Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bachelorarbeit

Entwicklung eines Browser-Plugins zur nutzerseitigen Filtration von Hate Speech in sozialen Netzwerken

vorgelegt von

Janik Schröder

MIN-Fakultät

Fachbereich Informatik

Language Technology Group

Studiengang: Wirtschaftsinformatik

Matrikelnummer: 6930372

Erstgutachter: Prof. Dr. Chris Biemann

Zweitgutachter: Dr-Ing. Gregor Wiedemann

Zusammenfassung

Hass und Hetze, insbesondere im Bezug auf soziale Netzwerke, stehen zunehmend im Fokus der Gesellschaft, der Politik und der Medien. Immer häufiger entflammt nach Straftaten mit öffentlichem Interesse die *Hate Speech*-Debatte. Im Rahmen dieser Arbeit wird das Phänomen Hate Speech auf seine Ursachen und Folgen, sowie auf die allgemeine Vielfältigkeit des Themas, untersucht. Dabei wird sich insbesondere auf die aktuellen Entwicklungen im deutschsprachigen Raum bezogen. Übergeordnetes Ziel dieser Arbeit ist die Planung und Entwicklung einer Browser-Erweiterung, welche auf der Seite des Nutzers Echtzeit-Analysen von Social Media-Posts vornimmt und potenzielle Hate Speech markiert. Damit sollen Nutzer nicht nur vor Hetze geschützt werden, es soll auch ein größeres Bewusstsein für das Vorhandensein von Hate Speech in sozialen Netzwerken geweckt werden. Darüber hinaus soll sich das Programm individuell an die Bedürfnisse und Toleranzgrenzen des Benutzers anpassen lassen.

Das im Rahmen dieser Arbeit entwickelte Plugin trägt nach den Ergebnissen eines User Experience-Experimentes erfolgreich dazu bei, Aufmerksamkeit für das Vorhandensein von Hass und Hetze in sozialen Netzwerken zu wecken. Dennoch wird deutlich, dass es bezüglich der automatisierten Erkennung von Hate Speech noch einige Verbesserungspotentiale gibt, insbesondere wenn andere Sprachen als Englisch untersucht werden.

Stichworte: Hetze, Hate Speech, Browser-Plugin, Soziale Netzwerke

Inhaltsverzeichnis

I. Abbildungsverzeichnis

II. Tabellenverzeichnis

1 Einführung

Im Rahmen dieses Kapitels wird die Motivation zur Entwicklung einer Browser-Erweiterung (*Plugin*) im Rahmen dieser Arbeit dargestellt. Darüber hinaus werden Ziele und Ansprüche an die Implementation definiert und Grenzen des Erwartbaren analysiert. Schließlich wird der Aufbau dieser Arbeit anhand des Inhaltes der verschiedenen Kapitel erläutert.

1.1 Motivation

Die Hate Speech-Thematik charakterisiert sich mitunter durch ihre Komplexität, Vielseitigkeit und Gefahr. Hetze ist in sozialen Netzwerken nicht immer als solche zu erkennen, kann durch Ironie oder Unwissenheit verschleiert werden und wird gezielt und organisiert als Mittel angewandt, um die Gesellschaft zu spalten und zu Gewalttaten zu motivieren (Kreißel et al., 2018). Setzt man sich diesem Klima in sozialen Netzwerken ohne Schutz auf Dauer aus, oder ist man direkt von Hass und Hetze betroffen, können fatale Schäden an psychischer und physischer Gesundheit die Folge sein (Sponholz, 2018, S. 35). Auch die demokratische Gesellschaft wird mit jedem Betroffenen, der sich ihr angehörig fühlt, durch Hetze attackiert. Medien und Politik haben die Ernsthaftigkeit des Themas erfasst, ausreichend wirksame Schutzmechanismen sind jedoch noch nicht in Sicht (Echikson & Knodt, 2018, S. 1ff.). Weitere Ausführungen zu diesen Erkenntnissen werden im Rahmen von Kapitel 2 getätigt.

Es stellt sich die Frage, welche Selbstschutzmaßnahmen vor den Folgen von Hass und Hetze jeder Einzelne für sich selbst treffen kann. An diesem Punkt soll diese Arbeit mithilfe der Entwicklung eines Browser-Plugins ansetzen, welches Nutzer bei der Identifizierung von Hetze und Hasskommentaren in sozialen Netzwerken unterstützt.

1.2 Ziel und Limitationen

Das übergeordnete Ziel dieser Arbeit ist die Entwicklung eines Browser-Plugins, welches während des Surfens in sozialen Netzwerken Posts und Kommentare identifiziert, mittels einer externen Schnittstelle, welche Techniken aus dem Gebiet des Natural Language Processings nutzt, auf das Vorhandensein von Hetze kategorisiert und anhand des Ergebnisses den Nutzer auf mögliche Hate Speech aufmerksam macht. Dadurch soll das Bewusstsein der Anwender für Hasskommentare und Hasspostings geweckt werden und ein Schutz vor den Folgen von Hetze eintreten. Anders als die politischen Ansätze und die

Maßnahmen der Betreiber der sozialen Netzwerke unterscheidet sich dieser Ansatz grundlegend darin, dass er ausschließlich auf der Seite der Nutzer agiert.

Bei der Implementation eines solchen Plugins sind bestimmte Limitationen unabwendbar. So wird im Rahmen von Kapitel 2 deutlich, dass Hate Speech sehr vielseitig auftreten kann und es sich selbst für Menschen als schwierig herausstellt, Hetze zweifelsfrei zu erkennen. Daher wäre es realitätsfern, an das zu entwickelnde Plugin einen Anspruch der fehlerfreien Erkennung von Hass und Hetze zu stellen. Ziel ist es hingegen mithilfe des Plugins Nutzer der sozialen Netzwerke für die Hate Speech-Thematik zu sensibilisieren und einen aufmerksamen Umgang mit Posts und Kommentaren zu fördern.

Außerdem soll das Plugin sich an die Bedürfnisse der Nutzer anpassen können. Da sich weder Experten aus der Politik noch aus dem Rechtswesen auf eine genaue Definition der Hate Speech-Problematik einigen können, liegt es auch nahe, dass Benutzer eines solchen Plugins unterschiedliche Ansprüche daran stellen, welche Posts als Hetze deklariert werden sollten und welche nicht. Somit sollte das Plugin individuell konfigurierbar sein.

1.3 Aufbau und Organisation

Im Rahmen dieser Arbeit soll die Entwicklung und Implementation des Browser-Plugins dokumentiert und begründet werden. Dazu werden in Kapitel 2 die fachlichen Grundlagen zu Hass und Hetze in sozialen Netzwerken, mit besonderem Augenmerk auf den deutschsprachigen Raum, analysiert. Der Inhalt dieses Kapitels erörtert die Grundlagen, um die Ernsthaftigkeit des Themas zu durchdringen, und zeigt auf, warum es wichtig ist, dass Nutzer von sozialen Medien für das Thema sensibilisiert werden und von ihrer Seite aus aktiv werden.

In Kapitel 3 wird ein Überblick darüber geliefert, welche Ansätze zur automatisierten Erkennung von Hate Speech bereits existieren.

Anschließend werden in Kapitel 4 die technischen Grundlagen für die Entwicklung des Plugins erklärt, die Planung des Plugins dargestellt und schließlich die finale Implementation mit den wichtigsten Kernkonzepten, welche zentrale Probleme lösen, erläutert.

Nach der Implementation des Plugins wird dieses einer Testgruppe im Rahmen eines User Experience-Experimentes zur Verfügung gestellt. Dadurch sollen sowohl Stärken des Plugins als auch etwaige Verbesserungspotenziale entdeckt und aufgezeigt werden.

Informationen zu dem Experiment, sowie die Evaluation desselbigen, sind in Kapitel 5 zu finden.

Abschließend werden im Rahmen von Kapitel 6 Schlussfolgerungen zu der gesamten Hate Speech-Thematik und der Entwicklung des Plugins gezogen. Darüber hinaus wird ein Ausblick auf mögliche Verbesserungen im Umgang mit Hate Speech und der automatischen Erkennung dieser gegeben. Auch weitere Möglichkeiten zur Erweiterung des Plugins werden im Rahmen des Fazits erwähnt.

2 Hate Speech in sozialen Netzwerken

Als Grundlage für die technische Betrachtung der Hate Speech-Erkennung ist es wichtig, die Komplexität und Vielseitigkeit der Thematik zu verstehen. Dafür werden im Rahmen dieses Kapitels Definitionsansätze von Hate Speech näher betrachtet. Außerdem wird analysiert in welcher Art und an welche Personen sich Hass und Hetze hauptsächlich richten. Im Anschluss wird erklärt, wie sich Gruppen in Deutschland gezielt organisieren, um Hate Speech in sozialen Netzwerken als Instrument für die Durchsetzung eigener Interessen zu nutzen. Die Wichtigkeit der Thematik wird deutlich, wenn im anschließenden Unterkapitel die Folgen von Hate Speech auf Individuen und die Gesellschaft betrachtet werden, sowie die aktuelle öffentliche Diskussion zu dem Thema.

2.1 Definition von Hate Speech

Der Begriff „Hate Speech“ gilt nicht als sehr streng definierte Bezeichnung und wird im allgemeinen Sprachgebrauch als Oberbegriff für öffentliche Äußerungen genutzt, die einzelne Menschen, insbesondere aber Menschengruppen, in verschiedener Weise diffamieren, beleidigen, manipulieren oder in sonstiger Weise angreifen (Sponholz, 2018, S. 31ff., 58ff.; Bundesamt für politische Bildung, 2017). Häufig wird sich zur Beschreibung von Hate Speech im deutschsprachigen Raum auf das Grundgesetz der Bundesrepublik Deutschland bezogen: „Die Würde des Menschen ist unantastbar“ (Grundgesetz, Art. 1, Abs. 1). Äußerungen, welche als Hate Speech kategorisiert werden, übertreten oder tangieren die Grenze der Meinungsfreiheit häufig, indem sie die Würde anderer Menschen angreifen (Bundesamt für politische Bildung, 2017; Landesanstalt für Medien NRW, AJS, 2019, S. 3; Meibauer, 2013).

Obwohl Hate Speech nicht nur im Internet präsent sein kann, wird der Begriff heutzutage in der Regel in diesem Kontext genutzt, da Hate Speech vor allem in sozialen Netzwerken auftritt. In Form von Hate Speech-Kommentaren werden üblicherweise Meinungen geäußert, die nur wenige Menschen teilen, oder dies offen zugeben würden (Landesanstalt für Medien NRW, AJS, 2019, S. 4). Durch die gefühlte oder auch reale Anonymität im Internet fühlen sich Autoren von Hass-Kommentaren und -Artikeln dazu motiviert, diese Meinung zu verbreiten (a.a.O., S. 4). Noch dazu kann man in den sozialen Netzwerken deutlich gezielter Menschengruppen mit ähnlichen Meinungen ansprechen und kann mit Minderheitsmeinungen dennoch ein größeres Publikum erreichen, als dies außerhalb des Internets der Fall wäre (a.a.O., S. 4). All dies sorgt dafür, dass Hate Speech in sozialen

Netzwerken auf einen fruchtbaren Boden stößt und damit häufig als Begriff diesem Kontext zugeordnet wird.

Hate Speech kann sehr direkt und offen geäußert werden, doch gerade im Kontext der Meinungsmache tritt Hate Speech auch sehr subtil formuliert auf und trägt zum Beispiel dazu bei, in sozialen Netzwerken ein Meinungsbild zu simulieren, welches nicht der tatsächlichen Allgemeinmeinung einer Bevölkerung entspricht (Meibauer, 2013, S. 1f.; Landesanstalt für Medien NRW, AJS, 2019, S. 4; Kreißel et al., 2018, S. 24).

Wenngleich im deutschsprachigen Raum auch der Begriff „Hassrede“ geläufig ist, wurde die englische Variante „Hate Speech“ in die deutsche Sprache übernommen und wird üblicherweise zur Beschreibung des Phänomens genutzt. Beide Begriffe werden jedoch als Synonym verwendet und sind im Allgemeinen geläufig. Die Gleichstellung und Unschärfe der Begriffe „Hate Speech“ und „Hassrede“ ist jedoch umstritten, so schreibt Dr. Liriam Sponholz (2018, S. 50) in ihrem Werk „Hate Speech in den Massenmedien“ Folgendes: „Anders als der Name nahelegt, ist Hate Speech weder notwendigerweise von Hass getrieben noch beschränkt es sich auf sprachliche Äußerungen. Aus diesem Grund ist die deutsche Übersetzung „Hassrede“ irreführend bzw. falsch. In der deutschen Alltagssprache kommt das Wort „Hetze“ (Duden: „feindliche Stimmungsmache“) dem Phänomen viel näher.“.

2.2 Verschiedene Ausprägungen von Hate Speech

Um die Vielseitigkeit der Hate Speech-Problematik zu erfassen, widmen sich die folgenden Unterkapitel der Frage, gegen welche Menschengruppen sich Hate Speech in der Regel richtet. Außerdem wird beschrieben, in welcher Form, aber auch mit welcher Intention Hetze dieser Art geäußert wird. Dabei wird auch der Fakt beleuchtet, dass sich ausgewählte Gruppen gezielt organisieren, um in sozialen Netzwerken Hass-Kampagnen zu starten, um ihre Ideologie und Themen in die mediale Öffentlichkeit zu bringen.

2.2.1 Häufige Opfergruppen

Wenngleich sich Hate Speech gegen verschiedene Menschen und Menschengruppen richten kann, kann ein großer Teil von Beiträgen, die als Hate Speech klassifiziert werden, einer engeren Auswahl von Opfergruppen und „Feindbildern“ zugeordnet werden. Dabei ist es auch möglich, dass ein einzelner Hasskommentar mehrere der genannten Menschengruppen attackiert (Bundesamt für politische Bildung, 2017).

Eine im deutschen Raum stark wahrnehmbare Form der Hate Speech, gerade in sozialen Netzwerken, sind rassistische und fremdenfeindliche Äußerungen. Insbesondere im Rahmen der sogenannten „Flüchtlingskrise“ im Jahr 2015 und der damit einhergehenden Stärkung von politischen Parteien innerhalb Deutschlands und Europas, die dem rechten politischen Spektrum zugeordnet werden (Schellenberg, 2018), ist ein deutlicher Anstieg von Äußerungen, die sich als fremdenfeindlich und rassistisch deklarieren lassen, wahrzunehmen (Kreißel et al., 2018, S. 8ff.; Landesanstalt für Medien NRW, AJS, 2019, S. 6). In welchem Ausmaß Hass-Kampagnen ganz gezielt von der rechten Szene organisiert werden, wird in Kapitel 2.2.3 näher betrachtet.

Auch die Hetze gegen andere Religionen, häufig antisemitisch oder antimuslimisch motiviert, Sexismus, sowie Homo- und Transphobie werden häufig als Hauptkategorien für die Klassifizierung von Hate Speech genannt (Sponholz, 2018, S. 31, 48, 56; Landesanstalt für Medien NRW, AJS, 2019, S. 6ff; Friesel, 2013). Das Bundesamt für politische Bildung (2017) listet in einem Artikel zum Thema Hate Speech außerdem Antiziganismus, Ableismus, Klassismus und Lookismus als Hate Speech-Kategorien. Zu erwähnen sind außerdem die Hetze gegen Politiker und weitere Personen, die öffentlich für bestimmte Werte eintreten, sowie gegen Menschen, die gegen Hate Speech-Kommentare argumentieren und in der Folge selbst von Hetze gegen ihre Person betroffen sind (Landesanstalt für Medien NRW, AJS, 2019, S. 11). Im Hinblick auf islamistisch motivierte Hetze lässt sich im Kontext sozialer Netzwerke ein deutlicher Rückgang feststellen (Kreißel et al., 2018, S. 7). Dies liegt daran, dass die großen Unternehmen zunehmende Gegenmaßnahmen gegen Posts dieser Art eingeleitet haben und radikale Islamisten somit vermehrt auf verschlüsselte Nachrichtendienste statt auf öffentliche Netzwerke zurückgreifen (ebd.).

2.2.2 Typische Formen der Äußerung von Hate Speech

Die Identifikation von Hate Speech ist häufig kein einfaches Unterfangen. Dies liegt daran, dass sich Hetze in ganz verschiedenen Formen äußern kann und dabei häufig nur sehr hintergründig auftritt. Zur Veranschaulichung dieses Sachverhaltes werden in diesem Hinblick vereinzelt Beispiele aufgeführt, von denen sich hiermit inhaltlich distanziert wird – sie dienen lediglich der Veranschaulichung der beschriebenen Hate Speech-Formen. Die gewählten Beispiele stammen teilweise aus Fällen öffentlicher oder von den Opfern öffentlich gemachter Hetzschriften und -kommentare, oder wurden in der Fachliteratur als Beispiel gewählt.

Zu den offensichtlichsten und primitivsten Formen der Hetze gehören direkte Beleidigungen von Menschen und Menschengruppen („miese GEZ Hure“[1]), das Aufrufen, Ankündigen oder die Unterstützung von Gewalttaten („Bereite […] dich […] auf deine Hinrichtung vor“[2]), sowie die Verbreitung von Falschaussagen („Die Flüchtlinge haben alle teure Handys“ (Landesanstalt für Medien NRW, AJS, 2019, S. 12)), welche mutwillig oder durch Uninformiertheit auftreten kann (a.a.O., S. 6, 10, 12).

Weniger direkt wird Hetze zum Beispiel geäußert, wenn sie in Verbindung mit Sarkasmus oder Humor verbreitet wird („Ich will auch ein neues Smartphone. Werd‘ ich im nächsten Leben halt Asylant.“ (a.a.O.., S. 12)). Gerade in sozialen Netzwerken können Anfeindungen auch abseits sprachlicher Mittel stattfinden, zum Beispiel in Form von Bildern, die rassistische Klischees erfüllen, oder durch die Nutzung von Symbolen, welche eine Haltung implizieren, die andere Menschen attackiert (Sponholz, 2018, S. 57).

Auch die verbale Trennung von Menschengruppen, in Form einer „Wir-Die-Rhetorik“ („*unsere* Frauen müssen vor *denen* geschützt werden“ (Landesanstalt für Medien NRW, AJS, 2019, S. 6)), ist ein typisches Stilmittel, um andere Menschen zu einer Gruppe zu mobilisieren und damit gegen andere Menschen zu hetzen. Festigen sich diese Haltungen, können sie sich zu einem stark verankerten und problematischen Weltbild entwickeln („Entweder wir oder sie“ (Sponholz, 2018, S. 56f.)).

Falsche oder eingeschränkte („Filter Bubble“) Weltbilder, häufig gespickt mit Verschwörungstheorien, können ein weiterer Grund dafür sein, warum Hetze auftritt und können auch Inhalt dieser Hetze sein. So wird zum Beispiel gegen Medien Stimmung gemacht, deren Berichterstattung nicht mit den eigenen Vorstellungen übereinstimmt, was zu Bezeichnungen als "Lügenpresse" oder anderem Niedermachen der Medien führt. Es werden Vergleiche gezogen, die jeder wissenschaftlichen Grundlage oder gesellschaftlichen Einordnung widersprechen (beispielsweise der Vergleich von Homosexualität und Pädophilie (Landesanstalt für Medien NRW, AJS, 2019, S. 6)) und mit Verschwörungstheorien ergänzt, die sich an mögliche Ängste der Zielgruppe richten und sich gerade im Internet rasant verbreiten können („die Homo-Lobby erzieht unsere Kinder um“ (ebd.)). Das Teilen einseitiger Berichterstattung zu komplexen Straftaten und das Verbreiten von Fake News trägt ebenfalls einen Teil zur Hetze bei, indem die öffentliche

[1] Zitiert nach https://twitter.com/dunjahayali/status/1204448344552169472
[2] Zitiert nach https://twitter.com/dunjahayali/status/1202184499553062912

Wahrnehmung von Themen beeinflusst und häufig auch gezielt manipuliert wird (ebd.; Sponholz, 2018, S. 58).

Zusammenfassend lässt sich feststellen, dass Hate Speech in verschiedener Art und Weise geäußert werden kann. Aufgrund der Vielfalt an Möglichkeiten, wie hetzende Kommentare aufgebaut sein können, kann es sich als schwierig herausstellen, sie zu identifizieren. Dies ist insbesondere der Fall, wenn die Hetze hintergründig als Stilmittel genutzt wird und sich nicht offensichtlich oder direkt formuliert äußert.

2.2.3 Hass-Kampagnen: Strategien und Organisation

In den vorherigen Kapiteln wurde bereits darauf verwiesen, dass Hetze in sozialen Netzwerken genutzt werden kann, um Meinungen von Minderheiten zu propagieren und das öffentliche Meinungsbild zu manipulieren. Dies ist nicht nur Wissenschaftlern bewusst, sondern ebenso Gruppierungen, die diesen Sachverhalt gezielt ausnutzen wollen und zu instrumentalisieren wissen. Im Folgenden soll aufgezeigt werden, dass es Gruppen gibt, die hochgradig organisiert sind und gezielte Hetz-Aktionen und Hass-Kampagnen starten, um die Diskussionen in sozialen Netzwerken zu manipulieren.

Zur Veranschaulichung dieses Sachverhaltes wird sich vorrangig auf die Studie „Hass auf Knopfdruck“ des Londoner *Institute for Strategic Dialogue* in Kooperation mit dem Verein *ichbinhier e.V.* (Kreißel et al., 2018) bezogen, da sie insbesondere die Situation im deutschsprachigen Raum aufgreift und die dargestellten Sachverhalte mithilfe ausführlicher eigener Analysen begründet.

Durch das Aufstellen eines Datensatzes an Kommentaren, die mit hoher Wahrscheinlichkeit Hasskommentare enthalten, konnte im Rahmen der Studie festgestellt werden, dass ein Großteil der Kommentare und der Interaktionen mit den Kommentaren auf eine auffallend kleine Gruppe an Accounts reduziert werden konnte, die jedoch sehr aktiv ist (a.a.O., S. 11f.). „Das aktivste Prozent der Nutzer generierte gar 25% der Likes für Hateful-Speech-Kommentare“ (a.a.O., S. 12). Infolge dieser Erkenntnis wurden die Accounts anhand ihrer vergebenen Likes analysiert, um zu untersuchen, mit welchen politischen Parteien die Accounts sympathisieren und so mögliche Gemeinsamkeiten bezüglich ihrer politischen Gesinnung aufzudecken. Dabei konnte festgestellt werden, dass mehr Likes von Accounts vergeben wurden, die mit der Partei „Alternative für Deutschland“ sympathisieren, als dies aufaddiert für alle anderen Parteien des Bundestags der Fall ist. Darüber hinaus wurde eine erhöhte Aktivität von Accounts festgestellt, die mit der „Identitären Bewegung“

sympathisieren (a.a.O., S. 13f.), welche vom Verfassungsschutz als rechtsextrem eingestuft ist (Zeit Online, 2019).

Da die Studie auf Basis der Datenanalyse gemeinsames Auftreten von Accounts feststellen konnte, wurde sich anschließend der Aufdeckung und Analyse von koordinierten Aktionen gewidmet. Dabei wurde festgestellt, dass insbesondere seit dem Wahlsieg Donald Trumps, den Troll-Organisationen mitunter sich zurechnen, rechtsextreme Gruppen auch in Europa versuchen, Wahlen zu ihren Gunsten zu manipulieren (Kreißel et al., 2018, S. 14ff.). Für diesen Zweck werden verschlüsselte Chatsysteme wie Discord und Telegram verwendet, in denen sich tausende Accounts sammeln (ebd.). Neben fremdenfeindlichem und nationalsozialistischem Gedankengut lassen sich in diesen Gruppen auch Anleitungen zum Bau von Waffen und Ähnlichem finden (ebd.).

Innerhalb der Gruppen konnte festgestellt werden, dass verschiedene Strategien bewusst und koordiniert angewendet werden, um unterschiedliche Wirkungen in sozialen Netzwerken zu erzielen. Dazu gehört zum Beispiel das Provozieren bekannter Menschen, um ihnen unpassende Statements zu entlocken, das Manipulieren von Hashtags wie #refugeeswelcome, damit darunter fast ausschließlich gegenteilige Meinungen zu finden sind, oder das Zuspammen von bestimmten Posts mit den eigenen Inhalten (a.a.O., S. 16f.). Außerdem wird versucht, eigene Hashtags in einem koordinierten Zeitraum, insbesondere vor Wahlen, so aktiv zu benutzen, dass sie beispielsweise auf Twitter in den Trends landen und so von vielen weiteren Personen gesehen werden (ebd.). „Diese Operationen sind häufig äußerst erfolgreich. So schafften es die rechtsextremen Aktivisten beispielsweise, 15 Tage vor der Bundestagswahl sieben von ihren Hashtags […] in den Top-20-Hashtags zu platzieren“ (a.a.O., S. 17). Anweisungen an die Gruppenmitglieder, wie man Accounts als unauffällig verschleiert und bekannt macht, zeigen weiter den hohen Organisationsgrad dieser Kampagnen, wie in Abbildung 1 zu sehen ist (a.a.O., S. 17).

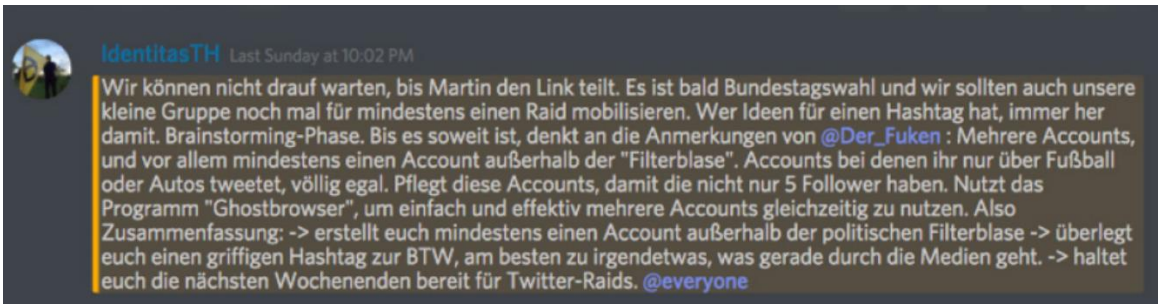

Abbildung 1- Mobilisierung von rechtsextremen Gruppenmitgliedern (Kreißel et al., 2018, S. 19)

Die Studie stellte außerdem fest, dass bekannte Politiker der AfD und verschiedene Medien dazu beitragen, die Hashtags und Themen, die in den Gruppen vorangetrieben werden, weiter in die Öffentlichkeit zu rücken und diese somit gebräuchlicher machen (a.a.O., S. 19f., 23, 25).

Zusammenfassend lässt sich feststellen, dass die Diskussion in sozialen Netzwerken bewusst manipuliert wird und in dieser Hinsicht insbesondere Rechtsextreme innerhalb Deutschlands eine Sonderrolle einnehmen. Um das Meinungsbild zu manipulieren, wird sich hochgradig koordiniert und gezielt Strategien entwickelt.

2.3 Die Auswirkungen von Hetze in sozialen Netzwerken

Nachdem aufgezeigt wurde, dass die Folgen von Hetze in sozialen Netzwerken bewusst instrumentalisiert und angewendet werden, wird in diesem Kapitel deutlich gemacht, wie vielseitig und problematisch ihre Auswirkungen sein können. Dafür werden Folgen für die Gesellschaft analysiert, welche auftreten können, wenn das Meinungsbild in sozialen Netzwerken über einen längeren Zeitraum manipuliert wird oder im Allgemeinen von Hetze geprägt ist. Anschließend werden auch die möglichen Folgen für Individuen, die von Hate Speech betroffen sind, herausgestellt.

2.3.1 Mögliche Folgen für die Gesellschaft

Durch den hohen Grad der Koordination von Hetzkampagnen, an denen sich häufig hunderte bis tausende Accounts beteiligen, kann der Eindruck von der Mehrheitsmeinung der Online-Community verfälscht werden (Landesanstalt für Medien NRW, AJS, 2019, S. 6). Fremdenfeindliche Äußerungen können ihren Charakter dadurch zum Beispiel verschleiern. Es entsteht zunehmend der Eindruck, dass es sich um harmlose Meinungen handelt, die nicht grundlos von so vielen Nutzern online unterstützt werden können (ebd.). Hass und Hetze können dadurch „salonfähig“ werden (ebd.).

Diese Veränderungen bezüglich der gesellschaftlichen Debatte bieten eine gefährliche Grundlage für weitere Eskalationen. Meinungsmache, Propaganda und das Hetzen gegen andere Bevölkerungsgruppen schufen in zahlreichen Fällen die Basis für schreckliche Verbrechen wie Völkermorde, Kriege, Terrorismus oder die Unterdrückung freier Meinungen (Marker, 2013, S. 61). Sowohl in der Vergangenheit als auch in der Gegenwart lassen sich Beispiele für Diktaturen finden, die ihre Positionen durch das Aufhetzen ihrer Unterstützer gegen Feindbilder festigten. So gehörten Propaganda und Hetze zu den grundlegenden Stilmitteln des NS-Regimes und waren damit Teil der Ursache für viele Verbrechen, die daraus resultieren konnten (ebd.). Heutzutage macht etwa der philippinische Präsident Rodrigo Duterte öffentlich Witze über die Vergewaltigung von Frauen und möchte das Problem seines Landes mit Drogenkonsumenten und -dealern dadurch lösen, dass er zur Ermordung ebenjener aufruft (Der Tagesspiegel, 2016). Um ihre Anhänger zu mobilisieren und radikalisieren, setzen terroristische Vereinigungen bewusst darauf, Hass und Wut in ihnen auszulösen (Marker, 2013, S. 61).

Auch in Deutschland lassen sich einige der am schärfsten kritisiertesten Gewalttaten aus den letzten Jahren auf diese Entwicklungen beziehen. Unterstützer der in Form von Hetze geäußerten Meinungen können sich durch Debatten, Propaganda, und das verfälschte Meinungsbild mobilisiert und motiviert fühlen – auf Gedanken können Taten folgen. Die Meinungsmache in den sozialen Medien trägt „zu einem gesellschaftlichen Klima bei, das rassistischen und rechtsextremen Personen und Gruppierungen das Gefühl gibt, im Sinne und als Sprachrohr einer schweigenden Mehrheit zu handeln“ (Landesanstalt für Medien NRW, AJS, 2019, S. 6).

Deutlich wurde dies beispielsweise im Jahr 2019, als ein bewaffneter Rechtsextremist versuchte in Halle eine Synagoge zu stürmen und zahlreiche Juden zu ermorden. Die Tat wurde als Livestream im Internet übertragen, der Täter bezog sich auf weitere Anschläge, wie das ebenfalls gestreamte Christchurch-Attentat, und machte in einem Manifest, sowie in dem Video deutlich, dass er sich im Internet radikalisiert hat und an Verschwörungstheorien glaubte (Käppner & Bovermann, 2019; Biermann et al., 2019; Gensing, Dokumente des Hasses, 2019). Durch die zunehmende Dokumentation solcher Taten können sich wiederum andere potenzielle Täter radikalisieren und ihre Vorgänger imitieren. Es handelt sich um eine gefährliche Eskalationsspirale, die Täter gezielt in Gang setzen.

Ein weiteres Beispiel aus dem Jahr 2019 zeigt, dass der Rechtsextremismus in Deutschland eine neue Form der Gewalt annimmt und eine gegenseitige Motivation stattfindet. Der CDU-Politiker Walter Lübcke wurde im Rahmen rechter Hasskampagnen dafür attackiert, dass er sich 2015 für den Schutz von Flüchtlingen einsetzte (Boeselager, 2019); im Jahr 2019 wurde er wohl aus diesem Grund ermordet. Die Rolle sozialer Netzwerke und der dort getätigten Hetze sind eindeutig. Während Rechtsextreme im Internet jahrelang gegen Lübcke hetzten und zum Mord gegen ihn aufriefen, wurde exakt dies später Realität. Auch nach der Tat wurden viele Kommentare, die den Mord befürworteten, veröffentlicht (Gensing, Rechtsextreme verhöhnen Getöteten, 2019).

Im Februar 2020 wurden sämtliche Zweifel daran beseitigt, dass der Rechtsextremismus in Deutschland eine neue Dimension erreicht hat. Erst wurde ein rechtsextremes Terror-Netzwerk ausgehoben, welches dabei war, Pläne zu konkretisieren, bewaffnet in zehn Moscheen verschiedener Bundesländer einzudringen und Betende zu ermorden (Spiegel, 2020). Als konkretes Ziel wurde das Herbeiführen bürgerkriegsähnlicher Zustände vermutet (Spiegel, 2020). Nur eine Woche später tötete ein vermutlicher Terrorist mindestens elf Menschen, darunter mindestens neun mit Migrationshintergrund, im hessischen Hanau. Der Generalbundesanwalt verkündete eine „zutiefst rassistische Gesinnung“ des Täters (Tagesschau, 2020). Dieser hatte zuvor ebenfalls ein Manifest und Videos über soziale Medien und eine eigene Website verbreitet (ebd.).

Die genannten Fälle haben in Deutschland zahlreiche Debatten angestoßen, einige davon auch im Bezug auf Hetze in sozialen Medien, die immer in direkten Kontext zu den Fällen gebracht wurde. Durch spätere Morddrohungen von Rechtsextremisten gegen Journalisten und insbesondere Lokalpolitiker, wurde deutlich, dass ein Motivationseffekt durch diese Taten entsteht, da sich darin auf vorherige Täter bezogen wird. So veröffentlichte die deutsche Journalistin Dunja Hayali im Dezember 2019 eine Morddrohung gegen sich. Darin heißt es unter Anderem „Lübcke, Hollstein, Reker waren nicht die letzten Politiker, sondern die Ersten.“ oder „Nehmt euch alle ein Beispiel an Stephan Balliet, Brenton Tarrant und David S.“[1], bei denen es sich um die Täter von Halle, Christchurch und dem rechtsextrem motivierten Amoklauf in München handelt (Wiedemann-Schmidt, 2019).

Die Strategie der Hetzer zeigt Wirkung: Ebenfalls noch im Jahr 2019 traten Lokalpolitiker aufgrund rechter Drohungen zurück, ein weiterer fühlte sich so sehr gefährdet, dass er sich

[1] Zitiert nach https://twitter.com/dunjahayali/status/1202184499553062912

bewaffnen möchte (Ziegler, 2020). Somit zeigt sich eine weitere Gefahr für die Gesellschaft durch Hetze: Durch die immer häufigeren Bedrohungen werden die gewählten Vertreter der Bevölkerung dazu provoziert ihr Amt niederzulegen, Gegenmeinungen werden nicht mehr öffentlich geäußert. Ein Klima der Angst vor Rechtsextremismus scheint in erschreckendem Tempo zu entstehen und kann gerade auf lokaler Ebene, wo Politiker ehrenamtlich arbeiten, zu deutlichen Änderungen in der Politik führen.

Doch nicht nur Gewalttaten und die Gefährdung innerer Sicherheit werden durch die Veränderung der gesellschaftlichen Debatte gefördert. Öffentliche Hetze führt zu einem allgemeinen Klima des Misstrauens und trägt zu Verfeindungen zwischen Bevölkerungsgruppen bei (Sponholz, 2018, S. 22). Daraus kann wiederum ein Effekt der Desintegration entstehen (Sponholz, 2018, S. 33). Ganze Bevölkerungsgruppen könnten sich isolieren und nicht mehr am öffentlichen Leben teilhaben oder ihre Möglichkeiten der politischen Mitbestimmung scheuen und somit dazu beitragen, dass die Hetzer ihre Positionen weiter etablieren können (Marker, 2013, S. 65). Die möglichen Folgen von öffentlicher Hetze werden zusammenfassend als demokratiefeindlich oder -schädlich betitelt, sie sei häufig „Vorbote antidemokratischer Entwicklungen“ (Sirsch, 2013, S. 165).

2.3.2 Mögliche Folgen für Individuen

Neben den Folgen, die das vermehrte Auftreten von Hate Speech in sozialen Medien auf die Gesellschaft haben kann, sind auch die Folgen auf die direkt betroffenen Individuen zu betrachten.

Die psychischen Folgen für Opfer von Hate Speech sind vielseitig. Da sich Betroffene verletzt fühlen können, wird sogar debattiert, ob die hervorgerufenen psychischen Folgen mit physischen Verletzungen gleichzusetzen sind, weil sie eine ähnliche Wirkung hervorrufen (Marker, 2013, S. 64).

Eine der Hauptauswirkungen auf Opfer von Hassrede ist die Flucht in die soziale Isolation, also das Vermeiden des Kontaktes mit anderen Menschen, insbesondere im öffentlichen Raum. Hervorgerufen wird diese Auswirkung durch den Einschüchterungsaspekt, der mit der Hetze einhergeht (Sponholz, 2018, S. 22). Im vorherigen Unterkapitel wurden die möglichen Folgen hieraus bereits aufgezeigt; ganze Menschengruppen könnten ihre Interessen nicht mehr öffentlich äußern und zum Beispiel ihr Wahlrecht nicht wahrnehmen. Die politische Debatte würde dadurch immer stärker von den Hetzern geprägt werden können (Marker, 2013, S. 65). Betroffene können durch Hetze außerdem unter

Angstzuständen leiden und das Selbstbewusstsein oder das generelle Vertrauen in andere Menschen verlieren (Sponholz, 2018, S. 22, 33ff.). Es können außerdem Depressionen und seelische Probleme hervorgerufen oder verstärkt werden, die bis zum Suizid führen könnten (Sponholz, 2018, S. 35; Sirsch, 2013, S. 170).

Neben der Isolation können Betroffene jedoch auch sehr gegensätzlich reagieren und ihre Verletzung in Form von Aggression verarbeiten (Marker, 2013, S. 65f.). Durch das hervorgerufene Misstrauen können sich feindliche Gedanken gegenüber anderen Menschen und Menschengruppen entwickeln, die schließlich in Gewalttaten enden können oder zu diesen motivieren (Sponholz, 2018, S. 22; Sirsch, 2013, S. 171).

Doch Hetze kann auf unterschiedliche Weise auch physische Folgen hervorrufen. So zeigte sich in verschiedenen Experimenten, dass rassistisch beleidigte Menschen im Durchschnitt häufiger rauchen, als dies eine Vergleichsgruppe tut (Sponholz, 2018, S. 34). Man kann dies also als physische Selbstverletzung ansehen, die wohl zur Beruhigung der psychischen Folgen vorgenommen wird. Durch die zuvor geschilderte Aggressionsentwicklung und die resultierenden Gewalttaten, können nicht nur die Betroffenen selbst körperlichen Schaden erleiden, sondern auch Unbeteiligte, oder nicht direkt Beteiligte. Schließlich kann das Betroffensein von Hetze aber auch ganz direkte körperliche Schäden verursachen, zum Beispiel eine Erhöhung des Blutdrucks oder des Pulses, sowie Kopfschmerzen. Diese Symptome können besonders längerfristig negative Folgen für den Betroffenen haben und weitere Schäden hervorrufen (Sirsch, 2013, S. 170).

Außerdem kann Hetze soziale Benachteiligung bedingen. So zeigte sich, dass Betroffene von rassistischer Hetze durch die psychischen Folgen in Leistungssituationen weniger gut abschneiden, als es ihnen theoretisch möglich wäre, da der Stress aus der Leistungssituation von dem dauerhaften Stress der Anfeindungen weiter erhöht wird (Sirsch, 2013, S. 170). Somit ruft Hetze indirekte Schäden für das Prinzip der Chancengleichheit hervor (ebd.). Auch demokratische Werte wie die Bewegungs- und Versammlungsfreiheit können durch Anfeindungen infolge von Hetze gestört werden (Sponholz, 2018, S. 33).

2.4 Hate Speech als Thema der öffentlichen Debatte

Die Diskussion um Hass und Hetze in und außerhalb von sozialen Netzwerken nimmt in Deutschland zuletzt immer weiter zu. Die Gründe dafür sind vielseitig. Neben den bereits erwähnten Straftaten, sorgen auch immer mehr Umfragen und Studien mit teilweise schwerwiegenden Ergebnissen zu der Thematik dafür, dass Medien über das Thema

berichten. So veröffentlicht beispielsweise die *Landesanstalt für Medien NRW* mittlerweile jährlich eine *forsa*-Befragung zum Thema Hate Speech. Sie kam unter Anderem zu dem Schluss, dass 78% der Befragten im Jahr 2018 angaben, schon persönlich Hasskommentare in sozialen Netzwerken gesehen zu haben – eine Steigerung von 8% innerhalb nur eines Jahres (Landesanstalt für Medien NRW, 2019). Auch die von Google bereitgestellten Suchtrends zeigen, dass das Interesse an den Themen „Hetze", „Hate Speech" und „Hassrede" seit der Mitte des Jahres 2015 im Durchschnitt höher liegt als je zuvor (Google, 2020).

Die mediale Berichterstattung und die Zunahme der öffentlichen Debatte setzen die Politiker zunehmend unter Druck, stärker gegen die Hate Speech-Problematik vorzugehen. Dabei kommt auch immer wieder die Frage auf, ob sich dem Thema nicht juristisch genähert werden könnte. Derzeit stellt sich die Situation in Deutschland so dar, dass das Verbreiten von Hasskommentaren kein eigenständiger Straftatbestand ist, jedoch andere Straftatbestände wie Volksverhetzung, Beleidigung oder Verleumdung erfüllen kann (Landesanstalt für Medien NRW, AJS, 2019, S. 7, 9). Die Grenze zwischen dem Grundrecht der freien Meinungsäußerung und diesen Straftatbeständen zu ermitteln, erweist sich dabei nicht immer als einfaches Unterfangen. Dies gilt jedoch auch für die Einführung eines eigenständigen Tatbestandes für das Verbreiten von Hate Speech. Tatsächlich gibt es bereits seit Jahrzehnten juristische Debatten zu dem Thema, doch es gelingt weder Wissenschaftlern noch Juristen, Hate Speech präzise und einheitlich zu definieren, ohne dabei die Meinungsfreiheit in Gefahr zu bringen (Sponholz, 2018, S. 43). So ist es auch nicht verwunderlich, dass selbst innerhalb der EU rechtlich verschiedene Definitionen zum Thema Hate Speech angewandt werden (Alkiviadou, 2017).

Als erster groß angesetzter politischer Schritt gegen Hetze im Internet trat am 1. Januar 2018 in Deutschland das *Netzwerkdurchsetzungsgesetz* (NetzDG) in Kraft. Es ermöglicht hohe Geldstrafen für die Betreiber sozialer Netzwerke, wenn sie als problematisch gemeldete Inhalte nicht rechtzeitig löschen (Echikson & Knodt, 2018, S. 1). Das NetzDG wird jedoch teilweise scharf kritisiert. Seine Wirkkraft sei zu schwach, es sei zu hastig aufgesetzt worden und decke daher nicht genügend Tatbestände ab. Außerdem sei die zeitliche Frist zur Löschung problematischer Inhalte in Anbetracht der Schnelllebigkeit sozialer Netzwerke viel zu hoch angesetzt (a.a.O., S. 4f.). Darüber hinaus richtet es sich nur an die besonders großen sozialen Netzwerke, die es teilweise extra aufwändig gestalten, Beschwerden einzureichen, um die Anzahl ebendieser gering zu halten (a.a.O., S. 7). Heiko Maas,

ehemaliger Justizminister und Autor des NetzDG, bezeichnete das Gesetz ebenfalls nur als einen ersten Schritt. Langfristig sollte eine europäische Lösung im Kampf gegen Hetze in sozialen Medien gefunden werden (a.a.O., S. 13).

Im Februar 2020 wurden weitere Richtlinien beschlossen, die unter anderem die Betreiber sozialer Netzwerke dazu verpflichten, relevante Fälle direkt an das Bundeskriminalamt zu melden (Marx, 2020). Auch diese Verschärfungen lösten jedoch, in erster Linie aus datenschutzrechtlichen Gründen, Kritik aus (ebd.).

Es ist also davon auszugehen, dass das Thema auch in den kommenden Jahren öffentliches und politisches Interesse hervorrufen wird. Der Kampf gegen Hate Speech setzt sich fort und in Anbetracht der analysierten Trends könnte die Diskussion auch noch deutlich größer werden. Maßnahmen mit starken Auswirkungen könnten noch Jahre auf sich warten lassen.

3 Verwandte Arbeiten

Im Rahmen dieses Kapitels sollen vorhandene Ansätze zur Bekämpfung von Hate Speech dargestellt werden. Der Fokus soll dabei auf die automatisierte Klassifizierung gelegt werden. Allerdings werden auch Initiativen anderer Art, die sich der Bekämpfung von Hetze in sozialen Netzwerken widmen, kurz vorgestellt.

3.1 Differenzierung von Hate Speech und Offensive Language

Davidson et al. (2017) widmen sich dem Problem, automatisiert Hetze von allgemeiner offensiver Sprache, also zum Beispiel dem einfachen Gebrauch von Schimpfwörtern und Ähnlichem, zu differenzieren. Dazu wurden Probanden gebeten, gesammelte Tweets, nach gegebenen Definitionen, zwischen *Hate Speech*, *Offensive Language* und *keine der beiden Kategorien* zu klassifizieren. Anschließend wurde ein automatisierter Klassifikator auf Basis dieser Einschätzungen trainiert und genutzt, um weitere Tweets, die zuvor von Menschen eingeschätzt wurden, zu bewerten. Die Ergebnisse des trainierten Klassifikators, welcher auf verschiedenen Modellen, die erfolgreich zur Klassifizierung von Hate Speech genutzt wurden, beruht, wurden anschließend mit denen lexikonbasierter Methoden verglichen. Dies bedeutet, dass die Tweets auf Basis einer lexikonähnlichen Sammlung von Schimpfwörtern und typischen hetzenden Formulierungen kategorisiert wurden.

Lediglich 5% der Tweets, welche durch die lexikonbasierten Methoden ermittelt wurden, wurden auch von den Probanden als Hate Speech deklariert. 76% der Tweets wurden hingegen mehrheitlich als Offensive Language eingeordnet. Durch dieses Experiment wurde somit deutlich, dass lexikonbasierte Methoden zur Erkennung offensiver Sprache sehr geeignet scheinen, jedoch daran scheitern, zwischen dieser Kategorie und konkreter Hetze zu unterscheiden.

Diese Unterscheidung gelang dem trainierten Klassifikator deutlich besser. Dem am erfolgreichsten untersuchten Modell zur Klassifizierung der Tweets gelang eine Übereinstimmung von 91% mit der Einschätzung der Probanden. Darüber hinaus ließ sich jedoch eine Tendenz erkennen, wonach das Modell dazu neigt, die Tweets im Vergleich zu den Probanden als weniger hetzend einzuordnen. Fehler sind insbesondere dann aufgefallen, wenn problematische Begriffe in einem positiven Kontext genutzt wurden.

Die Autoren schlussfolgern, dass dem Kontext eines Posts bei der Hate Speech-Erkennung in Zukunft mehr Aufmerksamkeit geschenkt werden muss und trainierte Modelle genauere Ergebnisse erzielen als lexikonbasierte Verfahren.

3.2 Aktuelle Ansätze zur automatisierten Erkennung von Hate Speech

Schmidt und Wiegand (2017) bieten einen Überblick über die bekanntesten aktuellen Ansätze zur automatisierten Erkennung von Hate Speech. Betrachtet wird unter anderem der *Bag-of-words*-Ansatz. Hierbei handelt es sich um eine der simpelsten Methoden, um Texte zu klassifizieren. Dabei werden sowohl Zusammenhänge zwischen den Wörtern als auch deren Reihenfolge vollständig ignoriert; es wird lediglich gezählt, wie häufig welches Wort in einem Text enthalten ist (McTear et al., 2016, S. 166f.). Dafür werden Füllwörter, welche nicht zum Inhalt des Textes beitragen, herausgefiltert. Über einen Prozess namens Lemmatisierung werden Wörter außerdem auf ihren Wortursprung zurückgeführt, sodass zum Beispiel deutlich wird, dass das Wort „gelaufen“ die Bedeutung der Grundform „laufen“ widerspiegelt (ebd.). Während es in anderen Anwendungsfällen von Vorteil ist, dass der Bag-of-words-Ansatz sehr simpel gestaltet ist, ist der Einsatz zur Klassifizierung von Hate Speech oft unzureichend. Allerdings kann die Kombination dieser Methode, sowie ähnlich simplen Ansätzen, mit fortgeschritteneren Methoden zu präziseren Ergebnissen führen (Schmidt & Wiegand, 2017, S. 2).

Ein weiterer von Schmidt und Wiegand erwähnter Ansatz ist die Generalisierung von Wörtern. Dabei wird die Bedeutung von Wörtern in Kategorien, oder in moderneren Ansätzen durch Vektoren, dargestellt. Dies hat zur Folge, dass Wörter mit einer ähnlichen Bedeutung eine ähnliche Repräsentation als Vektor erhalten. Außerdem kann die Bedeutung von ganzen Sätzen repräsentiert werden, indem zum Beispiel die Vektoren der einzelnen Wörter aufaddiert werden und somit einen neuen Vektor ergeben. Schmidt und Wiegand bezeichnen diesen Ansatz im Kontext der Hate Speech-Erkennung als unzureichend, allerdings kann er als Basis für fortgeschrittenere Techniken dienen (a.a.O., S. 2f.).

Zur Unterstützung von Hate Speech-Klassifikatoren erwähnen die Autoren außerdem den Einsatz von bestehenden Algorithmen, die die übermittelte Stimmung innerhalb eines Textes analysieren. Da Hate Speech in aller Regel in einem negativen Kontext verbreitet wird, kann eine solche Einordnung den Prozess der Identifikation von Hetze unterstützen (ebd.).

Auch der bereits erwähnte lexikonbasierte Ansatz zur Identifikation von Hate Speech wird in der Veröffentlichung als Methode erwähnt. Allerdings erziele dieser laut den Autoren nur als Ergänzung zu weiteren Modellen Erfolge (a.a.O., S. 3f.).

Ein weiterer wichtiger Ansatz zur Erkennung von Hate Speech ist das Einbeziehen der Linguistik. So kann die Kombination von gewissen Wörtern ein guter Indikator dafür sein, in welchem Kontext eine Aussage steht, ob sie also zum Beispiel beleidigend gemeint ist (a.a.O., S. 4). Die Autoren nennen hierfür ein Beispiel aus dem Bereich des Antisemitismus. Würde zum Beispiel das Wort „Jude“ mit einer wenig eindeutigen Beleidigung, zum Beispiel dem Wort „Schwein“, welches neben einem beleidigendem Sinn auch wertungslos das Tier beschreiben kann, im Kontext stehen, wäre es deutlich einfacher, die Gesinnung des untersuchten Textes festzustellen, als wenn das alleinige Vorhandensein der beiden Wörter analysiert werden würde. Dies wäre zum Beispiel bei dem Bag-of-words-Ansatz der Fall. Grundlage für diese Methode sind große Datensammlungen an Wortkombinationen und deren Bewertung (ebd.).

Der Einbezug von Wissen über gesellschaftliches und individuelles Verhalten sowie Werte und Normen ist ein weiterer Ansatz, welchen Schmidt und Wiegand erwähnen. Weiß ein Klassifikator zum Beispiel, welches Verhalten eher für männliche Personen und welches eher für weibliche Personen typisch ist, kann diese Information genutzt werden, um zum Beispiel Gender-Angriffe gezielter einzuordnen und zu klassifizieren. Dieser Ansatz ist noch weitgehend unerforscht, da das Wissen sehr spezifisch auf etwaige Diskriminierungsarten zusammengetragen und aufbereitet werden muss, was einen hohen Programmieraufwand zur Folge hat (a.a.O., S. 4f.).

Durch den Einbezug von Meta-Informationen, beispielsweise darüber, welche Art von Person welche Worte an welchen Typ von Person richtet, oder den Einbezug aktuell relevanter Themen in der weltweit öffentlichen Debatte, können Klassifikatoren bessere Einschätzungen darüber treffen, welche Absicht mit Aussagen einhergeht – allerdings wird die Identifikation von Hate Speech dadurch noch deutlich komplexer (a.a.O., S. 5f.). Gerade im Kontext der sozialen Netzwerke ist dies jedoch ein vielversprechender Ansatz. So lassen sich üblicherweise mit vergleichsweise geringem Aufwand Informationen zu den Nutzern über die Programmierschnittstellen der sozialen Netzwerke sammeln. Wenn ein User in der Vergangenheit auffällig häufig als Hetze einzuordnende Posts abgesetzt hat, ist die Chance, dass er dies wieder tut, zum Beispiel deutlich höher als bei einem User, dessen Post-

Vergangenheit kaum negativ einzuordnen ist. Weitere Argumente, die in eine Klassifikation einbezogen werden können, sind zum Beispiel das Geschlecht des Autors, die Anzahl abgesetzter Posts oder das Ursprungsland eines Users. Für den Erfolg dieser Methoden ist es sehr relevant, die einbezogenen Meta-Informationen korrekt einzuordnen und gezielt einzusetzen (ebd.).

Im Kontext sozialer Medien ist der Einbezug von multimedialen Daten in die Beurteilung von Hate Speech ebenfalls relevant. Ist ein Klassifikator in der Lage, die Bedeutung von Bildern, Videos und Audio zu analysieren, kann der Kontext von dazu geposteten Kommentaren präziser erfasst werden (ebd.). Außerdem kann Hate Speech auch direkt in anderer Form als Text, zum Beispiel durch Bilder und Videos, geäußert werden. Da die Klassifizierung durch diese Art von Analyse deutlich umfangreicher und komplexer wird, ist dieser Ansatz wenig erforscht, obwohl er als sehr vielversprechend gilt (a.a.O., S. 6).

Im Allgemeinen macht die Untersuchung deutlich, dass die Kombination verschiedener Ansätze zur Klassifikation von Hate Speech deren Erfolg deutlich verbessern kann (a.a.O., S. 2ff.).

Die Veröffentlichung von Schmidt und Wiegand zeigt auch aktuelle Probleme und zukünftige Herausforderungen für das Thema auf. So kritisieren die Autoren zum Beispiel, dass die meisten Ansätze bislang nur in englischer Sprache annähernd ausreichend getestet und angewandt wurden. Es würde sich erst zeigen, wie gut sich die erfolgreichen Modelle der englischen Sprache auch auf andere Sprachen anwenden lassen. Außerdem würden noch immer charakteristische Sonderfälle existieren, die von den aktuellen Ansätzen kaum abgedeckt werden können. Als Beispiel hierfür wird der Kommentar „*Kermit the frog called and he wants his voice back.*“ (Schmidt & Wiegand, 2017, S. 8) genannt, der sich über die Stimme einer Frau lustig macht. Allerdings bietet der Kommentar, ohne weiteren Kontext, automatisierten Methoden wenig Möglichkeiten, diesen „Angriff“ als solchen zu identifizieren. Auch Sarkasmus ist ein Gebiet, welches in seiner Vielfältigkeit bisher nicht von Klassifikatoren beherrscht wird (ebd.).

Insgesamt wird in der Untersuchung deutlich, dass die automatische Hate Speech-Erkennung ein Forschungsgebiet darstellt, welches noch sehr viel Potenzial in sich birgt, das insbesondere in internationaler Betrachtung ausgeschöpft werden könnte. Doch auch das Einbeziehen von Kontexten wie Nutzerinformationen, oder mit dem Kommentar zusammenhängende Bilder, sind vielversprechende Ansätze, welche in Zukunft weiter

erforscht werden sollten. Die Autoren versprechen sich außerdem Erfolg von einer großen, öffentlich zugänglichen Sammlung an allgemeinen Formulierungen, da bisherige Klassifikatoren oft auf Daten trainiert wurden, die besondere Eigenarten aufweisen könnten (zum Beispiel typische Formulierungen, die auf Twitter genutzt werden) (a.a.O., S. 8f.).

3.3 Projekt *Forum 4.0*

Bei *Forum 4.0* handelt es sich um ein Projekt der Universität Hamburg, der HAW Informatik und des Hans-Bredow-Instituts, in dessen Kontext auch diese Arbeit steht. Ziel des Projektes ist es, Vorteile herauszustellen, die Nutzerkommentare, zum Beispiel in sozialen Netzwerken, im Journalismus oder zur Entwicklung von Software, bieten. Kommentare mit hohem Nutzen durch Feedback und Kritik sollen zum Beispiel automatisiert von jenen Kommentaren getrennt werden, die weniger wichtigen Inhalt oder Hate Speech enthalten. Dadurch kann Moderatoren, Journalisten, Softwareentwicklern und weiteren potenziellen Nutzern die Arbeit erleichtert und neue Möglichkeiten geboten werden (Universität Hamburg, 2018).

Aus dem Projekt heraus resultierten bereits mehrere wissenschaftliche Veröffentlichungen. So stellten sich Wiedemann et al. (2018) unter Anderem der Herausforderung, dass die Erkennung offensiver Sprache bisher fast ausschließlich in englischer Sprache erforscht ist und entwickelten einen Klassifikator für deutschsprachige Twitter-Postings auf Basis des Konzeptes von Transfer Learning innerhalb von Machine Learning-Prozessen. Dabei wurden auch die Problembereiche falscher Rechtschreibung, die Bedeutung von Emoji und weiteren Zeichenkombinationen, sowie die fehlende Variabilität in der Nutzung von Datensätzen, welche für eine spezifische Aufgabe gesammelt wurden, analysiert. Die Untersuchung kam zu dem Schluss, dass die automatisierte Erkennung offensiver Sprache stark durch die Nutzung von Transfer Learning-Techniken, großen Datensätzen, auf denen die verschiedenen Modellebenen individuell trainiert werden, und dem Einbezug von markierten Nutzern, profitiert.

In einer weiteren Studie im Kontext des Projektes (Wiedemann et al., 2019) wurden die Ergebnisse von einem Klassifikator, welcher durch Überwachtes Lernen, also mittels eines Machine Learning-Modells mit dem Ziel ein vorgegebenes Ergebnis zu erreichen (Smola & S.V.N., 2008, S. 32), trainiert wurde, mit denen eines Klassifikators, der ohne Überwachtes Lernen trainiert wurde, verglichen. Dabei wurde deutlich, dass der Erfolg von dem Einsatz Überwachten Trainings von der Datenbasis und ihrem Umfang abhängig ist.

Das Projekt *Forum 4.0* geht neben der Klassifizierung von Hate Speech und Offensive Language auch weiteren automatisierten Ansätzen nach, beispielsweise der Frage, an welche Personen sich ein Kommentar richtet (Häring et al., 2018).

3.4 Fallbeispiel: Moderne Hate Speech-Klassifikation durch Maschinelles Lernen

Burnap & Williams (2015) bieten ein Beispiel für einen Hate Speech-Klassifikator, welcher auf modernen Ansätzen mithilfe des Maschinellem Lernens basiert. Dafür wurden Tweets, welche nach der Ermordung von Lee Rigby durch Islamisten in London, veröffentlicht wurden, gesammelt und sowohl von Probanden als auch einem trainierten Klassifikator auf das Vorhandensein von Hate Speech analysiert. Die Ergebnisse des Klassifikators wurden anschließend mit der menschlichen Einschätzung verglichen. Somit wurde auch in diesem Beispiel der Prozess des Überwachten Lernens angewendet (a.a.O., S. 223).

Zum Testen des Modells wurden 2.000 Tweets von Probanden auf das Vorhandensein von religiösen oder rassistischen Äußerungen bewertet. Umstrittene Ergebnisse in der Bewertung wurden anschließend entfernt. Dadurch entstand ein eindeutiger Datensatz von 1.901 Tweets, von denen 222 als Hate Speech eingeordnet wurden (a.a.O., S. 227f.).

Das entwickelte Modell basiert auf dem in Abschnitt 3.2 bereits erwähnten Generalisierungsprozess, bei dem alle Wörter in den Tweets als Vektoren repräsentiert werden. Darüber hinaus wurden alle Tokens, also alle relevanten Wörter, in eine kleingeschriebene Variante transformiert, um Unterschiede in der Groß- und Kleinschreibung, welche keine Auswirkungen auf die Bedeutung haben, zu ignorieren (a.a.O., S. 230). Zeichen, welche keine emotionale Bedeutung in sich tragen, anders als zum Beispiel Smileys oder Ausrufezeichen, wurden ebenso wie Füllwörter entfernt (ebd.). Auch der bereits erwähnte Prozess der Lemmatisierung, also das Zurückführen von verschiedenen Wortvarianten auf den Wortursprung, wurde als Technik angewandt (ebd.). Anschließend wurden die relevanten Tokens in sogenannte N-Gramme transformiert, indem zusammenstehende Tokens in der Anzahl zwischen 1 und 5 miteinander kombiniert wurden. Die N-Gramme enthalten somit zusammenstehende Formulierungen, deren Bedeutung anschließend interpretiert werden kann, zum Beispiel „send them home“ (ebd.).

Im Anschluss wurde der Datensatz in zwei Varianten weiterverarbeitet. In der ersten Variante wurden alle N-Gramme beibehalten, während im zweiten Ansatz mittels eines Lexikons, welches typisch rassistische Formulierungen sammelt, die Anzahl der N-Gramme

auf verdächtige Formulierungen reduziert wurde (a.a.O., S. 230f.). Mithilfe des bereits vorhandenen Stanford-Parsers wurden anschließend Zusammenhänge zwischen den N-Grammen festgestellt (ebd.). Durch diesen Schritt konnten die Beziehungen der verschiedenen N-Gramme innerhalb eines Satzes analysiert werden (de Marneffe & Man, 2008, S. 1f.).

Bei der Implementation der beschriebenen Techniken wurden sämtliche Tweets in eine Datenstruktur überführt, die sowohl die menschliche Einschätzung bezüglich des Vorkommens von Hetze als auch die analysierten N-Gramme und Beziehungen zwischen den N-Grammen enthielt. Das Modell wurde anschließend auf Basis dieser Daten auf vier verschiedenen Wegen trainiert (Burnap & Williams, 2015, 230ff.).

(1) Es wurde untersucht, welche Teile der Datenstruktur typisch für hetzfreie Tweets sind und welche häufig auftauchen, wenn der Tweet von Menschen als hetzend eingeordnet wurde.
(2) Mithilfe einer digitalen Repräsentation von Entscheidungsbäumen wurden Regeln zur Einordnung der Tweets generiert. Dadurch kann verschiedenen Anhaltspunkten unter anderem eine unterschiedliche Gewichtung im Entscheidungsprozess zugeteilt werden.
(3) Der Einsatz einer Support Vector Machine wurde zur möglichen Verbesserung der Varianten (1) und (2) untersucht. Eine Support Vector Machine ist in der Lage Klassen (Hetze / frei von Hetze) so zu bestimmen, dass eine möglichst deutliche Trennung zwischen den Klassen besteht (Suykens et al., 2003, S. 29ff.).
(4) Die Methoden (1), (2) und (3) werden angewandt, anschließend wird das eindeutigste Ergebnis ausgewählt.

Somit wurde dem Modell beigebracht, die Tweets nach dem Empfinden von Menschen auf das Vorhandensein von Hetze zu interpretieren. Anschließend an den Trainingsprozess ließ sich das Modell auf Tweets anwenden, welche zuvor nicht kategorisiert wurden (ebd.).

Dabei wurde festgestellt, dass sich die besten Resultate auf einem Datensatz erzielen lassen, der sowohl die analysierten Beziehungen zwischen N-Grammen als auch N-Gramme, die keine rassistischen Formulierungen aus dem angewandten Lexikon enthielt (a.a.O., S. 233). Dies verdeutlicht, dass Menschen bei der Klassifikation der hetzenden Tweets nicht nur auf bestimmte Begriffe und Formulierungen achten, sondern auch weniger auffällige Merkmale analysiert wurden (ebd.). Die Schlussfolgerungen zeigen außerdem, dass die besten

Ergebnisse mit Methode (4) erzielt wurden. Dies bedeutet wiederum, dass die Methoden (1), (2) oder (3) je nach getestetem Tweet das beste Ergebnis erzielt haben und eine Kombination aller Methoden ratsam ist. Mit Methode (4) konnten Präzisionswerte von 96% bei nicht-hetzenden Tweets, sowie 89% bei hetzenden Tweets erzielt werden (a.a.O., S. 233ff.).

3.5 Conversation AI

Bei der *Conversation AI* handelt es sich um einen gemeinsamen Forschungsansatz von *Jigsaw*, einem Expertenteam, welches sich der Herausforderung stellt, Menschen mithilfe von Technologie zu schützen (Jigsaw, 2019), und Googles *Counter Abuse Technology*-Team. Gemeinsam versuchen die Beteiligten die Möglichkeiten und Grenzen von Machine Learning im Kontext der Verbesserung von Online-Kommunikation zu erforschen (Conversation AI, o.J.).

Dafür werden im Rahmen der Conversation AI zum Beispiel große und hochqualitative Kommunikations-Datensätze erstellt und veröffentlicht[2]. Ein weiteres zentrales Projekt stellt die *Perspective API* dar, eine Online-Schnittstelle, an die Texte gesendet werden können, welche im Anschluss auf verschiedene problematische Kommunikationskategorien untersucht werden (Conversation AI, 2017). Die API antwortet anschließend mit einem Wahrscheinlichkeitswert für jede Kategorie, ob der Text in dieser Hinsicht problematisch ist (ebd.). Neben der englischen Sprache unterstützt die Perspective API mittlerweile auch die französische, spanische, deutsche, italienische und portugiesische Sprache in experimentellen Modi (ebd.). Im Rahmen von Kapitel 4.3.4 werden die Möglichkeiten der Perspective API näher betrachtet, da sie für die Implementation des Browser-Plugins verwendet wird.

Genaue Funktionsweisen der Perspective API sind nicht öffentlich bekannt, es wird lediglich kommuniziert, dass die API auf Maschinellem Lernen basiert und daher wahrscheinlich ähnliche Techniken verwendet werden, die in Kapitel 3.4 beschrieben sind. Die finalen Modelle wurden laut Angaben der Entwickler auf „hunderttausenden Kommentaren trainiert, die von Menschen bewertet wurden“ (Conversation AI, 2020), die experimentellen Modelle auf kleineren Datensätzen. Die Vielzahl an Modellen deutet darauf hin, dass die menschlichen Tester die Datensätze im Hinblick auf die verschiedenen Kategorien eingeordnet haben, sodass anschließend für jede Kategorie ein individueller Klassifikator entwickelt und trainiert werden konnte.

3.6 Vorhandene Browser-Erweiterungen zur Klassifikation von Hate Speech

Der Ansatz die Klassifikation von Hate Speech über Browser-Erweiterungen auf die Seite des Nutzers zu verlagern wurde bereits verfolgt. Mit Tune und Negator werden im Folgenden zwei dieser Ansätze näher betrachtet.

Bei Tune handelt es sich um eine Erweiterung, die anders als das im Rahmen dieser Arbeit entwickelte Plugin nicht für den Firefox-Browser, sondern für Chrome entwickelt wurde. Tune wurde von Jigsaw, der Unterfirma von Google, welche auch an der Perspective API mitarbeitet, entwickelt und nutzt daher ebenfalls ein Modell der API. Die Erweiterung unterstützt ausgewählte Websites wie Facebook, Twitter und Reddit, analysiert allerdings nur englischsprachige Kommentare und diese auch nur auf eines der vielen API-Modelle (Tune, 2019). Außerdem muss eine Verbindung zu einem Google-Konto hergestellt werden. Tune lässt sich über eine moderne und übersichtliche Benutzerschnittstelle konfigurieren, indem verschiedene Websites aktiviert und deaktiviert werden können und mittels eines Reglers ein Schwellenwert zum Ein- und Ausblenden von Kommentaren bestimmt werden kann.

Auch Negator ist eine Browser-Erweiterung für Chrome. Sie sendet den Inhalt von Webseiten an Server der Entwickler, wo diese mittels NLP-Techniken auf das Vorhandensein von Hate Speech analysiert werden. Anschließend werden Stellen, die Hetze enthalten, durch das Plugin zensiert. Auch bei Negator ist es möglich ausgewählte Websites zu aktivieren und zu deaktivieren, sowie einen Schwellenwert für die Zensur von Texten anzugeben. Neben einem klassischen Modell zur Hate Speech-Identifikation spezifischer Textstellen bewertet Negator auch die allgemeine Hate Speech-Intensität eines gesamten Textes (Jain & Deepali, 2020).

3.7 Gesellschaftliche Initiativen zur Bekämpfung von Hate Speech

Neben den technischen, politischen und juristischen Ansätzen zur Bekämpfung von Hate Speech in sozialen Medien gibt es auch Initiativen aus der Gesellschaft heraus, die meist auf freiwilliger Basis aktiv werden. So versucht zum Beispiel die Initiative #ichbinhier in Kommentaren unter Posts bekannter Medien, Hass und Hetze in die Richtung einer sachlichen Diskussion zurückzulenken (Kreißel et al., 2018, S. 9). Die *Landesanstalt für Medien NRW* listet #refugeeswelcome, #mundaufmachen, #heidepack und #NichtEgal als weitere prominente Beispiele für das Eintreten gegen Rassismus und Hetze und stellt fest

„Sich den Hassreden im Netz entgegenzustellen ist eine gesamtgesellschaftliche Aufgabe.“ (Landesanstalt für Medien NRW, AJS, 2019).

4 Entwicklung des Browser-Plugins

Im Rahmen dieses Kapitels wird die Vorgehensweise bei der Implementation des Browser-Plugins beschrieben. Zu diesem Zweck werden erst die wichtigsten technischen Grundlagen und anschließend der Ablauf der Implementation erklärt. Im Unterkapitel 4.3 werden die hauptsächlichen Funktionen und Problemlösungen im Detail analysiert und erläutert. Abschließend wird ein Zwischenfazit zur Implementation des Plugins gezogen.

4.1 Technische Grundlagen

Bei einem Browser-Plugin handelt es sich um Programmstrukturen, mit denen sich die Funktionalitäten von üblichen Web-Browsern erweitern lassen. Üblicherweise und auch in diesem konkreten Beispiel, werden Plugins in der Programmiersprache JavaScript implementiert, welche von allen modernen Browsern ausgeführt werden kann, ohne dass eine Kommunikation mit externen Servern nötig ist.

Da das geplante Browser-Plugin in der Lage sein soll, Posts und Kommentare zu markieren, welche als potenziell gefährdend einzuordnen sind, ist eine Modifikation am Aufbau der Website notwendig. Deren Struktur wird durch die Formatierungssprache HTML (*Hypertext Markup Language*) definiert und das Design der HTML-Elemente mithilfe der Sprache CSS (*Cascading Stylesheet*) festgelegt. JavaScript-Programme sind in der Lage, die HTML-Struktur zu verändern und zusätzlich das festgelegte Design zu modifizieren.

Um eine Klassifikation von Textausschnitten vornehmen zu können, wird das Plugin keine eigenständige Bewertung vornehmen, sondern die in Kapitel 3.5 erwähnte *Perspective API* nutzen. Bei einer API (*Application Programming Interface*) handelt es sich dabei um eine Programmierschnittstelle, die beispielsweise durch den Aufruf einer bestimmten URL angesprochen werden kann. Dabei lassen sich Daten an die Schnittstelle mitliefern. Die API antwortet anschließend auf die gestellte Anfrage in Form eines weiteren Datensatzes, der das Ergebnis für die Anfrage enthält. In diesem konkreten Fall werden die gefundenen Posts und Kommentare an die Perspective API geschickt, ebenso wie einige Metadaten, auf die der Text untersucht werden soll, beispielsweise die zu untersuchende Sprache. Als Antwort liefert die API einen Wahrscheinlichkeitswert zwischen 0 und 1. Dieser Wert gibt an, ob der Text im Hinblick auf verschiedene Hate Speech-Kategorien als problematisch gewertet werden kann (Conversation AI, 2017).

4.2 Planung und Organisation

Da Browser verschiedene Möglichkeiten bieten, um Plugins umzusetzen, die sich im Detail oft unterscheiden, musste vorab eine Entscheidung fallen, für welchen Browser das Plugin programmiert wird. Die Wahl ist dabei auf Mozillas Firefox-Browser gefallen, da dieser effiziente und einfache Möglichkeiten bietet, um Plugins einzuspielen, zu testen, zu konfigurieren und darüber hinaus eine ausführliche Dokumentation über Features des Browsers, sowie praktische Hilfestellungen und Anleitungen bezüglich der Plugin-Implementation zur Verfügung stellt (mozilla.org, 2020). Das Plugin besteht aus einer Sammlung von Dateien, die direkt im Firefox-Browser eingespielt werden können.

Um sich der finalen Implementation iterativ anzunähern, wurde der folgende Plan entworfen und abgearbeitet:

1. Grundlegender Aufbau der zentralen Dateien des Plugins, sowie deren Zusammenspiel erarbeiten
2. Implementation einer Funktion, um Änderungen an den Websites (zum Beispiel das Nachladen neuer Posts und Kommentare) festzustellen
3. Implementation einer „Dummy“-Version, die auf Basis einer Schimpfwortliste Posts kontrolliert und über eine erste Methode markiert
4. Hinzufügen eines Icons in der Kopfleiste des Browsers, inklusive eines Popups, welches die Aktivität bestätigt
5. Erstellen einer Einstellungsseite, mit der sich das Plugin konfigurieren lässt
6. Zugriff auf die Perspective API und Handhabung der Antworten implementieren
7. Verschiedene Verbesserungen und kleine Erweiterungen der Funktion, darunter auch Performance-Verbesserungen, vornehmen

Um den Überblick zu bewahren, wurden die einzelnen Schritte während des Programmierens jeweils noch detaillierter unterteilt. Dabei wurde die Funktionalität neuer Programmteile, immer direkt nach der Implementation über Ausgaben auf der Browser-Konsole, kontrolliert und getestet. Insbesondere nach den oben gelisteten Etappen wurde die Funktionalität ausführlich geprüft, um Fehler nicht mit in den nächsten grundlegenden Schritt zu übernehmen.

4.3 Kernfunktionen und -features

Innerhalb dieses Kapitels werden verschiedene zentrale Funktionen des Plugins näher beschrieben und erklärt. Die folgende Abbildung 2 gibt dabei einen allgemeinen Überblick

über das Zusammenspiel der verschiedenen Kernelemente des Plugins nach der ersten Initialisierung der Variablen.

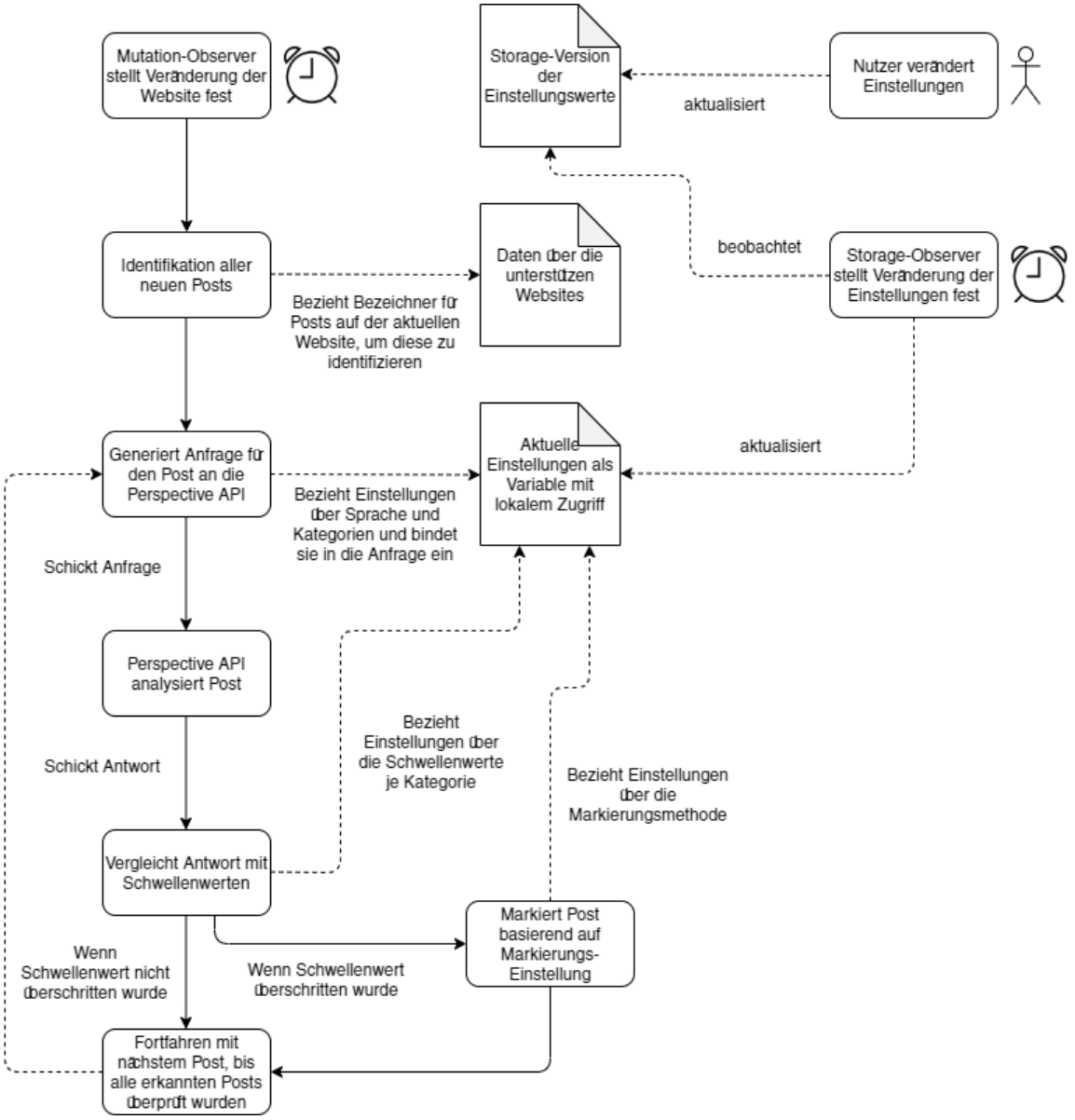

Abbildung 2 – Vereinfachter Programmablauf des Plugins (Eigendarstellung)

Wie das Diagramm zeigt, existieren insgesamt zwei Komponenten innerhalb des Plugins, die einen sequenziellen Prozess einleiten. Dies ist einerseits der MutationObserver, welcher reagiert, sobald Änderungen in der HTML-Struktur der Website festgestellt werden. Außerdem registriert ein Storage-Observer Änderungen an den Einstellungen und stellt diese lokal zur Verfügung. Sobald der MutationObserver eine Änderung festgestellt hat, werden alle Posts auf Basis einer Datei innerhalb des Plugins, welche die Namen der zu untersuchenden HTML-Klassen für jede Website enthält, gesammelt. Diese Posts werden nach einem Filter-Durchgang, der dafür sorgt, dass nur neue Posts weitergereicht werden,

nacheinander in eine Anfrage an die Perspective API konvertiert. Nachdem diese Anfrage abgesendet wurde, antwortet die API mit Wahrscheinlichkeitswerten für verschiedene Kategorien, ob Hetze in dem Post vorhanden ist. Diese Werte werden dann mit Schwellenwerten, die der Nutzer über die Einstellungsseite konfiguriert, verglichen. Liegt der Wahrscheinlichkeitswert in einer der Kategorien über dem vom Nutzer angegebenen Schwellenwert, wird der Post sichtbar auf der Webseite markiert. Dafür stehen verschiedene Markierungsmethoden in den Einstellungen zur Verfügung, zwischen denen der Nutzer wählen kann und welche sowohl im Storage als auch in der lokal lesbaren Kopie der Einstellungen gespeichert werden.

4.3.1 Grundaufbau des Plugins

Ein Plugin ist ein Programm, welches die Funktionalität eines Browsers erweitern kann. Dabei besteht es aus einer Sammlung von Dateien, die miteinander interagieren. Das implementierte Plugin umfasst insgesamt neun Dateien, von denen die vier Hauptdateien im allgemeinen Verzeichnis liegen und einige Hilfsdateien in passenden Ordnern, um die Übersicht zu gewährleisten. Die folgende Liste stellt die Ordner- und Dateistruktur des Plugins dar:

```
¬ icons
    ¬ plugin-icon.png
¬ json
    ¬ websites.json
¬ popup
    ¬ aktivitatBestatigung.css
    ¬ aktivitatBestatigung.html
    ¬ aktivitatBestatigung.js
¬ hatespeechfilter.js
¬ manifest.json
¬ options.html
¬ options.js
```

Bei `plugin-icon.png` handelt es sich um ein simples, selbstentworfenes Icon, welches im Browser und in einer Popup-Meldung angezeigt wird. Die Datei `websites.json` listet Daten der unterstützten Websites, darunter die HTML-Klassen, welche Posts und Kommentare auf den jeweiligen Websites definieren. Die Dateien im `popup`-Ordner sind zuständig für ein kleines Fenster, welches sich öffnet, wenn der Nutzer in der Browser-Kopfleiste auf das Icon des Plugins klickt. Es enthält eine Anleitung, die zur Einstellungsseite leitet. Die Datei `hatespeechfilter.js` beinhaltet den funktionalen Kern des gesamten Plugins und ist für die Beobachtung der Websites, das Filtern von Posts

und die Anfragen an die Perspective API zuständig. In der `manifest.json`-Datei werden grundlegende Konfigurationen, wie Berechtigungen und Metadaten, für den Browser angegeben, welche sie zu einem Pflichtbestandteil eines jeden Plugins macht. Die Dateien `options.html` und `options.js` gestalten die Einstellungsseite, die ebenfalls einer der Hauptbestandteile des Plugins ist, und steuern ihre Funktionalität.

4.3.2 Erkennen neuer Posts und Kommentare

Wie zuvor bereits kurz beschrieben, handelt es sich bei den Abläufen im Browser-Plugin nicht per se um einen fortlaufenden Prozess. Stattdessen wird eine sequenzielle Folge von Aktionen eingeleitet, sobald Änderungen an der Website festgestellt werden, also wenn sich deren HTML-Struktur verändert.

Dies wird über ein standardmäßig in JavaScript enthaltenes Konstrukt ermöglicht, einem MutationObserver (MDN web docs, 2019). Der MutationObserver wird mit verschiedenen Werten konfiguriert, beispielsweise dem zu beobachtenden Bereich der Website. Außerdem wird eine Methode angegeben, die aktiviert wird, sobald der MutationObserver eine Änderung auf der Website registriert. Der MutationObserver wird in Gang gesetzt, sobald das Plugin aktiviert ist.

4.3.3 Zugriff auf Dateien innerhalb des Plugins

Damit das Programm überprüfen kann, ob es auf der aktuell geöffneten Website Posts und Kommentare identifizieren soll, und um die Daten zu erhalten, wie auf der jeweiligen Website Posts erkannt werden können, muss die `hatespeechfilter.js`-Datei Zugriff auf `websites.json` erhalten. Dieser Zugriff ist nach der Installation des Plugins jedoch nicht standardmäßig gegeben. Stattdessen wird der Datei `websites.json` eine URL zugewiesen, über die anschließend darauf zugegriffen werden kann.

```
function createGlobalWebsiteInformation()
{
    // Daten aus websites.json
    var websitesURL = browser.runtime.getURL("json/websites.json"); // URL der websites.json
    var obj;

    fetch(websitesURL) // Asynchroner Zugriff auf die Daten
      .then(res => res.json()) // Website-Daten werden im JSON-Format eingelesen
      .then(data => obj = data)
      .then(() => getOptions())
      .then(() => writeToGlobal(obj)) // Website-Daten werden in lokale, globale Variable gespeichert
      .then(() => createPostlist())
      .then(() => startObserver());
}
```

Abbildung 3 – Zugriff auf Dateien des Plugins und Asynchronität (Code-Ausschnitt)

Wie in Abbildung 3 zu sehen ist, kann die URL von Dateien innerhalb des Plugins über die Methode `browser.runtime.getURL(dateipfad)` (MDN web docs, 2019) angefragt werden. Über die `fetch`-API (MDN web docs, 2020) wird der Inhalt schließlich abgerufen und als JSON-Code interpretiert. Hier wird ein weiteres Konzept von JavaScript deutlich, welches während der Implementation häufig eine entscheidende Rolle spielt. JavaScript ist eine asynchrone Programmiersprache und führt daher weiteren Code aus, obwohl die Durchführung vorheriger Methoden womöglich noch nicht abgeschlossen wurde. Da in diesem Beispiel das Interpretieren der Daten jedoch erst starten kann, wenn der Zugriff vollständig durchgeführt wurde, muss dies mithilfe des `then()`-Befehls (MDN web docs, 2019) verhindert werden. Hiermit wird der Browser gezwungen, die Antwort vorheriger Befehle abzuwarten, bevor weitere Programmschritte durchgeführt werden.

Nachdem die Daten aus der `websites.json`-Datei empfangen wurden, werden sie mithilfe der Methode `writeToGlobal(obj)` in einer globalen Variable gespeichert. Dadurch müssen die Daten nur einmalig abgefragt werden, stehen im Anschluss aber dauerhaft zur Verfügung. Neben der `websites.json`-Datei wird dieses Konzept auch genutzt, um innerhalb des Plugins Zugriff auf das Plugin-Icon zu erhalten.

4.3.4 Perspective API

Die Perspective API (Conversation AI, 2017) wird innerhalb des Plugins angewendet, um die Wahrscheinlichkeit des Vorkommens von Hetze in den identifizierten Posts und Kommentaren zu überprüfen. Dabei werden unterschiedliche „Modelle“ (ebd.) angeboten, auf denen die Texte untersucht werden können. Welche Modelle dabei auf den Text angewendet werden können, ist stark von der Sprache abhängig, welche untersucht werden soll. Insbesondere im Englischen sind die Modelle bereits ausführlicher getestet als in anderen Sprachen, wo sie sich häufig noch in einer experimentellen Phase befinden.

Da es während der Entwicklung des Plugins das Ziel war, dieses bestmöglich und individuell konfigurierbar zu gestalten, werden die Möglichkeiten, welche die API bietet, in vollem Umfang genutzt. Dementsprechend nutzt das Plugin ebenfalls alle Sprachen, welche von der API unterstützt werden. Außerdem werden alle Modelle für die einzelnen Sprachen angeboten, um Posts mithilfe ihrer Anwendung zu analysieren. Die folgende Tabelle 1 bietet einen Überblick darüber, welche Modelle für welche Sprachen verfügbar sind.

	en	fr	es	de	it	pt
TOXICITY	✓	✓	✓	×	×	×
TOXICITY_EXPERIMENTAL	×	×	×	✓	✓	✓
SEVERE_TOXICITY	✓	✓	✓	×	×	×
SEVERE_TOXICITY_EXPERIMENTAL	×	×	×	✓	✓	✓
IDENTITY_ATTACK	✓	×	×	×	×	×
IDENTITY_ATTACK_EXPERIMENTAL	×	✓	✓	✓	✓	✓
INSULT	✓	×	×	×	×	×
INSULT_EXPERIMENTAL	×	✓	✓	✓	✓	✓
PROFANITY	✓	×	×	×	×	×
PROFANITY_EXPERIMENTAL	×	✓	✓	✓	✓	✓
THREAT	✓	×	×	×	×	×
THREAT_EXPERIMENTAL	×	✓	✓	✓	✓	✓
SEXUALLY_EXPLICIT	✓	×	×	×	×	×
FLIRTATION	✓	×	×	×	×	×

Tabelle 1 – Verfügbarkeit der Modelle je Sprache

Da die API neben dem zu analysierenden Text auch Informationen darüber benötigt, auf welche Sprache und Modelle hin der Text untersucht werden soll, wird zuvor für jeden Post eine Datensammlung `data_dict` generiert. Die Konstante `requestedAttributes` speichert dafür alle Modelle, die für die jeweiligen Sprachen existieren. In der Variable `currentOptions` stehen dabei die aktuellen Einstellungen des Plugins zur Verfügung, wie zum Beispiel die Sprache, auf die die Texte untersucht werden sollen. Die tatsächliche Anfrage an die Perspective API wird anschließend abermals mit der `fetch`-API ausgeführt. Die Antwort, welche unter Anderem die Wahrscheinlichkeitswerte je Kategorie enthält, wird dann im Anschluss an die Methode `checkScore(apiresponse, post)` weitergeleitet. Liegt der Wert für eine der Kategorien über dem Grenzwert, der von dem Benutzer eingestellt wurde, wird eine Markierung des jeweiligen Posts durchgeführt.

```
// Überprüft Liste von Posts darauf, ob Hate Speech enthalten ist; stößt Markierung anschließend an.
function checkPosts(posts)
{
    for (const post of posts) // Iterien über die einzelnen Posts
    {
        var requestedAttributes = {};
        for (cat of languageCategories[currentOptions["language"]]) // Zusammenstellen der Modelle für Sprache
        {
            requestedAttributes[optionsToPerspective[cat]] = {};
        }
        const postcontent = post.innerText;
        const data_dict = { // Enthält weitere Daten, die mit der Anfrage an die API gesendet werden
            'comment': {'text': postcontent},
            'languages': [currentOptions["language"]],
            'requestedAttributes': requestedAttributes
        };
        fetch(url, {
          method: 'POST',
          body: JSON.stringify(data_dict)
        })
        .then(response => response.json())
        .then(data => checkScore(data,post))
        .catch(err => console.log(err));
    }
}
```

Abbildung 4 – Generieren der Anfrage an die Perspective API (Code-Ausschnitt)

4.3.5 Einstellungsseite

Da das individuelle Konfigurieren des Plugins eines der wichtigsten Ziele darstellt, handelt es sich bei der Einstellungsseite um einen weiteren Hauptbestandteil der Browser-Erweiterung, welche selbst eine Menge an Funktionalität enthält. Darüber hinaus ist sie die Hauptinteraktionsstelle zwischen dem Benutzer des Plugins und dem Plugin selbst.

Um die Möglichkeiten der Perspective API bestmöglich zu nutzen, sollten auf der Einstellungsseite sowohl die Sprache, auf die Posts analysiert werden, als auch die Schwellenwerte für jedes Modell, welches für die gewählte Sprache zur Verfügung steht, einstellbar sein. Um dies zu ermöglichen, wurde ein Dropdown-Menü implementiert, welches die Auswahl zwischen den sechs verfügbaren Sprachen ermöglicht. Erst auf Basis dieser Auswahl werden Slider für die unterstützten Modelle eingeblendet. Die Slider erlauben eine Einstellung zwischen den Werten 0 und 100, welche für den prozentualen Schwellenwert stehen, mit dem die Antwort von der Perspective API abgeglichen wird. Um es für Nutzer besser nachvollziehbar zu machen, welcher Wert aktuell eingestellt ist, wird das standardmäßige Slider-Element (w3schools.com, o.J.) um eine Skala mit den Werten 0, 25, 50, 75 und 100 erweitert.

Neben den für die Perspective API relevanten Werten kann über ein weiteres Dropdown-Menü eine von vier Methoden ausgewählt werden, die für die Markierung von Posts eingesetzt werden soll (siehe Abbildung 5). Dabei werden ein roter Rahmen um den Text,

ein Verblassen des Textes, die Markierung des Posts mit dem Plugin-Icon, aber auch das Ersetzen des Textes durch eine Meldung des Plugins als Alternativen angeboten.

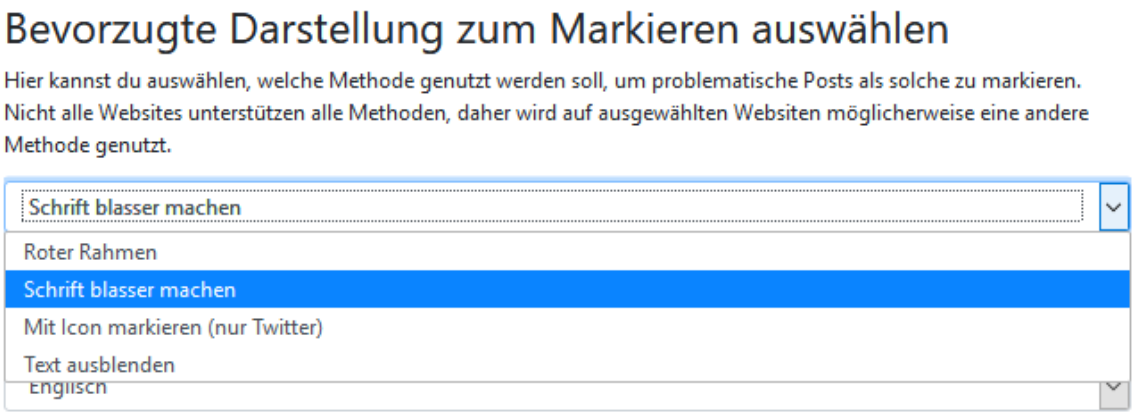

Abbildung 5 – Auswahl der Markierungsmethode (Screenshot)

Für die Gestaltung der Einstellungsseite wird neben eigenem Stylesheet auch das Bootstrap-Framework (Bootstrap Team, 2020) in der Version 4.4.1 eingebunden, welches Designs für eine Vielzahl typischer Anwendungsfälle enthält. Das optische Resultat der Einstellungsseite ist in Abbildung 5 und 6 zu sehen. Dieses Framework wird darüber hinaus in der Popup-Meldung des Plugins verwendet.

Abbildung 6 – Ausschnitt der Einstellungsseite (Screenshot)

Die eingestellten Werte werden intern in Form eines Dictionary gespeichert. Der Wert für jeden einzelnen Slider wird dabei als Integer-Wert dargestellt. Auch die ausgewählte

Methode wird als Zahl repräsentiert, so steht der Wert `1` zum Beispiel für die Möglichkeit des roten Rahmens. Die Sprache wird in Form eines einfachen Strings in Form des Sprachenkürzels gespeichert, für Englisch zum Beispiel "`en`".

Damit jede Änderung an den Einstellungen übernommen wird, wurde jedes der Einstellungselemente mit mindestens einem EventListener (w3schools.com, o.J.) ausgestattet. Sie stoßen die Ausführung einer Methode an, sobald eine Änderung an ihrem Wert festgestellt wurde. Dies ermöglicht nicht nur, dass nur die Slider angezeigt werden, welche zur gewählten Sprache gehören, sondern auch den Verzicht auf einen Button, durch den die Einstellungen gespeichert werden. Denn jede Änderung an den Einstellungen sorgt sofort dafür, dass eine aktuelle Kopie der Einstellungen in den Browser-Storage (MDN web docs, 2019) geladen werden. Dabei handelt es sich um einen Speicherplatz innerhalb des Browsers, auf dessen Daten nur das Plugin selbst zugreifen kann. Im Browser-Storage steht somit immer eine aktuelle Kopie der Einstellungen für den Hauptteil des Plugins zur Verfügung.

Mit einer weiteren Maßnahme wurde schließlich gewährleistet, dass Änderungen an den Einstellungen auch sofort von dem Plugin übernommen werden, selbst wenn es bereits im Einsatz ist. Dafür wird ein weiterer Observer für das Browser-Storage-Objekt implementiert. Sobald es eine Änderung feststellt, wird die globale Variable `currentOptions` aktualisiert.

Sollte es in einem der genannten Schritte zu unerwarteten Fehlern kommen, sind sowohl die Einstellungsseite, als auch das Plugin an sich in der Lage, sich auf ein vordefiniertes Set an Standardeinstellungen zurückzusetzen. Darüber hinaus verfügt die Einstellungsseite über einen Button, der diese Standardeinstellungen aktiviert.

4.3.6 Markierung der Posts

Die Markierung von kritischen Posts wird durch Änderungen an dem Stylesheet des Posts, oder an dessen HTML-Struktur vorgenommen. Dabei haben sich auch nach der Nutzung verschiedener Implementationen nicht alle vier Methoden als kompatibel für die fünf ausgewählten Websites erwiesen. Dies ist mitunter in einem sehr außergewöhnlichen Aufbau der Website mit selbst kreierten HTML-Tags, die Bearbeitungen zu verhindern scheinen, oder sich regelmäßig aktualisieren und Änderungen des Plugins damit überschreiben, begründet. Die folgende Tabelle gibt einen Überblick darüber, welche Websites mit welchen Methoden kompatibel sind. Sollte eine Einstellung nicht auf eine

Website anwendbar sein, wird stattdessen automatisch mithilfe eines roten Rahmens markiert, ohne dass die Einstellungen sich dafür verändern.

	Roter Rahmen	Blasser Text	Icon einblenden	Text ersetzen
Twitter	✓	✓	✓	✓
Facebook	✓	✓	×	✓
Instagram	✓	×	×	✓
YouTube	✓	×	×	✓
Reddit	✓	✓	×	✓

Tabelle 2 – Verfügbarkeit der Markierungsmethoden je Website

Im Rahmen der Implementation von den Markierungsmethoden wurde deutlich, dass sich die Methoden nicht wie erhofft auf alle möglichen Websites erweitern ließen. Darüber hinaus hätten HTML-Analysekenntnisse vorausgesetzt werden müssen, die es dem User ermöglichen könnten, das Plugin eigenständig auf weitere Websites zu erweitern. Pläne und versuchte Implementationen in diese Richtung wurden daher verworfen und die Benutzung des Plugins auf die fünf gelisteten Websites beschränkt.

4.3.7 Performance-Verbesserungen

Da das Plugin zwischenzeitlich spürbar den Browser verlangsamt hat, wurden ein paar Verbesserungen an der Performance implementiert. Dazu zählt zum Beispiel, dass die Website nur einmal je Sekunde auf neue Posts kontrolliert wird. Für den Benutzer macht dies kaum einen spürbaren Unterschied, die Perfomance des Browsers konnte dadurch jedoch deutlich stabilisiert und verbessert werden.

Darüber hinaus werden bereits getestete Posts gesammelt und aus neuen „Scans“ der Website aussortiert. Die Anzahl ausgeführter Programmschritte konnte dadurch deutlich reduziert werden, was ebenfalls deutliche Performance-Verbesserungen zur Folge hatte.

In der finalen Implementation ließen sich keine nennenswerten Verlangsamungen des Browsers feststellen. Es vergeht eine gewisse Zeit, bis neu geladene Posts und Kommentare markiert werden, allerdings ist diese Zeit in aller Regel kürzer als die Lesedauer für den Text.

4.4 Zwischenfazit zur Implementation

Die zuvor beschriebenen Implementationen haben ein funktionales Browser-Plugin hervorgebracht (siehe Abbildungen 7 und 8), welches sich anhand der in Kapitel 1.2 definierten Ziele messen lässt. Hierfür wird im Rahmen von Kapitel 5 ein User Experience-

Experiment durchgeführt, bei dem eine Testgruppe das Plugin ausprobiert und einen Fragebogen beantwortet.

Abbildung 7 – Anwendung des Plugins in den Kommentaren eines Youtube-Videos der AfD[1] (Screenshot mit anschließender Schwärzung der Profile)

Aus funktionaler Sicht wurden die definierten Ziele vollständig erreicht. Das Browser-Plugin agiert wie geplant auf der Seite des Benutzers. Darüber hinaus lässt es sich über die Einstellungsseite sehr individuell konfigurieren und Änderungen an den Einstellungen haben auch spürbare Auswirkungen in den sozialen Medien zur Folge. Darüber hinaus werden diese Einstellungen umgehend angewandt, ohne dass eine Neuinstallation oder auch nur das erneute Laden der Website notwendig ist. Lediglich die nutzerseitige Erweiterung an unterstützten Webseiten, welche zwischenzeitlich in Betracht gezogen wurde, konnte nicht so umgesetzt werden, dass sie für durchschnittliche Benutzer ohne Fachkenntnisse anwendbar gewesen wäre.

[1] Aufrufbar über https://www.youtube.com/watch?v=GWO-3o4o7-c, zuletzt aufgerufen am 17.03.2020

Abbildung 8 – Anwendung des Plugins in den Kommentaren eines Tagesschau-Tweets[1], welcher über Schüsse auf eine Shisha-Bar berichtet (Screenshot mit anschließender Schwärzung der Profile)

[1] Aufrufbar über https://twitter.com/tagesschau/status/1231335492676997126

5 User Experience-Experiment

Im Rahmen eines User Experience-Experiments wurde das entwickelte Plugin Testern zur Verfügung gestellt, um mithilfe eines Fragebogens Verbesserungspotenziale, aber auch Stärken des Programms zu ermitteln. In diesem Kapitel wird zunächst die Durchführung des Experiments beschrieben, um anschließend die Antworten der Tester zu evaluieren.

5.1 Durchführung des Experiments

Für die Durchführung des Experiments wurde das Plugin insgesamt 12 Testern zur Verfügung gestellt. Sie wurden darüber informiert, welche Ziele das Plugin verfolgt und darum gebeten, es lokal auf ihrem PC einzuspielen. Da das Plugin nicht etwa im Mozilla-Store zur Verfügung steht, muss das Plugin von den Testern temporär im Browser installiert werden. Da dieser Prozess kein Teil des eigentlichen Experiments darstellte, wurden bis zu diesem Punkt Hilfestellungen geleistet, falls sie benötigt wurden. Sobald das Plugin jedoch erfolgreich eingespielt wurde, gab es keine weitere Unterstützung, um die Umfrage im Hinblick auf die Verständlichkeit des Programms nicht zu verfälschen.

Die Tester wurden gebeten das Plugin auf den unterstützten Webseiten mit verschiedenen Einstellungen zu nutzen, bis sie sich einen guten Überblick über die vorhandenen Funktionen und deren Einsatz machen konnten. Anschließend füllten die Tester den aufgesetzten Fragebogen mit der Bitte aus, das Plugin intuitiv zu bewerten.

Der Fragebogen wurde dabei in zwei Hälften unterteilt. Im ersten, deutlich ausführlicheren Teil. wurden Fragen bezüglich der Benutzung des Plugins, also zum Beispiel darüber, wie verständlich und intuitiv die Konfiguration war, gestellt. Dafür wurde den Nutzern ein standardmäßiger Usability-Fragebogen, der sogenannte User Experience Questionnaire, zur Verfügung gestellt (Team UEQ, 2018). Mithilfe des Fragebogens konnten die Tester das Plugin in 26 Kategorien auf einer Skala von 1 bis 7 zwischen zwei gegensätzlichen Emotionen bewerten. Im Anschluss folgten sieben Fragen bezüglich der allgemeinen Funktionalität und Zielerfüllung der Hate Speech-Erkennung, die gezielt in Bezug auf das vorgelegte Plugin ausgewählt wurden. Der Fokus der Umfrage sollte bewusst auf dem ersten Teil liegen, da bezüglich der Funktionalität mit der Perspective API ein externer Dienst angesprochen wurde, auf dessen Einschätzungen das Plugin keinen Einfluss ausübt. Darüber hinaus ist die Entscheidung, welche Posts und Kommentare markiert wurden, stark von den getroffenen Einstellungen des jeweiligen Nutzers abhängig.

5.2 Evaluation

Für den ersten Teil der Umfrage, welcher Fragen zu der Benutzung des Plugins beinhaltet, lautete die Fragestellung immer gleich: „Die Benutzung und Steuerung des Plugins war...“. Die Tester konnten anschließend einen Wert zwischen 1 und 7 auswählen, der ihre Erfahrung in der jeweiligen Kategorie am zutreffendsten beschreibt. Diese Werte wurden für die Evaluation auf eine Skala zwischen -3 und +3 konvertiert. Ein mittelmäßiger Wert liegt somit bei exakt 0, alle größeren Werte stellen ein positives Ergebnis dar und alle kleineren Werte ein negatives Ergebnis.

Bedeutung Wert +3	Bedeutung Wert -3	Arithmetisches Mittel	Varianz	Standardabweichung
gut	schlecht	2,2	0,5	0,7
verständlich	unverständlich	2,0	0,9	1,0
übersichtlich	verwirrend	2,0	1,8	1,3
schnell	langsam	1,8	0,9	0,9
erwartungskonform	nicht erwartungskonform	1,8	0,6	0,8
aufgeräumt	überladen	1,8	3,7	1,9
interessant	uninteressant	1,7	0,8	0,9
sicher	unsicher	1,7	1,3	1,2
erfreulich	unerfreulich	1,6	0,4	0,7
leicht zu lernen	schwer zu lernen	1,6	1,9	1,4
angenehm	unangenehm	1,6	1,2	1,1
effizient	ineffizient	1,5	1,4	1,2
pragmatisch	unpragmatisch	1,5	1,0	1,0
sympathisch	unsympathisch	1,5	0,8	0,9
wertvoll	minderwertig	1,4	1,0	1,0
attraktiv	unattraktiv	1,4	1,7	1,3
unterstützend	behindernd	1,2	1,8	1,3
einfach	kompliziert	1,2	2,3	1,5
voraussagbar	unberechenbar	1,1	1,9	1,4
kreativ	phantasielos	1,0	0,5	0,7
spannend	langweilig	0,8	1,6	1,3
aktivierend	einschläfernd	0,8	2,0	1,4
anziehend	abstoßend	0,7	1,2	1,1
neuartig	herkömmlich	0,6	1,5	1,2
innovativ	konservativ	0,6	1,2	1,1
originell	konventionell	0,4	2,3	1,5

Tabelle 3 – Auswertung der Usability-Fragen

Der User Experience Questionnaire-Fragebogen teilt die verschiedenen Attribute darüber hinaus in Oberkategorien ein, die eine übersichtliche Analyse der Stärken und Schwächen des getesteten Programms, erlauben (Team UEQ, 2018).

Oberkategorie	Arithmetisches Mittel	Varianz
Durchschaubarkeit	1,688	1,04
Effizienz	1,646	0,52
Attraktivität	1,486	0,47
Steuerbarkeit	1,417	0,70
Stimulation	1,188	0,90
Originalität	0,646	0,79

Tabelle 4 – Auswertung der Usability-Fragen nach Oberkategorien

Bezüglich der Fragen hinsichtlich der Funktionalität und Zielerreichung des Plugins konnten verschiedene Thesen zwischen dem Wert 1 (stimme vollkommen zu) und 7 (stimme gar nicht zu) bewertet werden. Auch diese Ergebnisse wurden für die Evaluation auf eine Skala von -3 für negative Bewertungen bis +3 für positive Bewertungen konvertiert.

These	Arithmetisches Mittel	Varianz	Standardabweichung
Das Plugin bietet zufriedenstellende Einstellungsmöglichkeiten.	2,6	0,58	0,76
Ich kann das Plugin durch die Einstellungen auf meine persönlichen Präferenzen konfigurieren.	2,2	1,31	1,14
Das Plugin regt mich zu einem aufmerksameren Umgang mit Posts und Kommentaren in sozialen Netzwerken an.	1,8	1,47	1,21
Durch die Benutzung des Plugins wurde mir besser bewusst, dass Hass und Hetze in sozialen Netzwerken allgegenwärtig sind.	1,8	0,81	0,90
Die markierten Posts enthielten einen höheren Teil an Hass und Hetze als die nicht markierten Posts.	1,6	1,74	1,32
Die Auswahl der markierten Posts spiegelte die vorgenommenen Einstellungen wider.	1,3	1,86	1,36
Das Plugin gibt mir ein erhöhtes Sicherheitsgefühl bei der Nutzung der sozialen Netzwerke.	0,2	1,31	1,14

Tabelle 5 – Auswertung der Fragen zur Funktionalität und Zielerreichung

Bezüglich der Interpretation der Daten wird deutlich, dass sich im arithmetischen Mittel bei jeder Usability-Frage und in der Folge auch in jeder Usability-Oberkategorie ein positiver Wert ergibt. Durchschnittlich sind die Tester also hinsichtlich jeder Frage positiv eingestellt. Besonders gut wird die Benutzbarkeit des Plugins in Hinblick auf die Verständlichkeit und Übersichtlichkeit bewertet. Der beste Wert (2,2) wird sogar bei der sehr allgemeinen Bewertung zwischen „gut“ (+3) und „schlecht“ (-3) erzielt. Die Testgruppe zieht somit im Allgemeinen ein positives Fazit bezüglich der Benutzung und Steuerung des Plugins.

Die schwächsten Werte in den Usability-Fragen ergeben sich bei den Attributen spannend (0,8), originell (0,4), anziehend (0,7), neuartig (0,6), aktivierend (0,8) und innovativ (0,6). Es wird also deutlich, dass ein gewisses Innovationsgefühl hinsichtlich der Benutzung des Plugins ausbleibt, allerdings könnte dieser Fakt auch für die guten Bewertungen hinsichtlich

der Verständlichkeit und Übersichtlichkeit gesorgt haben. Insbesondere auf der Einstellungsseite wurde darauf geachtet, dass die vielseitigen Konfigurationen anschaulich dargestellt werden. Daher wurden Elemente genutzt, die den meisten Benutzern bekannt vorkommen. Dennoch sind auch diese schwächeren Werte deutlich im positiven Bereich angesiedelt und können damit als zufriedenstellende Resultate angesehen werden.

Bezüglich der Varianz in den Antworten sticht insbesondere der Wert in der Kategorie „aufgeräumt – überladen“ (3,7) heraus. Grund hierfür sind zwei besonders auffällige Antworten. Während alle anderen Tester hier zwischen den Werten +1 und +3 entschieden, bewertete ein Nutzer in dieser Kategorie mit dem Wert -3 und ein weiterer mit -1. Durch den Blick auf die Antworten der anderen Tester erscheint es möglich, dass die deutlich hervorstechende Antwort mit dem schlecht möglichsten Wert (-3) auf einen Fehler des Testers zurückzuführen sein könnte. Dagegen spricht jedoch, dass der Tester vergleichbare Kategorien wie etwa „übersichtlich – verwirrend“ mit einem ähnlichen Wert (-2) bewertet hat. Offenbar wird das Plugin in dieser Hinsicht also sehr unterschiedlich von verschiedenen Testern bewertet.

Betrachtet man die Bewertung der vom Fragebogen vorgegebenen Oberkategorien, wird die Analyse hinsichtlich der Originalität des Plugins bestätigt, hierbei handelt es sich um den schwächsten Wert im Durchschnitt (0,646). Am besten bewerteten die Tester in den Oberkategorien Durchschaubarkeit (1,688) und Effizienz (1,646). Da die Einstellungen des Plugins sehr vielfältig und damit auf den ersten Blick hin hätten verwirrend sein können, bestätigt vor allem das Ergebnis bezüglich der Durchschaubarkeit, dass die Entscheidungen, die Einstellungsseite mit ausführlichen Beschreibungen auszustatten und nur die Werte einzublenden, die zur gewählten Sprache passen, sinnvoll getroffen wurden.

Hinsichtlich der Daten zur Funktionalität und Zielerreichung fällt auf, dass sich im Durchschnitt aller Antworten auch hier für jede Frage ein positiver Wert ergibt. Die Ergebnisse machen deutlich, dass die Einstellungsseite des Plugins von den Testern sehr gut aufgenommen wurde und die Vielfalt an Einstellungen erreicht beinahe den möglichen Bestwert (2,6). Außerdem stimmen die Nutzer deutlich zu, dass sich das Plugin auf persönliche Präferenzen konfigurieren lässt (2,2). Die Funktionalität der API wurde ebenfalls bewertet, dabei stellen die Tester einen höheren Anteil an Hass und Hetze in den markierten Posts fest (1,6) und nehmen Änderungen an den Einstellungen auch anhand der Markierungen wahr (1,3). Mit besseren Werten konnte in dieser Hinsicht kaum gerechnet

werden, da auf der offiziellen Website der Perspective API offen kommuniziert wird, dass sie nicht fehlerfrei arbeitet (Google, 2020). Bezüglich der Sensibilisierung für Hass und Hetze in sozialen Netzwerken, sind der Umfrage gute Ergebnisse zu entnehmen (je 1,8); eine spürbare Erhöhung des Sicherheitsgefühls durch die Benutzung des Plugins konnte jedoch nicht festgestellt werden (0,2). Bezüglich der Varianz und Standardabweichung der Ergebnisse in dieser Kategorie liegen keine auffällig hohen Werte vor.

In der Gesamtbetrachtung lässt sich somit feststellen, dass das Plugin durchweg positiv bewertet wurde. Die Einstellungsmöglichkeiten wurden von den Testern dabei sehr positiv bewertet und die Funktionalität wurde im erwartbaren Raum bestätigt. Das Ziel, zu einer Sensibilisierung für Hass und Hetze in sozialen Netzwerken beizutragen, wird vom Plugin erfüllt. Die Bedienung des Plugins wurde darüber hinaus durchweg positiv bewertet und die durchschnittliche Originalität durch eine gute Durchschaubarkeit und Effizienz ausgeglichen. Die Ergebnisse der Umfrage können insgesamt als sehr zufriedenstellend bewertet werden.

6 Fazit und Ausblick

Im Rahmen dieser Arbeit wurde deutlich, dass Hate Speech nicht grundlos zu einem Dauerthema im politischen Diskurs und der öffentlichen Debatte wurde. Es handelt sich um ein derartig vielfältiges Thema, dass Juristen und Politiker sich mit der Entwicklung wirksamer Gegenmaßnahmen schwertun. Und dennoch sind Aufklärung, Sensibilisierung und Bekämpfung im Kontext von Hass und Hetze in Zeiten der sozialen Medien unverzichtbar, denn deren Folgen können in vielerlei Hinsicht gefährlich sein. Dass Hate Speech in sozialen Netzwerken bereits bewusst und kontrolliert eingesetzt wird, um Einfluss auf Politik und Gesellschaft zu nehmen, unterstreicht die Dringlichkeit diesbezüglich. Das gilt auch für die spürbare Zunahme an Straftaten, die im Kontext von Hass und Hetze stehen.

Mit der Entwicklung eines Browser-Plugins zur nutzerseitigen Analyse von Social Media-Posts auf Hate Speech wurde ein Ansatz im Rahmen des Natural Language Processings verfolgt, um zur Sensibilisierung für die Thematik beizutragen. Durch die Nutzung der Perspective API ist das entwickelte Plugin bereits in der Lage, erste Modelle in der internationalen Hate Speech-Erkennung anzuwenden. Dennoch gäbe es zahlreiche Möglichkeiten, um die Funktionalität des Plugins zu erweitern; dazu zählt zum Beispiel die verworfene Idee der individuellen Erweiterung um weitere Websites oder das Entwickeln verschiedener Nutzungsprofile mit maßgeschneiderten Funktionen. Damit könnte das Plugin neben einem privaten Nutzen auch gezielt auf die Moderation von Kommentaren, zur Datensammlung über Hetze in sozialen Medien und auf weitere Anwendungsfälle angepasst werden. Die im Rahmen dieser Arbeit angestrebten Funktionen wurden durch das User Experience-Experiment als zufriedenstellend bestätigt. Insbesondere die Anpassung an die individuellen Nutzerbedürfnisse wurde sehr positiv angenommen.

Während der Recherche und der Implementation wurde deutlich, dass bezüglich der automatisierten Erkennung von Hass und Hetze in sozialen Netzwerken noch zahlreiche Potenziale existieren, in deren Richtung weiter geforscht werden muss, um exaktere Resultate zu erzielen. Einer der Schwerpunkte sollte es dabei sein, vorhandene Techniken auf weitere Sprachen anzuwenden und sie durch neue Erkenntnisse zu verfeinern und zu verbessern. Insbesondere in den sozialen Medien sind auch das Einbeziehen von Metadaten, zum Beispiel aus dem Userprofil, oder die Analyse angehängter Bilder ein vielversprechender Ansatz, um die Intention von Posts besser einordnen zu können.

Literaturverzeichnis

Alkiviadou, N. (2017). Regulating Hate Speech in the EU. In S. Assimakopoulos, F. Baider, & S. Millar, *Online Hate Speech in the European Union - A Discourse Analytic Perspective* (S. 6-9). Cham: Springer.

Biermann, K., Hommerich, L., Musharbash, Y., & Polke-Majewski, K. (9. Oktober 2019). *Attentäter mordete aus Judenhass*. Von https://www.zeit.de/gesellschaft/zeitgeschehen/2019-10/anschlag-halle-helmkamera-stream-einzeltaeter/komplettansicht abgerufen, zuletzt am 17. Februar 2020.

Boeselager, M. (4. Juni 2019). *So hasserfüllt war die rechtsextreme Kampagne gegen den erschossenen CDU-Politiker*. Von https://www.vice.com/de/article/mb85bq/walter-luebcke-tot-so-hasserfullt-war-die-rechtsextreme-kampagne-gegen-erschossenen-cdu-politiker abgerufen, zuletzt am 17. Februar 2020.

Bootstrap Team. (2020). *Bootstrap*. Von https://getbootstrap.com abgerufen, zuletzt am 26. Februar 2020.

Bundesamt für politische Bildung. (12. Juli 2017). *Was ist Hate Speech?* Von https://www.bpb.de/252396/was-ist-hate-speech abgerufen, zuletzt am 17. Februar 2020.

Burnap, P., & Williams, M. (2015). *Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making.* In *Policy & Internet 7.2*. Malden, Oxford: Wiley Perdiodicals, Inc.

Conversation AI. (2017). *Perspective Comment Analyzer API documentation*. Von https://github.com/conversationai/perspectiveapi abgerufen, zuletzt am 28. April 2020.

Conversation AI. (27. März 2020). *Attributes*. Von https://github.com/conversationai/perspectiveapi/blob/master/2-api/models.md abgerufen, zuletzt am 28. April 2020.

Conversation AI. (kein Datum). *Conversation AI*. Von https://conversationai.github.io abgerufen, zuletzt am 24. März 2020.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language.* In *Eleventh International AAAI Conference on Web and Social Media*.

de Marneffe, M.-C., & Man, C. (2008). *Stanford typed dependencies manual.*

Der Tagesspiegel. (10. Mai 2016). *Rodrigo Duterte ist der Donald Trump der Philippinen*. Von https://www.tagesspiegel.de/politik/staatspraesident-mit-unflaetigen-spruechen-rodrigo-duterte-ist-der-donald-trump-der-philippinen/13576662.html abgerufen, zuletzt am 17. Februar 2020.

Echikson, W., & Knodt, O. (2018). *Germany's NetzDG: A key test for combatting online hate.* Brüssel: Counter Extremism Project.

Friesel, E. (2013). Juden-Hass gestern und heute: Ein historischer Blick auf 130 Jahre judeophobische Feindseligkeit. In J. Meibauer, *Hassrede/Hate Speech - Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (S. 17-27). Gießener Elektronische Bibliothek.

Gensing, P. (10. Oktober 2019). *Dokumente des Hasses*. Von https://www.tagesschau.de/inland/halle-taeter-107.html abgerufen, zuletzt am 17. Februar 2020.

Gensing, P. (4. Juni 2019). *Rechtsextreme verhöhnen Getöteten*. Von https://www.tagesschau.de/inland/rechtsextreme-regierungspraesident-101.html abgerufen, zuletzt am 17. Februar 2020.

Google. (17. Februar 2020). *Google Trends - Vergleichen.* Von https://trends.google.de/trends/explore?date=all&geo=DE&q=hate%20speech,hassrede,hetze abgerufen, zuletzt am 29. Februar 2020.

Google. (2020). *Perspective API - FAQs*. Von https://support.perspectiveapi.com/s/article/perspective-faqs abgerufen, zuletzt am 4. März 2020.

Grundgesetz. (1949). *Grundgesetz für die Bundesrepublik Deutschland Art 1*. Von https://www.gesetze-im-internet.de/gg/art_1.html abgerufen, zuletzt am 17. Februar 2020.

Häring, M., Loosen, W., & Maalej, W. (2018). *Who is Adressed in this Comment? Automatically Classifying Meta-Comments in News Comments.* Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 67.

Jain, S., & Deepali, K. (2020). *Hate Speech Detector: Negator.* Delhi: International Conference On Innovative Computing And Communication (ICICC-2020).

Jigsaw. (2019). *Wie kann man mit Technologie die Menschen auf der Welt schützen?* Von https://jigsaw.google.com/vision/ abgerufen, zuletzt am 16. Februar 2020.

Käppner, J., & Bovermann, P. (9. Oktober 2019). *Eine Stadt geht in Deckung*. Von https://www.sueddeutsche.de/politik/halle-synagoge-schuesse-1.4633708 abgerufen, zuletzt am 17. Februar 2020.

Kreißel, P., Ebner, J., Urban, A., & Guhl, J. (2018). *Hass auf Knopfdruck - Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz.* London; Washington DC; Amman; Beirut; Toronto: Institute for Strategic Dialogue.

Landesanstalt für Medien NRW. (2019). *Forsa-Befragung zur Wahrnehmung von Hassrede.* Von https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html abgerufen, zuletzt am 29. Februar 2020.

Landesanstalt für Medien NRW, AJS. (2019). *Hate Speech - Hass im Netz. Informationen für Fachkräfte und Eltern.* Düsseldorf: Landesanstalt für Medien NRW.

Marker, K. (2013). Know Your Enemy. Zur Funktionalität der Hassrede für wehrhafte Demokratien. In J. Meibauer, *Hassrede/Hate Speech - Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (S. 59-94). Gießener Elektronische Bibliothek.

Marx, I. (19. Februar 2020). *Verschärfte Gesetze gegen Hass und Hetze*. Von https://www.tagesschau.de/inland/hasskriminalitaet-internet-101.html abgerufen, zuletzt am 20. Februar 2020.

McTear, M., Callejas, Z., & Griol, D. (2016). *The Conversational Interface - Talking to Smart Devices.* Schweiz: Springer.

MDN web docs. (29. Oktober 2019). *MutationObserver*. Von https://developer.mozilla.org/de/docs/Web/API/MutationObserver abgerufen, zuletzt am 24. März 2020.

MDN web docs. (23. März 2019). *Promise.prototype.then()*. Von https://developer.mozilla.org/de/docs/Web/JavaScript/Reference/Global_Objects/Promise/then abgerufen, zuletzt am 24. März 2020.

MDN web docs. (18. März 2019). *runtime.getURL()*. Von https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/API/runtime/getURL abgerufen, zuletzt am 24. März 2020.

MDN web docs. (23. März 2019). *Web Storage API*. Von https://developer.mozilla.org/de/docs/Web/API/Web_Storage_API abgerufen, zuletzt am 24. März 2020.

MDN web docs. (6. März 2020). *Using Fetch*. Von https://developer.mozilla.org/en-US/docs/Web/API/Fetch_API/Using_Fetch abgerufen, zuletzt am 24. März 2020.

Meibauer, J. (2013). Hassrede - von der Sprache zur Politik. In J. Meibauer, *Hassrede/Hate Speech - Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (S. 1-16). Gießener Elektronische Bibliothek.

mozilla.org. (18. Februar 2020). *Browser Extensions*. Von https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions abgerufen, zuletzt am 20. Februar 2020.

Schellenberg, B. (18. Oktober 2018). *Rechtspopulismus im europäischen Vergleich – Kernelemente und Unterschiede*. Von https://www.bpb.de/politik/extremismus/rechtspopulismus/240093/rechtspopulismus-im-europaeischen-vergleich-kernelemente-und-unterschiede abgerufen, zuletzt am 17. Februar 2020.

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing.* Valencia: Association for Computational Linguistics.

Sirsch, J. (2013). Die Regulierung von Hassrede in liberalen Demokratien. In J. Meibauer, *Hassrede/Hate Speech - Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (S. 165-193). Gießener Elektronische Bibliothek.

Smola, A., & S.V.N., V. (2008). *Introduction to Machine Learning.* Cambridge: Cambridge University Press.

Spiegel. (19. Februar 2020). *Ermittler finden bei Terrorverdächtigem Chemikalien*. Von https://www.spiegel.de/politik/deutschland/mutmassliche-rechtsextreme-zelle-

ermittler-finden-chemikalien-bei-terrorverdaechtigem-a-605cf2c3-1329-48a9-9230-6a9f76f68048 abgerufen, zuletzt am 20. Februar 2020.

Spiegel. (15. Februar 2020). *Terrorverdächtige kommen in Untersuchungshaft*. Von https://www.spiegel.de/politik/deutschland/ermittlungen-gegen-rechtsextreme-terrorverdaechtige-kommen-in-untersuchungshaft-a-7d1cd5c3-4476-4a0b-a004-72acd87ba731 abgerufen, zuletzt am 20. Februar 2020.

Sponholz, L. (2018). *Hate Speech in den Massenmedien.* VS Verlag für Sozialwissenschaften.

Suykens, J., Van Gestel, J., De Brabanter, J., De Moor, B., & Vandewalle , J. (2003). *Least Squares Support Vector Machines.* Singapur: World Scientific Publishing Co. Pte. Ltd.

Tagesschau. (20. Februar 2020). *Was über den Anschlag in Hanau bekannt ist*. Von https://www.tagesschau.de/inland/faq-hanau-101.html abgerufen, zuletzt am 20. Februar 2020.

Team UEQ. (2018). *User Experience Questionnaire*. Von https://www.ueq-online.org abgerufen, zuletzt am 1. März 2020.

Tune. (17. Dezember 2019). *Tune (experimental)*. Von Chrome webstore: https://chrome.google.com/webstore/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp abgerufen, zuletzt am 28. April 2020.

Universität Hamburg. (2018). *Semi-automated analysis, aggregation, and visualization of user comments*. Von https://scan.informatik.uni-hamburg.de/forum40/ abgerufen, zuletzt am 18. Februar 2020.

w3schools.com. (kein Datum). *HTML <input type="range">*. Von https://www.w3schools.com/tags/att_input_type_range.asp abgerufen, zuletzt am 24. März 2020.

w3schools.com. (kein Datum). *JavaScript HTML DOM EventListener*. Von https://www.w3schools.com/js/js_htmldom_eventlistener.asp abgerufen, zuletzt am 24. März 2020.

Wiedemann, G., Ruppert, E., & Biemann, C. (2019). *UHH-LT at SemEval-2019 Task 6: Supervised vs. Unsupervised Transfer Learning for Offensive Language Detection.* In *Proceedings of the 13th International Workshop on Semantic Evaluation* (S. 782-787). Minneapolis: Association for Computational Linguistics.

Wiedemann, G., Ruppert, E., Jindal, R., & Biemann, C. (2018). *Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter.* In *Proceedings of GermEval Task 2018, 14th Conference on Natural Language Processing* (S. 85-94). Wien: KONVENS 2018.

Wiedemann-Schmidt, W. (25. Oktober 2019). *Polizei stuft Münchner Attentat doch als rechtsextrem ein*. Von https://www.spiegel.de/panorama/justiz/muenchen-ermittler-stufen-oez-attentat-doch-als-rechtsextrem-ein-a-1293401.html abgerufen, zuletzt am 17. Februar 2020.

Zeit Online. (11. Juli 2019). *Identitäre Bewegung ist eindeutig rechtsextremistisch*. Von https://www.zeit.de/politik/deutschland/2019-07/verfassungsschutz-identitaere-bewegung-rechtsextremismus-einstufung abgerufen, zuletzt am 26. März 2020.

Ziegler, J.-P. (11. Januar 2020). *Der Genosse, der sich bewaffnen will*. Von https://www.spiegel.de/panorama/gesellschaft/kamp-lintfort-christoph-landscheidt-ein-buergermeister-in-angst-a-25301960-1ba4-4a0d-b75c-991475a39b46 abgerufen, zuletzt am 20. Februar 2020.