Automatische Klassifikation offensiver deutscher Sprache in sozialen Netzwerken

Christoph Demus, Dirk Labudde, Jonas Pitz, Nadine Probol, Mina Schütz und Melanie Siegel

1 Einleitung

Die sozialen Medien wie Twitter, Facebook und auch die Kommentarspalten der Online-Präsenzen von Zeitungen und Radiosendern werden zunehmend von Menschen dominiert, die diffamieren, beleidigen und bedrohen. Automatisch generierte Nachrichten werden verwendet, um den Eindruck zu erwecken, dass diese extremen Meinungen in der Bevölkerung weit verbreitet sind, aber auch, um politische Gegner mundtot zu machen. Infolgedessen gelingt es vielen Betreibern von Social-Media-Webseiten nicht mehr, Nutzerbeiträge manuell zu moderieren, und das bedeutet für

C. Demus · D. Labudde
Digital Intelligence & Investigation, Fraunhofer SIT, Darmstadt, Deutschland
E-mail: cdemus@hs-mittweida.de; christoph.demus@sit.fraunhofer.de

D. Labudde
E-mail: dirk.labudde@sit.fraunhofer.de

J. Pitz · N. Probol · M. Schütz · M. Siegel (✉)
Forschungszentrum für Angewandte Informatik, Hochschule Darmstadt, Darmstadt, Deutschland
E-mail: melanie.siegel@h-da.de

J. Pitz
E-mail: jonas.pitz@h-da.de

N. Probol
E-mail: nadine.probol@stud.h-da.de

M. Schütz
E-mail: mina.schuetz@h-da.de

S. Jaki und S. Steiger (Hrsg.), *Digitale Hate Speech*,
https://doi.org/10.1007/978-3-662-65964-9_4

die Moderator*innen eine enorme psychische Belastung. Daher besteht ein dringender Bedarf an Methoden zur automatischen Identifizierung verdächtiger Beiträge. Erst in den letzten Jahren hat man damit begonnen, Methoden zur automatischen Klassifikation von Hatespeech auch auf deutschsprachige Texte anzuwenden. In diesem Beitrag stellen wir verschiedene Methoden zur Erkennung deutschsprachiger agressiver Textbeiträge vor. Grundlage unserer Darstellungen sind die Shared Tasks, die in den letzten Jahren zu diesem Thema stattgefunden haben.

Dabei gehen wir zunächst auf die Besonderheiten der deutschen Sprache ein, die es erforderlich machen, die Methoden der Forschungsliteratur für die Analyse von Hatespeech, die sich zunächst auf die englische Sprache bezogen haben, zu erweitern und zu verfeinern. Danach beschreiben wir eine Methode der Forschung zur Verarbeitung von Sprache, mit der Datensätze erzeugt und Forschungsansätze verglichen werden, die Shared Task. Anschließend werfen wir einen Blick auf die verfügbaren deutschsprachigen Datensätze, die zumeist im Kontext der Shared Tasks entstanden sind. Die Methoden zur automatischen Erkennung, die durch die Forschungsgruppen entwickelt wurden, werden danach kurz erklärt.

2 Deutschsprachige Hatespeech: Besonderheiten der Analyse

Die automatische Verarbeitung deutscher Sprache bringt einige Besonderheiten und Schwierigkeiten mit sich. Die am besten untersuchte Sprache im Natural Language Processing (NLP) ist die englische Sprache (Ortmann et al., 2019). Das ist vor allem darin begründet, dass Englisch die international am meisten verwendete Sprache ist, die nahezu überall – vor allem im wissenschaftlichen Kontext – verstanden wird und daher auch weltweit Bestandteil von Untersuchungen ist. Durch die große Verbreitung gibt es für viele NLP-Aufgaben große englische Datensätze. Deshalb ist es notwendig, bei der Abweichung von der englischen Sprache mit einem höheren Aufwand geeignete Datensätze in der entsprechenden Sprache zu finden. Eine weitere Folge der Verbreitung englischer Sprache ist, dass sehr viele Tools und Code-Bibliotheken auf die englische Sprache ausgerichtet sind und nicht immer direkt für die deutsche Sprache übernommen werden können. Ortmann et al. (2019) haben deshalb mehrere Tools für die Verarbeitung deutscher Sprache zusammengetragen und einheitlich auf Datensätzen verschiedener Domains evaluiert. Es wurden Tools für Satzsegmentierung, Tokenisierung, Part-of-Speech (POS) Tagging, morphologische Analyse, Lemmatisierung und Dependency Parsing getestet.

Neben Schwierigkeiten, die auf die geringere Ausbreitung der deutschen Sprache zurückzuführen sind, unterscheiden sich die grundlegenden Strukturen der deutschen von der englischen Sprache in einigen Punkten. Das führt dazu, dass andere Verarbeitungsschritte notwendig sind. Ein umfassender Vergleich der beiden Sprachen ist beispielsweise in Hawkins (2015) zu finden. Hier sollen jedoch nur einige Besonderheiten herausgegriffen werden, die für das NLP zum Zweck der Erkennung von Hatespeech relevant sind.

Beispielsweise ist es im Englischen meist ausreichend, ein Stemming durchzuführen, wohingegen im Deutschen Lemmatisierung aufgrund einer komplexeren Morphologie der Wörter besser geeignet ist.

Ein weiterer bedeutender Unterschied ist die Kompositabildung im Deutschen. Dadurch können lange Wörter entstehen, die jedoch nur sehr selten vorkommen, weil Substantive nahezu beliebig kombiniert werden können. Das bereitet Schwierigkeiten, da selten vorkommene Wörter nur schwer maschinell interpretierbar sind. Im Englischen werden dagegen die Wörter in der Regel getrennt geschrieben, z. B. bei „Kettenreaktion“ und „chain reaction“.

Einen Vorteil hat man im Englischen auch bei der Named-Entity Recognition, die oft als Schritt zum besseren Textverständnis genutzt wird. Im Gegensatz zum Englischen, wo nur Namen groß geschrieben werden, werden im Deutschen alle Substantive und Namen groß geschrieben werden. Die Named-Entity Recognition kann daher nicht auf die Großschreibung fokussieren wie im Englischen und muss andere Methoden einsetzen.

Für die Detektion von Hatespeech ist die Erkennung von Negationen ein wichtiger Bestandteil. Die Wortstellung in der englischen Sprache ist viel weniger variabel als in der deutschen Sprache. Dort fällt auf, dass die Negation eines Wortes nicht im nahen Umfeld stehen muss, sondern beispielsweise sogar durch Kommata getrennt von diesem stehen kann. Probleme treten insbesondere häufig mit dem Vorkommen von „dass“ auf, beispielsweise in dem Satz „Ich denke nicht, dass mir das Spaß macht.“ (Siegel & Alexa, 2020).

Zusammenfassend kann man sagen, dass die automatische Verarbeitung deutscher Sprache einige Schwierigkeiten im Vergleich zum Englischen mit sich bringt. Diese sind zum einen darauf zurückzuführen, dass Englisch weiter verbreitet ist und zum anderen, dass die deutsche Sprache einige sprachliche Besonderheiten aufweist, die dadurch weniger gut erforscht sind als die Besonderheiten der englischen Sprache.

3 Shared Task – eine Methode zur Datenerhebung und zum Vergleich von Klassifikationsansätzen

Im Bereich der Sprachverarbeitung sind sogenannte „Shared Tasks“ ein häufig erfolgreich eingesetztes Mittel, um vor allem zu neuen Fragestellungen und zu noch nicht untersuchten Sprachen Daten und Ressourcen aufzubauen und Methoden auszuprobieren. Hier sind vor allem die Shared Tasks der „SemEval“-Reihe zu nennen, die seit 1998 Wettbewerbe durchführen, die sich mit der Semantik von Sprache beschäftigen. Während sich die ersten Shared Tasks der SemEval-Reihe vor allem mit lexikalischer Semantik beschäftigten, sind in den letzten Jahren Themen wie Sentiment-Analyse, Question-Answering, Wissensextraktion und das Erkennen von Argumentstrukturen vorherrschend. Seit 2019 wird auch die automatische Klassifikation von offensiver Sprache in Shared Tasks der SemEval-Reihe untersucht.[1]

Bei einer Shared Task werden Sprachdaten zunächst gesammelt, dann annotiert, und anschließend wird der größere Teil der annotierten Daten (meist ca. 80 %) als Trainingsdaten öffentlich verfügbar gemacht. Internationale Forschungsgruppen entwickeln anhand dieser Trainingsdaten Systeme zur automatischen Klassifikation entlang der Annotationen. Die zurückbehaltenen Daten werden ohne Annotation als Testdaten diesen Forschungsgruppen gegeben. Die Forschungsgruppen wenden die entstandenen Systeme und Modelle auf diese Testdaten an und geben den Organisatoren der Shared Task ihre Ergebnisse (manchmal auch ihre Systeme und Modelle) zur Auswertung. Die Auswertung vergleicht die automatischen Klassifikationen mit den Annotationen und erstellt eine Rangliste der Systeme. Wichtig dabei ist, dass die beteiligten Forschungsgruppen ihre Methoden beschreiben und dann in einem Workshop miteinander vergleichen, so zu innovativen Kombinationen und erweiterten Sprachressourcen kommen und das Forschungsfeld damit vorantreiben.

Die Klassifikation von offensiver/aggressiver Sprache ist seit dem Jahr 2018 Gegenstand von Shared Tasks. In dem Jahr gab es Shared Tasks zu italienischer Hassrede in Twitter und Facebook (Bosco et al., 2018), zu italienischer und englischer Misogynie (Fersini et al., 2018), zur Erkennung von Aggression in Englisch und Hindi (Kumar et al., 2018) und auch schon zur automatischen Erkennung offensiver deutscher Sprache (Wiegand et al., 2018b). Im Jahr 2019 fand eine Neuauflage der GermEval mit erweiterten Daten statt (Struss et al., 2019).[2] Außerdem fan-

[1] (siehe https://semeval.github.io/SemEval2021/tasks.html, https://alt.qcri.org/semeval2020/index.php?id=tasks, https://alt.qcri.org/semeval2019/index.php?id=tasks).

[2] Im Jahr 2021 wird die GermEval-Reihe mit neuen Daten fortgeführt: https://germeval2021toxic.github.io/SharedTask/.

Tab. 1 Shared Tasks und Klassifikationsaufgaben

Jahr	Shared task	Subtasks	Sprachen
2018	HaSpeeDe	A – binär auf Facebook-Daten B – binär auf Twitter-Daten C – mit beiden Daten-Arten (Cross)	Italienisch
2018	EVALITA	A – Misogyny Identification (binär) B – Misogynistic Behaviour and Target Classification	Italienisch, Englisch
2018	TRAC	A – Overtly aggressive, covertly aggressive, non-aggressive	Hindi, Englisch
2018	GermEval	A – Offense or other (binär) B – Profanity, Insult, Abuse	Deutsch
2019	GermEval	A – Offense or other (binär) B – Profanity, Insult, Abuse C – explicit, implicit	Deutsch
2019	OffensEval	A – Binäre Klassifikation B – Offense types (Hassrede – Profanity) C – Offense target identification (Individual – Group – Other)	Englisch
2019	SemEval	A – binäre Klassifikation B – Ziel als individuell oder generisch	Spanisch, Englisch
2019	HASOC	A – binär B – Hate Speech, Offensive, Profane C – Targeted, Untargeted	Deutsch, Hindi, Englisch
2020	TRAC	A – Overtly, Covertly or Non-Aggressive B – gendered or non-gendered	Bengalisch, Hindi, Englisch
2020	OffensEval	A – Binäre Klassifikation B – Offense types (Hassrede – Profanity) C – Offense target identification (Individual – Group – Other)	Arabisch, Dänisch, Englisch, Griechisch, Türkisch
2020	HaSpeeDe	A – binär (Hate or Not) B – Stereotype Detection C – Nominal Utterance Detection	Italienisch
2020	HASOC	A – binär (Hate or Not) B – Hate, Profane and Offensive	Tamil, Malayalam, Hindi, Englisch, Deutsch
2021	GermEval	A – Binäre Klassifikation B – Engaging Comment Classification (binär, besonders gute und engagierte Kommentare) C – Fact-Claiming Comment Classification (binär)	Deutsch
2021	HASOC	A – Englisch und Hindi B – Dravidian Languages C – Arabic Misogyny Identification D – Urdu	Englisch, Hindi, Dravidian, Arabisch, Urdu

den 2019 Shared Tasks zur Klassifikation englischer offensiver Sprache (Zampieri et al., 2019b), zur mehrsprachigen englisch-spanischen Erkennung von Hassrede gegen Immigranten und Frauen (Basile et al., 2019) und zur Klassifikation von Hatespeech und offensivem Inhalt in Indo-Europäischen Sprachen (Mandl et al., 2019) statt. Im Jahr 2020 wurden die Shared Tasks der Reihen TRAC (Kumar et al., 2020) mit zusätzlichen Daten in bengalischer Sprache, OffensEval (Zampieri et al., 2020) mit zusätzlichen Daten in Arabisch, Dänisch, Englisch, Griechisch und Türkisch, HaSpeeDe (Sanguinetti et al., 2020) mit zusätzlichen Daten einer neuen Domäne und HASOC (Mandl et al., 2020) mit zusätzlichen Daten in Tamilisch, Malayalam, Hindi, Englisch und Deutsch fortgeführt. HASOC wurde auch 2021 mit neuen Sprachen (Englisch, Hindi, Dravidian, Arabisch, Urdu) durchgeführt, ebenso GermEval mit neuen Daten.[3]

Alle Shared Tasks haben als grundlegende Aufgabe die binäre Klassifikation der Texte als offensiv/aggressiv oder nicht. In den meisten Fällen kommen aber weitere Klassifikationen hinzu. Bei der GermEval 2018 war das eine feinere Klassifikation der offensiven Texte als *Abuse, Insult* oder *Profanity* (Ruppenhofer et al., 2018). EVALITA 2018 hatte in der Klassifikationsaufgabe für die feinere Klassifikation sogar fünf Klassen von Misogynie (Fersini et al., 2018). OffensEval 2019 unterschied in zwei Subtasks die Arten offensiver Sprache in *Hassrede* und *Profanity* sowie die Ziele von Hassrede in *Individuum – Gruppe – Andere* (Zampieri et al., 2019b). Andere Klassifikationsaufgaben betreffen den Ursprung der Daten (Bosco et al., 2018) oder auch die Unterteilung der offensiven Texte in explizite oder implizite Aussagen (Struß et al., 2019) (siehe Tab. 1). Die Gestaltung der Aufgaben bestimmt natürlich auch die Annotation der Daten.

Diese Aktivitäten führten dazu, dass große annotierte Datensätze für verschiedene Sprachen verfügbar sind, darunter auch Datensätze für die deutsche Sprache.

4 Daten für die Klassifikationsaufgabe

Eine wesentliche Grundlage für die Entwicklung von Methoden zur automatischen Klassifikation sind Datensätze mit annotierten Daten. Diese müssen in ausreichender Zahl vorhanden und qualitativ hochwertig sein. In diesem Abschnitt beschäftigen wir uns daher mit den deutschsprachigen Datensätzen.

[3] http://fire.irsi.res.in/fire/2021/hasoc, https://germeval2021toxic.github.io/SharedTask/.

4.1 Art der Datensammlung

Die weitaus häufigste Quelle für Hatespeech-Datensätze ist Twitter. Dies ist vor allem auf die gut zugängliche API von Twitter zurückzuführen. GermEval 2018 und 2019 nutzten Twitter-Daten, GermEval 2021 dagegen die Facebook-Seite einer politischen Talkshow. Facebook ist ein weiteres Netzwerk, das zur Datensammlung genutzt wird. Die Shared Task TRAC 2020 nutzte YouTube-Kommentare.

Wenn man einfach alle deutschsprachigen Tweets (oder auch Facebook Posts) in einem bestimmten Zeitraum sammeln würde, wäre das Datenset, das dabei herauskommt, äußerst schlecht balanciert: Es wären viel zu wenige Beispiele für Hatespeech darin enthalten. Wenn man auf so einem Datenset ein Modell maschinell lernen würde, dann würde dieses Modell am besten funktionieren, wenn es immer „No Hate“ klassifizieren würde. Die Fehlerrate wäre sehr gering, aber das gelernte Modell wäre nicht nutzbar, um Hatespeech zu klassifizieren. Das Problem wird ausführlich in Wiegand et al. (2019) beschrieben. Um die Daten zu verdichten, haben die Organisator*innen der Shared Tasks verschiedene Methoden entwickelt. Diese Methoden führen jedoch häufig zu Daten, die ein verzerrtes Bild liefern (auf Englisch „Biased Data“). Wenn man einfach nach Hass-Schlüsselwörtern suchen würde, dann macht man es der automatischen Klassifikation extrem leicht, die ebenfalls wieder nach diesen Schlüsselwörtern suchen muss. Ein solches Modell ist aber nicht auf neue Daten und weitere Schlüsselwörter übertragbar. Eine andere Möglichkeit ist, nach Themen zu suchen, zu denen häufig Hasskommentare gepostet werden. Das Problem dabei ist, dass sich diese Themen im Laufe der Zeit verändern. War es in den Jahren 2018 und 2019 vor allem das Thema „Flüchtlinge“, so hat sich das mit dem Aufkommen der Corona-Pandemie verlagert. Bei der GermEval 2018 und 2019 war es z. B. so, dass Tweets, die sich auf die Kanzlerin Merkel bezogen, vor allem Hassrede waren, sodass die Klassifikatoren falsche Schlüsse gezogen hätten. Daher wurden gezielt Tweets der CDU hinzugenommen, die Frau Merkel in einen positiven Kontext stellen. Vorsicht ist auch geboten, um zu verhindern, dass vor allem Hasskommentare von wenigen Autoren verwendet werden, wie Wiegand et al. (2019) beschreiben. In dem Fall könnte es passieren, dass sich die Klassifikatoren an der Sprache des Autors orientieren und die entstandenen Modelle wiederum auf neue Daten nicht anwendbar sind. Bei der GermEval 2018 und 2019 wurden zunächst mit Schlüsselwörtern, die auf Hass hindeuten, eine große Menge von Accounts identifiziert, von denen häufig in offensiver Sprache gepostet wurde. Aus diesen Accounts wurde ein Teil der Timeline extrahiert, anschließend wurden weitere Posts hinzugezogen, um „Biased Data“ zu vermeiden. Es wurde streng darauf geachtet, dass Trainings- und Testdaten aus verschiedenen Accounts stammten (Struß et al., 2019). Dennoch sind die so entstandenen Daten vor allem aus dem

Themengebiet „Flüchtlinge“, das zur Zeit der Datensammlung vorherrschend war. HASOC 2020 (Mandl et al., 2020) versuchte einen neuen Weg der Datensammlung: Aus dem kompletten Archiv von Twitter für Mai 2019 wurden Daten der beteiligten Sprachen herausgezogen. Auf den Daten der HASOC 2019 und GermEval 2018 wurde ein SVM-Modell trainiert, das einen F1-Score von etwa 0,5 hat. Alle Tweets, die damit als Hassrede klassifiziert wurden, wurden ins Datenset aufgenommen, zusätzlich 5 % der Daten, die nicht als Hassrede klassifiziert wurden. Diese Art der Datensammlung führte zu realistischeren Daten und machte die Aufgabe der automatischen Klassifikation extrem schwierig, sodass die Ergebnisse deutlich schlechter waren als die auf anderen Datensets.

4.2 Annotation der Daten

Die Annotation der Daten ist extrem zeitaufwändig und muss sehr sorgfältig gemacht werden, damit die Daten überhaupt für das automatische Training nutzbar sind. Es muss entschieden werden, welche Personen die Annotationen durchführen, nach welchen Standards annotiert wird und wie viele Personen jeweils einen Tweet bzw. Post annotieren.

Grundsätzlich gibt es drei Möglichkeiten, wer die Daten annotiert (Poletto et al., 2020). Im besten Fall werden die Daten von ausgewählten Fachexperten annotiert. Das ist aufgrund des hohen Aufwands jedoch nicht immer möglich, weshalb Amateure zur Annotation herangezogen werden. Das können wiederum ausgewählte Personen sein (z. B. Studierende), deren fachliche Herkunft bekannt ist. Die dritte Möglichkeit ist die Nutzung von Crowdsourcing Plattformen, wo die konkreten Annotator*innen im Vorfeld nicht bekannt sind. Letztere Methode ist aber nützlich, wenn Daten vor allem von vielen Personen annotiert werden sollen, um ein breites Bild der Gesellschaft zu erhalten.

Die GermEval 2018 (Wiegand et al., 2018b) hat zunächst Annotationsrichtlinien dafür aufgestellt, ebenso wie andere Shared Tasks es getan haben, z. B. in Zampieri et al. (2019a). Die Annotationen wurden zunächst von den drei Organisator*innen der Shared Task durchgeführt, nachdem sie sich nach vielen Tests sicher waren, dass sie eine gute Annotationsübereinstimmung erreicht hatten. Dann wurde aber jeder Tweet nur von einer Person annotiert. Andere Shared Tasks gingen einen anderen Weg und ließen die Daten von Personen annotieren, die dafür beauftragt wurden, z. B. Studierende, wie HASOC 2020. In dem Fall wurden aber alle Tweets/Posts von mehreren Personen annotiert und die Ergebnisse verglichen.

4.3 Datenqualität

Für das Training von Hatespeech-Klassifikatoren werden große Mengen annotierter Trainingsdaten benötigt. Die Beschaffung und Annotation dieser Daten ist in der Regel von hohem manuellem Aufwand, doch dieser ist am Ende für eine ausreichende Datenqualität und Datenquantität von zentraler Bedeutung.

Die Datenqualität (Abb. 1) kann unter drei Aspekten betrachtet werden: Interpretierbarkeit *(Interpretability),* Relevanz *(Relevancy),* Genauigkeit *(Accuracy)* (Kiefer, 2016). Die Interpretierbarkeit beschreibt die Erwartung des Konsumenten (Maschine/Algorithmus oder Mensch) an die Daten. Es müssen verschiedene Ansprüche erfüllt sein, damit die Daten überhaupt verarbeitet werden können. In der Detektion von Hatespeech ist beispielsweise zu überlegen, wie mit Texten in Bildern umgegangen wird, da NLP-Algorithmen nur Text verarbeiten können. Ein Bild wäre demzufolge nicht interpretierbar. Die Relevanz gibt an, wie geeignet die Daten zum Lösen des konkreten Problems oder der Fragestellung sind. In der Detektion von Hatespeech ist unter diesem Punkt die Auswahl der Daten anzusiedeln, d. h. es sollte ein gewisser Teil Hatespeech, aber nicht nur Hatespeech enthalten sein und es muss beachtet werden, dass der Datensatz keinen Bias enthält. Der dritte Punkt, die Genauigkeit, gibt schließlich an, inwieweit die Daten die Realität widerspiegeln. Da in der Regel nicht alle existierenden Daten genutzt werden können, weil es schlicht deutlich zu viele und nicht alle Daten (öffentlich) verfügbar sind, sollte der ausgewählte Datensatz trotzdem probieren, die realen Daten adäquat abzubilden.

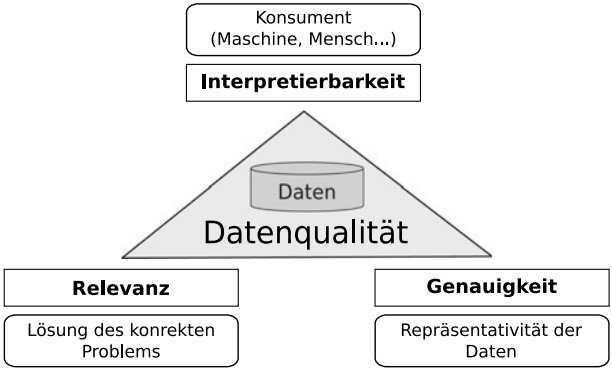

Abb. 1 Aspekte der Beurteilung von Datenqualität

Die bisher genannten Merkmale dienen der Einschätzung der Qualität der Daten selbst. Darüber hinaus spielt auch die Qualität der Annotation der Daten zur Nutzung für überwachte Lernverfahren eine wichtige Rolle. Ziel ist immer eine einheitliche und konsistente Annotation. Dafür sind detaillierte Annotationsrichtlinien notwendig, die eine klare Trennung der zu annotierenden Klassen (z. B. Hatespeech – Ja/Nein) vorgeben. Die Meinung und Erfahrung des Einzelnen sollten aus technischer Sicht keine Rolle spielen, da sonst auch das System keine klare Grenze lernen kann. In der Realität ist das im Bereich Hatespeech-Detektion ein großes Problem, da Hatespeech kaum genau definiert werden kann. Sehr häufig gibt es Grenzfälle, die auch durch umfassende Diskussion nicht eindeutig geklärt werden können. Ross et al. (2017) haben in ihrem Versuch zur binären Annotation von Hatespeech sogar gezeigt, dass das Inter-Annotator-Agreement höher sein kann, wenn keine Definition von Hatespeech vorgegeben wird.

Wie bereits angesprochen, wird in der Praxis häufig das Inter-Annotator-Agreement als Maß für die Güte der Annotationen herangezogen: Je größer die Übereinstimmung ist, desto besser die Annotationen. Als konkrete Werte dienen dabei meist Cohens Kappa oder Fleiss Kappa, bei denen die Übereinstimmung der Annotationen von zwei oder mehreren Annotator*innen gemessen werden (Struß et al., 2019; Bretschneider & Peters, 2017; Ross et al., 2017). Problematisch ist jedoch, dass die Werte nur bedingt vergleichbar sind, weil sie auf unterschiedliche Weisen erhoben wurden. In Struß et al. (2019) und Mandl et al. (2019) wurde beispielsweise jeweils nur eine geringe Menge Kommentare von mehreren Personen annotiert, um den Kappa-Wert zu bestimmen, und der Rest der Kommentare wurde jeweils nur noch von einer Person annotiert. Im Gegensatz dazu wurde in Bretschneider und Peters (2017) und Ross et al. (2017) der ganze Datensatz von mehreren Personen annotiert. Auch die Anzahl der Annotator*innen, über die das Inter-Annotator-Agreement berechnet wird, variiert.

Aufgrund der Probleme bei der Annotation von Daten haben Hanke et al. (2020) das Inter-Rater-Agreement-Learning vorgeschlagen, womit insbesondere die Verlässlichkeit von Annotator*innen bestimmt werden kann. Zur Beurteilung fließen zum einen Eigenschaften der Annotatoren ein, wie Vorerfahrung, Expertise im entsprechenden Gebiet und ggf. themenspezifische Merkmale. Zum anderen werden die Dauer der Annotation pro Text und die Konsistenz gemessen. Zur Messung der Konsistenz bekommt jede Person einige Texte doppelt zur Annotation, was einen Vergleich möglich macht. Werden gleiche Daten von einer Person oft (im Idealfall immer) gleich annotiert, deutet das auf eine hohe Konsistenz hin. Der Nutzen der Analyse besteht darin, dass die Annotationen der verschiedenen Annotator*innen anschließend gewichtet werden können und somit ein verlässlicherer Gold-Standard erstellt werden kann.

4.4 Datenquantität

Der Einfluss der Datenquantität – der Menge der vorhandenen Trainingsdaten – auf das Klassifikationsergebnis ist schwer einschätzbar. Es ist bekannt, dass klassische Machine-Learning-Modelle wie SVM und Naive Bayes in der Regel weniger Trainingsdaten benötigen als hochkomplexe Deep-Learning-Modelle wie Transformer oder Deep Neural Networks (Zampieri et al., 2019b; Kumar et al., 2020). Allerdings erreichen letztere bei ausreichender Menge vorhandener Trainingsdaten meist ein besseres Klassifikationsergebnis. Es stellt sich daher die Frage, ab welcher Datenmenge es sinnvoll ist, Deep-Learning-Modelle zu nutzen und bis zu welcher Menge vorhandener Daten klassische Modelle besser funktionieren. Zum aktuellen Zeitpunkt ist uns keine Untersuchung bekannt, die das Problem umfassend untersucht. In engem Zusammenhang mit dem Problem steht die Abschätzung der nötigen Trainingsdatenmenge (Sample Size Determination) (Figueroa et al., 2012), jedoch ist uns auch dabei noch kein Vergleich zwischen Deep-Learning und klassischen Modellen bekannt. Eine Analyse für SVM und Naive Bayes haben beispielsweise Riekert et al. (2021) vorgenommen. Doch obwohl einige Analysen in dieser Richtung existieren, sind Vergleiche zwischen diesen kaum möglich, weil sich die Ergebnisse je nach Domain, den konkreten Modellen, Daten und verwendeten Features unterscheiden.

Auch in dieser Hinsicht können Shared Tasks eine hilfreiche Möglichkeit sein, einen Eindruck zu bekommen, inwieweit die Menge der Trainingsdaten die Ergebnisse beeinflusst. Insbesondere die Ergebnisse der GermEval 2018 (Wiegand et al., 2018b) und der GermEval 2019 (Struß et al., 2019) können gut verglichen werden, weil 2019 der Trainingsdatensatz von 2018 um 7526 Kommentare (3994 neue Kommentare plus 3532 Testdaten von 2018) erweitert wurde. Die Ergebnisse haben gezeigt, dass sich trotz der mehr als doppelt so großen Trainingsdatenmenge der beste F1-Score nur um 0,0018 und damit unwesentlich verbessert hat (von 0,7677 auf 0,7695). Eine geringe Verbesserung (+0,0380) hat sich im Median der F1-Scores abgezeichnet, was vermutlich daran lag, dass 2019 schon bekannt war, welche Modelle 2018 gut funktioniert haben. Neben klassischen Modellen, die 2018 dominierten, wurden 2019 auch Transfomer-Modelle eingereicht. Jedoch konnten auch damit keine deutlichen Verbesserungen erzielt werden.

4.5 Deutschsprachige Datensätze

Eine Übersicht über verfügbare deutschsprachige Datensätze gibt Tab. 2. Der größte Datensatz ist der der GermEval Shared Task 2019 (Struß et al., 2019). Dieser

beinhaltet manuell annotierte Twitter-Kommentare. Für die Shared Task wurde der Datensatz der GermEval Shared Task 2018 erweitert. Insgesamt beinhaltet der Datensatz damit 15,567 Kommentare, die einerseits binär mit den Klassen OFFENSE oder OTHER und andererseits nach einer feineren Klassifikation mit ABUSE, INSULT und PROFANITY für die Klasse OFFENSE annotiert sind. Darüber hinaus gab es bei der GermEval 2019 eine Subtask zur Klassifikation der offensiven Tweets in explizit oder implizit. Dafür wurden Trainings- und Testdaten mit einem Gesamtumfang von 2888 annotierten Kommentaren zur Verfügung gestellt.

Beim GermEval Shared Task 2021 stand wie schon in vorherigen Jahren die Analyse von Kommentaren in sozialen Netzwerken im Vordergrund. Dafür wurde ein 4188 Kommentare umfassender Datensatz mit Kommentaren vom Facebookauftritt einer deutschen Talkshow von den Organisatoren bereitgestellt (Risch et al., 2021). Die Daten stammen aus dem Jahr 2019. Entsprechend der drei Subtasks wurde die Toxizität der Kommentare annotiert, ob ein Kommentar positiv zur Diskussion beiträgt (Engaging Comment Classification) und ob Kommentare Tatsachenbehauptungen enthalten (Fact-Claiming Comment Classification). Alle drei Annotationen sind binär.

Ein weiterer rund 5600 Facebook- und Twitterkommentare enthaltender Datensatz wurde von Bretschneider und Peters (2017) mit dem Ziel der Detektion von Hass gegen Ausländer erstellt. Dabei wurden Kommentare von drei ausländerfeindlichen Gruppen auf Facebook analysiert. Annotiert wurden jeweils, ob ein Kommentar Hatespeech enthält und wenn ja, wie ausgeprägt die Hatespeech ist (moderate oder clearly). Durch letztere Einschätzung können Grenzfälle von eindeutiger Hatespeech unterschieden werden. Darüber hinaus wurde das jeweilige Ziel der Hatespeech einer von sechs Gruppen (Targets) zugeordnet, sofern der Kommentar ein Target identifiziert. Die annotierten Gruppen sind u. a. Ausländer, Politiker, Medien und die Facebook-Community.

2019 wurde für die HASOC Shared Task ein Datensatz mit 4669 deutschsprachigen Kommentaren erstellt und annotiert (Mandl et al., 2019). Dabei gab es wie bei der GermEval Shared Task eine binäre Grobklassifizierung in Offensive und Non-Offensive und eine Feinklassifizierung in Hate, Offensive oder Profane. Mit der Feinklassifizierung wurden nur Kommentare klassifiziert, die in der Grobklassifizierung der Klasse Offensive zugeordnet wurden. Neben deutschen Kommentaren wurden für diese Shared Task auch annotierte Datensätze in Englisch und Hindi zur Verfügung gestellt, wodurch ein Vergleich zwischen verschiedenen Sprachen ermöglicht werden sollte.

Von Ross et al. (2017) wurde ein kleiner 470 Twitter-Kommentare umfassender Datensatz erstellt. Hierbei stand nicht das Training von Hatespeech-Klassifikatoren

im Vordergrund, sondern die Messung der Verlässlichkeit von Hatespeech-Annotationen unter Vorgabe unterschiedlicher Annotationsrichtlinien. Für die Erstellung des Goldstandards wurde jeder Kommentar von zwei Personen annotiert. Dabei wurde zunächst binär klassifiziert (Hatespeech oder nicht Hatespeech) und anschließend wurde die Stärke der Hatespeech auf einer Skala von 1 bis 6 bewertet.

5 Klassifikationsmethoden und Ergebnisse der Shared Tasks

In der Analyse natürlicher Sprache werden Methoden des maschinellen Lernens eingesetzt. Beim maschinellen Lernen von Klassifikationen geht es darum, aus Textdaten Modelle abzuleiten, mit denen neue Textdaten klassifiziert werden. Die Methoden lassen sich grob in überwachtes Lernen (Supervised Learning) und unüberwachtes Lernen (Unsupervised Learning) unterscheiden. Beim Supervised Learning stehen Dokumente zum Training zur Verfügung, die manuell klassifiziert sind, wie die Datensätze, die wir im Abschn. 4.5 beschrieben haben. Weil für das Supervised Learning große Mengen annotierter Daten benötigt werden, hat man nach Methoden gesucht, die mit Daten ohne Annotationen arbeiten. Auf der Basis von Daten, aber ohne Annotationen arbeitet daher das Unsupervised Learning. In den letzten Jahren kamen die Transformer als Verfahren dazu. Dabei werden die Textdaten mithilfe von Word Embeddings in numerische Vektoren überführt, wobei der Wortkontext berücksichtigt wird. Dieser Schritt fällt unter das Unsupervised Learning, denn er benötigt keine Annotationen. Anschließend kann mit einer kleineren Menge annotierter Daten das „Finetuning“ durchgeführt werden.

5.1 Supervised Learning

Eine Möglichkeit des Supervised Learnings ist das Lernen auf Merkmalen (Features). Dazu werden die Aspekte des Texts benannt, die einen Einfluss auf die Entscheidung haben könnten. Das können Wörter aus vorgegebenen Wortlisten sein, die im Text vorkommen, aber auch die Verwendung von Emojis, Satzzeichen, syntaktische Kategorien, Groß- und Kleinschreibung, Sentiment und andere. Die Aspekte werden als numerische Daten kodiert, also z. B. die Anzahl der Hasswörter im Text oder die Anzahl der Ausrufezeichen. Auch Metadaten wie Zeitangaben oder Autorenschaft können – falls vorhanden – als Features verwendet werden. Die Textdaten werden automatisch mit diesen Aspekten angereichert, sodass jeder Text als nume-

Tab. 2 Übersicht annotierter deutschsprachiger Datensätze

Datensatz	Quelle	Anzahl	Annotationen
Bretschneider und Peters (2017)	Facebook	5600	Grob: Hatespeech – Ja/Nein Ausprägung: moderate/clearly
Ross et al. (2017)	Twitter	470	Hatespeech – Ja/Nein Stärke: Skala 1–6
GermEval 2018 und 2019	Twitter	15,567	Grob: Offense, Other Fein: Abuse, Insult, Profanity
	Twitter	2888	Implizit, Explizit
HASOC 2019	Twitter, Facebook	4669	Grob: Offensive, Not Offensive Fein: Hate, Offensive, Profane
GermEval 2021	Facebook	4188	Toxic/Not toxic Engaging Comments Fact-Claiming Comments

risches Datum vorliegt. Es wird dann berechnet, welchen Einfluss welcher Aspekt auf die Klassifikation hat und damit ein Modell aufgebaut.

Eine andere Möglichkeit sind Wortlisten und andere lexikalische Methoden. Dazu zählen unter anderem Bag-of-words (BOW), N-Grams, Lemmatisierung und Stemming. Eine Hate-Wortliste enthält beispielsweise Wörter, die auf Hass schließen lassen. Diese Wortlisten entstehen auf unterschiedliche Arten, wobei sie häufig aus mehreren Projekten kompiliert werden. BOW funktioniert ähnlich, erstellt eine solche Liste jedoch automatisch aus den Trainingsdaten, indem die Wörter in den verschiedenen Klassen miteinander verglichen werden (Alrehili, 2019). Um herauszufinden, welche Wörter in einem Text besondere Bedeutung haben, kann die Term Frequency-Inverse Document Frequency (TF-IDF) verwendet werden.

In einigen Fällen werden sogenannte „N-Grams“ verwendet. Anstelle von Wörtern beim BOW treten bei N-Grams Ketten von Wörtern oder von Zeichen. N-Grams auf Wortebene sind Ketten von N (meist zwei oder drei) Wörtern, N-Grams auf Zeichenebene sind Ketten von N Zeichen (Buchstaben, Satzzeichen, Leerzeichen etc.). Das maschinelle Lernen lernt dabei die Häufigkeit des Vorkommens der N-Grams

in den einzelnen Klassen. N-Grams werden laut einer Studie von Alrehili (2019) am häufigsten genutzt. Diese können direkt als Features verwendet werden, wie zum Beispiel in Roy et al. (2020) oder Wiegand et al. (2018a).

Mithilfe von Part-of-Speech (POS) Tagging kann bestimmt werden, welcher syntaktischen Kategorie ein Wort angehört. Kombiniert man beispielsweise POS Tagging mit N-Grams, so lassen sich Rückschlüsse auf Wortart-Kombinationen schließen.

Zu den klassischen Machine-Learning-Methoden zählen unter anderem Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Decision Trees (DT), Gradient Boosting (GB) und k-nearest-neighbour (KNN). Roy et al. (2020) untersuchen verschiedene Methoden bei der Erkennung von Hatespeech und kommen zum Schluss, dass keine der angewandten Methoden konsistent besser als andere ist. Abhängig von den verwendeten Features (hier verschiedene N-Grams), erreichten sie mit RF, SVM oder KNN die besten Ergebnisse. Auffällig ist jedoch, dass die F1-Werte bei diesen Experimenten für „Nicht Hatespeech“ bei etwa 0,97 lagen, „Hatespeech“ dagegen bei gerade einmal 0,54 im besten Fall. Dies bedeutet, dass noch immer etwa die Hälfte aller „Hatespeech“-Daten falsch klassifiziert wurden. Die Autoren betonen jedoch die Unausgewogenheit der Datensätze (über 90 % der Daten waren „Nicht Hatespeech“), sodass nicht automatisch darauf geschlossen werden kann, dass die Methoden ungeeignet für bessere Klassifikationen von „Hatespeech“ sind.

Auch Kumar Sharma et al. (2018) erreichen mit SVM, RF, LR und GB Accuracy-Werte zwischen gerade einmal 0,523 (SVM) und 0,545 (RF). Durch das Trainieren mit zusätzlichen Features ist es Wiegand et al. (2018a) jedoch gelungen, deutlich höhere Werte zu erreichen.

5.2 Unsupervised Learning

Neben den überwachten Lernmethoden gibt es auch das unüberwachte Lernen (Unsupervised Learning). Dabei wird ein Modell auf nicht-annotierten Daten trainiert. Die zu lernenden Aspekte sind nicht in den Daten gekennzeichnet und der Lernalgorithmus versucht, Muster zu erkennen. Der wichtigste Vorteil davon ist, dass der Annotationsschritt wegfällt und damit potenziell mit weniger Aufwand größere Datenmengen verfügbar sind. Ein weitere Vorteil ist, dass das Modell auch Unterschiede lernen kann, die nicht in einem Feature für jedes Einzelbeispiel extrahiert worden sind. Der Nachteil ist jedoch, dass hierdurch die Erklärbarkeit der Ergebnisse erschwert wird. Im Vergleich zum überwachten Lernen werden für das

unüberwachte Lernen auch größere Datenmengen benötigt, um gute Ergebnisse erzielen zu können.

Zur Erkennung von Hatespeech mit Deep-Learning-Modellen werden die Textdaten in numerische Vektoren überführt. Hierzu verwendet man sogenannte „Word Embeddings". In diesen Vektoren ist für jedes Wort kodiert, mit welchen anderen Wörtern es im Kontext (mit welcher Wahrscheinlichkeit) auftreten kann. Es wird dabei der Kontext rechts wie auch links betrachtet. Dadurch kann man semantische Zusammenhänge zwischen Wörtern in den Trainingsdaten erkennen: Semantisch ähnliche Wörter, die in ähnlichen Kontexten auftreten, und semantisch zusammenhängende Wörter, die häufig gemeinsam auftreten.

Zu den bekanntesten Word Embeddings gehören word2vec, FastText,[4] Google Embeddings[5] und GloVe Embeddings.[6] Es ist jedoch auch möglich, durch die Deep-Learning-Modelle selbst Word Embeddings zu erstellen, deren Gewichtung in den einzelnen Trainingsdurchgängen trainiert und angepasst werden.

Die populärsten Deep-Learning-Verfahren im Bereich der Erkennung von Hatespeech basieren auf Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), Long-Short-Term-Memory (LSTM) und Recurrent Neural Networks (RNN). Bei der GermEval 2018 gab es viele Einreichungen mit LSTM (11), CNN (10) und GRU (6) (Wiegand et al., 2018b), ein Jahr später dagegen gab es keine Einreichungen mit Deep-Learning-Verfahren mehr (Struß et al., 2019). CNN können Beziehungen zwischen den benachbarten Wörtern gut erkennen, LSTM dagegen können längere Abhängigkeiten zwischen den Wörtern erkennen. Badjatiya et al. (2017) haben herausgefunden, dass Deep-Learning-Verfahren signifikant bessere Ergebnisse erreichen als klassische Machine-Learning-Verfahren. Laut einer Studie von Istaiteh et al. (2020) erreichen LSTM signifikant bessere Ergebnisse als CNN, jedoch kommen Badjatiya et al. (2017) zum gegenteiligen Ergebnis, dass CNN besser als LSTM performen. Die höchsten F1-Werte konnten jedoch mit LSTM, Word Embeddings mit zufälligen Startgewichten und Gradient Boosted Decision Trees (GBDT) erreicht werden (0,93). Das Verwenden von FastText-Embeddings oder GloVe-Embeddings hat nicht zu besseren Ergebnissen geführt. Dies kann daran liegen, dass diese vortrainierten Word Embeddings möglicherweise Wörter nicht enthalten, die im jeweiligen Kontext von Bedeutung sind, wie beispielsweise „Islamolunatic" (Pitsilis et al., 2018).

Ebenfalls gute Ergebnisse konnte Roy et al. (2020) mit einem Deep Convolutional Neural Network (DCNN) in Kombination mit einer Methode zur Adaption des

[4] https://fasttext.cc/docs/en/english-vectors.html.

[5] https://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/.

[6] https://nlp.stanford.edu/projects/glove/.

Modells auf ungesehene Daten, „K-Fold Cross-Validation“, erreichen (F1 = 0,92). Laut der Autoren ist eine K-Fold Cross-Validation bei unausgewogenen Datensätzen hochgradig performant. Bei einer K-Fold Cross-Validation wird der Datensatz in eine bestimmte Anzahl (K) Teile unterteilt. Ein Teil wird als Testset und ein anderer Teil als Validationsset deklariert. Die restlichen Teile werden als Trainingssets verwendet. Diese Aufteilung wird so oft wiederholt, bis alle möglichen Varianten durchgegangen sind.

5.3 Transformer-Modelle

Eine neue Entwicklung in der Sprachverarbeitung wurde durch sogenannte Transformer Modelle vorangetrieben. Diese unterscheiden sich von klassischen neuronalen Netzen dadurch, dass sie den sogenannten „Attention-Mechanismus“ verwenden. Der Attention-Mechanismus speichert die relevanteste Information eines verarbeiteten Satzes. Dadurch ist es möglich, die Kontextabhängigkeiten von Wörtern auch über Satzgrenzen hinweg während des Trainings zu speichern (Vaswani et al., 2017). Das hat auch dazu geführt, dass bidirektionale Sprachenmodelle verwendet werden können, die die Sätze nicht mehr nur von einer Richtung lesen, sondern von beiden Richtungen aus gleichzeitig. Eines dieser Modelle ist BERT (Bidirectional Encoder Representations from Transformers), welches in der ursprünglichen Version auf englischen Daten trainiert wurde (Devlin et al., 2019).

Solche Modelle werden auf einer großen Anzahl von generischen Daten (also Texten) vortrainiert und können damit für beliebige Arten von NLP-Aufgabenstellungen, wie Question Answering, Klassifikation, Named Entity Recognition u. a. verwendet werden (Devlin et al., 2019). Wenn vortrainierte Modelle auf spezifischen Daten weiter trainiert werden, wird dies auch „Fine-Tuning“ genannt. Dies führt dazu, dass sehr gute Vorhersagen bei Klassifikationen mit kleinen Datensätzen erreicht werden können. Mittlerweile gibt es vortrainierte monolinguale Transformer (Devlin et al., 2019; Liu et al., 2019) in vielen verschiedenen Sprachen oder auch als multilinguale Varianten (Conneau et al., 2019), die über 100 Sprachen beherrschen.

Dieser Ansatz wurde auch im Bereich der automatischen Erkennung von Hatespeech eingesetzt. Madukwe et al. (2020) haben verschiedene Experimente mit BERT und zwei Hatespeech-Datensätzen durchgeführt, um herauszufinden, welche Layer des Modells am Besten zur Klassifikation geeignet sind. Das beste Ergebnis hatte der Embedding Layer. Auch D’Sa et al. (2020) haben sich mit der Klassifikation von englischsprachiger Hatespeech, offensiver und toxischer Sprache befasst. Dabei vergleichen sie ein neuronales Netz mit FastText und BERT Embeddings und dem

klassischen Fine-Tuning von BERT. Bei allen Experimenten erzielt das Fine-Tuning mit Abstand die besten F1-Werte (0,97 F1). Aber auch bei der GermEval 2019 – die sich mit deutscher Sprache befasst hat – wurde BERT eingesetzt. Das Team bertZH (Graf & Salini, 2019) hat hierbei ein auf Deutsch vortrainiertes BERT-Modell als auch ein multilinguales verwendet, wobei die finale F1-Werte zwischen 0,43 und maximal 0,53 lagen. Ähnlich – für spanische und englische Tweet-Klassifikation – haben Stappen et al. (2020) ein eigenes Modell *AXEL* entwickelt und mit BERT und XLM (Conneau & Lample, 2019) verglichen. Safi Samghabadi et al. (2020) haben sich hingegen nur auf Misogynie und Aggressionserkennung (je drei Klassen) in Englisch, Hindi und Bengalisch fokussiert, welche auch als Sub-Kategorien bzw. Spezifizierungen im Bereich von Hatespeech sind. Als Modell wurde eine Kombination aus BERT, einem extra Attention- und Klassifikations-Layer verwendet, welches bei der Erkennung von Aggressionen einen F1-Wert von über 0,7 und bei Misogynie abhängig von der Sprache zwischen 0,8 bis über 0,92 erreichte (Safi Samghabadi et al., 2020). Mozafari et al. (2020) erforschten Rassismus in Hatespeech – mit Fokus auf Modelle, die Rassismus auf Basis der verwendeten Wörter einer sozialen Gruppe beim Training lernen. Dazu wurden mehrere Datensätze und Varianten von BERT getestet, wobei die F1-Werte auf die Testdaten zwischen 0,75 und 0,94 variierten.

Zudem wurde auch eine Studie zu Hatespeech (Florio et al., 2020) publiziert, die sich mit dem Vergleich von einer SVM und einem Transformer-Modell befasst hat. Dort wurde untersucht, wie viel Einfluss eine größere Datenmenge mit unterschiedlichen Zeitfenstern bei der Extraktion der Daten auf die jeweiligen Modelle haben. Das Ergebnis war, dass – durch die schnelle Veränderung der Themen in sozialen Netzwerken – Daten, die zeitlich näher beieinander liegen, die Klassifikationsergebnisse verbessern können, aber mehr Daten generell die Modelle weniger robust machen (Florio et al., 2020).

5.4 Methoden in den Shared Tasks

Bei den Methoden, die die Forschungsgruppen anwendeten, um die Klassifikationsaufgaben zu lösen, ist eine Entwicklung zu beobachten: Die Methodenvielfalt ist über den Zeitraum von zwei Jahren gesunken, zugunsten von neuronalen Netzen und auf BERT basierenden Transformern.

2018 wurden vor allem klassische Verfahren des maschinellen Lernens (ML) wie Support-Vector-Machines (SVM) eingesetzt, gefolgt von neuronalen Netzen (LSTM, RNN, CNN, BiLSTM, GRU) und lexikalischen Methoden wie TF-IDF, Bag-of-Words, Lexikon-Ressourcen für Hassrede und Sentiment, N-Grams und Word Embeddings. Die Gewinner-Systeme nutzten klassisches ML (SVM), lexikalische Methoden(TF-IDF) und neuronale Netze(LSTM, BiLSTM).

2019 wurden die klassischen ML-Verfahren nur noch selten eingesetzt, dafür mehr die neuronalen Netze. Die ersten Transformer-Modelle (BERT und ELMo) wurden zu den Wettbewerben eingereicht. Die Gewinner-Gruppen nutzten BERT, aber auch SVM.

Im Jahr 2020 schließlich nutzten die meisten Systeme Varianten der BERT-Transformer. Auf diesen basierten auch die Gewinner-Systeme, wobei in einem Fall ein neuronales Netz (BiLSTM) dazu kam.

Die F1-Werte der Gewinner-Systeme haben sich zwischen 2018 und 2019 nicht verbessert. Struß et al. (2019) stellen dies für die GermEval-Serie ebenfalls fest: „Compared to the previous year, this year's winning F-score is higher, but very slightly so (76,95 vs. 76,77)." Die Werte liegen 2018 zwischen 0,64 und 0,84 und 2019 zwischen 0,73 und 0,83. Im Jahr 2020 gab es zwei Besonderheiten: Der maximale F1-Wert, der bei HASOC erreicht wurde, lag lediglich bei 0,53. Mandl et al. (2020) führen das auf die neuartige Art der Datensammlung, aus der ein realistischeres Datenset entstanden ist, zurück. In der OffensEval konnte 2020 für die englische Sprache ein F1-Wert von 0,92 erreicht werden, indem das Transformer-Modell ALBERT genutzt wurde (Zampieri et al., 2019b).

6 Zusammenfassung und Ausblick

In diesem Kapitel haben wir uns mit Methoden zur automatischen Klassifikation von deutschsprachiger Hatespeech beschäftigt. Wir haben dargestellt, wie die Methoden, die für die Klassifikation englischer Sprache entwickelt worden sind, für die Verarbeitung deutscher Sprache angepasst werden müssen. Anhand einer Untersuchung von Shared Tasks zur automatischen Klassifikation von Hatespeech der letzten Jahre haben wir vielversprechende Methoden identifiziert und einen Trend von Standard-Machine-Learning-Methoden hin zu Transformer-Methoden festgestellt, wobei die Standard-Machine-Learning-Methoden nach wie vor ihre Berechtigung haben. Die meisten Forschungsgruppen beschäftigen sich jedoch ausschließlich mit binären Klassifikationen, wo schon recht gute Ergebnisse erzielt werden. Es wird sich in Zukunft zeigen, ob komplexere Klassifikationen (z. B. Art der Hatespeech, strafrechtliche Relevanz) andere Methoden benötigen. Da Daten eine wich-

tige Grundlage für Modelle zur Klassifikation sind, haben wir den öffentlich zugänglichen Datensätzen zur deutschsprachigen Hatespeech einen Abschnitt gewidmet. Die Datensätze stammen zumeist aus Twitter oder Facebook. Noch gibt es relativ wenige Datensätze für die deutsche Sprache. Durch die wachsende Zahl an Shared Tasks in dem Themenbereich ist jedoch zu erwarten, dass weitere Datensätze entstehen. Anschließend haben wir die Funktionsweise der wichtigsten aktuell eingesetzten Klassifikationsmethoden kurz dargestellt. Ein wichtiges Thema der nächsten Zeit wird die Erklärbarkeit der automatischen Klassifikationen sein, die vor allem im Kontext der Hatespeech-Klassifikation relevant ist. In diesem Bereich erwarten wir weitere Arbeiten in der nächsten Zeit, denn gerade im untersuchten Themenbereich können sich Anwender nicht mit Black-Box-Systemen zufrieden geben, deren Entscheidungen nicht nachvollziehbar sind.

Literatur

Alrehili, A. (2019). Automatic hate speech detection on social media: a brief survey. In *2019 IEEE/ACS 16th international conference on computer systems and applications (AICCSA)* (S. 1–6). http://dx.doi.org/10.1109/AICCSA47632.2019.9035228.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the international conference on world wide web (WWW), Perth, Australia* (S. 759–760).

Basile, V., Bosco, C., Fersini, E., Nozza, D., 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 *Proceedings of the 13th international workshop on semantic evaluation* (S. 54–63). https://iris.unito.it/retrieve/handle/2318/1723924/512658/S19-2007.pdf.

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. https://iris.unito.it/retrieve/handle/2318/1686264/465071/paper010.pdf.

Bretschneider, U., & Peters, R. (2017). Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii international conference on system sciences (2017)*. Hawaii International Conference on System Sciences. http://dx.doi.org/10.24251/hicss.2017.268.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Hrsg.), *Advances in neural information processing systems* (Bd. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR* abs/1911.02116. http://arxiv.org/abs/1911.02116.

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, (Long and Short Papers), Minneapolis, Minnesota*. Association for Computational Linguistics. (Bd. 1, S. 4171–4186). https://www.aclweb.org/anthology/N19-1423, http://dx.doi.org/10.18653/v1/N19-1423.

D'Sa, A. G., Illina, I., & Fohr, D. (2020). BERT and fastText embeddings for automatic detection of toxic speech. In *2020 international multi-conference on: „organization of knowledge and advanced technologies" (OCTA)* (S. 1–5). http://dx.doi.org/10.1109/OCTA49274.2020.9151853.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In *EVALITA evaluation of NLP and speech tools for Italian, 12* (S. 59). http://personales.upv.es/prosso/resources/FersiniEtAl_Evalita18.pdf.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making, 12*(1), 1–10.

Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: the challenge of time in hate speech detection on social media. *Applied Sciences, 10*(12). https://www.mdpi.com/2076-3417/10/12/4180, http://dx.doi.org/10.3390/app10124180.

Graf, T., & Salini, L. (2019). bertZH at GermEval 2019: fine-grained classification of German offensive language using fine-tuned BERT. In *KONVENS*.

Hanke, K. J., Ludwig, A., Labudde, D., & Spranger, M. (2020). Towards inter-rater-agreement-learning. In *IMMM 2020: the tenth international conference on advances in information mining and management*.

Hawkins, J. (2015). *A comparative typology of English and German*. Routledge. https://doi.org/10.4324/9781315687964.

Istaiteh, O., Al-Omoush, R., & Tedmori, S. (2020). Racist and sexist hate speech detection: literature review. In *2020 international conference on intelligent data science technologies and applications (IDSTA)* (S. 95–99). http://dx.doi.org/10.1109/IDSTA50958.2020.9264052.

Kiefer, C. (2016). Assessing the quality of unstructured data: an initial overview. In *LWDA* (S. 62–73). http://ceur-ws.org/Vol-1670/paper-25.pdf.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), Santa Fe, New Mexico, USA* (S. 1–11). Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-4401.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2020). Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (S. 1–5). https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/TRAC2book.pdf.

Kumar Sharma, H., Kshitiz, K., & Shailendra. 2018. NLP and machine learning techniques for detecting insulting comments on social networking platforms. In *2018 International conference on advances in computing and communication engineering (ICACCE)* (S. 265–272). http://dx.doi.org/10.1109/ICACCE.2018.8441728.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: a robustly optimized bert pretraining approach. http://arxiv.org/abs/1907.11692.

Madukwe, K. J., Gao, X., & Xue, B. (2020). A ga-based approach to fine-tuning BERT for hate speech detection. In *2020 IEEE symposium series on computational intelligence (SSCI)* (S. 2821–2828). http://dx.doi.org/10.1109/SSCI47803.2020.9308419.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC track at FIRE 2019: hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th forum for information retrieval evaluation* (S. 14–17). http://ceur-ws.org/Vol-2517/T3-1.pdf.

Mandl, T., Modha, S., Kumar, M. A., & Chakravarthi, B. R. (2020). Overview of the HASOC track at FIRE 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Forum for information retrieval evaluation* (S. 29–32). http://ceur-ws.org/Vol-2826/T2-1.pdf.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE, 15*(8), 1–26. https://doi.org/10.1371/journal.pone.0237861, http://dx.doi.org/10.1371/journal.pone.0237861.

Ortmann, K., Roussel, A., & Dipper, S. (2019). Evaluating off-the-shelf NLP tools for German. In *proceedings of the 15th conference on natural language processing (konvens 2019): Long Papers, Erlangen, Germany* (S. 212–222). German Society for Computational Linguistics & Language Technology. https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_55.pdf.

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence, 48*(12), 4730–4742.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. In *Language Resources and Evaluation* (S. 1–47). https://link.springer.com/content/pdf/10.1007/s10579-020-09502-8.pdf, http://dx.doi.org/10.1007/s10579-020-09502-8.

Riekert, M., Riekert, M., & Klein, A. (2021). Simple baseline machine learning text classifiers for small datasets. *SN Computer Science, 2*(3). http://dx.doi.org/10.1007/s42979-021-00480-4.

Risch, J., Stoll, A., Wilms, L., & Wiegand, M. (2021). Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 workshop on the identification of toxic, engaging, and fact-claiming comments : 17th conference on natural language processing KONVENS 2021*. https://netlibrary.aau.at/obvukloa/content/pageview/6435205.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In *Proceedings of NLP4CMC III: 3rd workshop on natural language processing for computer-mediated communication (Bochum), Bochumer Linguistische Arbeitsberichte, Sep 2016* (Bd. 17, S. 6–9). https://github.com/UCSM-DUE/IWG_hatespeech_public, http://arxiv.org/abs/1701.08118, http://dx.doi.org/10.17185/duepublico/42132.

Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access, 8,* 204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073.

Ruppenhofer, J., Siegel, M., & Wiegand, M. (2018). Guidelines for IGGSA shared task on the identification of offensive language. ms. https://projects.fzai.h-da.de/iggsa/.

Safi Samghabadi, N., Patwa, P., PYKL, S., Mukherjee, P., Das, A., & Solorio, T. (2020). Aggression and misogyny detection using BERT: a multi-task approach. In *Proceedings of the second workshop on trolling, aggression and cyberbullying, Marseille, France* (S. 126–131). European Language Resources Association (ELRA). https://aclanthology.org/2020.trac-1.20.

Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S, Stranisci, M., Bosco, C., Caselli, T., Patti, V., Russo, I., & Pisa, I. (2020). HaSpeeDe 2@ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In *Proceedings of seventh evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org*.

Siegel, M., & Alexa, M. (2020). *Sentiment-Analyse deutschsprachiger Meinungsäußerungen*. Wiesbaden: Springer Fachmedien. https://doi.org/10.1007/978-3-658-29699-5.

Stappen, L., Brunn, F., & Schuller, B. (2020). Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel. arXiv preprint http://arxiv.org/abs/2004.13850arXiv:2004.13850.

Struß, J. M., 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), Friedrich-Alexander-Universität Erlangen-Nürnberg* (S. 352–363). German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg. http://www.melaniesiegel.de/publications/2019_GermEval_overview.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Hrsg.), *Advances in neural information processing systems 30* (S. 5998–6008). Curran Associates, Inc. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018a). Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, New Orleans, Louisiana* (S. 1046–1056). Association for Computational Linguistics.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018b). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 workshop, Vienna, Austria. Austrian Academy of Sciences*. http://www.melaniesiegel.de/publications/2018_GermEval_Proceedings.pdf.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Bd. 1 (long and short papers)* (S. 602–608).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: tutorials (NAACL)*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). SemEval-2019 task 6: identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of SemEval@NAACL-HLT 2019*. https://arxiv.org/pdf/1903.08983.pdf.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, C. (2020). SemEval-2020 task 12: multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of SemEval 2020*. https://arxiv.org/pdf/2006.07235.pdf.