Entwicklung einer automatisierten Prüfung der Strafbarkeit von Hate Speech

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science des Fachbereiches Computerlinguistik
der Humanwissenschaftlichen Fakultät der Universität Potsdam

Celia Birle
Matrikel-Nr.: 792481

Erstgutachter: Prof. Dr. Manfred Stede, Universität Potsdam
Zweitgutachter: Silvio Peikert, Fraunhofer Institut FOKUS

Abgabe: 9. August 2022

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Motivation . . . 1
1.2 Fragestellung und Methodik . . . 2
1.3 Phänomen Hate Speech . . . 3
1.3.1 Gesellschaftliche Problematik . . . 3
1.3.2 Juristische Einordnung . . . 4

2 Übersicht des aktuellen Forschungsstandes **6**
2.1 Datensätze und weitere Ressourcen . . . 6
2.1.1 Annotationansätze . . . 6
2.1.2 Daten . . . 7
2.2 Klassifikationsmethoden . . . 8
2.3 Herausforderungen . . . 10
2.4 Juristische Ansätze . . . 11

3 Erstellen eines deutschsprachigen Referenzdatensatzes **13**
3.1 Technische Angleichung . . . 13
3.2 Annotation . . . 14
3.2.1 Angleichen der Annotation . . . 14
3.2.2 Zur Annotation von Straftatbeständen . . . 15
3.2.3 Volksverhetzung (§ 130 StGB) – Annotationsansatz . . . 17
3.2.4 Praktische Anmerkungen . . . 20

4 Eine automatisierte Prüfung der Strafbarkeit **21**
4.1 Methodik . . . 21
4.2 Klassifikation mit Logistischer Regression . . . 23
4.3 Klassifikation mit Transfer Learning auf Basis eines Sprachmodells 24
4.4 Zusammenführen der Teilklassifikatoren . . . 26
4.5 Evaluation . . . 27

5 Schlussbetrachtung **30**

Literaturverzeichnis **I**

Abbildungsverzeichnis **VII**

Tabellenverzeichnis **VII**

1 Einleitung

1.1 Motivation

Hate Speech ist in den letzten Jahren im Zuge der immer weiter verbreiteten Nutzung sozialer Medien zugleich als gesellschaftliches Problem und als Rechercheinteresse präsenter geworden. Hate Speech umfasst diverse Arten abwertender, angreifender und diskriminierender Äußerungen in sozialen Netzwerken. Das Spektrum dieser Art von Äußerungen ist breit – sie reichen von einfachen Beleidigungen bis zu menschenfeindlichen Beiträgen und finden sich in Form von Facebook-Kommentaren, Tweets oder Memes. Was genau Hate Speech konstituiert, ist jedoch nicht einheitlich definiert.

Soziale Netzwerke sind laut dem 2017 in Kraft getretenen Netzwerkdurchsetzungsgesetz[1] (NetzDG) dazu verpflichtet, strafbare Inhalte zu löschen, sobald sie gemeldet wurden – laut § 3 Abs. 1, 2 NetzDG bei offensichtlich strafbaren Inhalten innerhalb von 24 Stunden, andernfalls innerhalb von sieben Tagen. Das NetzDG verpflichtet soziale Netzwerke auch zur Einrichtung effektiver Beschwerdeverfahren und zu regelmäßigen Berichten über ihren Umgang mit Hassrede, Falschnachrichten und Weiterem.

Wenn Menschen im Internet zu Gewalt aufrufen oder Falschnachrichten und Schmähungen verbreiten, kann das sehr wohl reale Konsequenzen haben. Im deutschsprachigen Raum ging der Prozess Künast gegen Meta durch die Nachrichten. Die Grünen-Politikerin Renate Künast ist erfolgreich gegen Facebook vorgegangen, um zu klären, ob das hier abgebildete, rechtswidrige Meme mit einem ihr zugeordneten Falschzitat und darüber hinaus sinngleiche Inhalte von Facebook aktiv gesucht und beseitigt werden müssen, nachdem auf dessen Existenz aufmerksam gemacht worden war[2].

Abbildung 1: Künast-Meme, gekennzeichnet als Falschzitat;
abgerufen unter https://hateaid.org/klage-facebook-loeschpflicht/

Soziale Netzwerke moderieren nicht nur strafbare, sondern jegliche Art unerwünschter Inhalte. Angesichts der überwältigenden Anzahl neuer Inhalte, die täglich veröffentlicht werden, erfordert die Prüfung von Hand sehr viel Zeit, Geld und Menschen, deren natürliche Subjektivität vor allem in einem

[1] Netzwerkdurchsetzungsgesetz vom 1. September 2017 (BGBl. I S. 3352), das zuletzt durch Artikel 1 des Gesetzes vom 3. Juni 2021 (BGBl. I S. 1436) geändert worden ist, abgerufen unter https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html.

[2] S. https://hateaid.org/klage-facebook-loeschpflicht/ und https://hateaid.org/wp-content/uploads/2022/04/HateAid-Pressemitteilung-Grundsatzprozess-Kuenast-Facebook.pdf.

so potentiell aufwühlenden Bereich zu einem Mangel an Einheitlichkeit in der Moderation führt. Die Moderation allgemein und die Prüfung auf Strafbarkeit im Besonderen müssen daher langfristig mit Automatisierung unterstützt werden. Obwohl es einige erfolgreiche Ansätze gibt, die analysieren, wie die Untersuchung von Strafbarkeit in sozialen Medien automatisiert werden kann (Zufall et al., 2020; Zufall et al., 2019), konzentriert sich die überwältigende Mehrheit der Hate-Speech-Forschung auf allgemeinere Formen abwertender und anstößiger Inhalte. Diese Arbeit beschäftigt sich im Gegensatz dazu mit dem Teil von Hate Speech, der nicht nur anstößig, sondern sanktioniert ist – insbesondere wird dabei eine Strafbarkeit im Zusammenhang mit der Volksverhetzung gemäß § 130 des Strafgesetzbuchs (StGB) erforscht.

1.2 Fragestellung und Methodik

Ziel dieser Arbeit ist es daher zu untersuchen, bis zu welchem Grad die Prüfung von Hate Speech auf eine potentielle Strafbarkeit automatisiert werden kann. Grundsätzlich sind zur Klassifizierung von Hate Speech zunächst eine Definition derselben, ein entsprechend annotierter Datensatz und zuletzt ein adäquates Klassifikationsverfahren, das auf diesen Daten trainiert wird, notwendig. Der Stand der Technik für Klassifikationsverfahren im Bereich Hate Speech ist das Transfer Learning auf Grundlage von Transformer-Sprachmodellen; aber auch mit anderen Machine-Learning-Methoden sind für deutschsprachige Datensätze vergleichbare Ergebnisse erzielt worden (Mandl et al., 2020; Struß et al., 2019).

Zur Prüfung der strafrechtlichen Relevanz von Hate Speech sind demnach eine Arbeitsdefinition von Hate Speech im Allgemeinen sowie eine genauere Erläuterung der relevanten Tatbestände nötig. Der Fokus einschlägiger Studien liegt meist auf der grundsätzlichen Unterscheidung abwertender Äußerungen von allen anderen Inhalten, unabhängig von der Intensität dieser Abwertung. Manchmal werden zusätzlich Nuancen der Intensität von Hate Speech unterschieden: von Äußerungen, die Schimpfwörter enthalten, über Beleidigungen bis hin zu hasserfüllten, diskriminierenden Aussagen. Im Rahmen dieser Arbeit sollen lediglich unerwünschte von strafbaren Inhalten abgegrenzt werden. Daher werden konkret Straftatbestände annotiert. Aber nicht alle relevanten Straftaten können schon durch die Veröffentlichung eines einzelnen Tweets begangen werden. Exemplarisch wird daher für die Volksverhetzung (§130 StGB) untersucht, welche spezifischen Merkmale dafür sorgen, dass ein bestimmter Inhalt rechtswidrig ist, und wie diese annotiert werden können.

Im zweiten Schritt soll eine geeignete Datengrundlage geschaffen werden. Da die bisherigen Datensätze recht klein sind und nur wenige stark abwertende Posts enthalten, die potentiell strafbar sein könnten, gilt es öffentlich zugängliche Korpora zu einem Referenzdatensatz mit einheitlicher Annotation zusammenzuführen: Aus der Forschung zu Hate Speech in deutschsprachigen sozialen Medien sind einige Datensätze entstanden, zuletzt mit einem Fokus auf Covid-19 (Wich et al., 2021), weiterhin noch GermEval2018 (Wiegand et al., 2018), GermEval2019 (Struß et al., 2019) HASOC 2019 (Mandl et al., 2019) und HASOC 2020 (Mandl et al., 2020). Alle diese Korpora bestehen aus Daten von Twitter und haben zumindest eine binäre Annotation. Die GermEval-Daten und die HASOC-Daten, die jeweils auch detaillierte Annotation aufweisen,

sollen als Grundlage für den Referenzdatensatz genutzt werden.

Mit dem zu den Straftatbeständen ausgearbeiteten Annotationsschema einerseits und dem Referenzdatensatz andererseits kann dann ein Teil des Datensatzes entsprechend annotiert werden. Diese neu annotierten Daten sind geeignet, um im Folgenden als Trainings- und Testdaten für ein Klassifikationsverfahren zu dienen. Danach wird also ein automatisiertes Verfahren entwickelt, das die potentielle Strafbarkeit von Hate Speech einschätzt. Es werden zwei verschiedene Methoden genutzt – zum einen ein traditionelles Machine-Learning-Verfahren, logistische Regression, und zum anderen das Fine-Tuning eines Transformer-Sprachmodells. Es folgt eine vergleichende Evaluation der beiden Modelle zur Einordnung der Strafbarkeit von Hate Speech.

Diese Arbeit liefert zunächst eine Einordnung der juristischen Relevanz von Hate Speech und eine linguistisch-juristische Einordnung des Potentials auf automatisierte Prüfung. Weiterhin wird ein Annotationsschema und Klassifikationsverfahren für den §130 StGB entwickelt. Zusätzlich entsteht ein Referenzdatensatz, der – annotiert gemäß dem Annotationsschema – die Daten für die Klassifikationsverfahren liefert. Zuletzt werden mehrere Klassifikationsverfahren umgesetzt, getestet und vergleichend evaluiert.

1.3 Phänomen Hate Speech

1.3.1 Gesellschaftliche Problematik

Hate Speech ist komplex und entzieht sich einer einheitlichen Definition. Laut der Bundeszentrale für politische Bildung bezeichnet „Hate Speech alle Inhalte sozialer Medien, durch die einzelne Menschen oder Gruppen abgewertet, angegriffen oder gegen sie zu Hass und Gewalt aufgerufen wird“[3]. Hate Speech kennzeichnet sich durch die Veröffentlichung in sozialen Medien und ist damit z.B. vom Cyberbullying abzugrenzen. Oft wird Hate Speech zusammen mit verwandten Phänomenen wie Beleidigungen, Beschimpfungen und anstößigen Inhalten untersucht, ohne dazwischen klare Grenzen zu ziehen.

Hate Speech hat sich in den letzten Jahren immer weiter verbreitet und ist zugleich in den Fokus der Öffentlichkeit gerückt. Einseitige Diskussionsforen im Netz, die teils als sog. Echokammern fungieren, und die Möglichkeit, anonym und ohne Konsequenzen zu agieren, begünstigen die Verbreitung von Hassbotschaften. Laut einer deutschen Studie ist der Anteil derer, die schon (sehr) häufig persönlich Hasskommentare im Internet gesehen haben, zwischen 2016 und 2021 von 26% auf 39% gestiegen (Landesanstalt für Medien NRW, 2021). In einer EU-weiten Umfrage gaben 50% der jungen Erwachsenen an, selbst schon von Hass im Internet betroffen gewesen zu sein (HateAid GmbH, 2021). In Reaktion darauf hat der Europarat von 2014 bis 2017 eine Jugendkampagne, das No Hate Speech Movement[4], gefördert, die in diversen nationalen Organisationen weiter aktiv ist, und 2020 einen der Aufklärung dienenden Handlungsleitfaden herausgegeben (Keen et al., 2020). Seit 2020 ist die Organisation HateAid als „erste Beratungsstelle Deutschlands gegen Hass im Netz“ aktiv[5].

[3] S. https://www.bpb.de/252396/was-ist-hate-speech.

[4] S. https://www.coe.int/en/web/no-hate-campaign.

[5] U.a. mit Förderung des Staates; Näheres unter https://hateaid.org.

1.3.2 Juristische Einordnung

Hate Speech ist in vielen Fällen auch rechtswidrig: Die Ahndung bestimmter Formen von Hate Speech wie der Anstiftung zu Hass und Gewalt stützt sich auf das Diskriminierungsverbot in Art. 7 der Allgemeinen Erklärung der Menschenrechte der Vereinten Nationen[6] (AEMR); sie dient dem Schutz einer demokratischen Debattenkultur und sogar mittelbar der Genozidverhinderung. Gegen diesen Schutz muss aber immer das Recht auf freie Meinungsäußerung, Art. 19 AEMR, abgewogen werden. Konkret gilt im Völkerrecht Art. 4 der *International Convention on the Elimination of All Forms of Racial Discrimination*[7], die jedoch nicht gegen alle Arten von Hass schützt und nicht von allen Ländern umgesetzt wird.

In Deutschland ist die Anstiftung zu Hass und Gewalt als Volksverhetzung gemäß § 130 StGB strafbar. Die Verbreitung dieser und anderer Arten von Hate Speech im Internet stellt jedoch neue Herausforderungen. Zur Verbesserung der Rechtsdurchsetzung im Netz gilt in Deutschland deshalb seit 2017 das NetzDG: Die Betreiber größerer sozialer Netzwerke werden dazu verpflichtet, wirksame Beschwerdeverfahren einzurichten, die es Nutzer:innen ermöglichen, problematische Inhalte zu melden. Soziale Netzwerke trifft ein Bußgeld, wenn sie einen gemeldeten und tatsächlich rechtswidrigen Post nicht innerhalb von 24 Stunden löschen. Im §1 Abs. 3 NetzDG ist anhand einer Liste der relevanten Straftatbestände definiert, welche Inhalte genau gelöscht werden müssen. Die Tatbestände werden in Abschnitt 3.2.2 im Einzelnen erläutert. Mit dem NetzDG werden Betreiber sozialer Netzwerke zur Rechenschaft gezogen und die Nutzer:innenrechte geschützt; damit dient es in Europa als Vorbild.

In Europa hat es diverse Ansätze gegeben, um mit Hate Speech umzugehen: In einer Empfehlung des Ministerkomitees des Europarates von 1997[8] wurden Richtlinien zur Bekämpfung von Hassrede bereitgestellt und die Verurteilung jeder Form von Hassrede betont, weil sie „die demokratische Sicherheit, den kulturellen Zusammenhalt und den Pluralismus“ unterwandere. Seitdem ist 2004 ein Übereinkommen über Computerkriminalität (mit einer Erweiterung 2006) in Kraft getreten[9], das jedoch nicht alle Mitgliedstaaten unterzeichnet haben. Weitere Schritte gab es 2008[10], 2016[11] und 2017[12]. Im Dezember 2020 hat

[6] Allgemeine Erklärung der Menschenrechte, A/RES/217, UN-Doc. 217/A-(III), 1948, abrufbar unter https://www.un.org/depts/german/menschenrechte/aemr.pdf.

[7] International Convention on the Elimination of All Forms of Racial Discrimination, in Kraft getreten am 4.1.1969, als UN General Assembly resolution 2106 (XX), abrufbar unter https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial.

[8] Rec(97)20, 30.10.1997, zu "Hate Speech", abrufbar unter https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680505d5b.

[9] ETS No. 185, 23.01.2001, und ETS No. 189, 28.01.2003; abrufbar unter https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=185, https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=189.

[10] Rahmenbeschluss 2008/913/JI des Rates zur strafrechtlichen Bekämpfung bestimmter Formen und Ausdrucksweisen von Rassismus und Fremdenfeindlichkeit, 28.11.2008, abrufbar unter http://data.europa.eu/eli/dec_framw/2008/913/oj.

[11] The EU Code of conduct on countering illegal hate speech online; abrufbar unter https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_de.

[12] „Umgang mit illegalen Online-Inhalten; Mehr Verantwortung für Online-Plattformen“,

die Europäische Kommission schließlich einen Entwurf für einen Digital Services Act (DSA) und einen Digital Markets Act (DMA) vorgelegt[13], mit dem Ziel, eine einheitliche Richtlinie für den Schutz der Grundrechte in sozialen Netzwerken, Suchmaschinen und anderen Internetportalen sowie Maßnahmen zur Bewahrung eines fairen Marktes im Internet im EU-Raum zu verfassen. Sie beinhalten einige dem NetzDG sehr ähnliche Regelungen. Im März und April 2022 wurde politische Einigung über den DSA und DMA erreicht[14] und die beiden Entwürfe wurden vom Europäischen Parlament ratifiziert[15]. Alle weiteren Schritte, insb. die Ratifizierung durch den Rat der Europäischen Union, stehen noch aus[16].

Laut NetzDG können die Betreiber sozialer Medien genau dann sanktioniert werden, wenn sie strafbare Inhalte, auf die sie aufmerksam gemacht wurden, nicht löschen. Im Gegensatz dazu stellen die USA den Schutz der freien Meinungsäußerung deutlich höher als den Schutz vor Belästigung im Netz – Hate Speech per se ist nicht strafbar (Fisch, 2002) und Hostinganbietern wird auch im Falle von Hate Speech laut dem sog. Communications Decency Act[17] von 1996 Immunität gewährt. Die höhere Schutzfunktion, die dem NetzDG und in Zukunft dem DSA zukommen, ist daher insbesondere im Verhältnis zu den vermehrt in den USA angesiedelten Betreibern sozialer Netzwerke und anderer großer Portale von großer Wichtigkeit und kann potentiell auch im globalen Kontext als Vorbild dienen.

Kommunikation der Kommission; abrufbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52017DC0555.

[13]Vom 15.12.2020: Vorschlag für eine VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie 2000/31/EG, COM/2020/825 final, Dokument 52020PC0825, abrufbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=COM:2020:825:FIN und Vorschlag für eine VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES über bestreitbare und faire Märkte im digitalen Sektor (Gesetz über digitale Märkte), COM/2020/842 final, Dokument 52020PC0842, abrufbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52020PC0842.

[14]Pressemeldungen der Europäischen Kommission abgerufen unter https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545 und https://ec.europa.eu/commission/presscorner/detail/en/IP_22_197.

[15]Pressemeldung dazu abgerufen unter https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4313.

[16]S. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

[17]Section 230 des *Title 47 of the United States Code*, Teil des *Communications Decency Act* von 1996, abzurufen unter https://www.law.cornell.edu/uscode/text/47/230.

2 Übersicht des aktuellen Forschungsstandes

Hate Speech und andere anstößige, abwertende und angreifende Inhalte zu erkennen und von neutralen Äußerungen zu unterscheiden, kann als eine Aufgabe des Natural Language Processing (NLP) aufgefasst und automatisiert werden. Ziel kann dabei sowohl die linguistische Untersuchung dieser Äußerungen als auch deren potentielle Moderation sein. Diverse Workshops und Shared Tasks haben in den letzten Jahren wichtige Beiträge in diesem Bereich geleistet, insbesondere die Erstellung annotierter Datensätze (u.a. Akiwowo et al., 2020; Bhatia & Shaikh, 2020; Fišer et al., 2018; Kumar et al., 2020; Mandl et al., 2021; Mostafazadeh Davani et al., 2021; Roberts et al., 2019).

In Abschnitt 2.1 folgt eine Übersicht der vorhandenen Datensätze, insbesondere deutschsprachiger; dann werden in Abschnitt 2.2 die Klassifikationsmethoden und in Abschnitt 2.3 die Herausforderungen der Untersuchung von Hate Speech beleuchtet. Im Anschluss wird in Abschnitt 2.4 der Stand der Forschung juristischer Perspektiven auf Hate-Speech-Forschung erläutert.

2.1 Datensätze und weitere Ressourcen

Um Hate Speech zu untersuchen, wurden in den letzten Jahren viele Datensätze und andere Ressourcen wie z.B. Lexika von Schimpfwörtern oder diskriminierenden Begriffen erstellt. Die meisten Datensätze wurden aufgebaut, um als Grundlage eines Klassifikationsverfahrens zu dienen; die linguistische Analyse von Hate Speech ist dabei eher ein sekundäres Ziel. Detailliertere Übersichten zu den zahlreichen, in diversen Sprachen zur Verfügung stehenden Datensätzen und anderen Ressourcen wie z.B. Lexika sind bei Vidgen und Derczynski (2021)[18] , Poletto et al. (2021) und Schmidt und Wiegand (2017) zu finden.

2.1.1 Annotationansätze

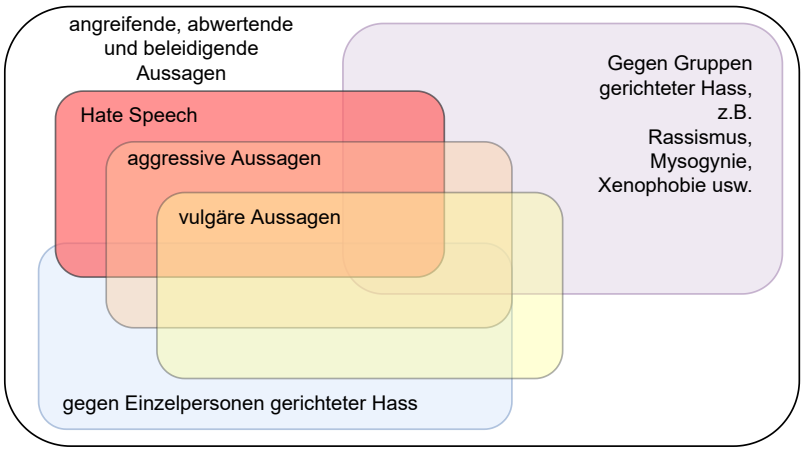

Abbildung 2: Mögliche Annotationsebenen von Hate Speech; angelehnt an Poletto et al. (2021, S. 482)

[18]Eine Übersicht der Datensätze findet sich auf https://hatespeechdata.com/.

Dass Hate Speech nicht einheitlich definiert wird, spiegelt sich auch in der Vielfalt der Annotationsschemata wider, anhand derer Datensätze erstellt werden: Ein Großteil der Arbeiten verwendet zwei Labels wie z.B. hateful / not hateful, um binär zwischen neutralen und in irgendeiner Form abwertenden Aussagen zu unterscheiden, z.B. Bohra et al. (2018), Mandl et al. (2019), Pavlopoulos et al. (2017) und Schäfer und Burtenshaw (2019); 18 der 44 von Poletto et al. (2021) untersuchten Datensätze sind so annotiert. Die abwertende Kategorie kann dabei, wie in Abb. 2 zu sehen ist, jedoch sehr verschiedene Arten anstößiger Kommunikation betreffen, z.B. aggressive Beleidigungen (Schäfer & Burtenshaw, 2019), aber auch spezifische Diskriminierungsformen wie Homophobie (Akhtar et al., 2019) oder Rassismus, Islamophobie und Antisemitismus (Oriol et al., 2019).

Außerdem wird zum Teil versucht zwischen spezifischen Arten von Hate Speech zu unterscheiden. Dabei zeigen die verschiedenen Annotationsansätze voneinander abweichende Verständnisse des Phänomens Hate Speech: Einige verstehen die Nutzung von Schimpfwörtern, Beleidigungen und diskriminierenden Äußerungen alle als Teil des gleichen Phänomens, unterschieden durch die Intensität des Hasses (Bretschneider & Peters, 2017; Vigna et al., 2017). Andere behandeln die Ausprägungen von Hate Speech als qualitativ verschiedene Phänomene und unterscheiden Kategorien wie z.B. obszön / beleidigend / beleidigend aber nicht obszön (Mubarak et al., 2017) oder auch Rassismus / Sexismus / beides / weder noch (Waseem & Hovy, 2016). Diese Annotation wird oft als zweite Ebene der binären Annotation behandelt.

Einige Arbeiten, die Formen des auf Gruppen bezogenen Hasses untersuchen, annotieren verschiedene Gruppen bzw. die Gruppenzugehörigkeit der angegriffenen Person, so z.B. Ousidhoum et al. (2019); manchmal wird auch nur Hass gegen eine bestimmte Gruppe erforscht, wie z.B. Ausländerfeindlichkeit (Roß et al., 2016). In einigen Arbeiten wird auch annotiert, ob es um implizite oder explizite abfällige Sprache geht (Palmer et al., 2020; Struß et al., 2019; Waseem et al., 2017). Diese verschiedenen Annotationsebenen überlappen sich zum Teil in komplexen Annotationsschemata.

2.1.2 Daten

Wer einen Hate-Speech-Korpus erstellt, steht vor der Schwierigkeit, dass viele für die Forschung relevante Inhalte bereits gelöscht wurden. Um Daten darüber zu sammeln, wie genau Hate Speech aussieht, muss zunächst Hate Speech gefunden werden. Die meisten sozialen Netzwerke löschen jedoch derartige Inhalte möglichst schnell. Die Forschung muss also entweder Zugang zu diesen gelöschten Inhalten erbitten oder innerhalb der öffentlich zugänglichen Inhalte Hate Speech suchen, die beim Löschen übergangen wurde. Die meisten Datensätze bestehen daher aus öffentlich zugänglichen Tweets sowie Kommentaren oder Posts von Facebook, Reddit, YouTube und Nachrichtenportalen (Poletto et al., 2021; Schmidt & Wiegand, 2017; Vidgen & Derczynski, 2021). Twitter wird, u.a. wegen seiner einfach zugänglichen API, am häufigsten verwendet. Es werden einzelne Posts oder Kommentare klassifiziert; nur wenige Ansätze setzen sich mit ganzen Diskussionsverläufen oder längeren Texten auseinander (Vidgen & Derczynski, 2021). Mittlerweile werden neben Texten z.B. auch

Memes als Form von Hate Speech untersucht (Mostafazadeh Davani et al., 2021; Oriol et al., 2019; Pramanick et al., 2021).

Zur Auswahl und Sammlung der relevanten Stichproben von Inhalten sozialer Netzwerke werden diverse sog. Samplingverfahren (*sample*: engl. Stichprobe) verwendet: Anhand von Listen (negativ konnotierter) Stichwörter werden große Mengen an Kommentaren gesammelt und dann weiter gefiltert (Poletto et al., 2021). Üblich ist es auch, Nutzer:innen zu identifizieren, die besonders viel Hate Speech posten und anhand dieser Posts, z.B. mit Hilfe von Diskussionsketten, auf weitere ähnliche Inhalte zu schließen. Alternativ werden relevante Foren ausgemacht und deren gesamte Chatverläufe ausgewertet (Kumar et al., 2018). Außerdem werden teilweise bereits existierende Klassifikatoren genutzt, um eine Vorauswahl zu treffen (Mandl et al., 2020).

Wenn auch die Mehrheit der vorhandenen Datensätze aus Inhalten in englischer Sprache besteht, gibt es dennoch auch einige Datensätze in anderen Sprachen: Poletto et al. (2021) haben diverse Korpora z.B. für das Italienische, Arabische, Hindi und weitere gefunden; https://hatespeechdata.com/ verzeichnet Korpora für das Portugiesische, Spanische, Dänische und viele weitere – insgesamt 25 verschiedene Sprachen. Für das Deutsche gibt es momentan sieben veröffentlichte Datensätze:

Paper	**Größe**	**grobe Annotation**	**präzise Annotation**
Roß et al. (2016)[19]	469	anti-refugee hate/none	NA
Bretschneider und Peters (2017)[20]	5836	offensive/other	offensive/severely offensive
GermEval 2018 (Wiegand et al., 2018)[21]	8541	offensive/other	profanity/insult/abuse
GermEval 2019 (Struß et al., 2019)[22]	7025	offensive/other	profanity/insult/abuse
HASOC19 (Mandl et al., 2019)[23]	4669	hate and offense/not	hate speech/offensive/profane
HASOC20 (Mandl et al., 2020)[24]	3400	hate and offense/not	hate speech/offensive/profane
Wich et al. (2021)[25]	4960	abusive/neutral	NA

Tabelle 1: Übersicht der deutschsprachigen Hate-Speech-Datensätze

2.2 Klassifikationsmethoden

Hate Speech zu erkennen und von normalen Äußerungen zu unterscheiden, ist ein Textklassifikationsproblem. Als Klassifikationsmethodik wird überwiegend überwachtes Machine Learning verwendet (Poletto et al., 2021; Schmidt & Wiegand, 2017). Grundlegende Herausforderungen bei der Klassifikation von Hate Speech liegen im Mangel einer allgemein akzeptierten Definition und Systematik, der linguistischen Komplexität (Fortuna & Nunes, 2018) sowie in der Knappheit qualitativ hochwertiger Daten (Poletto et al., 2021).

Davon ausgehend, dass Hate Speech abwertende Phrasen und Schimpfwörter enthält, besteht ein Ansatz zur Klassifikation in der Stichwortsuche anhand von Lexika (Schmidt & Wiegand, 2017). Es gibt einige Wortlisten, die im

[19]Datensatz abrufbar unter https://github.com/UCSM-DUE/IWG_hatespeech_public.
[20]Datensatz abrufbar unter http://ub-web.de/research/.
[21]Datensatz abrufbar unter https://github.com/uds-lsv/GermEval-2018-Data.
[22]Datensatz abrufbar unter https://projects.fzai.h-da.de/iggsa/data-2019/.
[23]Datensatz abrufbar unter https://hasocfire.github.io/hasoc/2019/dataset.html.
[24]Datensatz abrufbar unter https://hasocfire.github.io/hasoc/2020/dataset.html.
[25]Datensatz abrufbar unter https://github.com/mawic/german-abusive-language-covid-19.

Internet zur Verfügung stehen, unter anderem die mehrsprachigen Ressourcen Hatebase[26] und HurtLex[27] (Bassignana et al., 2018). Darüber hinaus gibt es auch Listen, die auf bestimmte Arten von Hate Speech ausgerichtet sind[28] oder solche, die ad hoc zusammengestellt werden, wie z.B. Njagi et al. (2015), die Verben sammeln, mit denen zu Gewalt aufgerufen wird. Lexika allein reichen zwar meistens nicht, um als Grundlage eines Klassifikationsverfahren zu dienen (Nobata et al., 2016), werden jedoch als Features für Machine-Learning-Verfahren oder ergänzend, z.B. bei der Datensammlung, genutzt (Poletto et al., 2021; Schmidt & Wiegand, 2017).

Weitere übliche Features sind Bag-of-Words-Darstellungen und N-Gramme. Bag-of-Words-Modelle speichern das Vokabular eines aus mehreren Texten bestehenden Korpus. Jedem Text wird ein Vektor der Größe des Vokabulars zugewiesen, in dem für jedes Wort angegeben wird, wie oft es in dem Text vorkommt (Jurafsky & Martin, 2009, S. 653; Manning & Schutze, 1999, S. 237). Wörter, die nur in anderen Texten des Korpus vorkommen, haben also die Frequenz `0`. Die Reihenfolge der Wörter im Text spielt dabei keine Rolle, deshalb haben solche Repräsentationen den Nachteil, dass sie keinen Wortkontext abbilden. N-Gramme sind die Verallgemeinerung dieses Prinzips – es werden nicht nur einzelne Wörter gezählt, sondern auch alle Wortkombinationen der Länge N oder anstatt von Wörtern die einzelnen Buchstaben und andere Zeichen (Jurafsky & Martin, 2009, Kap. 4). N-Gramme sind oft hilfreiche Features, wobei sie in einigen Arbeiten der Hate-Speech-Forschung auf Zeichenebene effektiver für die Klassifikation waren als solche auf Tokenebene (Mehdad & Tetreault, 2016; Schmidt & Wiegand, 2017). Außerdem werden die Anzahl von URL-Erwähnungen, die Zeichensetzung, die Länge der Kommentare oder der einzelnen Token, Groß- und Kleinschreibung und das Vorkommen unbekannter Wörter als Features benutzt (Schmidt & Wiegand, 2017).

Um damit umzugehen, dass Hate-Speech-Datensätze oft klein sind, wird mit Word Embeddings, einer Form der Wortverallgemeinerung, gearbeitet. Word Embeddings sind Vektoren mit fester Größe, die für jedes Wort abbilden, wie wahrscheinlich es mit anderen Wörtern auftritt. Weil ihre Grundlage das Lernen von Wortkontexten ist, geben sie semantische Informationen über die Wörter wieder. Diese Vektoren sind kürzer und dichter als die dem Bag-of-Words-Vektoren und in manchen Fällen auch besser als Features geeignet (Turian et al., 2010). Übliche Algorithmen zur Berechnung dieser Word Embeddings sind `Word2Vec` (Mikolov et al., 2013), `GloVE` (Pennington et al., 2014) und `FastText` (Bojanowski et al., 2016). Für das Sprachmodell BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) wird `Masked Language Modelling` verwendet. Sollen statt einzelner Wörter ganze Sätze oder Absätze abgebildet werden, kann dazu entweder der Durchschnitt der Word Embeddings oder ein gesondertes Verfahren für Absatz Embeddings verwendet werden (Le & Mikolov, 2014). In dieser Arbeit wird für die logistische Regression mit dem Bag-of-Words-Ansatz, für das Transfer Learning auf Basis von BERT mit Word Embeddings gearbeitet.

[26] https://hatebase.org/, 98 Sprachen.

[27] https://github.com/valeriobasile/hurtlex, mehr als 50 Sprachen und 17 verschiedene Kategorien.

[28] Z.B. http://www.rsdb.org/, rassistische Schimpfwörter im Englischen.

Einige Ansätze gehen davon aus, dass Hate Speech und negatives Sentiment gemeinsam auftreten und nutzen deshalb Sentiment-Klassifikation als Features oder Vorstufe einer Klassifikation (Rajamanickam et al., 2020; Schmidt & Wiegand, 2017). Auch linguistische Features werden genutzt – Part-of-Speech-Tags sind dabei weniger effektiv als z.B. Informationen zu Syntaxdependenzen (Chen et al., 2012; Xu et al., 2012). Ebenfalls werden Meta-Informationen über die Autor:innen der Kommentare verwendet: Wie oft sie posten, ob sie bereits Hasskommentare veröffentlicht haben, welches Geschlecht sie haben usw. Andere Arbeiten binden Kontextwissen und Ontologien in die Klassifikation ein (Schmidt & Wiegand, 2017).

Diese Features, d.h. die unterschiedlichen Arten, Text numerisch zu erfassen, dienen als Trainingsinput für mögliche Klassifikationsverfahren. Zur Klassifikation werden oft SVM und Random Forests eingesetzt (Fortuna & Nunes, 2018), andere (Mehdad & Tetreault, 2016; Wiegand et al., 2018) arbeiten mit rekurrenten neuronalen Netzen (RNN). Zur Textklassifikation werden außerdem Logistische Regression, Multinomial Naive Bayes und auch der K-Nearest-Neighbor-Algorithmus (kNN) verwendet (Kowsari et al., 2019). Seit sich Transformer-Archtitekturen (Vaswani et al., 2017) wie BERT (Devlin et al., 2018) in vielen NLP-Aufgaben behauptet haben, wird auch in der Hate-Speech-Forschung damit gearbeitet (Struß et al., 2019; Wiegand et al., 2018; Zufall et al., 2020). Diese Algorithmen liefern auf geringen Datenmengen unterschiedlich gute Ergebnisse und unterscheiden sich in ihren Trainingszeiten (Kowsari et al., 2019). Je nach Aufgabenstellung und verfügbaren Daten wird das jeweils adäquateste Modell genutzt. In dieser Arbeit wird zum einen mit logistischer Regression gearbeitet, weil diese mit schnellen Trainingszeiten und interpretierbaren Ausgaben als Vergleichsgrundlage dienen kann. Zum anderen wird ein BERT-Sprachmodell feingetunt, weil diese auch bei kleinen Trainingsmengen schon gute Ergebnisse bringen können. Mehr dazu folgt in Abschnitt 4.

2.3 Herausforderungen

Die Klassifikation von Hate Speech beinhaltet einige Hürden: Auf der sprachlichen Ebene untersuchen Palmer et al. (2020) diverse Formen linguistischer Komplexität, die vor allem falsch positive Klassifikationen verursachen können. Als problematisch identifizieren sie Ironie und Sarkasmus sowie eine ausgrenzende grammatikalische Konstruktion[29]. Manchmal werden als Hate Speech auch Beleidigungen benutzt, die jedoch zu Teilen von der angegriffenen Gruppe für den gruppeninternen Gebrauch zurückerobert wurden und in diesem Kontext nicht mehr als beleidigend gelten, wie z.B. „Kanake“. Palmer et al. (2020) erstellen einen Datensatz[30], indem sie in bereits als offensive / not annotierten Daten die benannten Phänomene suchen und annotieren. Damit testen sie bisherige Klassifizierer auf diese Schwachstellen und finden heraus, dass implizite und

[29]Im Englischen und teilweise auch im Deutschen gibt es aus Adjektiven gewonnene Gruppenbezeichnungen mit bestimmtem Artikel, die in ihrer sehr klaren Abgrenzung zwischen Sprecher:in und bezeichneter Gruppe eine diskursive Ausgrenzung bewirken. So zum Beispiel der Unterschied zwischen „black people“ und „the blacks“.

[30]Complex Offensive Language Data Set for English (COLD-EN), abrufbar unter https://github.com/alexispalmer/cold.

explizite Aussagen gleichermaßen korrekt zugeordnet werden, Aussagen mit Trotzwörtern jedoch mehrheitlich inkorrekt als offensive eingeschätzt werden.

Jenseits der linguistischen Komponente sind auch der Zugang zu und die Erstellung guter Daten durch die Materie erschwert. Hate Speech wird zu großen Teilen aus sozialen Medien gelöscht und zu den Sammlungen gelöschter Inhalte gibt es meistens keinen Zugriff. Die Samplingmethoden, auf die deshalb zurückgegriffen wird, bergen jedoch alle gewisse Risiken für Bias (Poletto et al., 2021; Wiegand et al., 2019); häufig gibt es z.B. einen Themenbias hin zu Politik, der nicht immer als solcher benannt wird. Alle aus bestimmten Diskussionsforen gewonnenen Datensätze sind in ihrem Themenspektrum begrenzt und können nicht ohne weiteres als generalisierbar dargestellt werden.

Hate Speech wird meist zum Zwecke der Moderation klassifiziert – diese Moderation muss begründet werden. Daher gibt es einige Ansätze zur Erklärung von Klassifikationsentscheidungen. Konkret wird versucht herauszufinden, welche Wörter in der Ausgabe ausschlaggebend für die Klassifikation sind. Bei einigen Methoden, z.B. Long Short-Term Memory (LSTM) mit Attention-Mechanismen, ist es möglich die Gewichtung der Wörter ausgeben zu lassen. Aber diese Gewichtung ist nur bedingt aussagekräftig, da oft ein kleiner Teil des Vokabulars klassenübergreifend als relevant gewichtet wird (Risch et al., 2020). Für andere Modelle können post hoc erfolgreich Schlagwörter identifiziert werden, indem für verschiedene Wortauslassungen im Input oder auch für synthetischen Input, der nur aus wenigen Wörtern besteht, Veränderungen im Output gemessen werden (Kennedy et al., 2020; Risch et al., 2020; Wang, 2018). Dies kann dabei helfen, Bias in den Modellen zu untersuchen, macht Modelle jedoch nicht vollständig interpretierbar (Risch et al., 2020; Wang, 2018).

2.4 Juristische Ansätze

Die juristischen Dimensionen von Hate Speech haben bisher nur wenige Arbeiten als Problem des NLP erforscht:

Fišer et al. (2017) setzen sich mit anstößigen Inhalten im Slowenischen auseinander. Sie erarbeiten ein Annotationsschema, das die slowenische Rechtslage widerspiegelt; insbesondere die Beleidigungstatbestände und die Volksverhetzung finden hier Parallelen. Anhand dessen überarbeiten sie die Annotationen eines von 2010 bis 2015 gesammelten Datensatzes mit insgesamt 13.000 problematischen Kommentaren[31]. Sie ordnen die Daten in vier Ebenen ein:

1. Keine problematischen Elemente (43 % der Daten),
2. Unangemessene Elemente (24 %),
3. Anstößige, angreifende Inhalte (16 %) und zuletzt
4. Hate Speech (nicht strafbar – 13 %, strafbar – 3 %).

Außerdem wird im Falle von Hate Speech auch das die Zielgruppe des Hasses auszeichnende Merkmal – ethnische Herkunft, "Rasse„, sexuelle Orientierung, politische Einstellung oder Religion – vermerkt. Ein eigenes Modell zur Klassifizierung wird nicht trainiert.

[31]Gesammelt von der Organisation Spletno Oko; die Mehrheit der Daten bilden Facebook-Posts und ihre Kommentare sowie Kommentare von Nachrichtenportalen.

Es konnten zwei Arbeiten gefunden werden, die die Klassifikation von deutschsprachigen Posts in Bezug auf ihre strafrechtliche Relevanz untersuchen: Zufall et al. (2020), Zufall et al. (2019) setzen sich bei mit deutschen Tweets auf der Grundlage des Hate-Speech-Korpus des GermEval2018 Shared Task (Wiegand et al., 2018) auseinander:

Zufall et al. (2019) erstellen einen Annotationsleitfaden für die Beleidigungstatbestände (§§ 185-187 StGB), der explizit für die Annotation durch juristische Laien angelegt ist – er besteht aus binären Entscheidungsfragen, aus denen nach der Annotation durch die Annotatoren die eigentliche Einordnung entschieden wird. Das Annotationsschema enthält nicht nur die Merkmale der Beleidigungstatbestände, sondern auch eine Abwägung der zu schützenden Rechtsgüter – das Recht auf den Schutz der persönlichen Ehre[32] einerseits und der Meinungsfreiheit[33] andererseits. Anhand dessen annotieren sie als Proof of Concept den als abusive annotierten Teil der GermEval2018-Daten und trainieren ein bidirektionales LSTM-Modell. Sie stellen fest, dass nur ein sehr geringer Anteil der Daten tatsächlich nach den Beleidigungstatbeständen strafbar ist und dass eine direkte Klassifikation strafbar/nicht strafbar sehr schlechte Ergebnisse erzielt. Stattdessen trainieren sie alle Schritte der Annotation, die ja jeweils binäre Entscheidungen sind, einzeln: In Bezug auf das Opfer der Beleidigung wird klassifiziert, ob es um eine (noch lebende) Einzelperson oder um eine bestimmte Gruppe geht; in Bezug auf die Tathandlung wird untersucht, ob eine Verunglimpfung, eine Tatsachenbehauptung, eine missbräuchliche Beleidigung oder eine missbräuchliche Kritik vorhanden ist und, ob die Äußerung von öffentlichem Interesse ist.

In Zufall et al. (2020) wird darauf aufbauend die Automatisierung einer Strafbarkeitskeitsprüfung von Hate Speech gemäß EU-Recht untersucht. Es wird ein Annotationsschema für die Strafbarkeit der Volksverhetzung gemäß EU-Recht, also der Anstiftung zu Hass und Gewalt ausgearbeitet. Es werden auch hier die als abusive annotierten Teile der GermEval2018-Daten annotiert, und auch hier ist ein Versuch der direkten Klassifizierung strafbar/nicht strafbar mit einem F1-Maß von 0,39 wenig erfolgreich. Daher wird die Prüfung zweigeteilt: Separat werden die target group des Tweets, also das Objekt des Hasses, und die target behaviour, also das propagierte Verhalten (Hass oder Gewalt) annotiert, trainiert und schließlich klassifiziert. In Kombination der Ergebnisse der Klassifizierung von target group und behaviour zur Feststellung der Strafbarkeit wird insgesamt eine Verbesserung der Ergebnisse zu F1-Maß von 0,42 erreicht. Es wurde mit dem GBERT-base-Modell gearbeitet.

[32]Abgeleitet von Art. 2(1) und Art. 1(1) des Grundgesetzes (GG); BVerfGE 35, 202; E 54, 148, 155.

[33]Art. 5 GG, § 193 StGB.

3 Erstellen eines deutschsprachigen Referenzdatensatzes

Für die vorliegende Arbeit zur automatischen Erkennung strafrechtlich relevanter Hate Speech wird zunächst ein deutschsprachiger Referenzdatensatz erstellt, der als Grundlage für alle weiteren Schritte dient[34]. Dazu werden vier aus Twitter zusammengestellte deutschsprachige Datensätze, deren Annotationen miteinander kompatibel sind, zu einem einzigen Datensatz kombiniert: Die GermEval-Datensätze von 2018 (Wiegand et al., 2018) und 2019 (Struß et al., 2019) und die deutschsprachigen Teile der HASOC-Datensätze 2019 (Mandl et al., 2019) und 2020 (Mandl et al., 2020). Die GermEval-Datensätze sind frei verfügbar[35]; die HASOC-Daten werden nur für Forschungszwecke freigegeben[36]. Danach wird ein Teil dieser Daten nach einem eigenen Schema neu annotiert.

Im Folgenden wird in Abschnitt 3.1 die technische Angleichung erläutert. In Abschnitt 3.2 wird erst die Angleichung der vorhandenen Annotationen und danach die potentielle Annotation von Straftatbeständen beschrieben. Schließlich wird das eigene Annotationsschema für den Tatbestand der Volksverhetzung gemäß § 130 StGB präsentiert und der Annotationsprozess dokumentiert.

3.1 Technische Angleichung

Zuerst werden die Daten einheitlich formatiert: Alle Zeilenumbrüche bzw. das in den GermEval-Datensätzen dafür eingesetzte Token `|LER|` werden mit einem Leerzeichen ersetzt – in der Annahme, dass sie in Tweets keine wichtige linguistische Funktion erfüllen. Weiterhin muss bei einem Teil der GermEval2019-Testdaten das fehlende erste Zeichen wiederhergestellt werden. `html`-Schnipsel wie z.B. Emojis, die im Format `<U+0001F643>` gespeichert sind anstatt als Emoji selbst, oder auch `&amp;` anstatt `&`, werden jeweils durch die `utf-8`-kodierte Version ersetzt. Außerdem werden alle Hyperlinks mit dem generischen Twitter-Link `https://t.co` ersetzt.

Dann werden angelehnt an Wich et al. (2021) im Falle sehr ähnlicher Tweets Duplikate entfernt. Dazu werden erst alle `@user`-Erwähnungen anonymisiert und dann für alle Tweetpaare der Jaccard-Index mit Multimengen (Leskovec et al., 2014, Kap. 3, S. 74-77; Broder, 1997) berechnet: Der Jaccard-Index ist ein Ähnlichkeitsmaß für Mengen mit Werten zwischen `0` und `1`. Auch Texte können damit verglichen werden, indem sie auf ihr Vokabular, d.h. die Menge ihrer Zeichen, abgebildet werden. Der Wert ergibt sich durch Division der Größe der Schnittmenge durch die Größe der Vereinigung. Jedoch werden damit auch recht unterschiedliche Texte als ähnlich eingeschätzt – so erhalten z.B. `„@user einigen wir uns doch darauf, dass linker antikommunismus ein nogo ist"` und `„@user ne mach doch was du willst, kann dem thema eh nicht ausm weg gehen"` einen Index von 0.95.

[34] Das Repository für diese und alle weiteren Arbeitsschritte ist unter https://github.com/cmbirle/legal-hate-speech zu finden.

[35] Unter https://github.com/uds-lsv/GermEval-2018-Data und https://projects.fzai.h-da.de/iggsa/data-2019/.

[36] Daher können die HASOC-Daten und der vollständige Referenzdatensatz nicht im Git-Repository hochgeladen werden.

Um dem vorzubeugen, kann statt mit Mengen mit Multimengen gerechnet werden. Sie können Elemente mehrmals enthalten, geben also die Frequenz der Zeichen im Text wieder, und beinhalten demnach mehr Informationen über sie. Bei der Berechnung wird statt durch die Vereinigung durch die Summe dividiert, und die genaue Berechnung des Durchschnitts weicht ab[37]. Der Jaccard-Index J für die Multimengen A und B zweier Texte beträgt dann:

$$J(A,B) = \frac{|A \cap B|}{|A \uplus B|} = \frac{|A \cap B|}{|A| + |B|} \tag{1}$$

Der Index liegt zwischen 0 (verschieden) und 0.5 (sehr ähnlich); für das oben genannte Paar z.B. 0.38. Für eine Reihe von Tweetpaaren, bei denen nur ein Wort oder nur die Länge unterschiedlich ist, wurde der Jaccard-Index berechnet und schließlich ein Grenzwert von 0,47 festgelegt, ab dem Tweets als Duplikate gelten. Die Duplikatpaare werden in insg. 101 Cluster einander gegenseitig ähnlicher Tweets zusammengefasst, von denen jeweils einer behalten und die anderen aus dem Datensatz entfernt werden (220 der insg. 23636 Tweets).

Schließlich werden alle Tweets bearbeitet und gesammelt im neuen Format gespeichert: `corpus_id \t tweet \t binarylabel \t finelabel` Anhand der neu vergebenen ID `corpus_id` kann später jeder Tweet auf den ersten Blick seinem ursprünglichen Datensatz zugeordnet werden[38].

3.2 Annotation

3.2.1 Angleichen der Annotation

In den Datensätzen gibt es zwei Annotationsebenen: Zum einen wird zwischen Tweets mit und ohne abwertender Aussage unterschieden. Zum anderen werden die abwertenden Tweets noch feiner annotiert. Diese feineren Unterscheidungen sind in GermEval und HASOC ähnlich definiert und daher miteinander kompatibel. Die Annotationslabels werden wie in Tabelle 2 dargestellt vereinheitlicht:

GermEval	**HASOC**	**Referenzdatensatz**
		binär
OFFENSE	HOF = hate and offense	NEG
OTHER	NOT	NOT
		detailliert
ABUSE	HATE (= hate speech)	HATE
INSULT	OFFN (= offensive)	INSOFF
PROFANITY	PRFN (= profane)	PRFN
OTHER	NONE	NONE

Tabelle 2: Vereinheitlichung der Annotation

[37] „Suppose that $A = \langle A, f\rangle$ and $B = \langle A, g\rangle$ are two multisets, then their intersection, denoted $A \cap B$, is the multiset $C = \langle A, h\rangle$, where for all $a \in A$: $h(a) = min(f(a), g(a))$.“, (Syropoulos, 2001).

[38] Die ID setzt sich aus einem Kürzel für jeden der vier Datensätze (GermEval2018: `01`, GermEval2019: `02`, HASOC2019: `03` und HASOC2020: `04`), der Zuordnung Trainingsdaten, `11`, und Testdaten, `22`, sowie der auf vier Stellen mit `0` aufgefüllten Zeilennummer im ursprünglichen Datensatz zusammen. Der Tweet in Zeile 150 der Testdaten des GermEval2018-Datensatzes erhält also die ID `01220150`.

Im Referenzdatensatz ergibt sich dadurch die in Abb. 3 gezeigte binäre Verteilung. Von insgesamt 23636 Tweets sind 6663, also 28%, als abwertend (NEG) eingestuft. Die GermEval-Daten enthalten davon einen größeren Anteil als die HASOC-Dateien. In Abb. 4 ist die Verteilung der detaillierten Annotation zu sehen: 1050 Tweets sind als PRFN, 2553 als INSOFF und 3060 als HATE gelabelt. HATE macht demnach 12.9% des gesamten Korpus aus.

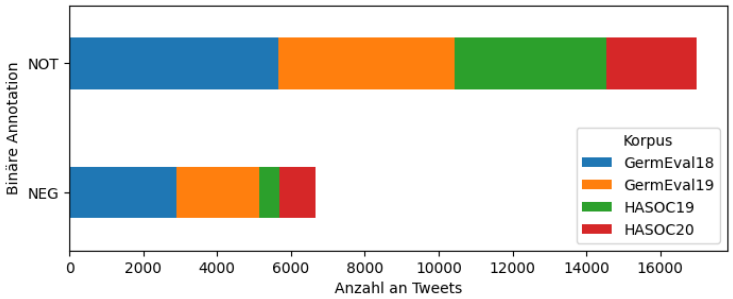

Abbildung 3: Verteilung der binären Annotation im Referenzdatensatz

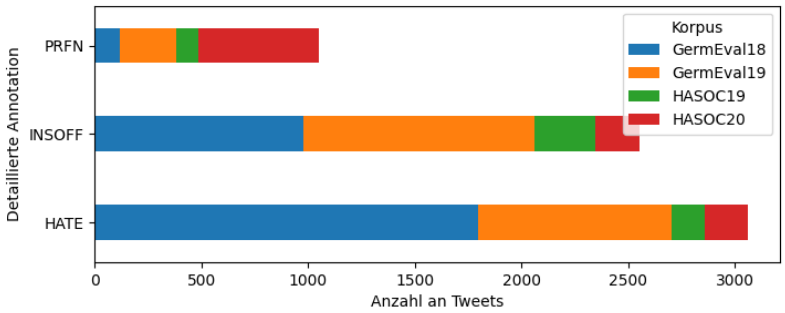

Abbildung 4: Verteilung der detaillierten Annotation im Referenzdatensatz

3.2.2 Zur Annotation von Straftatbeständen

Dieser Datensatz soll nun bezüglich seiner potentiellen Strafbarkeit weiter annotiert werden. Mit HATE sind die aggressivsten und daher auch wahrscheinlich am Ehesten strafbaren Tweets gelabelt, deshalb werden nur diese weiter annotiert. Aber wie genau kann eine potentielle Strafbarkeit annotiert werden? Wie kann sich jemand durch einen Tweet überhaupt strafbar machen?

Die gerichtliche Prüfung einer Straftat gliedert sich nach deutschem Recht in drei Komponenten: Tatbestand, Rechtswidrigkeit und Schuld. In der Tatbestandsprüfung wird erst der objektive Tatbestand geprüft – an Hand der vorliegenden Tatsachen nachvollzogen, ob die im StGB definierten Tatbestandsmerkmale erfüllt sind – und dann der subjektive Tatbestand, der sich auf die innere Welt der Täter:innen (z.B. Vorsatz) bezieht (Roxin & Greco, 2020, §10 F). Es folgen die Prüfung der Rechtswidrigkeit und Schuld – sie erfordert Wissen um die Umstände und den Kontext der Tat (Roxin & Greco, 2020, §§ 14, 19), die in einem isolierten Tweet nicht enthalten sind. Auch die subjektive Innenwelt des Täters kann schwerlich an einem Tweet abgelesen werden. In anonymisierten, aus dem Kontext gegriffenen Daten, so wie sie hier vorliegen,

und analog auch bei einer automatisierten Moderation vorlägen, können diese Komponenten nicht geprüft werden. Daher muss sich die Annotation, d.h. die Einschätzung, ob ein Tweet rechtswidrig ist, auf die Merkmale des objektiven Tatbestands beschränken.

Bei der Moderation sozialer Netzwerke wird von dem Vorliegen von Vorsatz, Rechtswidrigkeit und Schuld aber regelmäßig ausgegangen, denn nicht jede Löschung kann gerichtlich geprüft werden. Tatbestände, die die Verbreitung rechtswidriger Äußerungen sanktionieren, sind in verschiedenen Kontexten strafbar, u.a. durch Kundgebung im öffentlichen Raum, wozu auch die Veröffentlichung auf sozialen Netzwerken zählt[39] (Schäfer & Anstötz, 2021, § 130, Rn. 27). Im NetzDG werden die für soziale Medien potentiell einschlägigen Straftatbestände aufgeführt. Um zu beantworten, ob und wie eine Annotation möglich ist, können darunter drei Gruppen unterschieden werden: Tatbestände, die Textaussagen betreffen und ohne Kontext verständlich und strafbar sind; solche, die sich in den meisten Fällen auf Bildmaterial beziehen; und drittens Tatbestände, die erst im Kontext eingebunden klar strafbar sind – z.B. muss teilweise zwischen Tatsachenbehauptungen und Werturteilen unterschieden werden, oder die Begehung einer anderen Straftat ist vorausgesetzt. Tabelle 3 zeigt die Einordnung:

Betrifft Textaussagen; kein Kontext erforderlich
• §§ 86-86a StGB, Verbreitung und Verwendung der Propagandamittel und Kennzeichen verfassungswidriger und terroristischer Organisationen
• § 111 StGB, Öffentliche Aufforderung zu Straftaten
• § 126 StGB, Störung des öffentlichen Friedens durch Androhung von Straftaten
• **§ 130 StGB, Volksverhetzung**
• § 241 StGB, Bedrohung
Betrifft Textaussagen; Kontextwissen erforderlich
• § 89a StGB, Vorbereitung einer schweren staatsgefährdenden Gewalttat
• § 91 StGB, Anleitung zur Begehung einer schweren staatsgefährdenden Gewalttat
• § 100a StGB, Landesverrätischere Fälschung
• §§ 129-129b StGB, Bildung krimineller Vereinigungen
• § 140 StGB, Belohnung und Billigung von Straftaten
• § 166 StGB, Beschimpfung von Bekenntnissen, Religionsgesellschaften und Weltanschauungsvereinigungen
• §§ 185-187 StGB, Beleidigung, Üble Nachrede, Verleumdung
• § 189 StGB, Verunglimpfung des Andenkens Verstorbener
• § 269 StGB, Fälschung beweiserheblicher Daten
Betrifft Bildmaterial
• § 131 StGB, Verherrlichende Gewaltdarstellungen
• § 184 b StGB, Verbreitung, Erwerb und Besitz kinderpornographischer Inhalte
• § 201a StGB, Verletzung des höchstpersönlichen Lebensbereichs und von Persönlichkeitsrechten durch Bildaufnahmen

Tabelle 3: Hate Speech Tatbestände laut NetzDG

[39]BGH, 12.12.2000 - 1 StR 184/00.

In dieser Arbeit soll nur mit dem Text gearbeitet werden; Kontextwissen mit einzubinden, liegt nicht im Fokus. Weiterhin sollen Machine-Learning-Methoden erprobt werden. Damit scheiden die §§ 86-86a StGB, für die ein lexikalischer Ansatz geeignet ist, aus. Der § 111 StGB wird in der Praxis selten angeklagt, da häufig spezifischere Straftaten vorliegen. Er wird hier nicht untersucht. Obwohl auch die §§ 126, 241 StGB interessant wären, ist die Volksverhetzung (§ 130 StGB) die sich in den letzten Jahren immer mehr zu einem allgemeinen Diskriminierungstatbestand entwickelt hat (Stegbauer, 2021), öfter einschlägig und wird daher hier näher untersucht.

3.2.3 Volksverhetzung (§ 130 StGB) – Annotationsansatz

Aus dem Wortlaut der Volksverhetzung, s. Abb. 6, wird ein Annotationsansatz entwickelt: Grundsätzlich wird zwischen der allgemeinen Volksverhetzung i.S.d. Abs. 1-2 und der auf die nationalsozialistischen Gräueltaten bezogenen Volksverhetzung, Abs. 3-4 unterschieden. Für die allgemeine Volksverhetzung wird das Schema in Abb. 5 aufgestellt. Die Annotation ist in drei Ebenen gegliedert. Zuerst wird annotiert, ob in der vorliegenden Aussage eine Gruppe oder ein Mitglied dieser Gruppe erwähnt wird, deren Zugehörigkeit über eines der Merkmale Nationalität, "Rasse/ethnische Herkunft, Religion, politische Einstellung, Geschlecht, anderes Merkmal definiert wird. Zusätzlich wurde annotiert, um welche Gruppe es genau geht[40]. Zweitens wird annotiert, ob diese Gruppe durch eine der möglichen Tathandlungen, d.h. eine Aufstachelung zu Hass, eine Aufforderung zu Gewalt- oder Willkürmaßnahmen oder einen Angriff der Menschenwürde angegriffen wird. Werden sowohl eine Gruppe als auch eine Tathandlung gefunden, liegt ein Fall der allgemeinen Volksverhetzung vor, der als VVH-ALLG annotiert wird. Die zwei Ebenen Tathandlung und Gruppenmerkmal werden durchgehend annotiert, auch wenn eine Aussage nicht volksverhetzend ist, also die jeweils andere Ebene fehlt. Die Aufteilung in Ja/Nein-Fragen orientiert sich an Palmer et al. (2020) und Zufall et al. (2020). Die notwendige Eignung zur Störung des öffentlichen Friedens gemäß § 130 Abs.1 Satz 1 StGB gilt durch die Veröffentlichung von Aussagen im Internet als gegeben (Schäfer & Anstötz, 2021, § 130, Rn.22-26).

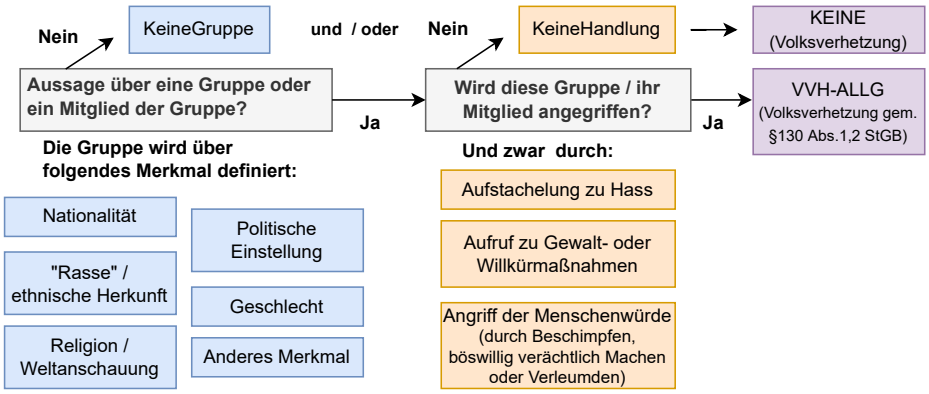

Abbildung 5: Annotationsschema (Annotationslabels farbig hinterlegt)

[40]Diese Gruppenlabels wurden im Laufe der Annotation festgelegt; ein Überblick findet sich in dem Annotationsbeispiel in Abb. 8 (Alle Labels von Die Grünen bis Syrer:innen).

(1) Wer in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören,

1. gegen eine nationale, rassische, religiöse oder durch ihre ethnische Herkunft bestimmte Gruppe, gegen Teile der Bevölkerung oder gegen einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung zum Hass aufstachelt, zu Gewalt- oder Willkürmaßnahmen auffordert oder
2. die Menschenwürde anderer dadurch angreift, dass er eine vorbezeichnete Gruppe, Teile der Bevölkerung oder einen Einzelnen wegen seiner Zugehörigkeit zu einer vorbezeichneten Gruppe oder zu einem Teil der Bevölkerung beschimpft, böswillig verächtlich macht oder verleumdet,

wird mit Freiheitsstrafe von drei Monaten bis zu fünf Jahren bestraft.

(2) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft, wer

1. einen Inhalt (§ 11 Absatz 3) verbreitet oder der Öffentlichkeit zugänglich macht oder einer Person unter achtzehn Jahren einen Inhalt (§ 11 Absatz 3) anbietet, überlässt oder zugänglich macht, der
   a) zum Hass gegen eine in Absatz 1 Nummer 1 bezeichnete Gruppe, gegen Teile der Bevölkerung oder gegen einen Einzelnen wegen seiner Zugehörigkeit zu einer in Absatz 1 Nummer 1 bezeichneten Gruppe oder zu einem Teil der Bevölkerung aufstachelt,
   b) zu Gewalt- oder Willkürmaßnahmen gegen in Buchstabe a genannte Personen oder Personenmehrheiten auffordert oder
   c) die Menschenwürde von in Buchstabe a genannten Personen oder Personenmehrheiten dadurch angreift, dass diese beschimpft, böswillig verächtlich gemacht oder verleumdet werden oder
2. einen in Nummer 1 Buchstabe a bis c bezeichneten Inhalt (§ 11 Absatz 3) herstellt, bezieht, liefert, vorrätig hält, anbietet, bewirbt oder es unternimmt, diesen ein- oder auszuführen, um ihn im Sinne der Nummer 1 zu verwenden oder einer anderen Person eine solche Verwendung zu ermöglichen.

(3) Mit Freiheitsstrafe bis zu fünf Jahren oder mit Geldstrafe wird bestraft, wer eine unter der Herrschaft des Nationalsozialismus begangene Handlung der in § 6 Abs. 1 des Völkerstrafgesetzbuches bezeichneten Art in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören, öffentlich oder in einer Versammlung billigt, leugnet oder verharmlost.

(4) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft, wer öffentlich oder in einer Versammlung den öffentlichen Frieden in einer die Würde der Opfer verletzenden Weise dadurch stört, dass er die nationalsozialistische Gewalt- und Willkürherrschaft billigt, verherrlicht oder rechtfertigt.

(5) Absatz 2 gilt auch für einen in den Absätzen 3 oder 4 bezeichneten Inhalt (§ 11 Absatz 3).

(6) In den Fällen des Absatzes 2 Nummer 1, auch in Verbindung mit Absatz 5, ist der Versuch strafbar.

(7) In den Fällen des Absatzes 2, auch in Verbindung mit den Absätzen 5 und 6, sowie in den Fällen der Absätze 3 und 4 gilt § 86 Absatz 4 entsprechend.

Abbildung 6: Volksverhetzung (§ 130 StGB) – Wortlaut[41]

Die benannte Gruppe muss entweder durch eines der in Abs. 1 gelisteten Merkmale definiert oder ein klar abgrenzbarer Teil der Bevölkerung sein (Schäfer & Anstötz, 2021, § 130, Rn. 27-35; Fischer, 2022, § 130, Rn. 4-6a). Institutionen wie z.B. die katholische Kirche gehören nicht dazu, erfasst sind

[41] Letzte Änderung 14.09.2021 (BGBl. I S. 4250), in Kraft getreten am 22.09.2021; https://www.gesetze-im-internet.de/stgb/__130.html.

z.B. „Migranten“, „in Deutschland lebende Ausländer“ oder auch „Polizisten“[42] (Schäfer & Anstötz, 2021, § 130, Rn. 34). Erfasst sind auch Merkmale wie Geschlecht[43] und sexuelle Orientierung[44]. „Invasoren Abschaum“ ist nicht klar abgrenzbar; „Die Linke“ ist es nur dann, wenn klar ist, dass es spezifisch um die Partei geht. Annotiert wird eine Gruppe nur dann, wenn sie angesprochen oder explizit eine Aussage über sie getroffen wird. Gruppen, die erwähnt werden, ohne im Zentrum der Aussage zu stehen, werden nicht annotiert; z.B. wird in „`@user Eher supremazistisch verschlagen. Wie sie fast alle. Europäischstämmige und Christen hassend. Gelebter Talmud`“ keine Aussage über Christen getroffen – die Gruppe, um die es eigentlich geht (vermutlich Juden) wird nicht explizit erwähnt und daher auch nicht annotiert. Die sich aus dieser Annotation ergebende Verteilung von Gruppenmerkmalen in den Tweets ist in Abb. 7 zu sehen. Besonders häufig vertreten sind Politische Einstellung, Anderes Merkmal, wozu hier u.a. Flüchtlinge gezählt werden, und Religion.

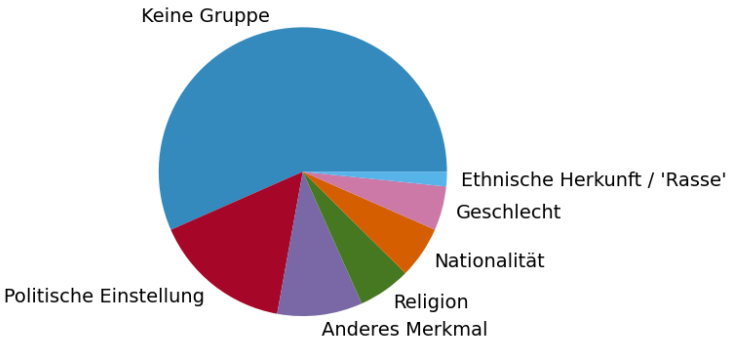

Abbildung 7: Verteilung der Gruppenmerkmale

Nun zu den Tathandlungen: Zu Hass stachelt auf, wer eine “über die bloße Ablehnung oder Verachtung hinausgehende, feindselige Haltung„[45] vertritt, während die Aussage „durch ein besonderes Maß an Gehässigkeit und Rohheit oder eine besonders gehässige Ausdrucksweise geprägt“[46] ist, so z.B. „`Zuwanderer? Was für eine Wortwahl, schon darin steckt Lügendefinition. Es sind Eindringlinge mit Unkulturhintergründen; passen nicht hier hin und müssen hinausgeleitet werden.`“ [sic] Tatsachenbehauptungen reichen nicht aus (Fischer, 2022, § 130 Rn. 8).

Das Auffordern zu Gewalt- oder Willkürmaßnahmen „setzt ein über bloßes Befürworten hinausgehendes, ausdrückliches oder konkludentes Einwirken auf andere voraus mit dem Ziel, in ihnen den Entschluss zu bestimmten Handlungen hervorzurufen.“[47] Willkürmaßnahmen können z.B. gewaltsame Abschiebungen wie in „`Macht die grenzen zu! Und schiebt alle Asylanten ab! Dann ist unser Deutschland sicher!!!`“ [sic] sein, oder auch der Ausschluss vom Ämtern (Fischer, 2022, § 130 Rn. 10). In vielen Tweets wird Gewalt zwar

[42] NStZ-RR 12, 77, Frankfurt NStZ-RR 00, 368.
[43] OLG Köln Urt. v. 9.6.2020 – III-1 RVs 77/20.
[44] OLG Köln Urt. v. 9.6.2020 – III-1 RVs 77/20.
[45] BGH 40, 102.
[46] BGH 3.4.2008 – 3 StR 394/07, BeckRS 2008, 06865 Rn. 18.
[47] BGH 3.4.2008 – 3 StR 394/07, BGHR StGB § 130 Nr. 1 Auffordern.

befürwortet, aber nicht zu ihr aufgefordert. Die Menschenwürde kann angegriffen werden, indem jemand beschimpft, böswillig verächtlich gemacht oder verleumdet wird – allerdings nur dann, wenn jemand dadurch im Kern seiner Existenz angegriffen bzw. verobjektiviert wird, z.B. durch die Bezeichnung als „Untermensch“ oder „krimineller Eindringling“[48].

Separat von der allgemeinen Volksverhetzung wird die NS-bezogene Volksverhetzung des § 130 Abs. 3-5 StGB annotiert: Werden die nationalsozialistische Gewaltherrschaft oder ihre Taten gebilligt, gerechtfertigt, geleugnet, verherrlicht oder verharmlost, folgt die Annotation VVH-NS sowie die genaue Angabe der eben genannten Tathandlungen.

3.2.4 Praktische Anmerkungen

Für die Annotation wurde mit dem Textannotationswerkzeug `doccano` (Nakayama et al., 2018) gearbeitet. Es wurden verschiedene Annotationswerkzeuge, unter anderem INCEpTION[49] (der Nachfolger von WebAnno), Prodigy[50], brat[51] und doccano[52] verglichen. Für die vorliegende Annotation war ein leicht zugängliches Werkzeug nötig, das die Metainformationen der eingespeisten Daten weiter mitführt, eine große Anzahl verschiedener Tags erlaubt und wo im Laufe der Annotation neue Tags hinzugefügt werden können. doccano erfüllt von den Genannten diese Bedingungen am Besten:

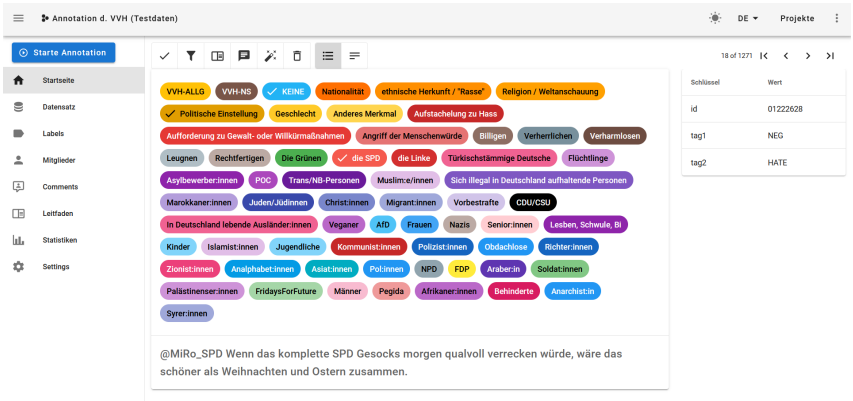

Abbildung 8: Annotation mit doccano

Nach abgeschlossener Annotation wurde, da dies nicht von `doccano` erzwungen wird, geprüft, ob alle Tags der Annotationslogik korrekt entsprechen – ob also nichts gleichzeitig als VVH-ALLG und KEINE annotiert wurde, dass für jedes spezifische Gruppenlabel ein entsprechendes Allgemeines existiert usw. Flüchtigkeitsfehler konnten so gefunden und korrigiert werden. Aus praktischen Gründen wurden die Daten von der Autorin als einziger Annotatorin annotiert, allerdings wurde zu einigen beispielhaften Zweifelsfällen Rücksprache mit einem Juristen gehalten. Die Annotatorin ist deutsche Muttersprachlerin, Studentin, hat linguistische Vorkenntnisse und ist juristischer Laie.

[48] OLG Rostock, Beschluss vom 25.05.2021 - 2 U 8/19.
[49] s. https://inception-project.github.io/.
[50] s. https://prodi.gy/docs/text-classification.
[51] s. https://brat.nlplab.org/index.html.
[52] s. https://github.com/doccano/doccano.

4 Eine automatisierte Prüfung der Strafbarkeit

4.1 Methodik

Auf Grundlage dieses annotierten Datensatzes soll ein automatisiertes Verfahren trainiert werden, das für eine beliebige Aussage einschätzt, ob sie volksverhetzend ist. Als erster Ansatz wurde ein BERT-Sprachmodell auf eben dieser binären Unterscheidung nachtrainiert. Erste Tests damit waren jedoch wenig erfolgreich. Der zweite Ansatz bestand darin, stattdessen die Tatbestandsmerkmale der Volksverhetzung zu erkennen und auf dieser Basis eine Einschätzung zu treffen. Es gibt daher zwei Teilprobleme: 1. zu erkennen, ob eine Gruppe oder ein Mitglied einer Gruppe benannt wird und 2. ob eine Tathandlung wie z.B. ein Angriff der Menschenwürde vorliegt. Kann sowohl 1. als auch 2. bejaht werden, muss folglich eine volksverhetzende Aussage vorliegen.

All diese Merkmale wurden annotiert. Ob eine Gruppe angesprochen ist oder nicht, leitet sich aus den Teilannotationen über die Erwähnung spezifischer Gruppen ab; genauso verhält es sich mit den Tathandlungen. Jedoch können in einem Tweet mehrere Gruppen auf einmal vorkommen und auch mehr als eine Tathandlung vorliegen, daher handelt es sich hierbei nicht mehr um ein binäres Klassifikationsproblem, sondern um eine Mehrlabelklassifikation. Nicht alle Klassifizierer sind für solche Probleme geeignet.

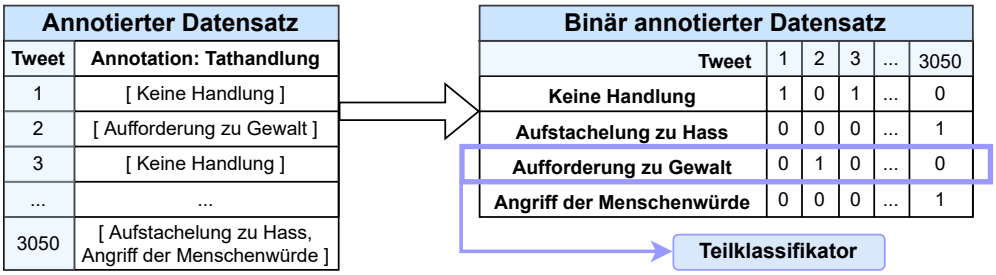

Abbildung 9: Binarisierung der Annotation

Das Problem kann durch die sog. Methode binärer Relevanz (Read et al., 2009) binarisiert werden: Wie in Abb. 9 zu sehen ist, wird die komplexe Annotation aufgespalten, indem für jede Klasse annotiert wird, ob sie vorhanden ist, 1, oder nicht, 0. Für jede Klasse wird so ein eigener binärer Klassifikator trainiert, in Abb. 9 beispielhaft für Aufforderung zu Gewalt. Es wird davon ausgegangen, dass die einzelnen Klassen voneinander unabhängig sind. Da auch KeineGruppe bzw. KeineHandlung zu den Klassen gehören, klassifziert eines der so entstandenen Modelle, ob überhaupt eine Gruppe oder Handlung vorhanden ist.

Für die Klassifikation werden zwei verschiedene Methoden genutzt und miteinander verglichen: Zum einen wird das Verfahren der Logistischen Regression genutzt, Abschnitt 4.2, zum anderen das Transfer Learning basierend auf einem vortrainierten deutschsprachigen Modell, Abschnitt 4.3. Das Training von BERT wird gewählt, weil damit schon anhand sehr weniger Trainingsdaten Probleme wie Textklassifikation gut gelöst werden können. Die logistische Regression

dient als Vergleichsgrundlage dafür: Sie ist einfach und schnell zu trainieren und kann schon bei einer geringen Menge an Trainingsdaten gute Ergebnisse liefern. Damit kann also untersucht werden, ob auch schon mit einem geringeren Rechenaufwand vergleichbare Ergebnisse erzielt werden können.

Wie in Abb. 7 zu sehen war, ist die Verteilung der Klassen im vollständigen Datensatz nicht ausbalanciert, weshalb diese Verteilung auch in den Trainings- und Testdaten bewahrt werden sollte. Daher werden die Daten stratifzierend in 70% Trainings- und 30% Testdaten aufgeteilt. Dafür mit der vollständigen, komplexen Annotation zu arbeiten ist schwierig, da jede Kombination der verschiedenen Gruppen bzw. Handlungen als eigene Klasse behandelt wird und sehr viele dieser kombinierten Klassen entstehen. Stattdessen wird eine Methode der Bibliothek `scikit-multilearn`[53] (Szymański & Kajdanowicz, 2017) genutzt, die mithilfe eines iterativen Stratifizierungsverfahrens dafür sorgt, dass alle Klassen ausbalanciert werden. Möglicherweise könnten noch bessere stratifizierte Aufteilungen gefunden werden, indem nur auf die jeweils relevante Teilmenge der Annotationslabels geachtet würde. Da jedoch zuletzt wieder alle Klassifikatoren kombiniert werden, um die Einschätzung Volksverhetzung – Ja/Nein? zu treffen, müssen die Trainings- und Testdaten für alle Klassifikatoren die selben sein.

Nach der Aufspaltung in Trainings- und Testdaten werden die Tweets einigen Vorverarbeitungsschritten unterzogen: Da nur wenige Trainingsdaten vorhanden sind, soll das Rauschen in den Daten reduziert und die Anzahl möglicher Features dadurch praktikabel gehalten werden. Deshalb werden die Tweets in Kleingeschriebenes umgewandelt und dann mithilfe des vortrainierten BERT-Tokenisierers bert-base-german-cased[54] tokenisiert. Mit diesem werden die Daten für das Training des BERT-Sprachmodells tokenisiert, und auch die Daten für das Feintuning müssen damit tokenisiert werden. Um die Vorverarbeitung vergleichbar zu halten, wird der BERT-Tokenisierer auch in der Vorverarbeitung für die logistische Regression verwendet: Der sog. WordPiece-Algorithmus (Schuster & Nakajima, 2012) unterteilt den Text, wie im Beispiel in Tabelle 4 zu sehen ist, in sinntragende Teile von Wörtern. Ausgehend von Tokens auf Zeichenebene werden iterativ je zwei Tokens dann zusammengefügt, wenn dieses neue Token in den Trainingsdaten des Tokenisierers mit höherer Wahrscheinlichkeit auftaucht als die beiden einzelnen. Dieser Prozess wird wiederholt, bis eine vorgegebene Vokabulargröße erreicht ist oder keine wahrscheinlicheren Kombinationen mehr gefunden werden. Auch unbekannte Wörter können so in sinnvolle Tokens unterteilt werden. Für die Weiterverarbeitung im Transfer Learning werden zusätzlich zur Tokenisierung alle Tweets mit Hilfe eines Padding-Tokens auf eine einheitliche Länge gebracht.

	@user			die	ganze	bande		muss	weg	!	
[CLS]	@	use	##r	die	ganze	ba	##nde	muss	weg	!	[SEP]

Tabelle 4: Beispiel für den Word-Piece-Tokenisierer

[53] http://scikit.ml/index.html.
[54] https://huggingface.co/bert-base-german-cased.

4.2 Klassifikation mit Logistischer Regression

Als erste Klassifikationsmethode wird das Verfahren der logistischen Regression genutzt. Für diese Klassifikation und für die vorangehende Vorverarbeitung und die Erstellung numerischer Features wurde mit der Bibliothek `sklearn` (Pedregosa et al., 2011) gearbeitet.

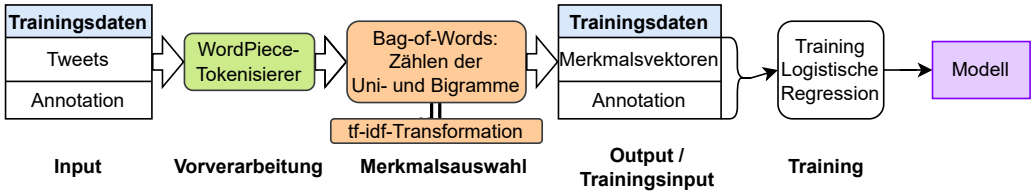

Abbildung 10: Klassifikation mittels logistischer Regression

Als Features werden für die Klassifizierung mittels logistischer Regression die in Abschnitt 2.2 besprochenen Bag-of-Words verwendet. Durch Versuche konnte gezeigt werden, dass zusätzlich zu Wortunigrammen die Verwendung von Wortbigrammen die Klassifikationsergebnisse verbessert. Zum Beispiel steigt der Macro-MCC für die Klassifikation der Gruppen von 0.46 auf 0.55. Darüber hinaus, also z.B. mit Trigrammen, wurden keine weiteren Verbesserungen erzielt. Da ein WordPiece-Tokenisierer genutzt wird, werden genaugenommen Teilwortuni- und bigramme gezählt. Es wurde zwar auch mit Zeichen-N-Grammen experimentiert, aber um die Vergleichbarkeit der beiden Methoden zu verbessern, wird letztlich mit (Teil)Wort-N-Grammen gearbeitet.

Zur Anpassung der Gewichtung der N-Gramme wird `term-frequency times inverse document-frequency (tf-idf)` (Jurafsky & Martin, 2009, S. 785) genutzt. Zusätzlich wird die Anzahl an Features, also an Uni- und Bigrammen im Korpus, von insgesamt 47249 auf die 3000 wichtigsten gestutzt, was im Experimentieren mit dem Verfahren bessere Ergebnisse geliefert hat und gleichzeitig die Trainingszeit verringert (Macro-MCC der Gruppen-Klassifikation bei 0.55 vs ohne Begrenzung bei 0.51). Außerdem werden die Daten skaliert: Die Werte jedes N-Gramms werden durch den maximalen Wert dieses N-Gramms im Datensatz geteilt, so dass die Werte nie über 1 liegen[55].

Für jede binäre Klassenunterscheidung wird ein eigenes Modell für die logistische Regression trainiert. Die Laufzeit für das Training aller Klassifikatoren beträgt insgesamt nur zwischen 2.6 und 8.1 Sekunden, je nach den detaillierten Einstellungen bei der Merkmalsauswahl.

In Tabelle 5 ist die Konfusionsmatrix der binären Klassifikation Gruppe zu sehen (eigentlich wird die Klasse KeineGruppe klassifiziert, um des besseren Verständnis willen wurden allerdings die Ergebnisse hier entsprechend umgekehrt). 712 Tweets werden korrekt klassifiziert, aber es gibt trotzdem recht viele falsch Positive, 119, und falsch Negative, 85. In Tabelle 6 finden sich die Konfusionsmatrizen der restlichen Gruppen. Für alle Gruppen außer Herkunft

[55] S. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html.

können viele Tweets korrekt als diese Gruppe klassifiziert werden; gleichzeitig gibt es in allen Fällen eine oft vergleichbare Anzahl an Tweets, die inkorrekt als falsch negativ klassifiziert werden. Falsch Positive sind, außer bei Politik, selten.

	Gruppe?	
	Korrekt	
Vorhergesagt	Nein	Ja
Nein	454	119
Ja	85	258

Tabelle 5: Gruppe binär

National.		**Herkunft**		**Religion**		**Politik**		**Geschlecht**		**Anderes**	
858	37	906	10	857	23	745	77	862	34	821	34
2	19	0	0	3	33	24	70	6	14	4	57

Tabelle 6: Einzelne Gruppenklassen

In Tabelle 7 ist das Ergebnis der binären Klassifikation der Tathandlungen zu sehen. In Tabelle 8 folgen die Ergebnisse für die Tathandlungen im Einzelnen, auch hier in Form der Konfusionsmatrizen. Die Tweets in allen Klassizierern werden konsequent jeweils der Kategorie Keine Handlung zugeordnet, es gibt also nur falsch Negative, aber keine falsch Positiven.

	Tathandlung?	
	Korrekt	
Vorhergesagt	Nein	Ja
Nein	888	28
Ja	0	0

Tabelle 7: Tathandlung binär

Hass		**Gewalt**		**Menschenwürde**	
913	3	901	15	906	10
0	0	0	0	0	0

Tabelle 8: Einzelne Tathandlungsklassen

4.3 Klassifikation mit Transfer Learning auf Basis eines Sprachmodells

Als zweites Klassifikationsverfahren wird Transfer Learning auf Grundlage von Sprachmodellen genutzt. Es wird verwendet, da es das State-of-the-Art-Verfahren für diverse Probleme im Bereich NLP ist[56]. Das entsprechend vortrainierte Sprachmodell, mit dem hier gearbeitet wird, ist `German BERT`[57] (Chan et al., 2020). Für dieses Verfahren wird mit der Huggingface Transformers-Bibliothek[58] (Wolf et al., 2020) gearbeitet. Das BERT-Modell wird geladen

[56]S. z.B. https://gluebenchmark.com/leaderboard.
[57]S. https://huggingface.co/bert-base-german-cased.
[58]S. https://github.com/huggingface/transformers.

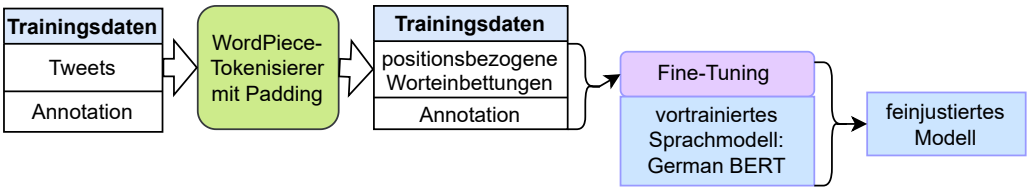

Abbildung 11: Klassifikation mittels Transfer Learning und BERT

und dann mit den annotierten Trainingsdaten feingetunt. Als Repräsentationen der Tweets werden dazu Word Embeddings genutzt, s. Abb. 11. Für das Feintuning müssen alle Eingaben die gleich Länge haben, daher werden im Zuge der Tokenisierung alle Eingaben auf eine maximale Länge gekürzt bzw. mit dem Token `[PAD]` auf die korrekte Länge verlängert. Das besondere an den Word Embeddings für BERT ist, dass sie Informationen über die Position der einzelnen Wörter im Text beinhalten. Die Hyperparameter für das Training sind wie folgt: Es wird eine Lernrate von 2e-5 verwendet, Batches der Größe 16 geformt, und für 10 Epochen trainiert. Die Laufzeit pro binärer Klasse beträgt bei 10 Epochen ca. 4 Minuten bei der Verwendung einer GPU.

Die Ergebnisse für die binäre Klassifikation der Gruppen ist in Tabelle 9 zu sehen. 775 Tweets werden korrekt vorhergesagt, aber auch 141 inkorrekt mit ähnlich vielen falsch Positiven wie falsch Negativen. Die Klassifikation der einzelnen Gruppen, s. Tabelle 10, ist in vielen Fällen erfolgreich. Wie schon bei der logistischen Regression gibt es mehr falsch Negative als falsch Positive.

	Gruppe?	
	Korrekt	
Vorhergesagt	Nein	Ja
Nein	463	76
Ja	65	312

Tabelle 9: Gruppe binär mit BERT

National.		**Herkunft**		**Religion**		**Politik**		**Geschlecht**		**Anderes**	
852	18	906	5	849	8	728	43	857	23	810	24
8	38	0	5	11	48	41	104	11	25	15	67

Tabelle 10: Einzelne Gruppenklassen mit BERT

Die Ergebnisse der Klassifikation der Tathandlungen sind in Tabellen 11 und 12 zu sehen. Im Gegensatz zur logistischen Regression werden immerhin 5 Tweets korrekt als Tathandlung zugeordnet. Auch hier überwiegen die falsch Negativen gegenüber den falsch Positiven.

	Tathandlung?	
	Korrekt	
Vorhergesagt	Nein	Ja
Nein	887	23
Ja	1	5

Tabelle 11: Tathandlung binär mit BERT

Hass		**Gewalt**		**Menschenwürde**	
913	3	897	13	906	9
0	0	4	2	0	1

Tabelle 12: Einzelne Tathandlungsklassen mit BERT

4.4 Zusammenführen der Teilklassifikatoren

Zuletzt gilt es, zurück zur Ausgangsfrage zu kommen: Ist die vorliegende Aussage potentiell volksverhetzend? Um diese Frage zu beantworten, werden die beiden Teilklassifikatoren wie bereits im Schema in Abb. 5 angelegt über einen Entscheidungsbaum kombiniert: Wird sowohl eine Gruppe als auch eine Tathandlung in der Aussage gefunden, gilt sie als volksverhetzend. Für jedes der sechs eine Gruppe definierenden Merkmale und für die drei möglichen Tathandlungen wurden einzelne Klassifikatoren trainiert, die vorhersagen, ob dieses Merkmal in einem Tweet vorhanden ist. Die Testdaten werden mit jedem dieser Klassifikatoren einzeln klassifiziert. Aus den gesammelten Ergebnissen wird dann ermittelt, ob die beiden Merkmale vorliegen. Als Klassifikatoren werden für diesen Entscheidungsbaum die feingetunten BERT-Modelle genutzt, da diese die besten Ergebnisse lieferten. Aus dieser Klassifikation ergibt sich die Konfusionsmatrix in Tabelle 13. Von den 23 Fällen von Volksverhetzung in den Testdaten werden nur 5 korrekt vorhergesagt, die restlichen 18 werden inkorrekt als nicht volksverhetzend klassifiziert. Die Mehrheit der Tweets wird korrekt als nicht volksverhetzend klassifiziert, zwei allerdings auch als falsch positiv.

	Volksverhetzung?	
	Korrekt	
Vorhergesagt	Nein	Ja
Nein	891	18
Ja	2	5

Tabelle 13: Klassifikation der Volksverhetzung mittels Entscheidungsbaum

4.5 Evaluation

Zur Evaluation der Ergebnisse werden hier für alle Teilprobleme vergleichend diverse Metriken für die Klassifikation mittels logistischer Regression und mit BERT gezeigt: Es werden jeweils `Precision` (die Genauigkeit), `Recall` (die Trefferquote) und das `F1-Maß` angegeben. Die Angaben beziehen sich immer auf die positive Ausprägung der binären Probleme (also z.B. auf die Klasse Nationalität und nicht auf die Klasse Keine Nationalität). Es werden nur die Metriken für die positive Ausprägung der Klassen angegeben, da zum einen die negativen Ausprägungen jeweils einen deutlich größeren Teil der Daten ausmachen und entsprechend mit Werten des F1-Maßes um 0.90 sehr viel besser vorhergesagt werden, und zum anderen, weil für die vorliegende Problematik die positiven Ausprägungen die eigentlich relevanten sind.

Zusätzlich wird der `Matthews-Korrelations-Koeffizient (MCC)` angegeben. Er beschreibt die Beziehung der korrekten und vorhergesagten Klassifikationen zueinander und kann direkt aus der Konfusionsmatrix berechnet werden (Matthews, 1975). Er eignet sich sehr gut für binäre Klassifikationsprobleme mit ungleicher Klassenverteilung (Boughorbel et al., 2017). Die Wertespanne liegt anders als bei den anderen angegebenen Metriken nicht zwischen 0 und 1, sondern zwischen -1 und 1, wobei 1 eine perfekte Klassifikation ist, 0 eine Leistung auf Zufallsniveau und -1 eine negative Beziehung zwischen Vorhersage und korrekter Klassifikation zeigt.

Die Ergebnisse der Klassifikation des Merkmals Gruppe und der einzelnen Gruppenkategorien finden sich in den Tabellen 14 und 15. Sowohl für die einzelnen Klassen als auch für die binäre Klassifikation werden mit BERT bessere Ergebnisse erzielt als mit der logistischen Regression. Bei einigen Klassen erreicht die logistische Regression sehr hohe Precision-Werte, aber im Vergleich dazu niedrigere Recall-Werte, so z.B. bei Geschlecht und Nationalität. Das Gruppenmerkmal ethnische Herkunft / 'Rasse' kann mittels logistischer Regression nicht erkannt werden – dies ist allerdings das am geringsten vertretene Merkmal, das in den Testdaten nur zwölfmal vorkommt. Für alle anderen Gruppen demonstrieren die MCC-Werte zwischen 0.5 und 0.75 auch mit der logistischen Regression eine positive Korrelation zwischen der Vorhersage und den korrekten Werten.

Alle Gruppen können mit BERT – laut den Werten des F1-Maßes zwischen 0.60 (Geschlecht) und 0.83 (Religion) – gut erkannt werden; der MCC liegt bei BERT zwischen 0.58 und 0.83. Auch die binäre Klassifikation ist erfolgreich. In den meisten Fällen liegen die Werte für die Precision höher als die des Recall, insbesondere bei der logistischen Regression. Da sowohl die logistische Regression und BERT als Klassifikation letztlich Wahrscheinlichkeiten ausgeben, wäre durch eine Veränderung des Schwellenwertes zwischen den Klassen eine Verbesserung des Recall möglich.

	Gruppe binär			
	Prec.	**Recall**	**F1**	**MCC**
LogReg	0.75	0.68	0.72	0.54
BERT	0.8	0.83	0.82	0.68

Tabelle 14: Klassifikation der Gruppen binär im Vergleich

	Nationalität				ethn. Herkunft / 'Rasse'			
	Prec.	**Recall**	**F1**	**MCC**	**Prec.**	**Recall**	**F1**	**MCC**
LogReg	0.90	0.34	0.49	0.54	0	0	0	0
BERT	0.83	0.68	0.75	0.74	1	0.5	0.67	0.70
	Religion				**Politische Einstellung**			
	Prec.	**Recall**	**F1**	**MCC**	**Prec.**	**Recall**	**F1**	**MCC**
LogReg	0.92	0.59	0.72	0.72	0.74	0.48	0.58	0.54
BERT	0.81	0.86	0.83	0.83	0.72	0.71	0.71	0.66
	Geschlecht				**Anderes Merkmal**			
	Prec.	**Recall**	**F1**	**MCC**	**Prec.**	**Recall**	**F1**	**MCC**
LogReg	0.70	0.29	0.41	0.44	0.93	0.63	0.75	0.75
BERT	0.69	0.52	0.60	0.58	0.82	0.74	0.77	0.75

Tabelle 15: Klassifikation der einzelnen Gruppen im Vergleich

Die Ergebnisse der Klassifikation des zweiten Merkmals, der möglichen Tathandlungen, sind in Tabelle 16 abgebildet. Die binäre Klassifikation kann nur einen F1-Wert von 0.29 erreichen. Der MCC von 0.38 zeigt, dass aber zumindest Vorhersagen getroffen werden, die besser als zufällige sind. Bei der Klassifikation der einzelnen Merkmale zeigt sich, dass das Merkmal Aufstachelung zu Hass überhaupt nicht klassifiziert werden kann (F1-Maß und MCC sind 0). Das Merkmal Aufforderung zu Gewalt- und Willkürmaßnahmen erreicht einen F1-Wert von 0.19 und einen MCC von 0.2, der Angriff der Menschenwürde einen F1-Wert von 0.18 und einen MCC von 0.32. Die Klassifikation gelingt also nicht zuverlässig. Dies hängt wahrscheinlich damit zusammen, dass die Menge an Beispielen in dem Datensatz sehr gering ist. Zusätzlich ist bereits während der Annotation aufgefallen, dass die Abgrenzung zwischen Aussagen, die lediglich beleidigend sind, und solchen, die tatsächlich offen zu Gewalt oder Hass auffordern, oft fließend ist. Zur Unterscheidung muss abgewogen werden, ob eine Aussage Gewalt nicht nur befürwortet, sondern andere direkt dazu auffordert. Denn nur dann gilt sie als volksverhetzend.

	Handlung binär			
	Prec.	**Recall**	**F1**	**MCC**
LogReg	0	0	0	0
BERT	0.83	0.18	0.29	0.38

Tabelle 16: Klassifikation der Tathandlung binär im Vergleich

Zuletzt ist der Vergleich zwischen der direkten Klassifikation der Volksverhetzung und der mittels Kombination der BERT-Modelle zum Entscheidungsbaum in Tabelle 17 zu sehen. Die Klassifikation mittels Entscheidungsbaum ist mit einem F1 von 0.3 mit der direkten Klassifikation mit 0.33 vergleichbar. Dafür könnte die Schwierigkeit bei der Klassifikation der Tathandlungen ausschlaggebend sein. Wie auch bei der Klassifikation der Gruppen ist die Precision deutlich höher als der Recall. Zufall et al. (2020) erreichen für die direkte Klassifikation der Volksverhetzung (mit etwas anderen Merkmalen, gemäß dem europäischen Recht) einen F1-Wert von 0.39, für die Klassifikation mittels Entscheidungsbaum einen F1-Wert von 0.42 und erzielen damit etwas bessere Ergebnisse. Allerdings arbeiten Zufall et al. (2020) mit einem um selbst erstellte Beispiele der Volksverhetzung erweiterten Datensatz, und verfügen daher vor allem in Bezug auf die Tathandlungsvarianten über eine breitere Datengrundlage.

	Volksverhetzung?				
	Prec.	**Recall**	**F1**	**Macro Avg F1**	**MCC**
LogReg direkt	1	0.04	0.08	0.54	0.21
BERT direkt	1	0.17	0.3	0.64	0.41
BERT + Entscheidungsbaum	0.71	0.22	0.33	0.66	0.39

Tabelle 17: Klassifikation der Volksverhetzung im Vergleich

Es ist möglich, automatisiert zu erkennen, ob eine durch ein bestimmtes Merkmal ausgezeichnete Gruppe oder ein Mitglied dieser Gruppe in einer Aussage erwähnt wird. Dabei liefert die Klassifikation mit einem feingetunten BERT-Modell deutlich bessere Ergebnisse als die Arbeit mit logistischer Regression. Einzuschätzen, ob eine Tathandlung vorliegt oder welche genau, kann hingegen, zumindest anhand so weniger Trainingsdaten, nicht automatisiert werden. Im ganzen Korpus gibt es von insgesamt 3050 Tweets nur 94 Positivbeispiele für Tathandlungen, von denen wiederum nur ca. 2/3 in den Trainingsdaten sind. Für eine zuverlässige Klassifikation wären mehr Trainingsdaten notwendig.

5 Schlussbetrachtung

In dieser Arbeit wurde untersucht, ob und wie es möglich ist, die Prüfung von Hate Speech auf ihre potentielle Strafbarkeit zu automatisieren. Dazu wurde die Plausibilität der Annotation von Straftatbeständen diskutiert und beispielhaft für die Volksverhetzung gemäß § 130 StGB ein Annotationsschema entwickelt. Ein Referenzdatensatz wurde aus bestehenden deutschsprachigen Datensätzen zusammengestellt und ein Ausschnitt dieses Datensatzes wurde anhand des neuen Annotationsschemas annotiert. Diese Daten dienten als Grundlage für das Training zweier Klassifikationsverfahren: Der logistischen Regression und dem Transfer Learning mit BERT. Anstatt Aussagen direkt als volksverhetzend zu klassifizieren, wurden separat zwei zentrale Tatbestandsmerkmale klassifiziert, die zusammen ein starkes Indiz dafür sind: Gruppe und Tathandlung.

Die durch verschiedene Merkmale definierten Gruppen konnten erfolgreich klassifiziert werden. Durch einen Mangel an Beispielen für die Tathandlungen konnten diese automatisiert jedoch nur ungenügend erkannt werden. Die Kombination der Klassifikatoren der beiden Merkmale zur Klassifikation der Volksverhetzung als Ganzes ist präzise, findet aber nur wenige korrekte Beispiele. Eine automatisierte Prüfung von Hate Speech auf ihre Strafbarkeit unter dem Tatbestand der Volksverhetzung ist also grundsätzlich möglich, für bessere Ergebnisse wären jedoch noch mehr annotierte Daten notwendig. Zwar konnten durch die Klassifikation mittels Entscheidungsbaum die Ergebnisse nur geringfügig verbessert werden, aber dieser Ansatz liefert Einsichten zu der Komplexität der Klassifikation.

Zufall et al. (2020) arbeiten zur Klassifikation Volksverhetzung laut europäischer Rechtslage und verwenden zur Aufstockung der Daten auch selbst verfasste Tweets. Auch sie spalten das Problem in die Klassifikation der einzelnen Tatbestandsmerkmale auf, so wie es in diesem Straftatbestand angelegt ist. Für die Klassifikation des Merkmals Gruppe werden vergleichbare Ergebnisse erzielt: Das F1-Maß beträgt respektive in dieser Arbeit und bei Zufall et al. (2020) für die binäre Unterscheidung 0.81 und 0.83, und für den Macro-Durchschnitt der einzelnen Gruppen in beiden Fällen 0.75. Bei dem Merkmal Handlung unterscheiden Zufall et al. (2020) nur zwischen der Aufstachelung zu Hass und der Aufforderung zu Gewalt. Genau wie in dieser Arbeit gelingt die Klassifikation des ersten Merkmals wegen eines Mangels an Daten nicht. Das zweite wird jedoch erfolgreich klassifiziert. Auch Zufall et al. (2020) kommen zu dem Ergebnis, dass die Klassifikation über die Kombination der Einzelmerkmale nur geringfügig effektiver ist als die direkte Klassifikation (F1-Maß: direkt – 0.39, über die Merkmale – 0.42). Diese Beobachtung kann in dieser Arbeit auch für die Interpretation nach deutschem Recht bestätigt werden.

Zuletzt wurde beobachtet, dass der Anteil strafbarer Aussagen innerhalb des Datensatzes trotz der Hate-Speech-Ausrichtung sehr gering ist. Daher stellt sich die Frage, ob diese Beobachtung repräsentativ für soziale Medien ist. In Zukunft könnten auch die anderen im NetzDG genannten Tatbestände untersucht werden. Die Klassifikation nach Gruppenmerkmalen kann auch für nicht auf Strafbarkeit fokussierte Arbeiten im Bereich Hate Speech hilfreich sein.

Literaturverzeichnis

Akhtar, S., Basile, V. & Patti, V. (2019). A New Measure of Polarization in the Annotation of Hate Speech. https://doi.org/10.1007/978-3-030-35166-3_41

Akiwowo, S., Vidgen, B., Prabhakaran, V. & Waseem, Z. (Hrsg.). (2020). *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics. https://aclanthology.org/2020.alw-1.0

Bassignana, E., Basile, V. & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, 51–56. https://doi.org/https://doi.org/10.4000/books.aaccademia.3085

Bhatia, A. & Shaikh, S. (Hrsg.). (2020). *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*. European Language Resources Association. https://www.aclweb.org/anthology/2020.stoc-1

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S. & Shrivastava, M. (2018). A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 36–41. https://doi.org/10.18653/v1/W18-1105

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Boughorbel, S., Jarray, F. & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, *12*. https://doi.org/10.1371/journal.pone.0177678

Bretschneider, U. & Peters, R. (2017). Detecting Offensive Statements towards Foreigners in Social Media. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2213–2222. https://doi.org/10.24251/HICSS.2017.268

Broder, A. (1997). On the resemblance and containment of documents. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, 21–29. https://doi.org/10.1109/SEQUEN.1997.666900

Chan, B., Schweter, S. & Möller, T. (2020). German's Next Language Model. *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598

Chen, Y., Zhou, Y., Zhu, S. & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Fisch, W. (2002). Hate Speech in the Constitutional Law of the United States. *The American Journal of Comparative Law*, *50*, 463–492. https://doi.org/10.1093/ajcl/50.suppl1.463

Fischer, T. (2022). StGB § 130 Volksverhetzung. *Strafgesetzbuch: StGB* (69. Aufl.). C. H. Beck oHG.

Fišer, D., Erjavec, T. & Ljubešić, N. (2017). Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene. *Proceedings of the First Workshop on Abusive Language Online*, 46–51. https://doi.org/10.18653/v1/W17-3007

Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z. & Wernimont, J. (Hrsg.). (2018). *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics. https://aclanthology.org/W18-5100

Fortuna, P. & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, *51*(4). https://doi.org/10.1145/3232676

HateAid GmbH. (2021). Grenzenloser Hass im Internet – Dramatische Lage in ganz Europa. https://hateaid.org/wp-content/uploads/2021/11/HateAid-Report-2021-DE.pdf

Jurafsky, D. & Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc.

Keen, E., Georgescu, M. & Gomes, R. (2020). Bookmarks - A manual for combating hate speech online through human rights education. https://www.coe.int/en/web/no-hate-campaign/bookmarks-connexions

Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M. & Ren, X. (2020). Contextualizing Hate Speech Classifiers with Post-hoc Explanation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5435–5442. https://doi.org/10.18653/v1/2020.acl-main.483

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4). https://doi.org/10.3390/info10040150

Kumar, R., Ojha, A. K., Lahiri, B., Zampieri, M., Malmasi, S., Murdock, V. & Kadar, D. (Hrsg.). (2020). *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.trac-1

Kumar, R., Ojha, A. K., Malmasi, S. & Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 1–11. https://aclanthology.org/W18-4401

Landesanstalt für Medien NRW. (2021). Ergebnisbericht - forsa-Befragung zu: Hate Speech 2021.

Le, Q. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In E. P. Xing & T. Jebara (Hrsg.), *Proceedings of the 31st International Conference on Machine Learning* (S. 1188–1196). PMLR. https://proceedings.mlr.press/v32/le14.html

Leskovec, J., Rajaraman, A. & Ullman, J. D. (2014). *Mining of Massive Datasets* (2. Aufl.). Cambridge University Press. https://doi.org/10.1017/CBO9781139924801

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. *Forum*

for Information Retrieval Evaluation, 29–32. https://doi.org/10.1145/3441501.3441517

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. *Proceedings of the 11th Forum for Information Retrieval Evaluation*, 14–17. https://doi.org/10.1145/3368567.3368584

Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schäfer, J., Ranasinghe, T., Zampieri, M., Nandini, D. & Jaiswal, A. K. (2021). Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. *CoRR*, *abs/2112.09301*. https://arxiv.org/abs/2112.09301

Manning, C. & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/https://doi.org/10.1016/0005-2795(75)90109-9

Mehdad, Y. & Tetreault, J. (2016). Do Characters Abuse More Than Words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 299–303. https://doi.org/10.18653/v1/W16-3638

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/ARXIV.1301.3781

Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V. & Waseem, Z. (Hrsg.). (2021). *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics. https://aclanthology.org/2021.woah-1.0

Mubarak, H., Darwish, K. & Magdy, W. (2017). Abusive Language Detection on Arabic Social Media. *Proceedings of the First Workshop on Abusive Language Online*, 52–56. https://doi.org/10.18653/v1/W17-3008

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y. & Liang, X. (2018). doccano: Text Annotation Tool for Human. https://github.com/doccano/doccano

Njagi, D., Zuping, Z., Hanyurwimfura, D. & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*, 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153. https://doi.org/10.1145/2872427.2883062

Oriol, B., Canton-Ferrer, C. & Giró-i-Nieto, X. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. In: *NeurIPS 2019 Workshop on AI for Social Good*. Vancouver, Canada, 2019, September.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y. & Yeung, D.-Y. (2019). Multilingual and Multi-Aspect Hate Speech Analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4675–4684. https://doi.org/10.18653/v1/D19-1474

Palmer, A., Carr, C., Robinson, M. & Sanders, J. (2020). COLD: Annotation scheme and evaluation data set for complex offensive language in English. *Journal for Language Technology and Computational Linguistics*, *34*(1), 1–28.

Pavlopoulos, J., Malakasiotis, P., Bakagianni, J. & Androutsopoulos, I. (2017). Improved Abusive Comment Moderation with User Embeddings. *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, 51–55. https://doi.org/10.18653/v1/W17-4209

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R. & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C. & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*, 477–523. https://doi.org/https://doi.org/10.1007/s10579-020-09502-8

Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M. S., Nakov, P. & Chakraborty, T. (2021). MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4439–4455. https://doi.org/10.18653/v1/2021.findings-emnlp.379

Rajamanickam, S., Mishra, P., Yannakoudakis, H. & Shutova, E. (2020). Joint Modelling of Emotion and Abusive Language Detection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4270–4279. https://doi.org/10.18653/v1/2020.acl-main.394

Read, J., Pfahringer, B., Holmes, G. & Frank, E. (2009). Classifier Chains for Multi-label Classification. In W. Buntine, M. Grobelnik, D. Mladenić & J. Shawe-Taylor (Hrsg.), *Machine Learning and Knowledge Discovery in Databases* (S. 254–269). Springer Berlin Heidelberg.

Risch, J., Ruff, R. & Krestel, R. (2020). Explaining Offensive Language Detection. *Journal for Language Technology and Computational Linguistics*, *34*(1).

Roberts, S. T., Tetreault, J., Prabhakaran, V. & Waseem, Z. (Hrsg.). (2019). *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics. https://aclanthology.org/W19-3500

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. *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (Bochum), Bochumer Linguistische Arbeitsberichte*, *17*, 6–9. https://doi.org/https://doi.org/10.48550/arXiv.1701.08118

Roxin, C. & Greco, L. (2020). *Strafrecht Allgemeiner Teil, Band I* (5. Aufl.). C. H. Beck oHG.

Schäfer, J. & Burtenshaw, B. (2019). Offence in Dialogues: A Corpus-Based Study. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1085–1093. https://doi.org/10.26615/978-954-452-056-4_125

Schäfer, J. & Anstötz, S. (2021). StGB § 130 Volksverhetzung. In V. Erb & J. Schäfer (Hrsg.), *Münchener Kommentar zum StGB* (4. Aufl., S. 744–794). C.H. Beck München.

Schmidt, A. & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Schuster, M. & Nakajima, K. (2012). Japanese and Korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079

Stegbauer, A. (2021). Rechtsprechungsübersicht zu den Propaganda- und Äußerungsdelikten. *NStZ*, *531*.

Struß, J. M., Wiegand, M., Siegel, M., Ruppenhofer, J. & Manfred, K. (2019). Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 352–363. https://doi.org/https://doi.org/10.5167/uzh-178687

Syropoulos, A. (2001). Mathematics of Multisets. In C. S. Calude, G. PĂun, G. Rozenberg & A. Salomaa (Hrsg.), *Multiset Processing* (S. 347–358). Springer Berlin Heidelberg.

Szymański, P. & Kajdanowicz, T. (2017). A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Turian, J., Ratinov, L.-A. & Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-Supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394. https://aclanthology.org/P10-1040

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Vidgen, B. & Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, *15*(12), 1–32. https://doi.org/10.1371/journal.pone.0243300

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M. & Tesconi, M. (2017). Hate Me, Hate Me Not: Hate Speech Detection on Facebook. *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 86–95.

Wang, C. (2018). Interpreting Neural Network Hate Speech Classifiers. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 86–92. https://doi.org/10.18653/v1/W18-5111

Waseem, Z., Davidson, T., Warmsley, D. & Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *Proceedings of the First Workshop on Abusive Language Online*, 78–84. https://doi.org/10.18653/v1/W17-3012

Waseem, Z. & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Wich, M., Räther, S. & Groh, G. (2021). German Abusive Language Dataset with Focus on COVID-19. *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 247–252. https://aclanthology.org/2021.konvens-1.26

Wiegand, M., Ruppenhofer, J. & Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 602–608. https://doi.org/10.18653/v1/N19-1060

Wiegand, M., Siegel, M. & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. *Proceedings of the GermEval 2018 Workshop - 14th Conference on Natural Language Processing (KONVENS 2018)*, 1–10. https://epub.oeaw.ac.at/0xc1aa5576_0x003a10d2.pdf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Xu, J.-M., Jun, K.-S., Zhu, X. & Bellmore, A. (2012). Learning from Bullying Traces in Social Media. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 656–666. https://aclanthology.org/N12-1084

Zufall, F., Hamacher, M., Kloppenborg, K. & Zesch, T. (2020). A Legal Approach to Hate Speech: Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task. https://doi.org/10.48550/ARXIV.2004.03422

Zufall, F., Horsmann, T. & Zesch, T. (2019). From legal to technical concept: Towards an automated classification of German political Twitter postings as criminal offenses. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1337–1347. https://doi.org/10.18653/v1/N19-1135

Abbildungsverzeichnis

1 Künast-Meme, gekennzeichnet als Falschzitat 1
2 Mögliche Annotationsebenen von Hate Speech 6
3 Verteilung der binären Annotation im Referenzdatensatz 15
4 Verteilung der detaillierten Annotation im Referenzdatensatz . . 15
5 Annotationsschema (Annotationslabels farbig hinterlegt) 17
6 Volksverhetzung (§ 130 StGB) – Wortlaut 18
7 Verteilung der Gruppenmerkmale 19
8 Annotation mit doccano 20
9 Binarisierung der Annotation 21
10 Klassifikation mittels logistischer Regression 23
11 Klassifikation mittels Transfer Learning und BERT 25

Tabellenverzeichnis

1 Übersicht der deutschsprachigen Hate-Speech-Datensätze 8
2 Vereinheitlichung der Annotation 14
3 Hate Speech Tatbestände laut NetzDG 16
4 Beispiel für den Word-Piece-Tokenisierer 22
5 Gruppe binär 24
6 Einzelne Gruppenklassen 24
7 Tathandlung binär 24
8 Einzelne Tathandlungsklassen 24
9 Gruppe binär mit BERT 25
10 Einzelne Gruppenklassen mit BERT 25
11 Tathandlung binär mit BERT 26
12 Einzelne Tathandlungsklassen mit BERT 26
13 Klassifikation der Volksverhetzung mittels Entscheidungsbaum . 26
14 Klassifikation der Gruppen binär im Vergleich 27
15 Klassifikation der einzelnen Gruppen im Vergleich 28
16 Klassifikation der Tathandlung binär im Vergleich 28
17 Klassifikation der Volksverhetzung im Vergleich 29

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne Zuhilfenahme anderer als der von mir angegebenen Quellen und Hilfsmittel angefertigt habe. Die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht. Die „Richtlinie zur Sicherung guter wissenschaftlicher Praxis für Studierende an der Universität Potsdam (Plagiatsrichtlinie) - Vom 20. Oktober 2010“, habe ich zur Kenntnis genommen.

Berlin, 9. August 2022
Celia Birle