) = \frac{\text{Häufigkeit des Begriffs i im Dokument j}}{\text{Gesamtwörter im Dokument j}}$$

$$IDF(i) = \frac{\text{Gesamtdokumente}}{\text{Dokumentananzahl mit Term i}}$$

$$TF - IDF = TF(i,j) * IDF(i)$$

Eine andere Methode, um Wörter zu repräsentieren, wäre der Einsatz von „Word Embeddings“. Word Embeddings stellen Wörter nicht nur mathematisch dar, sondern repräsentieren darüber hinaus die Bedeutung der Wörter im Kontext von anderen Wörtern. Dabei werden Begriffe anhand ihres Vorkommens im Textkorpus und der sie umgebenden Wörter so in einem multidimensionalen Raum angeordnet, dass Wörter, die im selben Kontext vorkommen, einen ähnlichen Vektor erhalten. Mittlerweile existieren vortrainierte Modelle wie etwa „fastText“ oder „GloVe“. Wenn genügend Daten vorhanden sind, können eigene Word Embeddings trainiert werden. Das folgende Beispiel zeigt Ähnlichkeitsmaße eines eigens trainierten Word-Embeddings-Modells:

- „Asylanten“ = „Asylforderer“ (0,73), „Flüchtlinge“ (0,72), „Krimigranten“ (0,70), „Invasoren“ (0,69), „Migranten“ (0,67)
- „KZ“ = „Räucherhaus“ (0,68), „Vernichtungslager“ (0,65), „Auschwitz“ (0,63)
- „ausrotten“ = „hassen“ (0,61), „abschlachten“ (0,58), „Zionisten“ (0,58), „töten“ (0,57)

Nachdem jedes Wort einen Tf-idf-Wert als Merkmal erhalten hat, wurde der KNN-Algorithmus trainiert. 70% der Rohdaten wurden als Trainings- und 30% als Testset verwendet. Auch hier wurde nach dem Training die euklidische Distanz genutzt, um die Performance des Algorithmus zu eruieren (k=5 Nachbarn). Die Accuracy liegt bei diesem Ansatz bei 0,82. Diese Ergebnisse zeigen die Effektivität der beiden Klassifikationsansätze.

6 Evaluierung

Um Hassrede automatisch zu identifizieren und zu klassifizieren, haben wir einen Textkorpus mit 25.296 manuell klassifizierten Twitternachrichten verwendet. Da die Klassenverteilung ungleich war, haben wir die kleinere Klasse mit domänenunabhängigen Tweets erweitert. Dieser Schritt war nötig, da KNN als Klassifikationsalgorithmus trainiert und evaluiert wurde. Beim KNN-Klassifikator sollte darauf geachtet werden, dass die Daten in den Klassen in etwa gleich verteilt sind, damit das Verfahren die stärker vorkommende Klasse bei der gewählten k-nächsten-Nachbarn Betrachtung nicht „bevorzugt“.

Um den KNN-Klassifikationsalgorithmus zu trainieren und dessen Performance anschließend zu untersuchen, haben wir die Frameworks NuPIC und Scikit-Learn verwendet. Für die Verwendung des NuPIC KNNs wurden als Feature-Vektoren Buchstabenzeichen binär in One-Hot-Vektoren umgewandelt. Bei der zweiten Verwendung mittels Scikit-Learn wurden Tokens nach dem Tf-idf-Maß gewichtet und als Datenrepräsentation eingesetzt. Nach dem Training wurde die euklidische Distanz genutzt, um die Performance des Algorithmus zu quantifizieren.

Die Ergebnisse der beiden Klassifikatoren werden in der in Tabelle 2 dargestellten Konfusionsmatrix zusammen mit den resultierenden Evaluierungsmaßen Accuracy (Acc.), Precison (P), Recall (R) und dem F_1-Score, dargestellt. Die Konfusionsmatrix stellt mit ihren vier möglichen Ausprägungen[11] die Grundlage für die Evaluierung eines Großteils der binären Klassifikationsverfahren dar.

Methode & Features	**Konfusionsmatrix**				**Performanzmaß**			
	TP	**TN**	**FP**	**FN**	**Acc.**	**P**	**R**	**F1**
NuPIC KNN One-Hot-Vekt.	4.875	3.253	2.487	869	0,71	0,66	0,85	0,74
Scikit-learn KNN Tf-idf	4.464	4.943	814	1.293	0,82	0,85	0,84	0,82

Tab. 2: Klassifikationsgüte der KNN-Klassifikatoren

Die besten Klassifikationsergebnisse erzielte der KNN Klassifikator, wenn die Tokens als Tf-idf-Vektoren repräsentiert wurden. Hierbei wurde eine Accuracy von 0,82 erzielt. Wurden die Daten als Buchstabenzeichen binär in One-Hot-Vektoren umgewandelt, wurde eine Klassifikationsgüte von 0,71 erreicht. Beim Vergleich der FP- und FN-Werte zeigt sich, dass beim Scikit-Learn-Verfahren mehr Hate Speech Tweets fälschlicherweise als neutrale Tweets klassifiziert wurden. Das könnte damit zusammenhängen, dass die Alltags- und Umgangssprache in sozialen Netzwerken eine starke Verwendung findet. Hierzu gehört auch der Gebrauch von Slang und Schimpfwörtern. Diese klassifiziert der Algorithmus fälschlicherweise als Hate Speech. Davidson et al. [Da17a] haben in diesem Zusammenhang angemerkt, dass die Wörter *„fag“, „bitch“* oder *„nigga“* sowohl Slang sein können, als auch eine Ausdrucksform von Rassismus, Fremdenfeindlichkeit, Sexismus oder Homophobie. Auch in den neutralen Tweets, die wir für Trainingszwecke verwendet haben, finden sich vermehrt Slang- und Schimpfwörter, die ebenfalls als Form der Intoleranz gegenüber Personen- und Personengruppen eingesetzt werden können (z.B. *„pussy“, „bitch“, „trash“*).

- *„Overdosing on heavy drugs doesn’t sound bad tonight. I do that pussy shit every day.“*
- *„a pissed lad past out. I would lick his dirty soles while he slept.“*
- *„welfare/government aid is claimed by white people. So y’all black slander is trash now.“*
- *„http://t.co/JOOsdSubIR he’s a bitch“*

7 Zusammenfassung und Ausblick

Radikalisierungs- und Diskriminierungsformen zeigen sich auf unterschiedliche Weise in sozialen Netzwerken. Hassrede liegt dabei aber nicht selten außerhalb des justiziablen Bereichs. Dennoch sind Hassaussagen problematisch, da sie beispielsweise mit falschen Fakten Menschen oder Gruppierungen radikalisieren können. Ohne eine automatisierte Erkennung ist deren Eindämmung kaum möglich.

In unserer Arbeit haben wir den KNN-Algorithmus eingesetzt, um Hate Speech in Tweets zu erkennen und zu klassifizieren. Zunächst haben wir die Begriffe Radikalisierung und Hate Speech sprachlich definiert und eingeordnet. Anschließend haben wir unterschiedliche Verfahren vorgestellt, wie Textdaten bereinigt und strukturiert werden können. Da im verwendeten Korpus Hate Speech über- und neutrale Tweets unterrepräsentiert waren, haben wir die Klasse mit neutralen Tweets erweitert, indem wir den Korpus mit Twittertexten angereichert haben, die keiner speziellen Domäne zugeordnet waren. Je Klasse wurden rund 19.200 Tweets verwendet. Die Klasse mit beleidigenden Tweets wurde

[11] True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)

nach einigen Testversuchen verworfen, da nicht genügend Trainingsmaterial zur Verfügung stand. Für KNN ist es essenziell, dass die Häufigkeit der Datenverteilung in den Klassen in etwa gleich ist.

Der Algorithmus KNN wurden auf zwei unterschiedliche Weisen eingesetzt - mittels des „NuPIC"- und des „scikit-learn"-Frameworks. Die besten Klassifikationsergebnisse hat KNN erzielt, wenn die Tokens als Tf-idf-Vektoren repräsentiert wurden. Hierbei wurde eine Accuracy von 0,82 erzielt. Wurden die Daten als Buchstabenzeichen binär in One-Hot-Vektoren umgewandelt, wurde eine Klassifikationsgüte von 0,71 erreicht.

Die Evaluierung hat ergeben, dass neutrale Tweets eher fälschlicherweise als Hate Speech klassifiziert werden, als umgekehrt. Das hängt mitunter damit zusammen, dass in sozialen Netzwerken Slang- und Schimpfwörter nicht nur als Ausdrucksform von Rassismus, Fremdenfeindlichkeit, Sexismus oder Homophobie verwendet werden. Untersuchungen haben gezeigt, dass rassistische und homophobe Tweets eher als Hassrede klassifiziert werden. Sexistische Tweets dagegen werden bevorzugt als beleidigend eingestuft. Für die zukünftige Forschung ist es deshalb essenziell einen hinreichend großen Korpus mit beleidigenden Texten zu labeln, um Verfahren zu trainieren, die nicht nur zwischen neutralen und Hasskommentaren unterscheiden können, sondern auch den Unterschied zwischen „nur" beleidigender Sprache und Hate Speech erkennen können. Zudem wollen wir mit anderen Features als Eingabevektoren experimentieren, wie beispielsweise mit den bereits erwähnten Word Embeddings oder indem die Wörter auf ihre Grundform reduziert werden.

Um auch radikale Inhalte maschinell klassifizieren zu können, bedarf es einer hinreichenden Menge von Trainingsdaten, welche zuvor von Experten händisch klassifiziert wurden. Hierfür muss allerdings im Vorfeld klar definiert werden, welche radikalen Aussagen bereits auf das Handeln vorbereiten. Die Textmuster werden anschließend für maschinelle Lernverfahren eingesetzt. In diesem Kontext muss klar abgegrenzt sein was noch als freie Meinungsäußerung gilt und was bereits rechtswidrig ist. Denn das Ziel der frühzeitigen Erkennung von Hate Speech und Radikalisierung ist es, Rechtswidrigkeit einzudämmen, nicht das subjektive Recht auf freie Rede sowie freie Meinungsäußerung einzuschränken.

Literaturverzeichnis

[Bu15] Bundeskriminalamt: Analyse der Radikalisierungshintergründe und -verläufe der Personen, die aus islamistischer Motivation aus Deutschland in Richtung Syrien oder Irak ausgereist sind. Wiesbaden, 2015.

[BW14] Burnap, Peter; Williams, Matthew Leighton: Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. 2014.

[Da17a] Davidson, Thomas; Warmsley, Dana; Macy, Michael; Weber, Ingmar: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media. 2017.

[Da17b] Davidson, Thomas; Warmsley, Dana; Macy, Michael; Weber, Ingmar: Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. ICWSM '17, S. 512–515, 2017.

[De17] Del Vigna12, Fabio; Cimino23, Andrea; Dell'Orletta, Felice; Petrocchi, Marinella; Tesconi, Maurizio: Hate me, hate me not: Hate speech detection on Facebook. 2017.

[Di80] Dieckmann, Walther: Sprache in der Politik? Die Rolle der Sprache in der Politik. Carl Hanser VerlG, 1980.

[GBH09] Go, Alec; Bhayani, Richa; Huang, Lei: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12):2009, 2009.

[Ha19] Hawkins, Jeff; Lewis, Marcus; Klukas, Mirko; Purdy, Scott; Ahmad, Subutai: A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex. Frontiers in Neural Circuits, 12:121, 2019.

[Ko15] Koch, Dominik: Verbesserung von Klassifikationsverfahren: Informationsgehalt der k-Nächsten-Nachbarn nutzen. BestMasters. Springer Fachmedien Wiesbaden, 2015.

[Me13] Meibauer, Jörg: Hassrede : von der Sprache zur Politik. In: Hassrede - hate speech : interdisziplinäre Beiträge zu einer aktuellen Diskussion. Gießener Elektronische Bibliothek, Gießen, S. 1–16, 2013.

[MM08] McCauley, Clark; Moskalenko, Sophia: Mechanisms of political radicalization: Pathways toward terrorism. Terrorism and political violence, 20(3):415–433, 2008.

[Ne13] Neumann, Peter: Radikalisierung, Deradikalisierung und Extremismus. Aus Politik und Zeitgeschichte, 63(29-31):3–10, 2013.

[Wi05] Wiktorowicz, Q.: Radical Islam Rising: Muslim Extremism in the West. G - Reference,Information and Interdisciplinary Subjects Series. Rowman & Littlefield, 2005.