KI-Verfahren für die Hate Speech Erkennung: Die Gestaltung von Ressourcen für das maschinelle Lernen und ihre Zuverlässigkeit

Thomas Mandl

1 Maschinelles Lernen zum Erkennen von Hate Speech

In sozialen Netzwerken treten aufgrund der fehlenden Moderation problematische Inhalte auf. Nutzer*innen posten häufig Hassbotschaften, aggressive Äußerungen, Beschimpfungen oder Desinformation, die dann online sichtbar und verfügbar bleiben. Aufgrund der schieren Menge von Nachrichten können solche problematischen Inhalte nur automatisch erkannt werden. Ausprägungen von Hassrede sowie deren negative Folgen werden in diesem Band ebenso erläutert (Jaki in diesem Band) wie die Problematik der politischen Regulierung solcher Inhalte (Schünemann und Steiger in diesem Band).

Methoden des maschinellen Lernens und der automatischen Sprachverarbeitung werden eingesetzt, um möglicherweise problematische Posts zu identifizieren (Schäfer in diesem Band). Solche Verfahren werden gemeinhin als Methoden der Künstlichen Intelligenz (KI) bezeichnet. Einen rechtlichen Rahmen dafür setzt u. a. das im Oktober 2019 ergangene Urteil des EUGH, das den Einsatz von automatisierten Verfahren sogar für notwendig befindet (Heldt, 2020).

Bei dieser Bewertung von menschlichen Texten durch Computer rückt die ethische Dimension und vor allem der schmale Grat zwischen Meinungsfreiheit und Zensur in den Fokus. Die Gesellschaft wird KI-Methoden nur akzeptieren, wenn das Vertrauen in ihre Ergebnisse sichergestellt werden kann (Kuhlen,

T. Mandl (✉)
Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim, Hildesheim, Deutschland
E-Mail: mandl@uni-hildesheim.de

S. Jaki und S. Steiger (Hrsg.), *Digitale Hate Speech*,
https://doi.org/10.1007/978-3-662-65964-9_6

1999). Marc Zuckerberg hat in seinen vier Thesen zur Regulierung des Internets gefordert, Hate Speech eindeutig zu definieren (Lewanczik, 2019). Jedoch gestaltet sich dies sehr schwierig, denn die Vorstellungen, welche Inhalte problematisch sind, variieren je nach Person sehr stark.

Das maschinelle Lernen konnte in den letzten Jahren erhebliche Fortschritte erzielen. Algorithmen suchen nicht nach einzelnen Wörtern oder anhand von manuell erstellten Regeln nach sprachlichen Mustern, sondern bauen aus vielen Beispielen Klassifikationsverfahren auf, die letztlich nach vergleichbaren Posts suchen (Schäfer in diesem Band). Ein Schema hierfür zeigt Abb. 1.

Diese Trainingsdaten bestehen aus Beispielen für unangemessene Inhalte und aus Gegenbeispielen für akzeptable Inhalte. Da die Themen innerhalb der hasserfüllten Inhalte äußerst heterogen sein können, sollten sie möglichst breit durch Trainingsdaten abgedeckt sein. Die für das Training verwendeten Texte und die Entscheidungen dazu sind für die Entwicklung von KI-Verfahren entscheidend. Ihre Zusammenstellung hat den größten Einfluss unter allen Design-Entscheidungen bei der Implementierung von Detektionssystemen für Hassrede.

Dieser Artikel stellt zunächst kurz einige wissenschaftliche Benchmarks vor, mit denen Verfahren trainiert und evaluiert werden. Dabei lassen sich neue Trends für die Gestaltung solcher Daten beobachten. Im Anschluss werden dann Herausforderungen besprochen, die zu Verzerrungen führen können, so dass die trainierten Systeme im Realbetrieb nicht die gewünschte Erkennungsleistung zeigen. Diese Gefahr droht, wenn Daten von den Trainingsdaten abweichen. Ein Weg, diese Übertragbarkeit auf reale Situationen zu erproben, besteht in der Anwendung mehrerer anderer Datenmengen für das Testen. Auch die Erhöhung der Transparenz im Betrieb für Moderator*innen oder Nutzer*innen kann Einblick in das Funktionieren der Systeme und damit auch die angestrebte Qualität der Daten bieten.

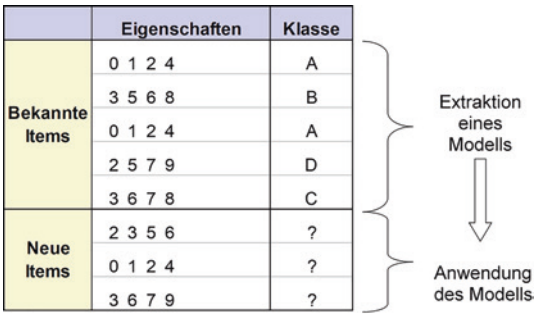

Abb. 1 Schematische Darstellung des maschinellen Lernprozesses bei der Klassifikation

2 Bestehende Benchmarks

Die Erstellung von Datensammlungen bildet ein zentrales Instrument für die Forschung zur Erkennung und Bekämpfung von Hassrede. Hierzu werden echte Tweets oder Posts aus sozialen Netzwerken systematisch gesammelt und zunächst von Menschen in zwei oder mehrere Klassen kategorisiert. Der Aufwand für die Erstellung solcher Daten ist hoch und kann nicht von allen Forscher*innen geleistet werden. Dementsprechend hat sich für diese Forschung das Prinzip der offenen Forschungsdaten etabliert. Einzelne Forschungsgruppen entwickeln Daten und stellen diese der Forschungs-Community zur Verfügung. Das führt zudem zu dem positiven Effekt, dass die verschiedensten Algorithmen von mehreren Forscher*innen anhand der gleichen Daten verglichen werden. Die Ergebnisse bei der Klassifikation sind direkt vergleichbar.

Diese Prinzipien haben beispielsweise auch bei der Forschung zum Information Retrieval zu erheblichen Fortschritten bei Suchalgorithmen geführt (Mandl, 2008; Womser-Hacker, 2013). Das Vorgehen wird oft als Organisation einer Shared Task oder als Aufbau eines Benchmarks bezeichnet. Dabei werden Daten nicht nur von der Fachwelt genutzt und gegebenenfalls auch später nachgenutzt, sondern auch die Qualität der Daten kann kritisch untersucht werden. Shared Tasks werden auch mit Daten zu Desinformation (Nakov et al., 2021), zu Bildern (Joly et al., 2020) oder Nutzer-Logfiles organisiert (Mandl et al., 2009).

So entstehen derzeit international zahlreiche Datensammlungen, um dem Problem der Hassrede begegnen zu können (siehe Madukwe et al., 2020 für einen Überblick). Auch für verwandte Themen wie Online-Extremismus entstehen solche Ressourcen (Gaikwad et al., 2021).

Im Folgenden werden der Benchmark GermEval für das Deutsche und die mehrsprachige Shared Task HASOC erläutert. Für das Deutsche wurden im Rahmen der GermEval-Initiative zwei Datensets entwickelt (Struß et al., 2019; Wiegand et al., 2018). Zugrunde lagen Twitter-Daten, welche die Organisator*innen annotieren ließen. GermEval definiert die Klassen ABUSE, INSULT und PROFANITY (Missbrauch und Beschimpfung, Beleidigung, Fluchen und Vulgarität). Die Systeme sollen als primäre Aufgabe alle Tweets in diese drei Klassen einteilen (De Smedt & Jaki, 2018). Insgesamt umfasste die Menge 2019 über 7000 Tweets. Dabei überwiegt laut den Organisatoren deutlich das extrem rechte politische Spektrum, denn 90 % der Inhalte fallen in diese Kategorie.

Neben dem Einordnen als Hassrede sollen die problematischen Tweets in zwei weitere Klassen, nämlich ‚implizit' vs. ‚explizit', sortiert werden. Gemessen

wird die Klassifikationsgenauigkeit der Systeme mit dem F1-Maß, das Recall und Precision zusammenfasst. Der Recall beschreibt, wie viel Hassrede-Beiträge gefunden wurden, und die Precision, ob dabei wirklich nur problematische Inhalte oder auch andere Inhalte zurückgeliefert wurden. Das beste System erreichte für die binäre Task (Inhalt problematisch oder nicht) ein F1-Maß von 0,76 (Struß et al., 2019).

Im Rahmen der Initiative Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC, hasocfire.github.io) wurden ebenfalls Daten für das Deutsche erstellt (Mandl et al., 2020; Modha et al., 2019). HASOC modelliert die Aufgabe zunächst als binäre Klassifikation. Die als problematisch erkannten Posts sollen dann im zweiten Schritt genauer in die folgenden Klassen unterteilt werden: HATE, OFFENSIVE und PROFANE (Hass gegen Gruppen, Aggression und Angriff gegen Einzelne, Fluchen und Vulgarität). Die Problematik der heterogenen Klassen bei Hate-Speech-Datensets diskutieren mehrere Forscher*innen (Fortuna et al., 2020) und sie wird unten erneut aufgegriffen.

Die Organisatoren von HASOC annotierten 2019 für das Deutsche 4600 Tweets und Facebook-Posts. Die Top Teams für das Deutsche nutzten überwiegend das BERT-System und liegen eng zusammen (Modha et al., 2019). Für HASOC 2020 wurden für das Deutsche über 3400 Tweets annotiert. Es zeigte sich, dass BERT und neuere Varianten wie RoBERTa zu sehr guten Ergebnissen führen, jedoch auch andere Systeme vergleichbare Werte erzielen (Mandl et al., 2020). Für den Durchgang im Jahr 2021 wurden Marathi, Hindi und Englisch angeboten und insgesamt über 10.000 Tweets annotiert (Modha et al., 2021).

Das Topic Model in Tab. 1 fasst die Themen im HASOC-2019-Datenset für das Deutsche zusammen. *Topic Modeling* versucht, die wichtigen thematischen Felder innerhalb einer Menge von Texten zu erkennen (Vayansky & Kumar, 2020). Dazu werden Wörter gesammelt, die häufiger zusammen vorkommen und zu einem Thema zusammengefasst. Diese Wörter werden der Reihenfolge ihrer Prominenz für das jeweilige Thema nach sortiert. Die Themen müssen interpretiert und mit einem Label als Überschrift versehen werden. Die Interpretation fällt nicht leicht, da auch häufig vorkommende Wörter wie Verben oder bei einem Twitter-Korpus die Namen von Nutzer*innen vorkommen.

Bei der Erstellung von Shared Tasks für Hassrede zeigen sich neben der Diversifizierung hinsichtlich der abgedeckten Sprachen noch weitere Trends. Neben der rein binären Klassifikation in problematische und unproblematische Inhalte werden weitere Ausprägungen von Hassrede berücksichtigt. So zielen beispielsweise drei Benchmarks spezifisch auf die Erkennung von Misogynie und Sexismus ab (Fersini et al., 2018; Guest et al., 2021; Ródriguez-Sánchez et al.,

Tab. 1 Topic Model aus den HASOC 2019 Trainings- und Testmengen für das Deutsche

Topic Label	Topic Wörter
Meinung	Gibt junge immer meinung land realjohr grünen kommen warum müssen
Polizei	Findbecci endlich finjafinte danke sucht polizei schwager deutschland ganze merkel
Chemnitz	Uwe_junge_mdl deutschland merkel hartes_geld chemnitz immer gehen sozialismus wählt angst
Nazis	Einfach nazis müssen berlin deutschland uwe_junge_mdl geld mainwasser migranten gewalt
Deutschland	Uwe_junge_mdl ralf69117 sagen immer deutschland dürfen gehen ekelwilfred ungebeten warum
Geflüchtete	Uwe_junge_mdl deutschland menschen deutschen deutsche welt heute flüchtlinge männer sagt
Wiltewka	Ekelwilfred wiltewka alias wilberg a…loch wilayawilanar wilke papa capitol merkel
Grüne	Ralf69117 frau polizei grünen gerade immer presse junge männer mutter

2021). Die Annotation für EXIST IberEval umfasst dabei detaillierte Unterklassen von problematischen Inhalten wie Stereotypisierung, sexuelle Gewalt und die Darstellung von Frauen als Objekt (Ródriguez-Sánchez et al., 2021).

Eine andere spezifische Verfeinerung besteht in der genauen Identifikation des Ziels von Hassrede. Dabei soll die Person oder Gruppe identifiziert werden, die das Target der Hassrede darstellt. Eine weitere Kollektion greift Ethnizität als ein Thema auf, das mit mehreren Klassen modelliert wird (Pronoza et al., 2021). Während andere Daten teilweise den Hass gegen Migrant*innen adressieren, stellen die Autor*innen spezifisch für die Situation in Russland als multiethnische Gesellschaft Daten zusammen, bei denen die betroffene Ethnie mit annotiert ist und erkannt werden soll. Diese Datensammlung entstand hauptsächlich aus dem sozialen Netzwerk Vkontakte und enthält Äußerungen zu über 190 Ethnien (Pronoza et al., 2021).

Im Rahmen der Shared Task *Profiling Hate Speech Spreaders on Twitter* im Rahmen der PAN Initiative (pan.webis.de) wurden nicht einzelne Botschaften als Hassrede, sondern Nutzer*innen als typische Verteiler und Sender von Hassbotschaften identifiziert (Bevendorff et al., 2021). Dabei besteht eine Schwierigkeit für die Klassifikation darin, dass diese Nutzenden auch harmlose Nachrichten verschicken. Ihr Verhalten muss ganzheitlich betrachtet werden.

Außerdem stellt Hate Speech in sozialen Medien ein multimodales Phänomen dar, das sehr häufig verbale und verschiedene nonverbale Elemente umfasst. Zum Beispiel können sich visuelle und verbale Elemente gegenseitig verstärken. Ebenso kann sich der Charakter von Hassrede nur durch das Zusammenspiel mehrerer Modalitäten ergeben, während jede Modalität alleine unverfänglich erscheint.

Besonders die erheblichen Fortschritte in der automatischen Bildanalyse der letzten Jahre können hier eingebracht werden. Systeme des sogenannten Deep Learning erlauben es, Bildinhalte teilweise zu erkennen und mit Texten zu verknüpfen. Auch die automatische Erkennung von Hate Speech hat die Problematik der Multimodalität aufgegriffen und dazu erste Datensets erstellt (z. B. Kiela et al., 2021). Für die gemeinsame Verarbeitung von Bild und Text liegen zwar noch deutlich weniger Arbeiten vor, jedoch erzielen Systeme jetzt schon Fortschritte durch die gemeinsame Verarbeitung. Meist werden noch Verfahren gewählt, die beide Modalitäten parallel verarbeiten und vor der Klassifikation oder dem letzten Schritt die Repräsentationen der beiden spezifischen Verarbeitungssysteme zusammenführen *(late fusion)*. So kann die Mächtigkeit der bestehenden Systeme für Text und Bild genutzt werden; die Beziehungen untereinander werden aber besser durch *Early-fusion*-Systeme ausgenutzt, welche beide Modalitäten parallel und unter Bezugnahme aufeinander analysieren.

Kollektionen für multimodale Hassrede umfassen z. B. Memes aus Facebook, bei denen das Bild und der darin eingebettete Text bereitgestellt wird. Beim Vergleich mehrerer Systeme konnte für das Datenset der Hateful Memes Challenge eine Accuracy bis zu 0,7 erreicht werden (Kiela et al., 2021). Mit der Kollektion MultiOFF konnten nur F1-Werte von ca. 0,5 erzielt werden (Suryawanshi et al., 2020). Auch für das Tamilische liegt eine multimodale Kollektion vor (Suryawanshi & Chakravarthi, 2021).

3 Kontext und Konversationsanalyse

Eine der großen Herausforderungen bei der Annotation von Hassrede für den Aufbau von Trainingsdaten ist der fehlende Kontext. In einem sozialen Netzwerk steht jede Äußerung in einem kommunikativen Zusammenhang und wird von den Leser*innen unter Einbezug eventuell vorhergehender Äußerungen interpretiert. Aufgrund der Kürze von Texten auf Online-Plattformen werden gerade dort Bezüge nicht explizit genannt, sondern der Sender vertraut auf den gegebenen Kontext. Betrachtet man nun jede Äußerung für sich, können völlig unterschiedliche Interpretationen entstehen.

Naturgemäß sind Betroffene von Hassäußerungen im Moment des Empfangs besonders belastet. In einer Interview-Studie befragten Forscher*innen sowohl die Sender als auch die Empfänger von problematischen Inhalten. Es zeigte sich, dass die beiden Gruppen die jeweilige Äußerung völlig unterschiedlich einstuften. Während die Empfänger diese teilweise als psychologisch sehr belastend wahrnahmen, hielten Sender ihre Botschaften für eher unproblematisch und empfindliche Reaktionen darauf für überzogen und nicht nachvollziehbar (Jhaver et al., 2018).

Aber auch unabhängig von der direkten Betroffenheit werden Inhalte sehr unterschiedlich wahrgenommen, wobei ebenfalls der Kontext eine große Rolle spielt. Eine positive und zustimmende Aussage mag für sich allein stehend unverfänglich sein, wenn man diese allerdings z. B. als Reaktion auf eine rassistische Beschimpfung liest, könnte sie ebenfalls als Hassrede gelten. Diese Problematik adressieren die oben erläuterten Shared Tasks im Forschungsfeld bisher nicht, denn Botschaften sollen darin in aller Regel nur aufgrund ihres Inhalts klassifiziert werden. Damit weichen sie allerdings auch von der Realität auf Plattformen ab. KI-Systeme im realen Einsatz können sehr wohl auf die vorherigen Botschaften zugreifen und den Kontext einbeziehen. Die Erstellung derartiger Datensets erfordert allerdings deutlich mehr Aufwand. Die Analyse der genauen Struktur von Kommunikation anhand der Abfolge von Nachrichten ist beispielsweise für Twitter nicht trivial.

In einem Ansatz für die Erkennung problematischer Inhalte haben Pavlopoulos et al. (2020) Kontext-Information eingesetzt. Ihre Definition von problematischen Inhalten greift auf den Begriff der Toxizität zurück und ist somit mit einer breiten Definition von Hassrede vergleichbar. Anwendungsdomäne für dieses Experiment waren die sogenannten Wikipedia Talk Pages, auf denen Autor*innen und Editor*innen über mögliche Verbesserungen der Online-Enzyklopädie diskutieren. Dazu wurden zunächst 20.000 Nachrichten extrahiert. Diese Datenmenge wurde über Crowd-Work annotiert.

Kontext wurde allerdings lediglich dadurch erzeugt, dass die Forscher*innen die ursprüngliche Nachricht und den Titel des Threads der Diskussion miterfassten (Pavlopoulos et al., 2020). Dabei ist die Ursprungs-Nachricht möglicherweise nicht die relevanteste für die Entscheidung über die eigentliche Bedeutung eines Beitrags. Der Anteil von Hassbotschaften innerhalb der Datensammlung liegt mit ca. 6 % im Vergleich mit anderen Benchmarks relativ niedrig. Dies zeigt, dass die Autoren eher zufallsgesteuert vorgegangen sind. Damit schaffen sie einen realistischen Eindruck von der Häufigkeit von Hassrede in realistischen Szenarien und davon, wie oft Nutzer*innen tatsächlich derartigen

Botschaften ausgesetzt sind. Allerdings stellt dieser niedrige Anteil eine erhebliche Schwierigkeit für das maschinelle Lernen dar.

Die Autoren verweisen darauf, dass die Hälfte der Daten ohne Kontext annotiert wurde (Pavlopoulos et al., 2020). Die Genauigkeit bei der Klassifikation bei beiden Mengen war vergleichbar. Daraus ziehen die Autoren den Schluss, dass Kontext für die Erkennung von Hassrede nicht hilfreich sei. Diese Folgerung erscheint jedoch als zu weitreichend, denn bei zwei Datensets ist auch bei sonst gleichen Bedingungen grundsätzlich damit zu rechnen, dass Systeme unterschiedliche Erkennungsraten erzielen (Fortuna et al., 2021).

Ein weiterentwickeltes Modell, das bei der Hate-Speech-Erkennung Kontext berücksichtigt, setzen Menini et al., (2021) um. Ihre Definition greift auf den Begriff des Missbrauchs zu und fällt damit wieder in eine weite Definition problematischer Inhalte (Menini et al., 2021). Um den Aufwand für die Erstellung der Benchmark-Daten zu erleichtern, wurde eine bereits bestehende Datenmenge genutzt. Diese Tweets wurden auf Basis des Texts erneut automatisch auf den sozialen Plattformen gesucht und im Erfolgsfall wurden die vorherigen Tweets extrahiert. Dadurch bestehen sehr unterschiedliche Größen des identifizierten Kontexts. Die Autoren berichten, dass etwa 45 % der hasserfüllten Tweets einen vorangehenden Tweet als Kontext besaßen und für weitere 45 % zwischen zwei und fünf vorangegangene Tweets gefunden wurden. Nur bei etwa 10 % standen mehr als fünf Tweets als Kontext zur Verfügung.

Bei der Annotation mit und ohne Kontext zeigte sich, dass fast 50 % der als Hassbotschaften ausgewiesenen Nachrichten mit Kontext nicht mehr als solche betrachtet wurden. Der umgekehrte Effekt war geringer (Menini et al., 2021).

Diese Problematik des Kontexts greift auch die HASOC *Contextual Subtask* 2021 auf. Dabei sollte auf eine weitgehend einheitliche Größe des Kontexts Wert gelegt werden und auch alle Kontext-Tweets sollten einheitlich annotiert werden. Zudem war ein weiteres Ziel, den Anteil der Hassrede relativ hoch zu halten, um ihn für maschinelles Lernen angemessen zu gestalten (Satapara et al., 2021).

Die erstellte Kollektion ICHCL *(Identification of Conversational Hate-Speech in Code-Mixed Languages)* besteht aus ca. 100 Tweets, ca. 2500 Antworten auf diese und ca. 1200 Reaktionen auf einige der Antworten (Replies). Jeder der Ursprungs-Tweets wurde mit ca. zehn Antworten und Replies erfasst. Alle Tweets, Antwort-Tweets und Reply-Tweets wurden annotiert und enthielten insgesamt ca. 50 % als Hassrede oder aggressive Posts. Im Datenset wird die Struktur und Abfolge der Tweets deutlich gekennzeichnet. Somit konnten Systeme bei der Auswertung darauf zugreifen. Eine Vorab-Analyse mit Baseline-Systemen zeigte, dass die Kontext-Information für die Verbesserung der Genauigkeit der Hate-Speech-Erkennung hilfreich ist (Satapara et al., 2021).

4 Vorgehen beim Aufbau von Trainingsdaten

Algorithmen für die Erkennung von Hassrede erzielen ihre Ergebnisse scheinbar völlig objektiv. Jedoch stehen hinter ihrer Gestaltung zahlreiche Entscheidungen. Beispielsweise ist der Aufbau von Trainingsmengen eine soziale Konstruktion, die in einem bestimmten Kontext unter Rahmenbedingungen und Zwängen erfolgt. Die Entwickler*innen von Daten treffen bei der Gestaltung der Trainingsdaten bewusst oder unbewusst Entscheidungen, die sich auf die Daten auswirken und somit auch die Wirksamkeit der KI-Verfahren beeinflussen (siehe auch Demus et al. in diesem Band).

Die Repräsentation von Texten für die Verarbeitung in Klassifikationssystemen erfolgte traditionell auf der Basis des Auftretens von Wörtern. Zunehmend gelangen auch verteilte semantische Repräsentationen zum Einsatz, welche nur kurze Vektoren mit ca. 100 Dimensionen benötigen (Mandl, 2020). Damit schreiten Systeme von symbolischen Repräsentationen zur subsymbolischen Ebene fort, bei der interne Strukturen nicht mehr eindeutig interpretiert werden können. Dokumente werden durch eine Reihe von Zahlen repräsentiert, so dass ähnliche Wörter ähnliche Vektoren besitzen. Somit können diese innovativen Verfahren des *Deep Learning* nicht nur auf lexikalischer Ebene entscheiden, sondern sie können die Bedeutung von Wörtern durch die Ähnlichkeiten zwischen diesen besser abbilden (siehe Schäfer in diesem Band). Gleichwohl können auch solche Verfahren mit der ironischen oder metaphorischen Verwendung von Begriffen (siehe Jaki in diesem Band) noch Schwierigkeiten haben.

Bei einer Trainingsmenge sollten möglichst viele verschiedene Beispiele als Repräsentanten von heterogenen Formen von Hate Speech präsent sein. Wie dieser Merkmalsraum vollständig abgedeckt werden kann, bleibt jedoch völlig offen. Innovative Formen der Hate Speech, die durch kreative sprachliche Muster entstehen oder auch Hate Speech zu neu aufkommenden Themen lässt sich so also nur schlecht erkennen.

Die Trainingsmenge sollte repräsentativ für die bekannten Formen der problematischen Inhalte sein, wobei aber aufgrund der Vielfältigkeit sprachlicher Ausdrucksformen unklar ist, wie diese Repräsentativität erreicht oder erkannt werden kann. Somit kann lediglich der Prozess der Erstellung von Hate-Speech-Datenmengen betrachtet und bewertet werden. Er besteht üblicherweise aus den folgenden Schritten (Vidgen & Derczynski, 2020):

1. Erstellung einer Strategie zur Vorauswahl von Inhalten aus sozialen Netzwerken;

2. Umsetzung der Strategie mit Werkzeugen und Extraktion von Posts aus großen Mengen Text aus sozialen Netzwerken;
3. Annotation einer Vorauswahl durch Menschen.

Die Strategie besteht zum einen oft im Auswählen von Begriffen, die für Hate Speech typisch sein könnten (GermEval) oder auch im Erstellen eines Vorab-Klassifizierers (Mandl et al., 2020). In beiden Fällen können bestimmte Inhalte präferiert werden. So spielen bei der Auswahl von Begriffen notwendigerweise Vorkenntnisse bzw. Annahmen über Hate Speech eine erhebliche Rolle. Ganze Komplexe problematischer Inhalte könnten übersehen werden, wenn lediglich bereits bekannte Themen in die Daten mit einbezogen werden.

Die Auswahl von bestimmten Hate-Speech-Beispielen durch manuelles Suchen kann dazu führen, dass diese Beispiele immer von einigen Autor*innen stammen, während die neutralen Äußerungen von anderen Autor*innen stammen. Das kann sogar darin münden, dass ein Klassifikationssystem letztlich eine Autorenerkennung durchführt. Eine solche Erkennung des individuellen Stils erkennt dann evtl. ganz andere Merkmale und ist nicht in der Lage, in einer realen Umgebung eine gute Erkennungsqualität für Hassbotschaften zu liefern (Arango et al., 2020). Deswegen sollten von jedem Profil in sozialen Netzwerken immer mehrere Posts gesammelt werden, um pro Autor*in Beispiele für problematische und unproblematische Inhalte einzubauen. Dies wurde z. B. im Rahmen von GermEval berücksichtigt (Struß et al., 2019).

Verzerrungen können auch beim technologischen Sammeln der Inhalte entstehen. Tools zur Suche oder die APIs für den Zugriff mit Programmierwerkzeugen können schon in den Plattformen Präferenzen für bestimmte Inhalte abbilden, die unerkannt bleiben. Retrieval-Systeme sind beispielsweise anfällig für die Länge von Texten als versteckter Einflussfaktor (Roelleke, 2013). In sozialen Netzwerken könnten Beiträge populärer Nutzer*innen bevorzugt werden oder andere Ranking-Kriterien implementiert sein. Selbst wenn die Strategien genau aufgezeichnet würden, könnten die erzielten Mengen nicht nachvollzogen werden, da z. B. die Twitter-Suche zu jedem Zeitpunkt andere, aktuellere und vielleicht auch personalisierte Ergebnisse liefert. Zudem können intern schon Methoden eingebaut sein, die problematische Inhalte detektieren und ihnen niedrige Ranking-Positionen zuweisen oder dafür sorgen, solche Inhalte weniger häufig anzuzeigen.

Eine wichtige Frage bei der Erstellung von Benchmarks ist auch der Umfang bzw. Anteil der jeweiligen Klassen. Es ist davon auszugehen, dass Hate Speech in realer Kommunikation weniger als 1 % der Inhalte ausmacht. Bei einer Zufalls-

auswahl und Annotation dieser Daten weisen Analysen darauf hin, dass Werte in dieser Größenordnung zu erwarten sind (Vidgen & Derczynski, 2020).

Alle Trainingsmengen weisen jedoch deutlich höhere Anteile auf, da es schwierig ist, mit nur einem geringen Anteil von Beispielen eine Klassifikation zu trainieren. Zudem liefern die Algorithmen bessere Ergebnisse, wenn die Klassen vergleichbar oft vorkommen. Aus dieser Perspektive spiegeln die Trainingsmengen in keiner Weise die Realität wider.

Auch in der letzten Phase, der Annotation der identifizierten Menge von Inhalten durch Menschen, können Schwierigkeiten auftreten. Die unvermeidbare Subjektivität wird im folgenden Abschnitt behandelt. Jedoch alleine die Richtlinien für die Annotation, welche den Menschen mitgegeben werden, variieren sehr stark. Es stellt sich die Frage, inwiefern hier überhaupt die gleiche Aufgabe bearbeitet wird. Bestehende Annotations-Guidelines nennen allein schon so verschiedene Überbegriffe wie *Hate, Open or Covert Aggression, Toxicity, Racism, Obscene* und *Inappropriate.* In dieser Vielfalt und Unschärfe spiegelt sich die Schwierigkeit, Hate Speech klar zu definieren. Diese verschiedenen Konzepte vergleichen Fortuna und Kollegen ausführlich (Fortuna et al., 2020). In Zukunft sollte intensiver erforscht werden, wie diese heterogenen Definitionen sinnvoll zur Erkennung von Hate Speech beitragen können.

5 Messung von Verzerrungen

Die Zuverlässigkeit von Datensets kann mit einigen Methoden überprüft werden. Die Sprachmodelle der Trainingsmenge sowie gegebenenfalls auch der Testmenge können untereinander und mit dem allgemeinen Sprachmodell im Korpus verglichen werden. Dazu kann z. B. das Maß *Mutual Information* eingesetzt werden. Es wurde schon beobachtet, dass bestimmte Begriffe in den Hate-Korpora häufiger vorkommen als allgemein. Dies führt teils zu guten Klassifikationsergebnissen bei der Entwicklung (Wiegand et al., 2019), die aber unter realen Einsatzbedingungen nicht erreicht werden.

Mit dem Ansatz des *Topic Modeling* kann überprüft werden, ob die gleichen Themen auch ähnlich häufig auftreten. Für HASOC 2019 konnte für das Englische ein Topic Model aus vier Themen identifiziert werden. Es zeigte sich, dass diese Themen sowohl in der Trainings- als auch der Testmenge über die problematischen Inhalte und die unproblematischen Inhalte relativ gleichmäßig repräsentiert waren. Diese Verteilung zeigt Abb. 2. Für Systeme reichte es also nicht aus, lediglich ein Thema zu erkennen, zu dem immer hasserfüllt kommentiert wurde.

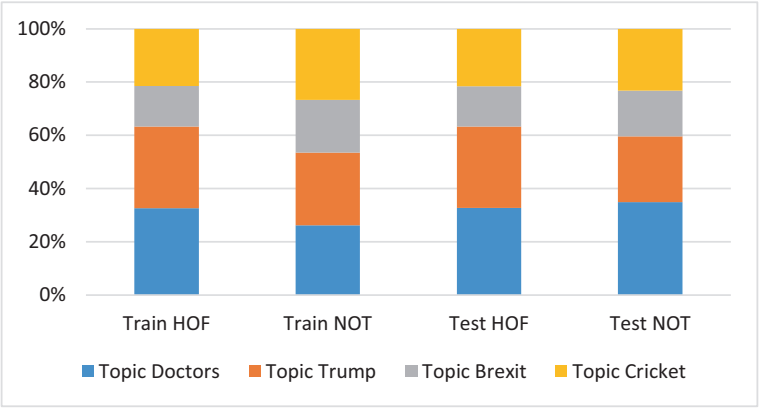

Abb. 2 Die Verteilung von vier Topics in Test- und Trainingsmenge aus HASOC 2019

Auch ein zu hoher Anteil an politisch extrem rechten oder extrem linken Posts in der Trainingsmenge kann in einer geringeren Genauigkeit bei der Erkennung resultieren (Wich et al., 2020).

Nach oder während der Annotation kann die interne Validität überprüft werden. Dazu können einige Texte mehrfach von verschiedenen Personen annotiert werden, was natürlich den Aufwand und die Kosten erhöht. Das sogenannte *Interrater Agreement* zeigt an, inwieweit die Bewertungen übereinstimmen. Niedrige Werte weisen darauf hin, dass die Annotation sehr subjektiv geprägt ist.

Annotationseffekte entstehen aber auch durch den Rahmen der Präsentation (Voorhees, 2000). Wer als Annotator*in schon zahlreiche hasserfüllte Botschaften gesehen hat, wird bei Grenzfällen etwas weniger streng. Einen Bezug zu demographischen Faktoren bei den Annotierenden versuchen Al Kuwatly et al. (2020) herzustellen. Weitere Studien haben gezeigt, dass Vertrautheit mit Sprachregistern starken Einfluss auf Annotationsentscheidungen hat (Sap et al., 2019).

Die Übereinstimmung von vier Annotierenden bei GermEval wurde für ein Sample von 300 Tweets gemessen und erreichte einen Kappa-Wert von 0,59, was als *moderate agreement* gilt (Struß et al., 2019). Hier zeigt sich, wie schwierig selbst das Finden gemeinsamer Maßstäbe ist.

Für HASOC 2019 wurde anhand der zweimal annotierten Tweets die Übereinstimmung der Annotator*innen gemessen. Tab. 2 zeigt, dass sie für die zweite Task mit der genaueren Einteilung von problematischen Inhalten sinkt. Eine Ana-

Tab. 2 Interrater-Statistik für HASOC 2019

	Anzahl der zweimal annotierten Tweets	Interrater Agreement
English sub-task 1	5389	74 %
English sub-task 2	5389	64 %
Hindi sub task 1	4122	80 %
Hindi sub task 2	4122	62 %
German sub task 1	1159	88 %
German sub task 2	1159	86 %

lyse von Ross et al. (2016) konnte sogar zeigen, dass selbst schriftliche Richtlinien keinen hohen Einfluss auf die Übereinstimmung zwischen Annotierenden haben.

Grenzfälle, die sich schwer einordnen lassen, weisen eine deutlich höhere Abweichung auf (Salminen et al., 2019). Dies bedeutet, dass ein gutes Interrater Agreement auch bedeuten kann, dass lediglich sehr klare Fälle in der annotierten Menge vorkommen. Die Häufigkeit von solchen Grenzfällen im Graubereich in den gesammelten Daten ist vorab ja nicht bekannt und lässt sich auch nicht steuern. Wenige oder unsicher bewertete Grenzfälle können es den Algorithmen danach allerdings zusätzlich erschweren, die Grenze deutlich zu ziehen. Somit muss selbst das Interrater Agreement als Qualitätsmerkmal auch hinterfragt werden. Das Vorgehen bei Fällen im Graubereich oder bei abweichenden Meinungen von Annotator*innen wird nicht einheitlich gehandhabt.

6 Transfer über Kollektionen hinweg

Die Genauigkeit der Vorhersage von Hate Speech variiert meist stark je nach Datenmenge. Dies weist auf die Bedeutung der Trainingsdaten hin. Für einen realen Einsatz ist natürlich viel wichtiger, wie gut ein System bei völlig anderen Daten unter echten Bedingungen funktioniert. Dann stellt sich die Frage, ob die Systeme aus der Forschung robust genug für einen Dauerbetrieb wären. Dies lässt sich transparent und mit Daten, die außerhalb von kommerziellen Plattformen zur Verfügung stehen, nur schwer überprüfen. Als gängigste Methode wird hierfür mit den Trainingsdaten eines Benchmarks ein Modell trainiert und dann mit den Testdaten anderer Benchmarks getestet. So kann die Messung des möglichen Bias durch Experimente über mehrere Datensets hinweg erfolgen.

Wenn ein Klassifikationssystem mit einer Menge trainiert wird und dann auf eine andere angewendet wird, zeigt sich zu einem gewissen Maß, ob diese das gleiche Konzept abbilden bzw. Verzerrungen in den Daten vorliegen. Die bisherigen Ergebnisse bei solchen Cross-Validitäts-Studien zeigen, dass teils deutlich niedrigere Trefferquoten erzielt werden (Wiegand et al., 2019).

Die umfangreichen Experimente von Fortuna et al. (2021) zeigen, dass die Performanz für einen Datensatz um über 30 % Genauigkeit schwanken kann, je nachdem mit welchem anderen Datensatz trainiert wurde. Gründe hierfür und mögliche Lösungsansätze werden in einem Überblicksaufsatz diskutiert (Yin & Zubiaga, 2021).

7 Erklärbarkeit und Nachvollziehbarkeit

Ein häufig genanntes Problem von Künstlicher Intelligenz und insbesondere von mächtigen Deep Learning-Verfahren besteht in der mangelnden Nachvollziehbarkeit der Entscheidungen. Diese Verfahren können ihre Ergebnisse nicht erklären und sehen sich daher dem Vorwurf der fehlenden Transparenz ausgesetzt. Ein eigener Forschungszweig unter der Bezeichnung *Explainable Artificial Intelligence* (XAI) befasst sich mit Möglichkeiten, solche Erklärungen zu generieren und Systeme oder ihre Entscheidungen besser verständlich zu machen. Ein Überblick über das Thema zeigt, dass es gerade für die Erklärung von Entscheidungen für die Textklassifikation im Vergleich zur Bilderkennung nur sehr wenige Ansätze gibt (Guidotti et al., 2018).

Vor allem zählen dazu die Verfahren *Shapley Values* und *Local Interpretable Model-Agnostic Explanations (LIME).* Sie setzen beide nach der Erstellung eines Modells an und versuchen, nachträglich den Beitrag von Wörtern zu einer Entscheidung zu messen und darzustellen. Für Hate Speech wurde LIME beispielsweise von Mahajan et al. (2021) umgesetzt.

Auch hier stellt sich die Frage, ob es wirklich darum gehen kann, die Verfahren des maschinellen Lernens an sich transparent darzustellen, oder ob nicht eine Offenlegung der Trainingsdaten etwa anhand von Beispielen einen besseren Eindruck von einem KI-System zur Hate-Speech-Erkennung liefert. Wenn beispielsweise verdeutlicht werden kann, dass zu einem Tweet keine ähnlichen Trainingsbeispiele vorliegen, könnten Nutzer*innen besser verstehen, dass das System gar nicht in der Lage ist, eine gute Entscheidung zu treffen. Somit könnten Ansätze zur Sicherung der Transparenz auch ein Mittel sein, Rückschlüsse auf die Qualität der Trainingsdaten zu erlauben.

Die meisten Systeme mit Ansätzen zur Erklärbarkeit gehen allerdings noch anders vor und bedienen meist sehr heterogene Nutzungsszenarien. Das System von Modha et al. (2020) erlaubt es, mit einem Browser-Plugin einen Account in sozialen Medien zu überwachen. Andere Systeme zielen auf die Überwachung ganzer Plattformen ab. So schlägt etwa Bunde (2021) ein Dashboard vor, das einzelne Beiträge präsentiert, die Entscheidungen dazu erklärt und auch die Aktivitäten von einzelnen Nutzer*innen überwacht. Das System *Hatemeter* möchte Moderator*innen Hilfestellung geben, indem es aktuelle Themen darstellt, die besonders viel Hassrede auf sich ziehen (Laurent, 2020).

Einen anderen Ansatz verfolgt das System von Sontheimer et al. (2022). Es unterstützt die Informationelle Autonomie von Bürger*innen und lässt sie mit den aktuellsten Algorithmen online experimentieren. Die Nutzer*innen können in dem System kurze Nachrichten eingeben und bevor sie in einer Plattform hochgeladen werden, prüfen lassen, ob ein System zur Hate-Speech-Klassifikation diese als problematisch einstuft. Dadurch soll nicht primär das Verständnis von Algorithmen gefördert werden, sondern durch ähnliche Beispiele und die Möglichkeit des Ausprobierens wird ein Erkennen der Wirksamkeit des Systems verdeutlicht.

8 Fazit und Ausblick

Ziel der Hate-Speech-Erkennung ist das Training robuster Verfahren, die auch im realen Einsatz erfolgreich sind. Die Gestaltung von Trainingsmengen entscheidet über die Leistungsfähigkeit von KI-Algorithmen. Diese sind zwar intransparent, aber die Gestaltung der Trainingsdaten spielt womöglich eine noch bedeutendere Rolle bei der Feinjustierung der Algorithmen. Somit ist ein transparenter Einblick in diese Daten und den Erstellungsprozess sehr wichtig und kann zur Durchschaubarkeit mehr beitragen als die Nachvollziehbarkeit von Algorithmen.

Zur Messung der Qualität von Trainingsdaten gibt es derzeit keine überzeugenden Methoden und gleichzeitig entstehen leicht Verzerrungen durch das Vorgehen beim Sammeln, Auswählen oder Annotieren. Deutlich mehr vergleichende Forschung zu Methoden der Erstellung wäre notwendig. Diese Forschung sollte nicht nur hinter den verschlossenen Türen der Internet-Plattformen stattfinden, sondern in offenen Foren. Nur so kann die Leistungsfähigkeit der Algorithmen angemessen diskutiert und eine breite gesellschaftliche Akzeptanz erzielt werden.

Literatur

Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms* (S. 184–190). https://doi.org/10.18653/v1/2020.alw-1.21.

Arango, A., Pérez, J., & Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 101584.

Bevendorff, J., Chulvi, B., Peña Sarracén, G. L., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska , M., & Zangerle, E. (2021). Overview of PAN 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection. *International conference of the cross-language evaluation forum for European languages*, 419–431. Springer, Cham. https://doi.org/10.1007/978-3-030-85251-1_26.

Bunde, E. (2021). AI-assisted and explainable hate speech detection for social media moderators–A design science approach. In *Proceedings of the 54th Hawaii international conference on System Sciences* (S. 1264).

De Smedt, T., & Jaki, S. (2018). Challenges of automatically detecting offensive language online: Participation paper for the germeval shared task 2018 (HaUA). *14th conference on natural language processing KONVENS*. https://doi.org/10.1553/0x003a105d.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *EVALITA Evaluation of NLP and Speech Tools for Italian, 12*, 59.

Fortuna, P., Soler, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings 12th Language Resources and Evaluation Conference* (LREC) (S. 6786–6794).

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management, 58*(3), 102524. https://doi.org/10.1016/j.ipm.2021.102524

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access, 9*, 48364–48404. https://doi.org/10.1109/ACCESS.2021.3068313

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (S. 1336–1350).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR), 51*(5), 1–42. https://doi.org/10.1145/3236009

Heldt, A. (2020). Pflicht zu weltweiter Löschung: Konsequente oder ausufernde Auslegung?–Anmerkung zum Urteil des EuGH v. 3.10. 2019, Rs. C-18/18 (Glawischnig-Piesczek). *EuR Europarecht, 55*(2), 238–245. https://doi.org/10.5771/0531-2485-2020-2-238.

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbe, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction* (TOCHI) 12. https://doi.org/10.1145/3185593.

Joly, A., Goëau, H., Kahl, S., Deneu, B., Servajean, M., Cole, E., Picek, L., Ruiz de Castañeda, R., Bolon, I., Durso, A., & Lorieul, T., (2020). Overview of LifeCLEF 2020: A system-oriented evaluation of automated species identification and species distribution Prediction. *International Conference of the Cross-Language Evaluation Forum for European Languages*, 342–363. https://doi.org/10.1007/978-3-030-58219-7_23.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C. A., Bull, P., Lipstein, G., Nelli, T., Zhu, R., Muennighoff, N., Velioglu, R., Rose, J., Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., Yannakoudakis, H., Sandulescu, V., Ozertem, U., Pantel, P., Specia, L., & Parikh, D. (2021). The hateful memes challenge: Competition report. NeurIPS 2020 Competition and demonstration track. In *Proceedings of Machine Learning Research* (S. 344–360).

Kuhlen, R. (1999). *Die Konsequenzen von Informationsassistenten: Was bedeutet informationelle Autonomie oder wie kann Vertrauen in elektronische Dienste in offenen Informationsmärkten gesichert werden?* Suhrkamp.

Laurent, M. (2020). Project Hatemeter: Helping NGOs and Social Science researchers to analyze and prevent anti-Muslim hate speech on social media. *Procedia Computer Science, 176*, 2143–2153. https://doi.org/10.1016/j.procs.2020.09.251

Lewanczik, N. (2019). *Datenschutz durch Dritte? Zuckerbergs Idee vom global regulierten Internet*. https://onlinemarketing.de/news/datenschutz-dritte-zuckerbergs-global-reguliertes-internet.

Madukwe, K., Gao, X., & Xue, B. (2020). In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the fourth workshop on online abuse and harms* (S. 150–161). https://www.aclweb.org/anthology/2020.alw-1.18.

Mahajan, A., Shah, D., & Jafar, G. (2021). Explainable AI approach towards toxic comment classification. *Emerging Technologies in Data Mining and Information Security*, 849–858.

Mandl, T. (2008). Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. *Informatica, 32*(1). https://www.informatica.si/index.php/informatica/article/viewFile/174/170.

Mandl, T. (2020). Die Erkennung unangemessener Inhalte im Internet: KI Verfahren, Evaluierung und Herausforderungen. *Bibliotheksdienst, 54*(3/4), 214–226. https://doi.org/10.1515/bd-2017-0083.

Mandl, T., Agosti, M., Di Nunzio, G. M., Yeh, A., Mani, I., Doran, C., & Schulz, J. M. (2009). LogCLEF 2009: The CLEF 2009 multilingual logfile analysis track overview. *Working Notes for CLEF 2009 Workshop*. Corfu, Greece, September 30–October 2. http://ceur-ws.org/Vol-1175/CLEF2009wn-LogCLEF-MandlEt2009.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 *Proceedings of the 12th annual meeting of the Forum for Information Retrieval Evaluation (FIRE)*, ACM. https://doi.org/10.1145/3441501.3441517.

Menini, S., Aprosio, A. P., & Tonelli, S. (2021). Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection. *arXiv preprint* arXiv:2103.14916.

Modha, S., Mandl, T., Majumder, P., & Patel, D. (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European Languages. In *Proceedings of the 11th annual meeting of the forum for information retrieval evaluation* (S. 167–190). http://ceur-ws.org/Vol-2517/.

Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Systems With Applications, 161*, 113725. https://doi.org/10.1016/j.eswa.2020.113725

Modha, S., Mandl, T., Shahi, G.K., Madhu, H., Satapara, S., Ranasinghe, T., & Zampieri, M. (2021). Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan Languages and conversational hate speech. *FIRE 2021: Forum for Information Retrieval Evaluation*, Virtual Event, 13th–17th December, ACM.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., … Kartal, Y. S. (2021). Overview of the CLEF–2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. *International Conference of the Cross-Language Evaluation Forum for European Languages* (S. 264–291). Springer, Cham. https://doi.org/10.1007/978-3-030-85251-1_19.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N, & Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (S. 4296–4305). https://www.aclweb.org/anthology/2020.acl-main.396/.

Pronoza, E., Panicheva, P., Koltsova, O., & Rosso, P. (2021). Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management, 58*(6), 102674. https://doi.org/10.1016/j.ipm.2021.102674

Ródriguez-Sánchez, F., de Albornoz, J. C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021). Overview of EXIST 2021: Sexism identification in social networks. *Procesamiento del Lenguaje Natural, 67*, 195–207.

Roelleke, T. (2013). *Information retrieval models: Foundations and relationships*. Synthesis Lectures on Information Concepts, Retrieval, and Services 5(3). https://doi.org/10.2200/S00494ED1V01Y201304ICR027.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*.

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2019). Online hate ratings vary by extremes: A statistical analysis. In *Proceedings Conference on Human Information Interaction and Retrieval,* (CHIIR) ACM (S. 213–217). https://doi.org/10.1145/3295750.3298954.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (S. 1668–1678). https://www.aclweb.org/anthology/S.19-1163.pdf.

Satapara, S., Modha, S., Mandl, T., Madhu, H., & Majumder, P. (2021). Overview of the HASOC subtrack at FIRE 2021: Conversational hate speech detection in code-mixed language. *Working Notes of FIRE 2021 – Forum for Information Retrieval Evaluation*. CEUR, 2021.

Sontheimer, L., Schäfer, J., & Mandl, T. (2022). Enabling Informational Autonomy through Explanation of Content Moderation: UI Design for Hate Speech Detection. In *UCAI 2022: Workshop on User-Centered Artificial Intelligence. Mensch und Computer 2022* – Workshopband 04.-07. September 2022, Darmstadt.

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) Nürnberg/Erlangen. https://doi.org/10.5167/uzh-178687.

Suryawanshi, S., & Chakravarthi, B. R. (2021). Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (S. 126–132). https://www.aclweb.org/anthology/2021.dravidianlangtech-1.16/.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, TRAC (S. 32–41).

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one, 15*(12). https://doi.org/10.1371/journal.pone.0243300.

Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management, 36*(5), 697–716. https://doi.org/10.1016/S0306-4573(00)00010-8

Wich, M., Bauer, J., & Groh, G. (2020). Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (S. 54–64). https://doi.org/10.18653/v1/2020.alw-1.7.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing* (KONVENS) Wien, Sept. 21. https://www.zora.uzh.ch/id/eprint/178687/1/GermEvalSharedTask2019Iggsa.pdf.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: The problem of biased datasets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (S. 602–608). https://doi.org/10.18653/v1/N19-1060.

Womser-Hacker, C. (2013). Evaluierung im Information Retrieval. In R. Kuhlen, W. Semar, & D. Strauch (Hrsg.), *Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und –praxis* (6. Aufl., S. 396–410). De Gruyter. https://doi.org/10.1515/9783110258264.396.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science, 7*, e598. https://doi.org/10.7717/peerj-cs.598