Detektion von Directed Hate Speech, Online Harassment und Cyberbullying in Online Communities

Dissertation

zur Erlangung des Grades

Doktor der Wirtschaftswissenschaft (Dr. rer. pol.)

der Juristischen und Wirtschaftswissenschaftlichen Fakultät
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

M.Sc. WI Uwe Bretschneider

Halle (Saale),

März 2017

Danksagung

Bei meinem Doktorvater Herrn Prof. Dr. Ralf Peters möchte ich mich für das entgegengebrachte Vertrauen und die Unterstützung bei der Bearbeitung dieses Forschungsthemas bedanken. Besonders dankbar bin ich für die Diskussionsbereitschaft und sowie die moralische Unterstützung bei der Erstellung der Arbeit. Im Besonderen möchte ich mich für die Freiheiten bedanken, die mir bezüglich der Ausgestaltung und Durchführung des Forschungsprojekts gewährt wurden.

Mein besonderer Dank gilt auch Herrn Prof. Dr. Stefan Sackmann für die Bereitschaft, als Gutachter dieser Promotionsarbeit zu fungieren. Weiterhin möchte ich mich recht herzlich für die Diskussionen und Ratschläge bedanken, die mir im Laufe meines Promotionsstudiums immer wieder zu neuen Ideen und Impulsen verholfen haben.

Weiterhin möchte ich mich bei meinen Kollegen, Dr. Thomas Wöhner und Sebastian Köhler für das sehr angenehme Arbeitsklima bedanken. Die Hilfs- und Diskussionsbereitschaft meiner Kollegen haben dazu beigetragen, dieses Forschungsprojekt voranzutreiben. Darüber hinaus möchte ich mich bei meinen Kollegen vom Lehrstuhl für Betriebliches Informationsmanagement für die angenehme und inspirierende Zusammenarbeit bedanken.

Schließlich möchte ich meinen Dank an meine Familie richten, die mich stets in dem Bemühen, diese Dissertationsarbeit fertigzustellen, unterstützt hat. Einen besonderen Dank richte ich an meine Lebensgefährtin Anja. Vielen Dank für den Rückhalt und die Ermutigung mich an Dinge zu wagen, die zunächst viel zu groß erscheinen.

Zusammenfassung

Online Communities sind Plattformen im Web, die es Nutzern ermöglichen, miteinander zu kommunizieren, digitale Inhalte verschiedenster Art auszutauschen und sich untereinander zu vernetzen. Digitale Inhalte verbreiten sich in *Online Communities* aufgrund der Netzwerkstruktur schnell und mit hoher Reichweite. Zudem sind sie häufig öffentlich einsehbar und können unter anonymen oder pseudonymen Zugängen erstellt werden. Aufgrund einer fehlenden Prüfung dieser Inhalte besteht die Gefahr des Missbrauchs der Plattformen für beleidigende oder hasserfüllte Kommunikation. Derartige Kommunikationsformen werden unter dem Sammelbegriff *Hate Speech* subsummiert. *Hate Speech* stellt ein zunehmend relevantes Problem in *Online Communities* dar. Der steigende Umfang von *Hate Speech* veranlasst *Online Communities* zur Erstellung von Policies zur Sicherstellung eines normgerechten Umgangs unter den Nutzern. Die Einhaltung der Policies wird jedoch typischerweise nicht automatisiert sichergestellt. Stattdessen setzen die Plattformen Personal ein oder verlassen sich auf die Nutzer selbst, um Regelverstöße zu erkennen. Aufgrund der Vielzahl an Nachrichten ist diese manuell durchgeführte Aufgabe arbeits- und kostenintensiv.

In der Forschungsliteratur existieren Lösungsvorschläge zum Umgang mit *Hate Speech* auf Basis von Systemen der Informationstechnik, die derartige Inhalte automatisch mithilfe von Klassifikatoren detektieren. Existierende Klassifikatoren aus der Forschung vernachlässigen die Detektion der referenzierten Opfer von *Hate Speech*. Dadurch weisen sie bezüglich der Klassifikationsgüte Schwächen auf. Sie stützen sich auf die Präsenz beleidigender Wörter als zentrales Klassifikationsmerkmal und vernachlässigen die Analyse des sprachlichen Kontextes. Die Präsenz von beleidigenden Wörtern ist jedoch nicht hinreichend, um *Hate Speech* präzise zu erkennen, da diese Wörter in einem Kontext stehen und gegen Individuen oder Gruppen gerichtet werden. Das Ziel der Dissertation besteht deshalb darin, Verfahren zur automatischen Detektion von *Hate Speech* in *Online Communities* einschließlich der referenzierten Opfer unter Berücksichtigung des sprachlichen Kontextes zu konzipieren und zu implementieren. Dabei werden diejenigen Formen von *Hate Speech* fokussiert, die sich gegen menschliche Ziele richten.

Im Rahmen der vorliegenden Dissertation wurden drei aufeinander aufbauende Verfahren konzipiert, implementiert und evaluiert, um *Hate Speech* gegenüber menschlichen Zielen in englischen und deutschen Texten automatisch zu detektieren. Die entwickelten Verfahren basieren auf einem *Sequenzmodell* zur Strukturierung von Texten und einem Pattern-basierten Ansatz zur Detektion von *Hate Speech* einschließlich der referenzierten Opfer. Die

Verwendung des *Sequenzmodells* erlaubt die Berücksichtigung des sprachlichen Kontextes, da es die Reihenfolge der Wörter erhält. Dadurch lassen sich Texte hinsichtlich syntaktischer Verbindungen zwischen für *Hate Speech* typischen Wörtern und Referenzen zu Individuen oder Gruppen untersuchen. Diese Verbindungen werden durch Patterns modelliert, die im Rahmen der Klassifikation abgeglichen werden, um Textpassagen mit *Hate Speech* zu detektieren.

Das erste Verfahren fokussiert *Online Harassment*. *Online Harassment* bezeichnet den einmaligen Versand einer elektronischen Nachricht mit dem Ziel, psychischen Schaden bei einem Individuum auszulösen. Mithilfe der Patterns untersucht das Verfahren Texte hinsichtlich der Präsenz von beleidigenden Wörtern und syntaktischen Verbindungen zu referenzierten Individuen. Durch die Markierung des referenzierten Opfers im Text ist eine nachgelagerte automatische Auswertung möglich. Das zweite Verfahren greift diesen Aspekt auf und identifiziert die markierten Opfer anhand eindeutiger Merkmale in *Online Communities*. Obwohl die Literatur bereits auf die Bedeutung der Erkennung der referenzierten Opfer hinweist, ist dieser Aspekt in der Mehrzahl existierender Arbeiten vernachlässigt. Dies ist jedoch insbesondere für die Erkennung von *Cyberbullying* entscheidend, da sich diese Form durch wiederholtes *Online Harassment* von demselben Autor gegenüber demselben Opfer auszeichnet. Demnach ist eine Re-Identifikation des referenzierten Opfers über mehrere Nachrichten hinweg notwendig, um eine korrekte Zuordnung zu gewährleisten. Schließlich ist die Identifikation der Opfer eine Voraussetzung, um *Directed Hate Speech* zu erkennen, die sich gegen Gruppen von Menschen richtet. Darunter fallen beispielsweise rassistische Äußerungen, die sich typischerweise gegen Nationalitäten oder Religionen richten. Das dritte entwickelte Verfahren greift diesen Aspekt auf, indem die Erkennung der referenzierten Opfer erweitert wird, sodass es zusätzlich Referenzen zu Gruppen detektiert.

Die Evaluationsresultate zeigen eine Verbesserung der Klassifikationsgüte gegenüber existierenden Verfahren, insbesondere im Bereich der *Online Harassment* und *Cyberbullying* Erkennung. Die erzielten Resultate haben Einfluss auf die praktische Anwendbarkeit der vorgestellten Verfahren. Sie unterstützen Personal voll- oder halbautomatisch vor dem Hintergrund der typischerweise großen Nachrichtenmenge in *Online Communities* bei der arbeitsintensiven Aufgabe, *Hate Speech* zu moderieren. Vollautomatische Systeme arbeiten ohne die Beteiligung menschlicher Kontrollinstanzen und skalieren gegenüber halbautomatischen Verfahren besser mit der großen Nachrichtenmenge in *Online Communities*. Mit einem proaktiven Einsatz verhindern sie die Publikation von *Hate Speech*, um psychischen Schaden bei Individuen oder die Gefährdung der Öffentlichkeit zu vermeiden. Die

Anforderungen an die Klassifikationsgüte derartiger Systeme sind hoch, insbesondere bezüglich der Fehlerrate von falsch positiven Resultaten. Sowohl das vorgestellte Artefakt zur Detektion von *Online Harassment* als auch das Artefakt zur Detektion von *Cyberbullying* zeichnen sich durch sehr geringe falsch-positiv Raten von weniger als 15% aus. Dies führt zu entsprechend wenigen Fehklassifikationen im Rahmen eines vollautomatischen Einsatzes. Halbautomatische Systeme markieren in einem ersten Schritt detektierte *Hate Speech* Nachrichten, um sie in einem zweiten Schritt einem Moderator zu präsentieren. Der Moderator prüft die Nachrichten, um Einzelfallentscheidungen zu treffen. Gegenüber vollautomatischen Systemen skalieren diese Ansätze aufgrund der Beteiligung von Menschen schlechter mit dem Nachrichtenaufkommen in *Online Communities*. Zudem vermeidet dieser Ansatz keine Schäden aufgrund der Publikation von *Hate Speech*, falls die Nachrichten bis zur erfolgten Moderation in der *Online Community* sichtbar sind. Hybride IT-Systeme mindern die wechselseitigen Nachteile, indem sie in Abhängigkeit der Intensität von *Hate Speech* ein voll- oder halbautomatisches Vorgehen wählen. Derartige Systeme sind mit einem Schwellwert konfiguriert, sodass eine vollautomatische Verarbeitung nur bei einer hohen Wahrscheinlichkeit einer korrekten Klassifikation stattfindet, während ein Moderator alle verbleibenden Fälle manuell prüft.

Inhaltsverzeichnis

Danksagung i

Zusammenfassung ii

Abbildungsverzeichnis vii

Tabellenverzeichnis viii

Abkürzungsverzeichnis ix

1 Einleitung 1

1.1 Problemstellung 1

1.2 Zielstellung 3

1.3 Methodik und Struktur 4

2 Hate Speech in Online Communities 5

2.1 Directed Hate Speech, Online Harassment und Cyberbullying 5

2.2 IT-Systeme zur Verarbeitung von Hate Speech in Online Communities 8

3 Verarbeitung unstrukturierter Texte im Rahmen des Text Mining 12

3.1 Text Mining 12

3.2 Modelle zur Strukturierung von Texten 14

3.3 Verfahren des Text Mining 16

3.4 Evaluation von Verfahren zur Klassifikation von Texten 20

4 Stand der Forschung und Forschungsdesign 22

4.1 Methodik zur Ermittlung des Standes der Forschung 22

4.2 Analyse existierender Hate Speech Klassifikationsverfahren 23

4.3 Analyse von Verfahren zur Detektion der referenzierten Ziele 26

4.4 Forschungsdesign 28

5 Vorstellung der Publikationen 31

5.1 Detektion von Online Harassment 31

5.1.1 Artefakt zur Detektion von Online Harassment 31

5.1.2 Evaluation 33

5.1.3 Diskussion und Ausblick 34
5.2 Detektion von Cyberbullying 36
5.2.1 Artefakt zur Detektion von Cyberbullying 36
5.2.2 Evaluation 38
5.2.3 Diskussion und Ausblick 39
5.3 Detektion von Directed Hate Speech 41
5.3.1 Artefakt zur Detektion von Directed Hate Speech 41
5.3.2 Evaluation 43
5.3.3 Diskussion und Ausblick 44
6 Anwendung der Forschungsergebnisse in anderen Domänen 47
6.1 Limitationen 47
6.2 Anwendung in anderen Domänen 48
7 Schlussbetrachtung 51
Literaturverzeichnis 54
Anhang A: Erläuterungen zu den Co-Autorenschaften 62
Anhang B: Publikation: Detecting Online Harassment in Social Networks 63
Anhang C: Publikation: Detecting Cyberbullying in Online Communities 78
Anhang D: Publikation: Detecting Offensive Statements towards Foreigners in Social Media 93

Abbildungsverzeichnis

Abbildung 1: Aufbau von Design Science Studien nach Gregor und Hevner (2013) 4
Abbildung 2: Differenzierung des Begriffs Hate Speech 7
Abbildung 3: IT-Systeme zur Verarbeitung von Hate Speech 9
Abbildung 4: Aspekte des Text Mining 12
Abbildung 5: bag-of-words und n-gram Textmodell 14
Abbildung 6: Dependency Graph 16
Abbildung 7: Trainings- und Klassifikationsphase bei überwachtem 17
Abbildung 8: Artefakte und Publikationen der kumulativen Dissertation 28
Abbildung 9: Architektur des Artefakts zur Detektion von Online Harassment 31
Abbildung 10: Anwendung des is-a Patterns im Sequenzmodell 32
Abbildung 11: Architektur des Artefakts zur Detektion von Cyberbullying 36
Abbildung 12: Harassment Graph 37
Abbildung 13: Architektur des Artefakts zur Detektion von Directed Hate Speech 41

Tabellenverzeichnis

Tabelle 1: Konfusionsmatrix für binäre Klassifikationsprobleme 20
Tabelle 2: Berechnungsvorschriften für typische Evaluations-Metriken 20
Tabelle 3: Identifizierte Quellen im Rahmen der Literaturrecherche 23
Tabelle 4: Ergebnisse der Literaturanalyse 27
Tabelle 5: Evaluationsergebnisse der Online Harassment Klassifikation 33
Tabelle 6: Evaluationsergebnisse der Cyberbullying Klassifikation 39
Tabelle 7: Evaluationsergebnisse der Directed Hate Speech Klassifikation 43
Tabelle 8: Evaluationsergebnisse der Klassifikation des referenzierten Ziels 44

Abkürzungsverzeichnis

fn	false negative
fp	false positive
ICCPR	Internationaler Pakt über bürgerliche und politische Rechte
IT	Informationstechnik
StGB	Strafgesetzbuch
tn	true negative
tp	true positive

1 Einleitung

1.1 Problemstellung

Online Communities sind Plattformen im Web, die es Nutzern ermöglichen, miteinander zu kommunizieren, digitale Inhalte verschiedenster Art auszutauschen und sich untereinander zu vernetzen.[1] Der Begriff *Online Community* subsummiert verschiedenste Web-Applikationen, wie beispielsweise Foren, Blogs, soziale Netzwerke oder Diskussionsplattformen.[2] Der Erfolg und die Popularität dieser Plattformen spiegeln sich in den beliebtesten Websites weltweit wider, unter denen sich *Online Communities* wie beispielsweise Facebook, YouTube, Twitter und Reddit finden.[3] Insbesondere unter Teenagern sind derartige Plattformen beliebt.[4] Aktuelle Studien zeigen, dass über 90% der Teenager täglich online sind, um auf dem neuesten Stand zu bleiben.[5] Die Nutzer erstellen vielfältige Inhalte auf den Plattformen. In sozialen Netzwerken, wie Facebook und Twitter, legen die Teilnehmer Profile an, veröffentlichen Bilder und versenden Textnachrichten untereinander. Die Nutzer stellen Videos auf YouTube ein, die wiederum von anderen Nutzern kommentiert werden.[6] Diskussionsplattformen, wie Reddit, laden die Nutzer zur öffentlichen Diskussion von Themen aller Art ein. Soziale Medien ermöglichen den Nutzern die Publikation von Inhalten in einer Netzwerkstruktur, wodurch sich diese mit großer Reichweite und hoher Geschwindigkeit verbreiten lassen.[7]

Die Plattformen orientieren sich am Recht auf freie Meinungsäußerung, sodass typischerweise die Möglichkeit besteht, nutzergenerierte Inhalte weitestgehend ohne Restriktionen und Inhaltsprüfung zu erstellen.[8] Darüber hinaus erlauben viele *Online Communities* die Publikation von Inhalten unter anonymen oder pseudonymen Zugängen, wodurch gegenüber der persönlichen Kommunikation die Hemmschwelle geringer ist, Inhalte jeglicher Art zu veröffentlichen.[9] Dadurch entsteht die Gefahr, dass Nutzer die Plattformen für Gewaltaufrufe sowie beleidigende oder sonstige hasserfüllte Kommunikation missbrauchen.[10] Aufgrund der mannigfaltigen Formen werden derartige Inhalte unter dem Sammelbegriff *Hate Speech*

[1] Vgl. Kraut und Resnick (2012), S. 1.
[2] Vgl. Kraut und Resnick (2012), S. 1.
[3] Vgl. Alexa (2015).
[4] Vgl. Lenhart (2015), S. 2.
[5] Vgl. Lenhart (2015), S. 2.
[6] Vgl. Alby (2007), S. 105f.
[7] Vgl. King et al. (2014), S. 170.
[8] Vgl. Bernstein et al. (2011), S. 51.
[9] Vgl. Tokunaga (2010), S. 279; Englander und Muldowney (2007), S. 84; Li (2007), S. 1786; Patchin und Hinduja (2006), S. 154.
[10] Vgl. Tokunaga (2010), S. 279; Patchin und Hinduja (2006), S. 154.

subsummiert.[11] Der zunehmende Umfang von *Hate Speech*[12] veranlasst viele Plattformen, wie beispielsweise Facebook und YouTube, Policies zur Sicherstellung eines normgerechten Umgangs miteinander zu erstellen.[13] Die *Online Communities* verlassen sich häufig auf die Nutzer, um Verstöße gegen die Policies zu erkennen und zu melden.[14] Eine Meldung erfordert jedoch die Einsichtnahme der Nachricht, was bei den Opfern bereits zu psychischer Schadwirkung führen kann.[15] Personal überprüft die gemeldeten Fälle manuell, was aufgrund der Vielzahl an Nachrichten arbeits- und kostenintensiv ist.[16] Zudem können Administratoren zwar die Nachrichten und den Account des betreffenden Täters im Nachhinein löschen, dieser kann jedoch neue Accounts anlegen, um die Kommunikation fortzuführen.[17] Somit können sich die Opfer in dem digitalen Umfeld nur eingeschränkt verteidigen. Da ein großer Anteil der betroffenen Nutzer in *Online Communities* sehr jung oder sogar minderjährig ist[18], sind die Folgen von *Hate Speech* besonders problematisch.[19] Die möglichen Auswirkungen reichen von Depressionen, Verschlechterung der schulischen Leistungen sowie verringertem Selbstwertgefühl bis hin zu Suizidgedanken.[20] In besonders schweren Fällen tragen die psychischen Belastungen durch *Hate Speech* dazu bei, Suizid tatsächlich zu begehen. Derartige Fälle erlangen häufig erst aufgrund ihrer Tragweite mediale Aufmerksamkeit, wie der Fall eines 14-jährigen Mädchens aus Italien verdeutlicht.[21] In der Literatur wird deshalb auf die zentrale Bedeutung der Erkennung der referenzierten Opfer hingewiesen, um diese gegebenenfalls durch Dritte unterstützten zu können.[22]

Aktuelle Bestrebungen auf politischer Ebene zur Bekämpfung hasserfüllter und rassistischer Inhalte in sozialen Netzwerken im Rahmen der Flüchtlingskrise zeigen eine andere Facette von *Hate Speech*, die zunehmend an Bedeutung gewinnt.[23] In *Online Communities* und insbesondere in sozialen Medien sind derartige Inhalte oft öffentlich einsehbar und erreichen daher eine große Anzahl an Nutzern.[24] Es besteht die Befürchtung, dass radikale Gruppierungen

[11] Vgl. Williams und Burnap (2016), S. 2f; Chen et al. (2012), S. 72.
[12] Vgl. Williams und Burnap (2016), S. 3; Aponte und Richards (2013), S. 18:4; Tokunaga (2010), S. 281; Li (2007), S. 1779f; Campbell (2005), S. 70.
[13] Vgl. Facebook (2016); YouTube (2016).
[14] Vgl. Chen et al. (2011), S. 72.
[15] Vgl. Cohen et al. (2014), S. 55.
[16] Vgl. Chen et al. (2011), S. 71; Wise et al. (2006), S. 30.
[17] Vgl. Patchin und Hinduja (2006), S. 154.
[18] Vgl. Statista (2014); Patchin und Hinduja (2010), S. 198.
[19] Vgl. Arseneault et al. (2010), S. 721f.
[20] Vgl. Cohen et al. (2014), S. 52f; Hinduja und Patchin (2013), S. 712; Arseneault et al. (2010), S. 721f; Tokunaga (2010), S. 277; Li (2007), S. 1779.
[21] Vgl. BBC News (2014).
[22] Vgl. Cohen et al. (2014), S. 53.
[23] Vgl. BBC News (2015).
[24] Vgl. Williams und Burnap (2016), S. 213; King et al. (2014), S. 170.

Hate Speech nutzen, um die Öffentlichkeit gegen Ausländer aufzuhetzen und neue Mitglieder zu rekrutieren.[25] Aktuelle Studien zeigen, dass tatsächlich ein Zusammenhang zwischen dem Ausmaß an rechtsgerichteten Aktivitäten in sozialen Netzwerken und Ereignissen in der realen Welt in Bezug auf Ausländer besteht.[26] Als Reaktion auf die aktuellen Entwicklungen hat beispielsweise die deutsche Regierung in Kooperation mit dem sozialen Netzwerk Facebook eine Task Force gegründet, die aus Mitgliedern von *Online Communities*, Parteien und dem deutschen Justizministerium besteht. Sie verfolgt im Wesentlichen die Aufgabe, *Hate Speech* gegen Ausländer und Flüchtlinge in Facebook zu identifizieren und zu moderieren.[27]

In der Literatur existieren Vorschläge zum Umgang mit *Hate Speech* in *Online Communities*, die auf dem Einsatz von Systemen der Informationstechnik (IT) basieren.[28] Diese IT-Systeme verwenden Klassifikatoren, die derartige Nachrichten automatisch identifizieren und von neutralen Nachrichten abgrenzen. Existierende Klassifikatoren aus der Forschung vernachlässigen die Detektion der referenzierten Opfer von *Hate Speech.*[29] Dadurch weisen sie bezüglich der Klassifikationsgüte jedoch Schwächen auf, wodurch sie nur bedingt für diesen Einsatzzweck geeignet sind.[30] Sie stützen sich auf die Präsenz beleidigender Wörter als zentrales Klassifikationsmerkmal und vernachlässigen die Analyse des sprachlichen Kontextes. Die Präsenz von beleidigenden Wörtern ist jedoch nicht hinreichend, um *Hate Speech* präzise zu erkennen, da diese Wörter in einem Kontext stehen und gegen Individuen oder Gruppen gerichtet werden. Schließlich konzentriert sich ein Großteil existierender Arbeiten auf die isolierte Betrachtung von *Hate Speech* Nachrichten. Die Verarbeitung von wiederholter und somit zusammenhängender Interaktion zwischen identischen Nutzern wird nicht weiter betrachtet.[31]

1.2 Zielstellung

Das Ziel der Dissertation besteht darin, Verfahren zur automatischen Detektion von *Hate Speech* in Online Communities einschließlich der referenzierten Opfer unter Berücksichtigung des sprachlichen Kontextes zu konzipieren, zu implementieren und zu evaluieren. Dabei werden diejenigen Formen von *Hate Speech* fokussiert, die sich gegen menschliche Ziele richten.

[25] Vgl. Glaser et al. (2002), S. 188.
[26] Vgl. Williams und Burnap (2016), S. 22.
[27] Vgl. BBC News (2015).
[28] Vgl. Cohen et al. (2014), S. 53ff; Aponte und Richards (2013), S. 18:6f.
[29] Vgl. Bretschneider et al. (2014), S. 9f.
[30] Vgl. Bretschneider und Peters (2016), S. 3f; Nahar et al. (2013), S. 51; Sood et al. (2012a), S. 1484f.
[31] Vgl. Bretschneider und Peters (2016), 3f; Rafiq et al. (2015), S. 617.

1.3 Methodik und Struktur

Zur Erreichung des gestellten Ziels wird die Forschungsmethodik der *Design Science* verwendet. Zur Strukturierung der Arbeit wird den Empfehlungen von Gregor und Hevner (2013) gefolgt, die den in Abbildung 1 dargestellten idealtypischen Aufbau von *Design Science* Studien vorschlagen.

Abbildung 1: Aufbau von Design Science Studien nach Gregor und Hevner (2013)[32]

Die Problemstellung, das Ziel der Arbeit sowie die Relevanz werden in der Einleitung diskutiert. Die Vorstellung der für die Arbeit relevanten Begriffe und eine detaillierte Darstellung des praktischen Bezugs finden, abweichend von Gregor und Hevner (2013), nicht in Kapitel 1 statt[33], sondern werden aufgrund des Umfangs gesondert in Kapitel 2 behandelt. In Kapitel 3 wird das Text Mining als theoretischer Hintergrund der Forschungsstudie vorgestellt. In Kapitel 4 wird eine strukturierte Literaturrecherche nach vom Brocke et al. (2009) durchgeführt, um existierende Ansätze im Bereich der *Hate Speech* Erkennung zu identifizieren. Die Ergebnisse der Literaturanalyse und insbesondere die ermittelten Forschungslücken werden im Anschluss diskutiert. Darauf aufbauend wird das Forschungsdesign in Hinblick auf die drei entwickelten und publizierten Software-Artefakte zur Erkennung von *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* im Rahmen dieser kumulativen Dissertation vorgestellt. Es wird insbesondere auf den Zusammenhang der Artefakte und die jeweiligen Zielstellungen eingegangen. Entgegen der vorgeschlagenen Vorgehensweise von Gregor und Hevner (2013), die Beschreibung des Artefakts, die Evaluation und die Diskussion der Ergebnisse in jeweils einem Hauptkapitel durchzuführen, werden diese Aspekte in Kapitel 5 jeweils für die drei Artefakte überblicksartig in Unterkapiteln vorgestellt. Eine detaillierte Darstellung der Artefakte ist in den im Anhang angefügten Publikationen enthalten. In Kapitel 6 werden die Limitationen der Artefakte in einem Gesamtzusammenhang kritisch reflektiert. Zudem findet eine Diskussion bezüglich der Übertragung der Ergebnisse in andere Anwendungsdomänen statt. Schließlich werden die wesentlichen Erkenntnisse in Kapitel 7 zusammengefasst.

[32] Vgl. Gregor und Hevner (2013), S. 350.
[33] Vgl. Gregor und Hevner (2013), S. 350.

2 Hate Speech in Online Communities

In diesem Abschnitt werden eine Arbeitsdefinition von *Hate Speech* und eine Begriffsabgrenzung von *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* vorgestellt. Im Anschluss werden Lösungskonzepte im Umgang mit *Hate Speech* auf Basis von IT-Systemen diskutiert.

2.1 Directed Hate Speech, Online Harassment und Cyberbullying

In der Literatur besteht aufgrund der vielfältigen Formen von *Hate Speech* kein Konsens über eine Definition.[34] Deshalb wird in diesem Abschnitt eine Arbeitsdefinition ausgearbeitet. Als Ausgangsganspunkt wird ein Definitionsansatz des Ministerkomitees des Europarats verwendet, der als Orientierung für Europäisches Case Law dient und nicht den Anspruch einer vollständigen Definition erhebt.[35] Demnach werden unter *Hate Speech* jegliche Formen von Ausdrücken mit rassistischen, xenophoben, antisemitischen oder anderen Inhalten basierend auf Hass und Intoleranz verstanden.[36] Somit fokussiert dieser Definitionsansatz Ausdrücke, die typischerweise gegen Gruppen von Menschen mit bestimmter Volksangehörigkeit oder Religion gerichtet sind. In Ergänzung dazu wird ein weiterer Definitionsansatz verwendet, um die sonstigen auf Hass und Intoleranz basierenden Inhalte näher zu spezifizieren und die Persönlichkeitsrechte von Individuen zu adressieren. Williams und Burnap (2016) verstehen *Hate Speech* als Sammelbegriff für Ausdrücke, die durch eines oder mehrere der folgenden vier Kriterien gekennzeichnet sind:[37]

(1) Ausdrücke, die die referenzierten Opfer psychisch verletzen
(2) Ausdrücke, die Gewalt provozieren
(3) Ausdrücke, die Anstoß bei Dritten erregen
(4) Ausdrücke, die sich gegen Mitglieder von Gemeinschaften oder ihre sozialen Beziehungen untereinander in herabwürdigender, beleidigender oder verleumdender Weise richten

Die Synthese der genannten Kriterien beider Definitionsansätze bildet die Arbeitsdefinition für diese Arbeit. Diese Kriterien stammen aus einem juristischen Kontext und sind gleichzeitig ein Indikator dafür, ob derartige Ausdrücke in Konflikt zu geltendem Recht stehen.[38] *Hate Speech* kann Persönlichkeitsrechte verletzten oder in Konflikt mit dem Schutz nationaler Sicherheit

[34] Vgl. Williams und Burnap (2016), S. 213; Chen et al. (2011), S. 72.
[35] Vgl. Weber (2009), S. 3.
[36] Vgl. Weber (2009), S. 3.
[37] Vgl. Williams und Burnap (2016), S. 213.
[38] Vgl. Williams und Burnap (2016), S. 213.

oder der Erhaltung der öffentlichen Ordnung stehen.[39] Die Verbreitung von *Hate Speech* stellt gemäß Art. 19 Abs. 3 des „Internationalen Pakts über bürgerliche und politische Rechte (ICCPR)“ eine Verletzung von Persönlichkeitsrechten dar.[40] Demzufolge fällt die Publikation von *Hate Speech* nicht unter das Recht auf freie Meinungsäußerung, das den Empfang und die Vermittlung von Informationen und Ideen aller Art zusichert.[41] In Deutschland kann *Hate Speech* beispielsweise einen Straftatbestand in Form von Beleidigungen (§ 185 StGB), übler Nachrede (§ 186 StGB) sowie Verleumdung (§ 187 StGB) gegenüber Individuen darstellen. Darüber hinaus wird in § 111 StGB die öffentliche Aufforderung zu Straftaten und in § 130 StGB die Volksverhetzung als Straftatbestand festgelegt. Somit kann die Publikation von *Hate Speech* zu juristischen Konsequenzen für die Autoren führen. Im digitalen Umfeld von *Online Communities* sind diese Inhalte in Schriftform dokumentiert. Ein Autor kann dennoch nicht immer eindeutig ermittelt werden, beispielsweise aufgrund von anonymer Publikation. Deshalb können auch für die *Online Community* Plattform juristische Konsequenzen entstehen, wie ein Urteil des Europäischen Gerichtshofs für Menschenrechte verdeutlicht.[42] Im betreffenden Fall wurde ein Schadenersatzanspruch gegenüber einer *Online Community* Plattform aufgrund der Publikation von *Hate Speech* Kommentaren durch anonyme Nutzer bestätigt.[43] Obwohl diese Kommentare durch die Plattform gelöscht wurden, ist bereits ein Schaden durch die öffentliche Sichtbarkeit entstanden, der auszugleichen ist.[44]

Der Begriff *Hate Speech* wird im Rahmen dieser Arbeit als zusammenfassender Begriff für die nachfolgend vorgestellten Formen verwendet. *Hate Speech* umfasst mehrere Formen hasserfüllter Ausdrücke, die sich nicht notwendigerweise gegen ein Ziel richten. Ein Konflikt zwischen dem Recht auf freie Meinungsäußerung und individuellen Rechten liegt aber nur dann vor, falls tatsächlich ein Individuum adressiert wird. Analog dazu richten sich rassistische und herabwürdigende oder beleidigende Ausdrücke typischerweise gegen menschliche Ziele in Form von Gruppen.[45] In Abbildung 2 ist deshalb eine Begriffsabgrenzung mithilfe eines Unified Modeling Language Klassendiagramms visualisiert, um diese Formen von dem undifferenzierten Begriff *Hate Speech* abzugrenzen. In dieser undifferenzierten Form umfasst *Hate Speech* sämtliche Ausdrücke, die die in der Arbeitsdefinition aufgeführten Kriterien

[39] Vgl. Office of the High Commissioner for Human Rights (o. J.).
[40] Vgl. Office of the High Commissioner for Human Rights (o. J.).
[41] Vgl. Office of the High Commissioner for Human Rights (o. J.).
[42] Vgl. Europäischer Gerichtshof für Menschenrechte (2015).
[43] Vgl. Europäischer Gerichtshof für Menschenrechte (2015); Scott (2015).
[44] Vgl. Europäischer Gerichtshof für Menschenrechte (2015); Scott (2015).
[45] Vgl. Gitari et al. (2015), S. 217.

erfüllen. Die Vererbungsbeziehungen stellen Spezialisierungen dieses Begriffs dar, die sich durch zusätzliche Merkmale auszeichnen.

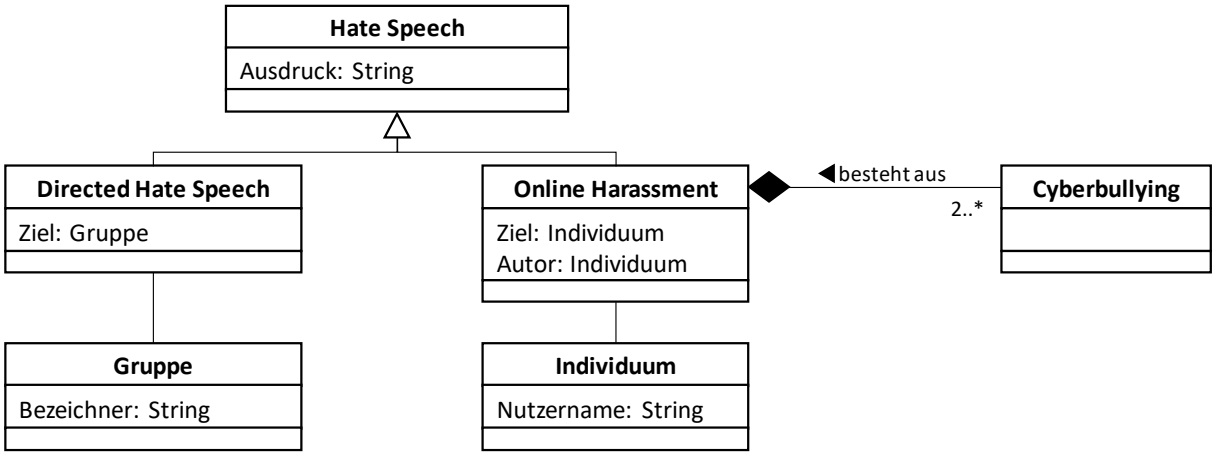

Abbildung 2: Differenzierung des Begriffs Hate Speech

Directed Hate Speech ist ein im Rahmen dieser Arbeit eingeführter Begriff, der sich an einer Begriffsabgrenzung von *Hate Speech* nach Burnap und Williams (2015)[46] sowie Gitari et al. (2015)[47] orientiert. *Directed Hate Speech* zeichnet sich durch Ausdrücke aus, die sich gegen Gruppen richten. Gruppen von Menschen lassen sich auf vielfältige Weise referenzieren, beispielsweise über die Volks- beziehungsweise Religionszugehörigkeit oder die Mitgliedschaft in einer politischen Partei.[48] Tokunaga (2010) unterscheidet darüber hinaus Formen von *Hate Speech*, die sich ausschließlich gegen Individuen richten. Darunter fällt *Online Harassment*, das den einmaligen Versand einer elektronischen Nachricht mit dem Ziel, psychischen Schaden bei einem Opfer auszulösen, bezeichnet.[49] *Online Harassment* ist somit eine spezielle Form von *Hate Speech*, die von einem Autor ausgeht und gegen ein Opfer gerichtet ist. Darauf aufbauend zeichnet sich *Cyberbullying* durch wiederholtes *Online Harassment* ausgehend von demselben Autor gegenüber demselben Opfer aus.[50] *Cyberbullying* ist demnach keine Spezialisierung von *Hate Speech*, sondern eine Komposition aus mehreren *Online Harassment* Ausdrücken. Zur Erkennung dieser Form ist es somit notwendig, den Täter und das Opfer wiederholt zu identifizieren, um die *Online Harassment* Ausdrücke dem *Cyberbullying* Fall korrekt zuzuordnen.

[46] Vgl. Burnap und Williams (2015), S. 231.
[47] Vgl. Gitari et al. (2015), S. 217.
[48] Vgl. Gitari et al. (2015), S. 217.
[49] Vgl. Tokunaga (2010), S. 278.
[50] Vgl. Tokunaga (2010), S. 278.

2.2 IT-Systeme zur Verarbeitung von Hate Speech in Online Communities

In diesem Abschnitt werden Lösungsvorschläge auf Basis von IT-Systemen zur Verarbeitung von *Hate Speech* in *Online Communities* diskutiert. Einführend werden Policies als formale Grundlage für die Detektion von *Hate Speech* vorgestellt.

Online Communities, wie beispielsweise Facebook[51] und YouTube[52], reagieren auf die Zunahme von *Hate Speech* mit der Einführung von Policies beziehungsweise Community Standards. Diese Policies definieren Richtlinien für ein zivilisiertes und normgerechtes Miteinander und sind neben gesetzlichen Vorgaben die formale Grundlage für den Umgang mit *Hate Speech.*[53] Administratoren beziehungsweise Moderatoren, aber auch Mitglieder der *Online Community* selbst, stellen die Einhaltung der Policies sicher. Während die Mitglieder typischerweise Verstöße gegen die Policies melden, prüfen Moderatoren die Fälle manuell und leiten gegebenenfalls weiterführende Schritte als Reaktion darauf ein.[54] Eine manuelle Prüfung durch Moderatoren setzt jedoch voraus, dass die Fälle tatsächlich gemeldet oder durch die Moderatoren detektiert werden. Eine Meldung solcher Fälle kann speziell für die Opfer von *Online Harassment* und *Cyberbullying* entscheidend sein, damit Dritte gegebenenfalls intervenieren und den Opfern bei der Verarbeitung der Vorfälle beistehen können.[55]

In der Literatur werden Vorschläge zur Unterstützung der Moderation von *Hate Speech* in *Online Communities* mithilfe von IT-Systemen diskutiert. Aponte und Richards (2013)[56] sowie Cohen et al. (2014)[57] schlagen diesbezüglich die Entwicklung von IT-Systemen vor, die derartige Inhalte automatisch blockieren oder markieren. Sie basieren auf einem Klassifikator, der *Hate Speech* für eine voll- oder halbautomatische Weiterverarbeitung detektiert. Abbildung 3 verdeutlicht die Funktionsweise der IT-Systeme anhand des Beispiels eines sozialen Netzwerkes. Die IT-Systeme sind anhand ihres Automatisierungsgrades und dem Zeitpunkt der Publikation von *Hate Speech* systematisiert. Unter *Hate Speech* Nachrichten werden im Folgenden Texte verstanden, die *Hate Speech* enthalten und als abgeschlossene Einheit in *Online Communities* publiziert werden.

[51] Vgl. Facebook (2016).
[52] Vgl. YouTube (2016).
[53] Vgl. Aponte und Richards (2013), S. 18:5.
[54] Vgl. Cohen et al. (2014), S. 56; Aponte und Richards (2013), S. 18:6; Dinakar et al. (2012), S. 18:2.
[55] Vgl. Cohen et al. (2014), S. 53.
[56] Vgl. Aponte und Richards (2013), S. 18:6.
[57] Vgl. Cohen et al. (2014), S. 53ff.

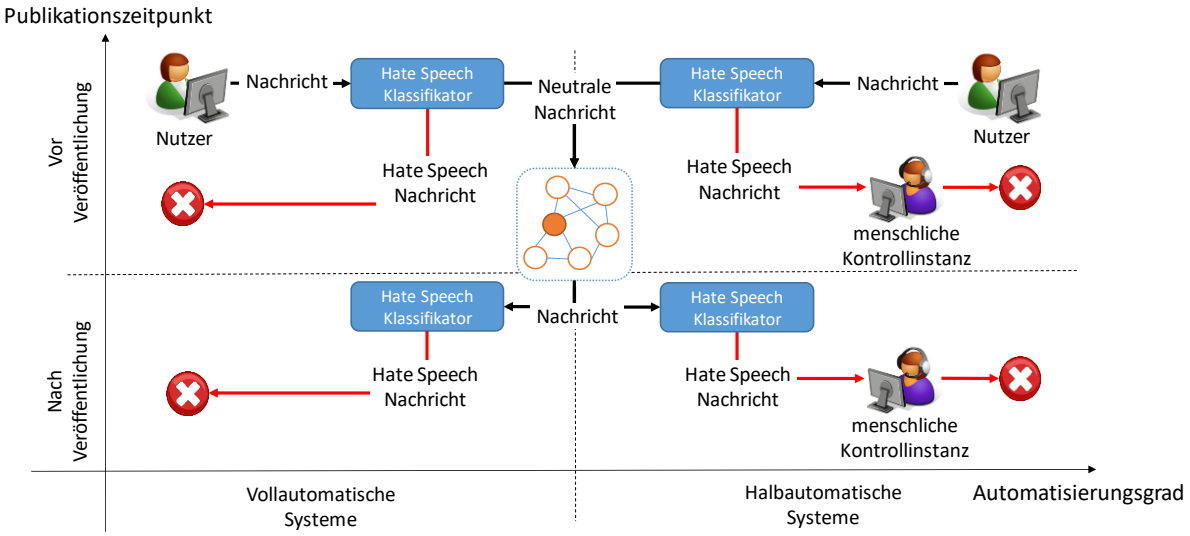

Abbildung 3: IT-Systeme zur Verarbeitung von Hate Speech

Vollautomatische IT-Systeme arbeiten proaktiv oder reaktiv ohne die Beteiligung menschlicher Kontrollinstanzen. Bei einem proaktiven Einsatz blockiert das System durch den Klassifikator identifizierte *Hate Speech* Nachrichten vor der Publikation, um sie automatisch zu löschen oder zu archivieren.[58] Im Gegensatz dazu analysiert das System bei einem reaktiven Einsatz bereits publizierte Nachrichten. Da vollautomatische Systeme ohne menschliche Kontrollinstanzen arbeiten, sind sie im Vergleich zu manueller Moderation und vor dem Hintergrund des typischerweise hohen Nachrichtenaufkommens von *Online Communities* kostengünstiger skalierbar.[59] Darüber hinaus stellt die Möglichkeit der Verarbeitung in Echtzeit unter Berücksichtigung der hohen Verbreitungsgeschwindigkeit und Reichweite von Nachrichten in *Online Communities* einen zentralen Vorteil dar.[60] Ein proaktiver Einsatz vermeidet insbesondere bei *Online Harassment* und *Cyberbullying* psychischen Schaden, da die Opfer keine Einsicht in die Nachrichten nehmen.[61] Analog dazu verhindert die Blockierung von *Directed Hate Speech* mit rassistischen Inhalten eine Aufhetzung der Öffentlichkeit.[62] Tätern ist es zudem nicht möglich, das System durch die Erstellung anonymer Zugänge zu umgehen, da der Klassifikator inhaltsbasiert arbeitet. Eine Archivierung der detektierten *Hate Speech* Nachrichten erlaubt, sie im Nachhinein auszuwerten, um beispielsweise Verstöße gegen die Policy der *Online Community* nachzuweisen.[63] Eine Archivierung von Verstößen ermöglicht

[58] Vgl. Cohen et al. (2014), S. 55.
[59] Vgl. Bretschneider et al. (2014), S. 10.
[60] Vgl. Chen et al. (2011), S. 71.
[61] Vgl. Cohen et al. (2014), S. 55.
[62] Vgl. Kim und Douai (2012), S. 176f.
[63] Vgl. Cohen et al. (2014), S. 55.

darüber hinaus, Nutzer erst ab einer bestimmten Anzahl an erkannten Fällen zu reglementieren.[64]

Vollautomatische Systeme stellen hohe Anforderungen an den Klassifikator. Typischerweise werden Klassifikatoren im Kontext der Spam Erkennung als Vergleich herangezogen[65], die geringe Raten von falsch blockierten Nachrichten von weniger als 10% aufweisen.[66] Falsch blockierte Nachrichten führen bei vollautomatischen Systemen zu anwendungsspezifischen Kosten.[67] In *Online Communities* entstehen beispielsweise Opportunitätskosten durch Nutzer, die als Reaktion auf falsch blockierte Nachrichten die Plattform verlassen.[68] Zudem greift das automatische Blockieren von Inhalten möglicherweise in das Recht auf freie Meinungsäußerung ein. Eine Einzelfallentscheidung durch Moderatoren, um festzustellen, ob derartige Inhalte tatsächlich mit dem Recht auf freie Meinungsäußerung oder der Policy der *Online Community* in Konflikt stehen, ist somit nicht möglich. Darüber hinaus verhindert das Löschen von Nachrichten eine nachträgliche Prüfung durch Moderatoren. Dies ist beispielsweise bei Gewaltandrohungen problematisch, die ein Eingreifen durch Dritte erfordern, um die Situation zu deeskalieren.[69]

Demgegenüber erfordern halbautomatische Systeme das Eingreifen von menschlichen Kontrollinstanzen, die in einem nachgelagerten Schritt die Ergebnisse des Klassifikators einsehen und überprüfen. Entweder werden diese Nachrichten unter Vorbehalt publiziert oder solange von der Publikation zurückgestellt, bis die Prüfung durch die menschliche Kontrollinstanz erfolgt ist. Dadurch sind präzise Einzelfallentscheidungen durch Moderatoren möglich, die eine Abwägung konkurrierender Rechte durchführen. Ein weiterer Vorteil besteht in der Möglichkeit, die Opfer von *Online Harassment* und *Cyberbullying* gezielt zu unterstützten. Einige Opfer sind nicht im Stande, externe Hilfe zu erbitten und melden demnach derartige Vorfälle nicht.[70] Eine halbautomatische Meldung ist in diesen Fällen hilfreich, damit Dritte gegebenenfalls eingreifen können.[71] In *Online Communities* übernehmen Moderatoren typischerweise diese Aufgabe und entscheiden nach der Einsichtnahme der Nachrichten über weitere Schritte. Darüber hinaus können Eltern oder Freunde des Nachrichtenempfängers die Rolle der menschlichen Kontrollinstanz im Rahmen sogenannter „Parental Control Systems“

[64] Vgl. Cohen et al. (2014), S. 57.
[65] Vgl. Dinakar et al. (2012), S. 18:2.
[66] Vgl. Goh und Singh (2015), S. 439.
[67] Vgl. Miner et al. (2012), S. 889; Witten et al. (2011), S. 163f.
[68] Vgl. Newell et al. (2016), S. 5.
[69] Vgl. Cohen et al. (2014), S. 55.
[70] Vgl. Tokunaga (2010), S. 281; Li (2007), S. 1787.
[71] Vgl. Cohen et al. (2014), S. 55.

übernehmen.[72] Schließlich kann auch der Autor selbst diese Rolle übernehmen, indem er die Option erhält, die Publikation der vom IT-System als *Hate Speech* markierten Nachrichten zu überdenken.[73] Auf diese Weise können überstürzte oder emotionale Handlungen verhindert werden, falls der Autor sich der problematischen Formulierung nicht bewusst ist.

Dadurch, dass eine menschliche Kontrollinstanz die markierten Nachrichten in einem nachgelagerten Schritt überprüft, sind die Anforderungen an den Klassifikator gegenüber der vollautomatischen Verarbeitung niedriger. Der Klassifikator dient dazu, den Arbeitsaufwand für menschliche Kontrollinstanzen zu reduzieren, indem er möglichst viele der tatsächlichen *Hate Speech* Nachrichten erkennt und sich möglichst wenige falsch klassifizierte neutrale Nachrichten in der Ergebnismenge befinden.[74] Die Beteiligung einer menschlichen Kontrollinstanz verhindert jedoch die Verarbeitung derartiger Nachrichten in Echtzeit, da die Reaktionszeit eines Menschen Verzögerungen verursacht. Falls eine vorläufige Publikation möglicher *Hate Speech* Nachrichten zugelassen ist, kann Schadwirkung durch die Einsichtnahme der Nachricht von anderen Nutzern eintreten. Darüber hinaus zeigt die zuvor diskutierte Entscheidung des Europäischen Gerichtshofs für Menschenrechte, dass nur eine proaktive Vorgehensweise vor juristischen Konsequenzen schützt.[75]

Schließlich lassen sich voll- und halbautomatische Systeme mithilfe von hybriden Strategien kombinieren, um wechselseitige Nachteile teilweise aufzuheben. Ein hybrides IT-System verarbeitet *Hate Speech* Nachrichten voll- oder halbautomatisch in Abhängigkeit eines Wahrscheinlichkeitsschwellwertes. Eine vollautomatische Verarbeitung findet statt, falls eine hohe Wahrscheinlichkeit der korrekten Klassifikation besteht. Andernfalls verarbeitet das IT-System die Nachricht halbautomatisch. Eine zentrale Herausforderung bei derartigen Ansätzen besteht somit in der Ermittlung der Klassifikations-Wahrscheinlichkeit und des Schwellwertes. Ein Einflussfaktor auf diese Wahrscheinlichkeit ist beispielsweise die Ausdrucksstärke der verwendeten Wörter in *Hate Speech* Nachrichten.[76]

[72] Vgl. Cohen et al. (2014), S. 55.
[73] Vgl. Bretschneider et al. (2014), S. 11.
[74] Vgl. Bretschneider et al. (2014), S. 11.
[75] Vgl. Europäischer Gerichtshof für Menschenrechte (2015); Scott (2015).
[76] Vgl. Chen et al. (2012), S. 75.

3 Verarbeitung unstrukturierter Texte im Rahmen des Text Mining

In diesem Abschnitt wird die Einordnung der *Hate Speech* Detektion in unstrukturierten Texten in das Forschungsfeld des *Text Mining* vorgenommen. In einem nächsten Schritt werden Verfahren des *Text Mining* und Evaluations-Metriken vorgestellt.

3.1 Text Mining

Die zuvor diskutierten Lösungsvorschläge zur Verarbeitung von *Hate Speech* basieren auf Verfahren, die unstrukturierte nutzergenerierte Textinhalte analysieren. Zur Einordnung dieser Verfahren in einen theoretischen Rahmen, wird in diesem Abschnitt das Forschungsfeld des *Text Mining* vorgestellt, das nach Miner et al. (2012) die computergestützte Verarbeitung von unstrukturierten und strukturierten Texten umfasst.[77] In der Literatur sind für die automatische Verarbeitung und Analyse von natürlicher und damit unstrukturierter Sprache ferner die Begriffe *Speech and Language Processing*[78] sowie *Computational Linguistics*[79] gebräuchlich. Aufgrund der Begriffsvielfalt gestaltet sich das Zusammentragen und trennscharfe Einordnen der Verfahren und Problemstellungen des *Text Mining* schwierig.[80] Deshalb wird die vereinfachte Darstellung nach Miner et al. (2012) in Abbildung 4 verwendet, um die für die Arbeit relevanten Aspekte des *Text Mining* vorzustellen.

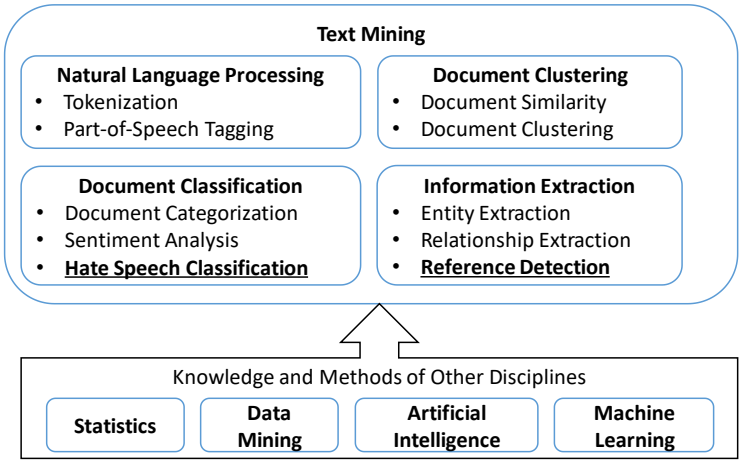

Abbildung 4: Aspekte des Text Mining[81]

Nach Miner et al. (2012) werden im *Text Mining* Methoden des *Natural Language Processing* als Vorverarbeitungsschritt genutzt, um unstrukturierte Texte lexikalisch zu analysieren und

[77] Vgl. Miner et al. (2012), S. 30.
[78] Vgl. Jurafsky und Martin (2009), S. 35.
[79] Vgl. Kay (2005), S. xvii.
[80] Vgl. Miner et al. (2012), S. 30.
[81] Vgl. Miner et al. (2012), S. xxiv; Miner et al. (2012), S. 33f.

somit in ein strukturiertes Textmodell zu überführen.[82] Mit Methoden aus angrenzenden Disziplinen, wie dem *Data Mining*, der *Künstlichen Intelligenz* und insbesondere dem *Machine Learning*, werden Textdokumente hinsichtlich verschiedener Problemstellungen verarbeitet.[83] Das *Document Clustering* gruppiert und ordnet Dokumente anhand ihrer Ähnlichkeit.[84] Das Gebiet der *Document Classification* widmet sich Problemstellungen, die die Einordnung von Textdokumenten in vorher festgelegte oder erlernte Klassen umfassen.[85] Darunter fallen insbesondere die Forschungsgebiete der *Sentiment Analysis* beziehungsweise dem häufig synonym verwendeten *Opinion Mining*.[86] Die *Sentiment Analysis* stellt Verfahren zur Einordnung von Textdokumenten hinsichtlich festgelegter Klassen zur Repräsentation von Stimmungen und Meinungen bereit.[87] Es bestehen Parallelen zwischen der *Sentiment Analysis* und der Detektion von *Hate Speech*. *Hate Speech* beinhaltet im weiteren Sinne besonders negative Stimmung, die gegen spezielle Bezugsobjekte gerichtet sein kann. Zudem wird häufig die *Sentiment Analysis* als verwandtes Forschungsgebiet in Arbeiten aus dem Bereich der *Hate Speech* Detektion referenziert, diesem jedoch nicht untergeordnet.[88] Die Autoren weisen darauf hin, dass die Verfahren der *Sentiment Analysis* nicht direkt auf die Problemstellung der *Hate Speech* Klassifikation übertragbar sind.[89] Vielmehr erfordern die Besonderheiten von *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* eine Modifikation der Verfahren, um das referenzierte Bezugsobjekt als wichtiges Merkmal in die Klassifikation einzubeziehen. Verfahren der *Sentiment Analysis* berücksichtigen diesen Aspekt typischerweise nicht weiter.[90] Die Verarbeitung von Bezugsobjekten ist Gegenstand der *Information Extraction*, die sich der Markierung von Entitäten und der Analyse von Beziehungen zwischen diesen Entitäten in Texten widmet.[91] Abbildung 4 beinhaltet deshalb entsprechend die Problemstellung der *Hate Speech Classification* und der *Reference Detection*. Die *Reference Detection* beinhaltet die automatische Erkennung des referenzierten Ziels von *Hate Speech*.

[82] Vgl. Miner et al. (2012), S. 32ff; Pustejovsky und Stubbs (2012), S. 4.
[83] Vgl. Miner et al. (2012), S. 36.
[84] Vgl. Miner et al. (2012), S. 32.
[85] Vgl. Miner et al. (2012), S. 32; Pustejovsky und Stubbs (2012), S. 5.
[86] Vgl. Pang und Lee (2008), S. 5; Tang et al. (2009), S. 10760f; Montoyo et al. (2012), S. 676.
[87] Vgl. Tsytsarau und Palpanas (2012), S. 481; Pang und Lee (2008), S. 6.
[88] Vgl. Gitari et al. (2015), S. 218; Burnap und Williams (2015), S. 225f; Sood et al. (2012b), S. 271; Yin et al. (2009), S. 2.
[89] Vgl. Gitari et al. (2015), S. 222; Sood et al. (2012b), S. 271.
[90] Vgl. Sood et al. (2012b), S. 272.
[91] Vgl. Miner et al. (2012), S. 32.

3.2 Modelle zur Strukturierung von Texten

Verfahren des *Text Mining* zur *Document Classification* verarbeiten strukturierte Eingaben.[92] Nutzergenerierte Texte sind jedoch typischerweise unstrukturiert. Deshalb findet zunächst eine Überführung von unstrukturierten Texten in strukturierte Textmodelle mithilfe von Verfahren des *Natural Language Processing* statt.[93] Es existieren verschiedene Textmodelle, die sich hinsichtlich ihrer Komplexität und insbesondere der Fähigkeit unterscheiden, grammatische Beziehungen und den Kontext zu erfassen.[94] Der Kontext enthält für das *Text Mining* wertvolle Informationen[95] und spielt insbesondere für die Detektion von zielgerichteter *Hate Speech* eine zentrale Rolle, da diese Ausdrücke aus mehreren grammatisch verbundenen Textbestandteilen bestehen.[96] Deshalb werden im Folgenden verschiedene Textmodelle vorgestellt und kritisch gewürdigt.

Das strukturell einfachste Textmodell ist das *bag-of-words* Modell[97]. Dieses Modell bildet einen Text als Multimenge (bag) ab.[98] In Abbildung 5 ist exemplarisch der Satz „This movie is not a masterpiece“[99] aus einer Filmkritik und dessen Überführung in ein *bag-of-words* Modell dargestellt. Jedes Wort entspricht einem Element der Multimenge. Die Anzahl des Auftretens eines Elements wird vermerkt, da Elemente in einer Multimenge mehrfach vorkommen können. Im Beispiel tritt jedes Wort genau einmal auf.

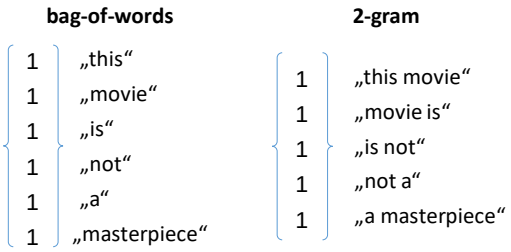

Abbildung 5: bag-of-words und n-gram Textmodell

Ein Nachteil dieses Modells besteht in dem Verlust des Kontextes der Wörter. Beispielsweise lässt das Wort „masterpiece“ eine positive Filmkritik vermuten. Die vorangestellte Negation „not“ verändert jedoch die Semantik des Wortes. Dieser Zusammenhang geht durch die Überführung in ein *bag-of-words* Modell verloren. Als Erweiterung des *bag-of-words* Modells

[92] Vgl. Miner et al. (2012), S. 46.
[93] Vgl. Miner et al. (2012), S. 46.
[94] Vgl. Jurafsky und Martin (2009), S. 117.
[95] Vgl. Bird et al. (2009), S. 230.
[96] Vgl. Bretschneider et al. (2014), S. 7f.
[97] Vgl. Miner et al. (2012), S. 45.
[98] Vgl. Witten et al. (2011), S. 387.
[99] Das Beispiel ist der Arbeit von Zuhang et al. (2006), S. 46 entnommen.

greift das *n-gram* Modell diesen Aspekt auf, indem es n aufeinanderfolgende Wörter in der Multimenge zur Erhaltung eines Teils des Kontextes abbildet.[100] Im *Text Mining* haben sich das *2-gram* und das *3-gram* Modell vor dem Hintergrund der maschinellen Verarbeitbarkeit etabliert.[101] In Abbildung 5 ist exemplarisch die Transformation in ein *2-gram* Modell dargestellt. Grammatische Beziehungen erstrecken sich typischerweise über mehrere Wörter.[102] Zur Modellierung derartiger Beziehungen reichen demnach kleine Werte für n nicht aus. Um beispielsweise die Negation „not" in Bezug zu dem Substantiv „masterpiece" zu setzen, ist bereits ein *3-gram* Modell notwendig.

Sequenzmodelle greifen die Idee des *n-gram* Modells auf, den Kontext durch die Beibehaltung der Reihenfolge der Wörter zu erhalten. Ein *Sequenzmodell* bildet einen unstrukturierten Text in Form von Listen ab.[103] Dieses Modell entsteht im Rahmen der Vorverarbeitung, die einen Text in Sätze aufspaltet und jeweils in eine Liste überführt.[104] Es zeichnet sich durch die vollständige Erhaltung des Kontextes und seinen einfachen Aufbau aus. Zur Überführung von Texten in ein *Sequenzmodell* sind einfache Verfahren des *Natural Language Processing* hinreichend. Mit dem Einsatz fortgeschrittener Verfahren erlaubt es darüber hinaus die Annotation der Elemente der Liste, beispielsweise mit „part-of-speech" Tags. Die „part-of-speech" Tags repräsentieren den grammatischen Typ von Wörtern, beispielsweise Substantive, Verben oder Adjektive.[105] Ein Nachteil dieses Modells besteht darin, dass keine grammatischen Zusammenhänge zwischen den Elementen erfasst werden.

Bäume und Graphen bilden die Grundlage für fortgeschrittene Textmodelle, die sowohl den Kontext der Wörter erhalten als auch grammatische Beziehungen abbilden.[106] Am Beispiel des Satzes „This movie is not a masterpiece" zeigt Abbildung 6 den resultierenden *Dependency Graph*. Das Beispiel beinhaltet verschiedene Informationen bezüglich der Wörter und deren grammatische Relationen. In jedem Knoten ist das Wort selbst sowie sein grammatischer Typ („part-of-speech") enthalten. Schließlich werden die grammatischen Relationen in Form von Kanten visualisiert.[107] Beispielsweise umfasst die Relation „nsubj(movie, is)" das syntaktische Subjekt des Teilsatzes inklusive der Verbindung zum Prädikat.[108] Das Subjekt wird im Beispiel

[100] Vgl. Murphy (2012), S. 591.
[101] Vgl. Chen et al. (2012), S. 72; Jurafsky und Martin (2009), S. 146f.
[102] Vgl. Chen et al. (2012), S. 76f.
[103] Vgl. Pustejovsky und Stubbs (2012), S. 160.
[104] Vgl. Jurafsky und Martin (2009), S. 103f.
[105] Vgl. Zhang und Liu (2014), S. 7f; Bird et al. (2009), S. 179f.
[106] Vgl. Zhang und Liu (2014), S. 8; Bird et al. (2009), S. 310.
[107] Vgl. Zuhang et al. (2006), S. 46.
[108] Vgl. Marneffe und Manning (2008), S. 7.

mit einer Stimmung bewertet. Im Kontext der *Sentiment Analysis* sind derartige grammatische Relationen für die Analyse zentral, da sie das Bezugsobjekt der Aussage enthalten.[109] Darüber hinaus zeigt das Beispiel den Zusammenhang zwischen grammatischen Relationen und der Semantik von stimmungsbehafteten Aussagen. Das grundsätzlich positive Substantiv „masterpiece" erhält eine negative Stimmung durch die vorangestellte Negation „not". Zur Erkennung dieses Zusammenhangs ist die grammatische Relation zwischen der Negation und dem Prädikat auszuwerten.

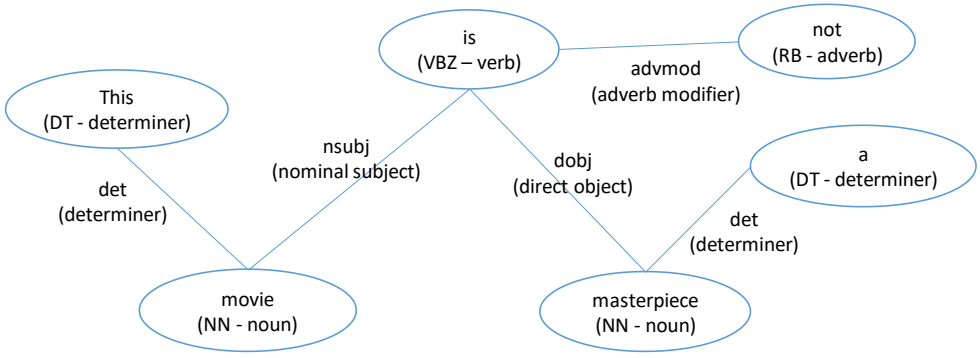

Abbildung 6: Dependency Graph[110]

Im Gegensatz zu den zuvor vorgestellten Textmodellen ist die Ermittlung des grammatischen Typs bei dem *Dependency Graph* nicht optional, sondern eine Voraussetzung zur Identifikation der grammatischen Beziehungen.[111] Demnach stellt die Überführung in Baum- und Graphen-Modelle höhere Anforderungen an Verfahren des *Natural Language Processing*.[112] Nutzergenerierte Texte zeichnen sich jedoch oft durch eine fehlerhafte Rechtschreibung und grammatische Struktur aus.[113] Dadurch ist die korrekte Überführung in dieses Modell fehleranfällig.[114]

3.3 Verfahren des Text Mining

Verfahren des *Text Mining* bauen auf den zuvor vorgestellten Textmodellen auf. In diesem Abschnitt werden Verfahren aus dem Bereich der *Document Classification* als übergeordnetes Gebiet der *Hate Speech* Erkennung vorgestellt und hinsichtlich einsetzbarer Textmodelle diskutiert. Während moderne *Text Mining* Anwendungen häufig auf Verfahren des *Machine Learning* basieren[115], haben sich insbesondere im Bereich der *Sentiment Analysis* sogenannte

[109] Vgl. Zuhang et al. (2006), S. 46.
[110] Zuhang et al. (2006), S. 46.
[111] Vgl. Cer et al. (2010), S. 2.
[112] Vgl. Zhang und Liu (2014), S. 7f.
[113] Vgl. Feldman (2013), S. 89; Sood et al. (2012a), S. 1484.
[114] Vgl. Miner et al. (2012), S. 48; Cer et al. (2010), S. 3; Jurafsky und Martin (2009), S. 191ff.
[115] Vgl. Miner et al. (2012), S. 14.

Lexikon Verfahren etabliert.[116] Darüber hinaus existieren Regel-basierte Systeme, die sich sowohl für die Klassifikation von Dokumenten als auch zur Ermittlung von Bezugsobjekten eignen.[117]

Lexikon Verfahren analysieren einen gegebenen Text hinsichtlich der Präsenz von Wörtern, die in einem Lexikon spezifiziert sind.[118] Ein Lexikon Eintrag besteht aus einem Wort und einem numerischen Wert, der die Klassenzugehörigkeit des Wortes abbildet. Der numerische Wert ermöglicht eine Gewichtung von Wörtern. Das Gewicht entspricht der Wahrscheinlichkeit, ob aufgrund der Präsenz des zugehörigen Wortes ein gegebener Text zu einer bestimmten Klasse gehört.[119] Eine Scoring-Funktion verknüpft schließlich die numerischen Werte aller identifizierten Wörter und bildet sie auf einen aggregierten Wahrscheinlichkeitswert ab, der die Entscheidungsgrundlage für die Klassifizierung darstellt.[120] Da die Präsenz bestimmter Wörter für die Klassifizierung hinreichend ist und der Kontext somit eine untergeordnete Rolle spielt, verwenden Lexikon Verfahren typischerweise *bag-of-words* Modelle zur Abbildung von Texten.[121]

Machine Learning Verfahren unterteilen sich in Verfahren des überwachten und nicht überwachten Lernens.[122] Im Bereich des *Text Mining* haben sich für Klassifikationsprobleme mit bekannten Klassen Verfahren des überwachten Lernens etabliert, weshalb diese Verfahren im Folgenden fokussiert werden.[123]

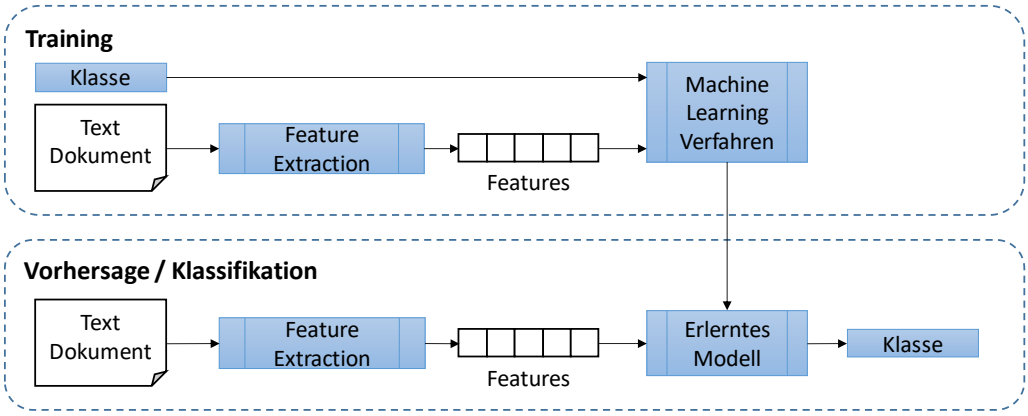

Abbildung 7: Trainings- und Klassifikationsphase bei überwachtem[124]

[116] Vgl. Feldman (2013), S. 83f; Tsytsarau und Palpanas (2012), S. 486f; Pang und Lee (2008), S. 27f.
[117] Vgl. Miner et al. (2012), S. 882; Witten et al. (2011), S. 69ff.
[118] Vgl. Tsytsarau und Palpanas (2012), S. 486f; Pang und Lee (2008), S. 27f.
[119] Vgl. Tsytsarau und Palpanas (2012), S. 486f; Pang und Lee (2008), S. 27f.
[120] Vgl. Tsytsarau und Palpanas (2012), S. 486f; Pang und Lee (2008), S. 27f.
[121] Vgl. Kontostathis et al. (2013), S. 196f; Sood et al. (2012b), S. 277.
[122] Vgl. Miner et al. (2012), S. 17; Murphy (2012), S. 2.
[123] Vgl. Miner et al. (2012), S. 17; Murphy (2012), S. 3; Tsytsarau und Palpanas (2012), S. 484ff.
[124] Vgl. Bird et al. (2009), S. 222.

Verfahren des überwachten Lernens beruhen auf einer Trainings- und einer Vorhersage- beziehungsweise Klassifikationsphase, die in Abbildung 7 dargestellt sind.[125] Mithilfe der Trainingsdaten ermitteln *Machine Learning* Verfahren eine Klassifikationsregel auf Basis eines mathematischen Optimierungsproblems, die zu möglichst wenigen Fehlentscheidungen beziehungsweise geringen Kosten gemessen an der Zielfunktion führt.[126] Schließlich wird die ermittelte Klassifikationsregel in ein mathematisches Modell überführt, um sie auf unbekannte Instanzen in der folgenden Phase anzuwenden.[127] Das mathematische Optimierungsproblem basiert auf numerischen Variablen. Deshalb findet im Rahmen der Vorverarbeitung eine Transformation der Textdokumente in *Features* statt.[128] Diese *Features* repräsentieren verschiedene Merkmale eines Textes, die auf diskrete oder kontinuierliche Werte abgebildet werden.[129] Die Gesamtheit der definierten *Features* bildet den sogenannten *Feature Space*, ein n-dimensionaler Vektorraum, in dem jede Dimension einem *Feature* entspricht.[130] Ein *Feature* bildet typischerweise die Präsenz eines bestimmten Wortes oder einer Kategorie von Wörtern in Texten ab, weshalb *bag-of-words* Modelle bei der Ermittlung dieser Features Anwendungen finden.[131] Eine Ausnahme bilden spezielle *Machine Learning* Verfahren für Sequenzdaten. Diese Verfahren sind bei entsprechenden Trainingsdaten in der Lage, Sequenzen zu klassifizieren oder deren Elemente zu annotieren.[132] Aufgrund der großen Menge an möglichen Sequenzen aus Sprachdaten unterliegen diese Verfahren vielen Annahmen und sind nur eingeschränkt für die Klassifikation einsetzbar.[133]

Die Wahl der *Features* ist anwendungsspezifisch.[134] Im Kontext der *Hate Speech* Detektion stellen beispielsweise die Präsenz oder die Anzahl von beleidigenden Wörtern im untersuchten Textdokument *Features* dar, während alle anderen Wörter im Rahmen der Klassifikation keine Berücksichtigung finden.[135] Durch diese Art der *Feature* Modellierung bleibt jedoch ein Vorteil von *Machine Learning* Verfahren ungenutzt: Die Identifikation neuer und vorher nicht bekannter Wörter zum Ausdruck von *Hate Speech*. Dies ist aufgrund der ausschließlichen Berücksichtigung bereits definierter Wörter nicht möglich.[136] Bei der Berücksichtigung aller

[125] Vgl. Bird et al. (2009), S. 222.
[126] Vgl. Miner et al. (2012), S. 14.
[127] Vgl. Murphy (2012), S. 2; Bird et al. (2009), S. 222.
[128] Vgl. Domingos (2012), S. 78.
[129] Vgl. Domingos (2012), S. 78.
[130] Vgl. Miner et al. (2012), S. 1024; Murphy (2012), S. 2; Hamel (2009), S. 40.
[131] Vgl. Miner et al. (2012), S. 15; Murphy (2012), S. 5.
[132] Vgl. Pustejovsky und Stubbs (2012), S. 160.
[133] Vgl. Pustejovsky und Stubbs (2012), S. 161.
[134] Vgl. Domingos (2012), S. 82f; Miner et al. (2012), S. 14.
[135] Vgl. Chen et al. (2012), S. 72f.
[136] Vgl. Murphy (2012), S. 1; Witten et al. (2011), S. 4f.

Wörter entstehen demgegenüber hochdimensionale *Feature Spaces*[137], deren Anzahl an Dimensionen der Anzahl der unterschiedlichen Wörter im gesamten Korpus entspricht.[138] Diese hochdimensionalen *Feature Spaces* führen zu dem Problem des „curse of dimensionality".[139] Durch die Zunahme an Dimensionen und somit der Variablen des mathematischen Optimierungsproblems, ist die Lösung des Problems durch computergestützte Verfahren in annehmbarer Zeit schwieriger.[140] Zudem ist die Beschaffung von Trainingsdaten aufwändiger, die eine adäquate Abdeckung der Merkmale der gewünschten Zielklasse gewährleisten. Ein hochdimensionaler *Feature Space* erfordert deshalb eine große Anzahl an Trainingsdaten.[141] Im Kontext der *Hate Speech* Detektion ist es aufgrund der geringen Verfügbarkeit besonders aufwändig, Trainingsdaten zu beschaffen.[142]

Sowohl Lexikon als auch *Machine Learning* Verfahren basieren auf *bag-of-words* oder *n-gram* Modellen. Regel-basierte Verfahren erlauben demgegenüber die Verwendung komplexerer Textmodelle, wodurch grammatische Beziehungen in die Klassifikationsentscheidung einfließen können.[143] Dazu werden Entscheidungsregeln genutzt, die *Features* in einem Regelwerk zueinander in Beziehung setzen und anhand der Ausprägungen der *Features* eine Klassifikationsentscheidung treffen.[144] Eine Regel bildet für die Zielklasse charakteristische grammatische Zusammenhänge ab, anhand deren Präsenz die Klassifikationsentscheidung fällt.[145] Einige Autoren bezeichnen diese Ansätze deshalb als Pattern-basierte Verfahren.[146] Diese grammatischen Zusammenhänge können manuell definiert oder durch den Einsatz von *Machine Learning* Verfahren erlernt werden.[147] Das automatische Erlernen erfordert ein Textmodell, in dem bereits grammatische Beziehungen ermittelt sind. Demnach sind nur Baum- oder Graphen-Modelle geeignet. Demgegenüber erlaubt die manuelle Definition der grammatischen Zusammenhänge den Einsatz von *Sequenzmodellen.*[148]

[137] Vgl. Miner et al. (2012), S. 15.
[138] Vgl. Miner et al. (2012), S. 15.
[139] Vgl. Miner et al. (2012), S. 15; Murphy (2012), S. 18.
[140] Vgl. Miner et al. (2012), S. 15; Murphy (2012), S. 18.
[141] Vgl. Domingos (2012), S. 81f.
[142] Vgl. Kontostathis et al. (2013), S. 201; Delort et al. (2011), S. 14.
[143] Vgl. Gitari et al. (2015), S. 223f; Chen et al. (2012), S. 74f; Spertus (1997), S. 1063.
[144] Vgl. Zhang und Liu (2014), S. 7; Witten et al. (2011), S. 69ff.
[145] Vgl. Gitari et al. (2015), S. 223f.
[146] Vgl. Gitari et al. (2015), S. 223f; Bretschneider et al. (2014), S. 3f.
[147] Vgl. Zhang und Liu (2014), S. 7f.
[148] Der Einsatz Pattern-basierter Verfahren in Kombination mit dem Sequenzmodell wird in Kapitel 5 näher erläutert.

3.4 Evaluation von Verfahren zur Klassifikation von Texten

Die Evaluation von Verfahren zur Klassifikation von Texten dient der Messung der Klassifikationsgüte anhand geeigneter Evaluations-Metriken. In diesem Abschnitt werden verschiedene Evaluations-Metriken vorgestellt, um in der anschließenden Literaturauswertung existierende Verfahren hinsichtlich ihrer Klassifikationsgüte einordnen zu können. Zudem sind mithilfe der Evaluations-Metriken die existierenden Verfahren und die im Rahmen dieser Arbeit vorgestellten Verfahren miteinander vergleichbar. Die Evaluations-Metriken orientieren sich an einer Zielklasse, der sogenannten positiven Klasse.[149] Im Kontext dieser Arbeit umfasst die positive Klasse somit *Hate Speech* und die negative Klasse neutrale Nachrichten.

		Ermittelte Klasse durch das Verfahren	
		Positive Klasse	Negative Klasse
Tatsächliche Klasse	Positive Klasse	true positive (tp)	false negative (fn)
	Negative Klasse	false positive (fp)	true negative (tn)

Tabelle 1: Konfusionsmatrix für binäre Klassifikationsprobleme[150]

Die Metriken berechnen sich aus der Konfusionsmatrix, die die ermittelten Klassen des Verfahrens in Relation zu den tatsächlichen Klassen der Instanzen setzt.[151] Tabelle 1 zeigt eine Konfusionsmatrix für binäre Klassifikationsprobleme. Die gebräuchlichsten Evaluations-Metriken sind *Precision*, *Recall*, *F1* und *Accuracy*, deren Berechnungsvorschriften in Tabelle 2 zusammengefasst sind.[152]

Evaluations-Metrik	Berechnungsvorschrift	Evaluations-Metrik	Berechnungsvorschrift
Precision	$\frac{tp}{tp + fp}$	*Recall*	$\frac{tp}{tp + fn}$
F1	$2 * \frac{precision * recall}{precision + recall}$	*Accuracy*	$\frac{tp + tn}{tp + fp + tn + fn}$

Tabelle 2: Berechnungsvorschriften für typische Evaluations-Metriken[153]

Die *Precision*-Metrik stellt den Anteil der korrekt klassifizierten Instanzen in Relation zu allen Instanzen, die durch das Verfahren als positive Klasse klassifiziert sind. *Precision* ist somit ein Maß für die Korrektheit des Klassifikators. Die *Recall*-Metrik stellt hingegen den Anteil der korrekt klassifizierten Instanzen in Relation zu allen Instanzen, die tatsächlich der positiven Klasse entsprechen. *Recall* ist folglich ein Maß für die Abdeckung des Verfahrens. Die *F1*-Metrik stellt das harmonische Mittel zwischen *Precision* und *Recall* dar. Die *Accuracy*-Metrik

[149] Vgl. Sokolova und Lapalme (2009), S. 428f.
[150] Vgl. Sokolova und Lapalme (2009), S. 429.
[151] Vgl. Sokolova und Lapalme (2009), S. 429.
[152] Vgl. Russel und Norvig (2010), S. 869; Sokolova und Lapalme (2009), S. 429f.
[153] Vgl. Sokolova und Lapalme (2009), S. 430.

ist bei Problemen mit ausgewogenen Klassenverhältnissen gebräuchlich. Sie setzt die insgesamt korrekt erkannten Instanzen der positiven und negativen Klasse ins Verhältnis zur Gesamtzahl der Instanzen.[154] *Accuracy* wird im Rahmen dieser Arbeit nicht verwendet, da der Anteil an positiven Instanzen im Kontext der *Hate Speech* Erkennung gegenüber der Gesamtzahl an Nachrichten sehr gering ist und somit ein unausgewogenes Klassenverhältnis besteht.[155]

[154] Vgl. Sokolova und Lapalme (2009), S. 430.
[155] Vgl. Bretschneider und Peters (2016), S. 8f; Kontostathis et al. (2013), S. 201; Sood et al. (2012a), S. 1484.

4 Stand der Forschung und Forschungsdesign

In diesem Abschnitt wird zunächst der aktuelle Stand der Forschung aufgearbeitet. Darauf aufbauend werden Forschungslücken identifiziert, an denen sich die Zielstellung der Arbeit und somit das Forschungsdesign orientieren.

4.1 Methodik zur Ermittlung des Standes der Forschung

Zur Aufarbeitung des aktuellen Standes der Forschung bezüglich der automatischen Detektion von *Hate Speech*, *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* wird eine repräsentative Menge an Fachbüchern sowie Konferenz- und Journalbeiträgen mithilfe einer strukturierten Literaturrecherche in Anlehnung an die Methodik nach vom Brocke et al. (2009) ausgewertet. Für die Suche nach wissenschaftlich hochwertigen Quellen werden die folgenden Suchmaschinen verwendet: ACM Digital Library, AIS Electronic Library, IEEE Xplore Digital Library, Science Direct (Elsevier), Springerlink und Wiley Online Library. Darüber hinaus wird Google Scholar verwendet, insbesondere um weitere Konferenzbeiträge zu finden. Die folgenden vier Suchanfragen werden gestellt:

- („profanity“ OR „harassment“) AND („online communities“ OR „social networks“ OR „social media“)
- “cyber bullying” OR “cyberbullying” OR “digital bullying”
- “online harassment”
- “hate speech” OR “cyberhate”

Es wird eine Volltextsuche durchgeführt und die Ergebnisse werden nach Relevanz sortiert. Falls mehr als 500 Ergebnisse anfallen, wird die Suche mit der zusätzlichen Schlagwortkombination („detecting“ OR „detection“) präzisiert[156], um die Ergebnissemenge auf Verfahren zur automatischen Erkennung einzugrenzen. Schließlich wird die Suche nach 500 Ergebnissen abgebrochen.

Nach einer Beurteilung der ermittelten Resultate anhand des Titels und des Abstracts in einem ersten Schritt[157], verbleiben 37 potentiell relevante Publikationen. Weitere 11 potentiell relevante Publikationen sind das Ergebnis einer Vorwärts- und Rückwärtssuche mithilfe von Google Scholar.[158] In einem zweiten Schritt wird der Inhalt der ermittelten Quellen analysiert. Dieser Schritt dient der Identifikation relevanter Quellen, die sich an dem Forschungsparadigma der *Design Science* orientieren und Verfahren zur automatischen

[156] Vgl. vom Brocke et al. (2009), S. 9.
[157] Vgl. vom Brocke et al. (2009), S. 8.
[158] Vgl. vom Brocke et al. (2009), S. 9.

Detektion von *Hate Speech* in deutschen und englischen Texten vorstellen. Des Weiteren werden Diskussionsbeiträge und nicht evaluierte Verfahren von der weiteren Literaturanalyse ausgeschlossen. Im Ergebnis verbleiben 15 relevante Quellen.

4.2 Analyse existierender Hate Speech Klassifikationsverfahren

Die zuvor ermittelten Publikationen werden in diesem Abschnitt vorgestellt und eingeordnet. Da die Publikationen die Begriffe *Hate Speech*, *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* nicht einheitlich verwenden, werden die in Kapitel 2.1 eingeführten Definitionen angewandt, um sie hinsichtlich der adressierten Problemstellung zu systematisieren. Diese in Tabelle 3 dargestellte Einordnung verfolgt das Ziel, einen Vergleich der in den Publikationen vorgestellten Software-Artefakte hinsichtlich der Evaluationsergebnisse gemessen am *F1*-Wert durchzuführen. Zusätzlich werden das verwendete Klassifikationsverfahren sowie das Textmodell analysiert.

Autoren	Klassifikationsverfahren	Textmodell	F1-Wert
Detektion von undifferenzierter Hate Speech			
Spertus (1997)	Regel-basiert	Sequenzmodell	64%
Delort et al. (2011)	Machine Learning	bag-of-words/n-gram	-
Chen et al. (2012)	Machine Learning	Dependency Graph	96%
Sood et al. (2012a)	Lexikon	bag-of-words/n-gram	46%
Sood et al. (2012b)	Machine Learning	bag-of-words/n-gram	54%
Detektion von Directed Hate Speech			
Burnap und Williams (2015)	Machine Learning	Dependency Graph	77%
Gitari et al. (2015)	Regel-basiert	Dependency Graph	65%
Ausschließliche Detektion von Online Harassment			
Yin et al. (2009)	Machine Learning	bag-of-words/n-gram	35%
Dinakar et al. (2011)	Machine Learning	bag-of-words/n-gram	-
Dinakar et al. (2012)	Machine Learning	bag-of-words/n-gram	-
Kontostathis et al. (2013)	Lexikon	bag-of-words/n-gram	57%
Nahar et al. (2013)	Machine Learning	bag-of-words/n-gram	35%
Nahar et al. (2014)	Machine Learning	bag-of-words/n-gram	59%
Ausschließliche Detektion von Cyberbullying			
Rafiq et al. (2015)	Machine Learning	bag-of-words/n-gram	69%
Detektion von „bullying traces“			
Xu et al. (2012)	Machine Learning	bag-of-words/n-gram, Sequenzmodell	77%

Tabelle 3: Identifizierte Quellen im Rahmen der Literaturrecherche

In Einklang mit den Ergebnissen von Rafiq et al. (2015) zeigt die Literaturanalyse, dass entgegen der in den jeweiligen Publikationen verwendeten Termini sich ein Großteil der Arbeiten mit der Problemstellung der undifferenzierten Detektion von *Hate Speech*

beschäftigt.[159] Demgegenüber zeichnet sich die Problemstellung der Detektion von *Directed Hate Speech* durch die Identifikation von Ausdrücken aus, die gegen Gruppen gerichtet sind. Burnap und Williams (2015) sowie Gitari et al. (2015) betrachten beispielsweise ausschließlich Gewaltaufrufe und rassistische Äußerungen gegen soziale Gruppen.[160] Die Problemstellung der ausschließlichen Detektion von *Online Harassment* betrachtet nur *Hate Speech* Ausdrücke gegenüber Individuen. Schließlich umfasst die Problemstellung der *Cyberbullying* Detektion, die nachrichtenübergreifende Analyse von *Online Harassment* durch denselben Autor an dasselbe Opfer. Nur die Publikation von Rafiq et al. (2015) widmet sich dementsprechend der Erkennung von *Cyberbullying*. Darüber hinaus stellt die Arbeit von Xu et al. (2012) eine Besonderheit dar. Im Rahmen der Studie werden sogenannte „bullying traces" detektiert. Die Autoren verstehen unter „bullying traces" Nachrichten in sozialen Medien, die Erfahrungen von Personen im Kontext von *Online Harassment* oder *Cyberbullying* enthalten.[161] Das zugehörige Artefakt ist als einziges in der Lage, die beteiligten Rollen solcher Vorfälle zu markieren und wird deshalb im folgenden Abschnitt diesbezüglich ausgewertet.

In einem nächsten Schritt werden die Unterschiede hinsichtlich der Klassifikationsgüte zwischen der Detektion von undifferenzierter *Hate Speech*, *Directed Hate Speech* und *Online Harassment* sowie der Detektion von *Cyberbullying* diskutiert. Die erzielten *F1*-Werte der Verfahren unterscheiden sich deutlich. Der beste erreichte *F1*-Wert entspricht 96% und bezieht sich auf ein Verfahren aus der Kategorie „Detektion von undifferenzierter *Hate Speech*". Für diese Problemstellung ist es aufgrund der oft weitgefassten Definition[162] hinreichend, eindeutige *Hate Speech* Ausdrücke unabhängig vom referenzierten Ziel zu erkennen. Chen et al. (2012) zeigen, dass dieses naive Vorgehen bereits zu guten Klassifikationsergebnissen führt.[163] Sie verbessern die Klassifikationsgüte gegenüber den anderen Verfahren jedoch deutlich, indem sie gegen menschliche Ziele gerichtete Ausdrücke berücksichtigen. Dazu unterteilen sie beleidigende Wörter in zwei Kategorien: eindeutige oder besonders ausdrucksstarke Wörter und weniger ausdrucksstarke Wörter.[164] Ein Text wird nur dann als *Hate Speech* klassifiziert, wenn mindestens ein starkes Wort oder eine Kombination aus einem weniger starken Wort und einer Personenreferenz enthalten ist.[165] Die Verfahren, die ausschließlich *Directed Hate Speech* oder *Online Harassment* detektieren, erzielen mit 77%

[159] Vgl. Rafiq et al. (2015), S. 617.
[160] Vgl. Burnap und Williams (2015), S. 231; Gitari et al. (2015), S. 217.
[161] Vgl. Xu et al. (2012), S. 657.
[162] Vgl. Sood et al. (2012a), S. 1481; Chen et al. (2012), S. 72; Delort et al. (2011), S. 11.
[163] Vgl. Chen et al. (2012), S. 77.
[164] Vgl. Chen et al. (2012), S. 75.
[165] Vgl. Chen et al. (2012), S. 75.

respektive 59% deutlich niedrigere *F1*-Werte. Da nur ein Artefakt in der Kategorie „Detektion von *Cyberbullying*" existiert, sind die Ergebnisse nur bedingt vergleichbar. Der erzielte *F1*-Wert von Rafiq et al. (2015) ist mit 69% vergleichsweise hoch. Dieses Ergebnis relativiert sich jedoch aufgrund der Annahme, dass sämtliche Nachrichten, die mindestens ein beleidigendes Wort unabhängig vom referenzierten Ziel enthalten, bereits als *Online Harassment* gelten.[166] Dadurch ist das Verfahren eher mit denen der undifferenzierten *Hate Speech* Erkennung vergleichbar.

Abschließend wird untersucht, inwiefern die verwendeten Klassifikationsverfahren und Textmodelle mit der Klassifikationsgüte zusammenhängen. Beleidigende Wörter sind zwar typischerweise Bestandteil von *Hate Speech*, ihre Bedeutung hängt jedoch vom jeweiligen Kontext ab.[167] Die Mehrzahl der Verfahren verwendet *bag-of-words* oder *n-gram* Modelle, wodurch der Kontext nur mit Einschränkungen abbildbar ist. Die Arbeiten von Sood et al. (2012a) und Kontostathis et al. (2013) basieren beispielsweise auf einem *bag-of-words* Modell und einem Lexikon Verfahren.[168] Diese Kombination erzielt vergleichsweise moderate *F1*-Werte. In Verbindung mit *bag-of-words* Modellen erreichen Verfahren des überwachten *Machine Learning* geringfügig bessere Ergebnisse, wie beispielsweise das von Rafiq et al. (2015) eingesetzte Verfahren.[169] Software-Artefakte, die den Kontext von beleidigenden Wörtern in die Klassifikation einbeziehen, erzielen deutlich bessere *F1*-Werte. Als eine der ersten Arbeiten wendet Spertus (1997) beispielsweise ein *Sequenzmodell* in Kombination mit manuell definierten Regeln an, um den Kontext von detektierten Schimpfwörtern zu untersuchen.[170] Das Ziel besteht darin, mithilfe der Regeln typische grammatische Muster von *Hate Speech* zu modellieren und für die Klassifikation zu nutzen.[171] Die 47 spezifizierten Regeln sind jedoch sehr feingranular.[172] Dadurch besteht die Gefahr, dass die Regeln an den Datensatz überangepasst (overfitting) sind und nicht auf andere Anwendungsdomänen übertragbar sind.[173] Ein ähnliches Vorgehen wählen Chen et al. (2012), Gitari et al. (2015) sowie Burnap und Williams (2015), die anhand der grammatischen Muster des *Dependency Graph* Regeln definieren, die in *Features* überführt werden.[174] Die Verfahren sind jedoch anfällig für verrauschte Texte, da die korrekte Ermittlung des komplexen Textmodells eine

[166] Vgl. Rafiq et al. (2015), S. 618.
[167] Vgl. Sood et al. (2012a), S. 1484; Dinakar et al. (2012), S. 18:21.
[168] Vgl. Kontostathis et al. (2013), S. 196f; Sood et al. (2012a), S. 1484f.
[169] Vgl. Rafiq et al. (2015), S. 621.
[170] Vgl. Spertus (1997), S. 1058.
[171] Vgl. Spertus (1997), S. 1058.
[172] Vgl. Spertus (1997), S. 1063.
[173] Vgl. Hamel (2009), S. 154f.
[174] Vgl. Burnap und Williams (2015), S. 229f; Gitari et al. (2015), S. 223; Chen et al. (2012), S. 75.

Voraussetzung für die Identifikation der grammatischen Beziehungen und somit der *Features* darstellt.[175] Verrauschte Texte enthalten Slang, Abkürzungen sowie Rechtschreib- und Grammatikfehler und sind typisch für nutzergenerierte Inhalte.[176] Chen et al. (2012) weisen deshalb auf die Bedeutung einer entsprechenden Textvorverarbeitung hin.[177] Zudem besteht das Problem der Beschaffung von manuell klassifizierten Trainingsdaten.[178] Durch die fehlende Verfügbarkeit adäquater Trainingsdaten ist es aufwändig, diese Verfahren für den praktischen Einsatz einzurichten.[179]

Zusammenfassend wird die vergleichsweise niedrige Klassifikationsgüte von Verfahren zur Detektion von *Directed Hate Speech*, *Online Harassment* und somit *Cyberbullying* festgehalten. In diesen Kategorien fehlt es an Verfahren, die den Kontext einbeziehen, um die Klassifikationsgüte zu verbessern. Zudem besteht bei Verfahren auf Basis von überwachtem *Machine Learning* die Problematik der Beschaffung adäquater Trainingsdaten. Somit ergibt sich die erste Forschungslücke: die fehlende Einbeziehung des Kontextes durch geeignete Textmodelle und Klassifikationsverfahren im Rahmen der Erkennung von *Directed Hate Speech*, *Online Harassment* und *Cyberbullying*. Existierende Verfahren zeichnen sich zudem durch eine moderate Klassifikationsgüte aus, die durch die Berücksichtigung des Kontextes Verbesserungspotential birgt.

4.3 Analyse von Verfahren zur Detektion der referenzierten Ziele

In diesem Abschnitt werden die identifizierten Verfahren hinsichtlich der Fähigkeit untersucht, das referenzierte Opfer zu markieren und zu identifizieren. Diese Fragestellung leitet sich aus der Zielstellung der Arbeit ab, die die Detektion von *Hate Speech* einschließlich des referenzierten Opfers umfasst. Auf Basis der Ergebnisse werden weitere Forschungslücken abgeleitet. In Tabelle 4 sind alle Verfahren dargestellt, die die referenzierten Ziele markieren beziehungsweise identifizieren können. Die Markierung der Opfer umfasst die Annotation der Wörter in einem Textmodell, die das Ziel eines *Hate Speech* Ausdrucks beschreiben. Im Fall von *Directed Hate Speech* entspricht dies menschlichen Zielen in Form von Gruppen. Demgegenüber ist das Ziel von *Online Harassment* stets ein Individuum. Obwohl in der Literatur bereits auf die Bedeutung der Erkennung der referenzierten Opfer hingewiesen wurde[180], wird dieser Aspekt in der Mehrzahl der Arbeiten vernachlässigt.

[175] Vgl. Spertus (1997), S. 1063.
[176] Vgl. Sood et al. (2012a), S. 1484.
[177] Vgl. Chen et al. (2012), S. 76.
[178] Vgl. Delort et al. (2011), S. 14.
[179] Vgl. Kontostathis et al. (2013), S. 201.
[180] Vgl. Cohen et al. (2014), S. 53.

Autoren	Opfer-Markierung	Opfer-Identifikation
Ausschließliche Detektion von Online Harassment		
Dinakar et al. (2012)	nein	ja, implizit
Detektion von Cyberbullying		
Rafiq et al. (2015)	nein	ja, implizit
Detektion von „bullying traces“		
Xu et al. (2012)	ja	nein

Tabelle 4: Ergebnisse der Literaturanalyse

Die Verfahren von Chen et al. (2012), Burnap und Williams (2015) sowie Gitari et al. (2015) sind nicht enthalten, da sie zwar Wörter als *Features* in die Klassifikation einbeziehen, die typischerweise Gruppen referenzieren, diese jedoch weder im Text markieren noch weiterverarbeiten.[181] Einzig die Arbeit von Xu et al. (2012) untersucht identifizierte „bullying traces“ hinsichtlich der beteiligten Akteure. Dazu werden die erkannten Texte in ein *Sequenzmodell* überführt, das die Reihenfolge der Wörter in einer Listenstruktur erhält. Dieses Modell wird mithilfe eines für Sequenzen geeigneten und zuvor trainierten *Machine Learning* Ansatzes analysiert, um die Bestandteile der Sequenz mit festgelegten Labeln zu annotieren.[182] Xu et al. (2012) sind dadurch in der Lage, die Opfer zu markieren. Diese Erkennung gelingt jedoch nur mit moderatem Erfolg.[183] Darüber hinaus werden spezielle Trainingsdaten für diesen Einsatzzweck benötigt, deren Beschaffung aufwändig ist.[184]

Somit ergibt sich aus der Analyse die zweite Forschungslücke: die fehlende Fähigkeit der Verfahren, die referenzierten Ziele von *Directed Hate Speech*, *Online Harassment* sowie *Cyberbullying* zu markieren.

Neben der Verbesserung der Klassifikationsgüte durch die Einbeziehung von Personenreferenzen, ist die Markierung dieser Referenzen und die anschließende Identifikation der Opfer eine Voraussetzung für die Erkennung von *Cyberbullying*. Einige Arbeiten erreichen über Annahmen eine implizite Identifikation des Opfers. Rafiq et al. (2015) nehmen vereinfachend an, dass sich alle Kommentare zu Videos auf der Plattform „Vine“ auf den Autor des Videos beziehen, der anhand der Metadaten ablesbar ist.[185] Dinakar et al. (2012) untersuchen Kommentare zu YouTube Videos und nehmen ebenfalls implizit an, dass sich alle Kommentare auf den Autor des Videos beziehen. Dadurch entfällt die Identifikation und wiederholte Zuordnung des Opfers. Eine derartige implizite Identifikation ist in vielen *Online*

[181] Vgl. Burnap und Williams (2015), S. 236f; Gitari et al. (2015), 225f; Chen et al. (2012), S. 75; Spertus (1997), S. 1063.
[182] Vgl. Xu et al. (2012), S. 659f.
[183] Vgl. Xu et al. (2012), S. 661f.
[184] Vgl. Xu et al. (2012), S. 660.
[185] Vgl. Rafiq et al. (2015), S. 619.

Communities, wie beispielsweise Foren, Diskussionsplattformen und sozialen Medien, nicht möglich, da die Textbeiträge sich typischerweise nicht ausschließlich auf einen Nutzer beziehen, sondern die Nutzer sich auch untereinander referenzieren. Demnach leitet sich die dritte Forschungslücke ab: die bisher vernachlässigte Identifikation der Opfer von *Directed Hate Speech* und *Online Harassment* in unstrukturierten Texten sowie die damit einhergehende wiederholte Identifikation der Opfer über mehrere Nachrichten hinweg, die eine Voraussetzung zur Erkennung von *Cyberbullying* darstellt.

4.4 Forschungsdesign

Auf Basis der zuvor identifizierten Forschungslücken werden in diesem Abschnitt die entwickelten Software-Artefakte mit den zugehörigen Zielstellungen und den resultierenden Publikationen dieser kumulierten Dissertation vorgestellt.[186] Abbildung 8 stellt dazu die Artefakte einschließlich der jeweiligen Publikation sowie die Beziehungen zwischen den Artefakten vor.

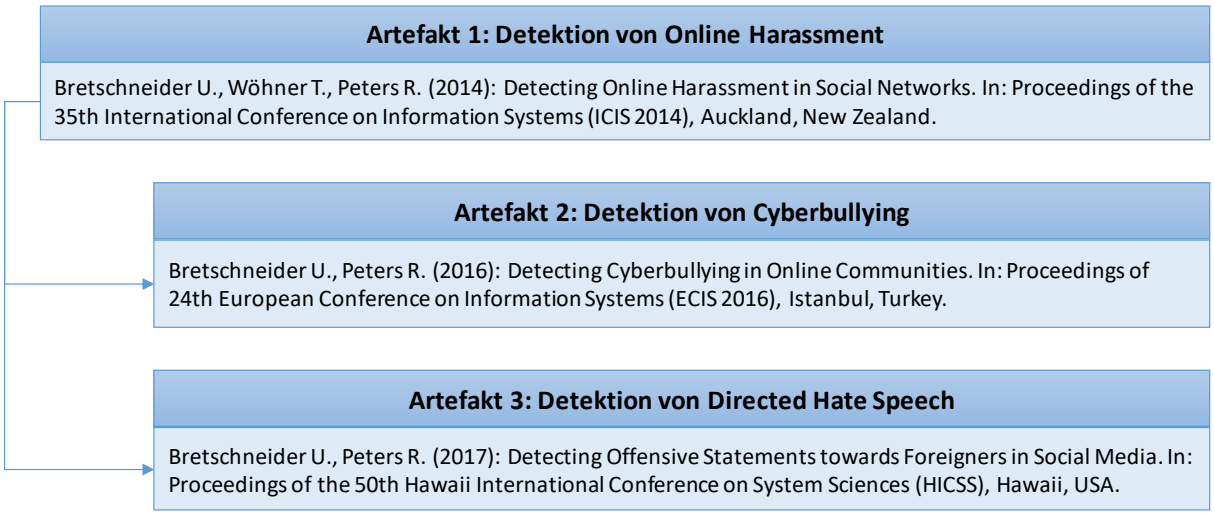

Abbildung 8: Artefakte und Publikationen der kumulativen Dissertation

Artefakt 1 bildet die Grundlage des Forschungsprojekts und umfasst ein Pattern-basiertes Verfahren zur Detektion von *Online Harassment*. Es adressiert die folgenden Ziele, die sich aus den ersten beiden zuvor vorgestellten Forschungslücken ergeben:[187]

1. Konzeption und Entwicklung eines Vorverarbeitungsschrittes zur Detektion von Personenreferenzen
2. Konzeption und Entwicklung eines Pattern-basierten Verfahrens zur Detektion von *Online Harassment*

[186] Vgl. Gregor und Hevner (2013), S. 349.
[187] Vgl. Bretschneider et al. (2014), S. 2.

Das Artefakt basiert auf einem *Sequenzmodell* zur Strukturierung von Texten, um den Kontext der Wörter zu erhalten.[188] Im Gegensatz zu bisherigen Verfahren identifiziert ein neu eingeführter Vorverarbeitungsschritt sprachliche Bezüge zu Individuen und markiert sie für die weitere Verarbeitung.[189] Mithilfe von Sprachmustern, den *Harassment Patterns*, analysiert das Artefakt Verbindungen zwischen diesen Bezügen und beleidigenden Wörtern, um *Online Harassment* effektiv detektieren zu können.[190] Eine Nachricht kann mehrere *Online Harassment* Ausdrücke enthalten, die sich gegen verschiedene Ziele richten. Ein zentraler Vorteil des Verfahrens besteht in der Markierung dieser Passagen einschließlich dem referenzierten Opfer. Mit diesem Vorgehen soll gleichzeitig die Klassifikationsgüte gegenüber bestehenden Verfahren verbessert werden.

Online Harassment ist nach Tokunaga (2010) durch den einmaligen Nachrichtenversand gekennzeichnet, wodurch eine isolierte Betrachtung von Nachrichten hinreichend ist. Wie in der dritten Forschungslücke dargestellt, zeichnet sich *Cyberbullying* hingegen durch den wiederholten Versand von Nachrichten zwischen denselben Nutzern aus, wodurch eine zusammenhängende Betrachtung der Nachrichten notwendig ist.[191] Dazu ist es erforderlich, sowohl den Autor als auch das referenzierte Opfer über mehrere Nachrichten hinweg zu identifizieren. Artefakt 2 verfolgt daran anknüpfend die folgenden Zielstellungen:[192]

1. Konzeption und Entwicklung eines Moduls zur Identifikation von Nutzern in unstrukturierten Texten
2. Konzeption und Entwicklung einer Datenstruktur zur Erfassung von wiederholter Kommunikation zwischen identischen Nutzern
3. Konzeption und Entwicklung eines Verfahrens zur Detektion von *Cyberbullying*

Das Pattern-basierte Verfahren aus Artefakt 1 dient zur Detektion von *Online Harassment* und stellt die Basis für Artefakt 2 dar. Im Rahmen von Artefakt 2 wird die Personenerkennung zu einer Personenidentifikation erweitert, um für jeden detektierten *Online Harassment* Ausdruck den Autor und das referenzierte Opfer zu identifizieren.[193] Dazu verwendet das Verfahren eindeutige Identifikationsmerkmale der *Online Community* Plattform, beispielsweise eindeutige Nutzernamen oder Identifikationsnummern. Mithilfe eines gerichteten Graphen werden alle erkannten *Online Harassment* Fälle einschließlich des Autors und des

[188] Vgl. Bretschneider et al. (2014), S. 4.
[189] Vgl. Bretschneider et al. (2014), S. 4.
[190] Vgl. Bretschneider et al. (2014), S. 7f.
[191] Vgl. Bretschneider und Peters (2016), S. 2.
[192] Vgl. Bretschneider und Peters (2016), S. 2.
[193] Vgl. Bretschneider und Peters (2016), S. 4.

referenzierten Opfers abgebildet, um wiederholte Vorfälle zwischen denselben Akteuren und somit *Cyberbullying* zu erkennen.[194]

Schließlich fokussiert Artefakt 3 *Directed Hate Speech*, die sich gegen Gruppen richtet. Somit wird ein weiterer Aspekt der zweiten identifizierten Forschungslücke adressiert. Artefakt 3 verfolgt daran anknüpfend die folgenden Ziele:[195]

1. Konzeption und Entwicklung eines Vorverarbeitungsschrittes zur Erkennung von Gruppenreferenzen
2. Konzeption und Entwicklung eines Verfahrens zur Erkennung von *Directed Hate Speech*

Die Markierung von Personenreferenzen aus Artefakt 1 und die Personenidentifikation aus Artefakt 2 sind in angepasster Form in Artefakt 3 enthalten. Artefakt 3 markiert Individuen und Gruppenreferenzen im Rahmen der Vorverarbeitung.[196] Somit wird zusätzlich zu *Online Harassment* auch *Directed Hate Speech* gegen Gruppen detektiert. Eine Erweiterung der Personenidentifikation dient schließlich der Identifizierung der Opfer. Das Artefakt ordnet die identifizierten Opfer in vordefinierte Klassen ein, um die referenzierten Ziele von *Directed Hate Speech* zu aggregieren. Da das Artefakt in einem deutschsprachigen Kontext angewendet wird und das ursprüngliche Verfahren in Artefakt 1 für englische Texte konzipiert ist, findet eine Anpassung der Harassment Patterns für die deutsche Sprache statt.[197]

[194] Vgl. Bretschneider und Peters (2016), S. 6.
[195] Vgl. Bretschneider und Peters (2017), S. 2214.
[196] Vgl. Bretschneider und Peters (2017), S. 2217.
[197] Vgl. Bretschneider und Peters (2017), S. 2217.

5 Vorstellung der Publikationen

In diesem Abschnitt werden die Publikationen dieser kumulativen Dissertation vorgestellt. Zu jeder Publikation werden die gesetzten Zielstellungen und das daraus abgeleitete Artefakt vorgestellt. Anschließend werden die Evaluation beschrieben sowie die Ergebnisse insbesondere in Hinblick auf die praktische Anwendbarkeit diskutiert.

5.1 Detektion von Online Harassment

5.1.1 Artefakt zur Detektion von Online Harassment

Das vorgestellte Artefakt verfolgt das Ziel, *Online Harassment* auf Basis von in Texten enthaltenen Personenreferenzen zu detektieren. Dazu werden die in Abbildung 9 dargestellten Schritte durchgeführt, die in einer vereinfachten Architekturdarstellung visualisiert sind. Die Architektur umfasst die drei zentralen Module des Artefakts und stellt über gerichtete Kanten deren Zusammenhänge im Verarbeitungsprozess der Textdokumente dar.

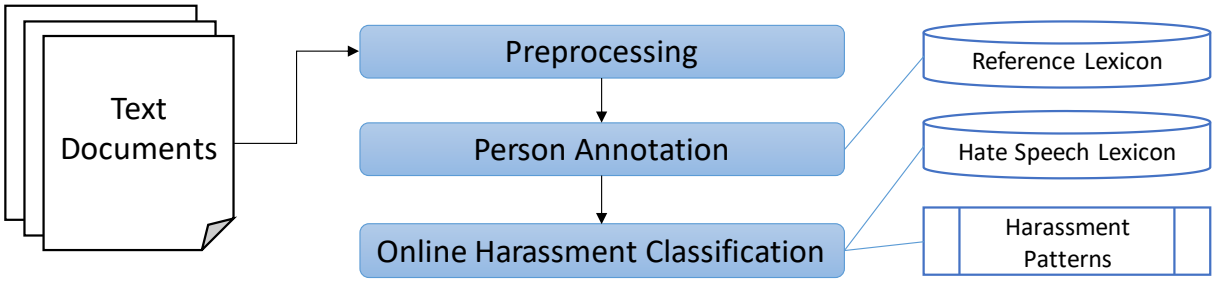

Abbildung 9: Architektur des Artefakts zur Detektion von Online Harassment[198]

Der Vorverarbeitungsschritt („preprocessing“) zerlegt unstrukturierte Textdokumente in seine atomaren Bestandteile und überführt sie in ein Textmodell. Das Artefakt verwendet ein *Sequenzmodell*, um gegenüber *bag-of-words* Modellen die Reihenfolge und damit den Kontext der Wörter zu erhalten. Darüber hinaus wird im Rahmen der Vorverarbeitung jedem Wort sein grammatischer Typ („part-of-speech“) zugeordnet sowie eine Normalisierung durchgeführt.[199] Nutzergenerierte Inhalte enthalten typischerweise Abkürzungen, Slang oder Tippfehler, wodurch die automatisierte Verarbeitung und somit der Klassifikationsprozess erschwert werden.[200] Eine Normalisierung überführt diese Elemente in ihre kanonische Form und trägt somit zur Verbesserung der Klassifikationsergebnisse bei.[201] Das „Person Annotation“ Modul erkennt Personenbezüge in Texten und markiert sie für die weitere Verarbeitung, sodass sie als

[198] Vgl. Bretschneider et al. (2014), S. 4.
[199] Vgl. Bretschneider et al. (2014), S. 3f.
[200] Vgl. Feldman (2013), S. 89; Sood et al. (2012a), S. 1484.
[201] Vgl. Bretschneider et al. (2014), S. 9; Sood et al. (2012a), S. 1484.

Klassifikationsmerkmal in den Prozess einfließen.[202] Nach Chen et al. (2012) verbessert die Beachtung dieses Merkmals die Klassifikationsgüte deutlich.[203] Personenbezüge werden direkt oder indirekt formuliert. Im Kontext von *Online Communities* drücken Nutzer direkte Referenzen typischerweise über Nutzernamen aus, während sie für indirekte Referenzen häufig Personalpronomen verwenden. Das Modul ist in der Lage, beide Arten von Referenzen zu erkennen.[204]

Nachdem die Personenbezüge markiert sind, analysiert das Pattern-basierte Verfahren Texte hinsichtlich der Präsenz von *Online Harassment*. Patterns stellen in diesem Kontext sprachliche Verbindungen zwischen Personenbezügen und Beleidigungen dar, die typischerweise für *Online Harassment* genutzt werden.[205] Es resultieren sieben Patterns aus der Analyse eines Trainingsdatensatzes unter der Zielsetzung einer möglichst allgemeingültigen Anwendbarkeit.[206] In Abbildung 10 ist exemplarisch das „is-a Pattern" dargestellt.

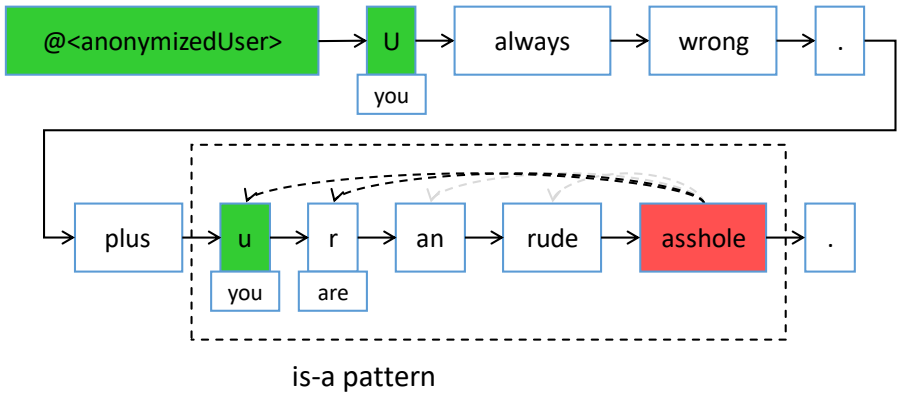

Abbildung 10: Anwendung des is-a Patterns im Sequenzmodell

In dem Beispiel sind die durch den Vorverarbeitungsschritt markierten Personenbezüge grün hervorgehoben. Es sind zwei indirekte Referenzen über Personalpronomen und eine direkte Referenz über einen Nutzernamen enthalten. Das Pattern-basierte Verfahren durchläuft den Text und sucht nach beleidigenden Wörtern, die in einem Lexikon hinterlegt sind. Sobald ein solches Wort gefunden ist, werden die damit verknüpften *Harassment Patterns* abgeglichen. Im Beispiel ist das rot hinterlegte Wort der Ausgangspunkt für den Abgleich des „is-a Patterns", das bei beleidigenden Substantiven Anwendung findet. Dieses Pattern klassifiziert genau dann eine Textpassage als *Online Harassment*, falls sich die notwendigen Bestandteile im Kontext des Schimpfwortes befinden und nur erlaubte Bestandteile dazwischenstehen. Die notwendigen

[202] Vgl. Bretschneider et al. (2014), S. 3f.
[203] Vgl. Chen et al. (2012), S. 75.
[204] Vgl. Bretschneider et al. (2014), S. 3f.
[205] Vgl. Bretschneider et al. (2014), S. 3f.
[206] Vgl. Bretschneider et al. (2014), S. 7f.

Bestandteile sind eine Personenreferenz und eine Form von „to be“, erlaubte Bestandteile sind beispielsweise Adjektive. Da diese Elemente vorhanden sind, markiert das Verfahren die Passage als *Online Harassment*. Die notwendigen und erlaubten Bestandteile sind Pattern-spezifisch definiert und in der Publikation von Bretschneider et al. (2014) erläutert.

5.1.2 Evaluation

Die Evaluation des Artefakts ist in der zugehörigen Publikation detailliert vorgestellt. In diesem Abschnitt werden die zentralen Ergebnisse präsentiert. Die Evaluation wird auf Basis von manuell annotierten Twitter Daten in zwei voneinander unabhängig erhobenen Datensätzen durchgeführt.[207] Zur Messung der Klassifikationsgüte werden die für Klassifikationsprobleme typischen Metriken *Precision*, *Recall* und *F1* ermittelt.[208] In Tabelle 5 sind die Resultate der Evaluation zusammengefasst.

	Datensatz 1			Datensatz 2		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	35%	70%	46,67%	43,63%	70,62%	53,94%
Baseline mit Erkennung von Personenbezügen	58,75%	64,09%	61,3%	70,05%	67,53%	68,77%
Pattern-basiert (ausbalanciert)	73,45%	71,82%	72,64%	79,01%	65,98%	71,91%
Pattern-basiert (restriktiv)	94,74%	32,73%	48,65%	91,67%	22,68%	36,36%

Tabelle 5: Evaluationsergebnisse der Online Harassment Klassifikation[209]

Ein Baseline-Klassifikator basierend auf einem Lexikon-Verfahren und einem *bag-of-words* Modell dient als Vergleichsbasis.[210] Um den Einfluss der Erkennung von Personenbezügen auf die Klassifikationsresultate zu messen, wird der Baseline Klassifikator in zwei Varianten implementiert.[211] Die erste Variante entspricht einer naiven Strategie, die nur typische beleidigende Wörter in die Klassifikation einbezieht. Anhand des erzielten *Recall*-Wertes wird ersichtlich, dass diese Variante einen großen Teil der im Datensatz vorhandenen *Online Harassment* Nachrichten erkennt. Demgegenüber steht jedoch ein niedriger *Precision*-Wert, der aus falsch klassifizierten Nachrichten resultiert, die beleidigende Wörter enthalten, aber kein *Online Harassment* darstellen.[212] Im Resultat entsteht ein moderater *F1*-Wert. Der Einfluss der Erkennung von Personenbezügen auf die Klassifikationsgüte ist durch die Erweiterung des

[207] Vgl. Bretschneider et al. (2014), S. 4f.
[208] Vgl. Sokolova und Lapalme (2009), S. 429f.
[209] Vgl. Bretschneider et al. (2014), S. 9.
[210] Vgl. Bretschneider et al. (2014), S. 9.
[211] Vgl. Bretschneider et al. (2014), S. 9f.
[212] Vgl. Bretschneider et al. (2014), S. 9f.

Baseline-Klassifikators ersichtlich. Der *Precision*-Wert erhöht sich deutlich bei gleichzeitiger Reduzierung des Anteils an falsch klassifizierten Nachrichten.

Das Pattern-basierte Verfahren ist in zwei Konfigurationen evaluiert, die sich hinsichtlich der verwendeten Patterns unterscheiden. Die ausgewogene Konfiguration umfasst alle sieben Patterns und erzielt hinsichtlich des *F1*-Wertes die besten Resultate in beiden Datensätzen.[213] Gegenüber dem Baseline Verfahren erreicht der Ansatz deutlich höhere *Precision*-Werte. Gleichzeitig bleibt der *Recall*-Wert nahezu konstant, sodass nicht wesentlich weniger *Online Harassment* Nachrichten erkannt werden. Im Resultat ist der *F1*-Wert wesentlich höher. Die zweite Konfiguration besteht aus Patterns, die bezüglich der erlaubten Sprachelemente im Kontext von beleidigenden Wörtern restriktiver parametrisiert sind.[214] Dadurch erfasst das Verfahren tendenziell eindeutige *Online Harassment* Nachrichten, was sich in den erreichten *Precision*-Werten widerspiegelt. Demgegenüber werden weniger der insgesamt vorhandenen *Online Harassment* Nachrichten erkannt, was sich in den niedrigeren *Recall*-Werten zeigt. Somit stellt die Wahl der Konfiguration einen Trade-off zwischen *Precision* und *Recall* dar. In beiden Datensätzen erzielen die Verfahren ähnliche Ergebnisse. Dies deutet darauf hin, dass die Patterns nicht dem Phänomen des „overfittings", d. h. der Überanpassung der Regeln an einen Trainingsdatensatz, unterliegen.[215]

5.1.3 Diskussion und Ausblick

Das vorgestellte Artefakt stellt die Basis für die in Kapitel 2.2 vorgestellten Ansätze zum computergestützten Umgang mit *Online Harassment* dar. Diese Ansätze dienen einerseits dazu, den Arbeitsaufwand für Moderatoren im Kontext der typischerweise großen Nachrichtenmenge in *Online Communities* zu reduzieren und andererseits dazu, die Opfer vor psychischen Schäden zu bewahren.

Im Rahmen eines vollautomatischen und proaktiven Ansatzes blockiert das IT-System vom Klassifikator detektierte *Online Harassment* Nachrichten automatisch, sodass keine Publikation stattfindet und somit kein psychischer Schaden entsteht. Allerdings ist in diesem Prozess keine menschliche Kontrollinstanz beteiligt, die die Klassifikation überprüft. Bei vollautomatischen IT-Systemen ist demnach die Vermeidung von falsch positiven Resultaten wichtig, die durch den *Precision*-Wert erfasst werden. Das vorgestellte Artefakt zeichnet sich durch einen hohen *Precision*-Wert aus, insbesondere unter Verwendung von Patterns, die eindeutige *Online Harassment* Fälle abdecken. Dadurch erzielt das Verfahren *Precision*-Werte von über 90%,

[213] Vgl. Bretschneider et al. (2014), S. 10.
[214] Vgl. Bretschneider et al. (2014), S. 10.
[215] Vgl. Hamel (2009), S. 154f.

sodass nur jede zehnte Nachricht fälschlicherweise als *Online Harassment* klassifiziert ist. Im Hinblick auf die gesetzte Zielstellung und im Vergleich zu existierenden Verfahren aus dem Bereich der *Online Harassment* Detektion, ist der erreichte *Precision*-Wert wesentlich höher, wodurch ein vollautomatisierter Einsatz möglich ist.[216] Trotz des damit einhergehenden geringen *Recall*-Wertes blockiert das Verfahren unter Berücksichtigung des hohen Nachrichtenaufkommens in *Online Communities* eine hohe Anzahl an *Online Harassment* Nachrichten.

Halbautomatische IT-Systeme markieren *Online Harassment* Nachrichten, um sie in einem nachgelagerten Schritt einer menschlichen Kontrollinstanz zu präsentieren. Die Systeme unterstützten Personal der *Online Community*, um den Arbeitsaufwand zu reduzieren. Dazu filtert der Klassifikator potentielle *Online Harassment* Nachrichten, um die zu analysierende Nachrichtenmenge zu reduzieren. Damit möglichst viele der tatsächlichen *Online Harassment* Nachrichten, aber möglichst wenige falsch positive Resultate in dieser vorgefilterten Menge enthalten sind, ist sowohl ein hoher *Precision*-Wert als auch ein hoher *Recall*-Wert und somit einer hoher *F1*-Wert von Vorteil. Der vorgestellte Ansatz erfüllt diese Anforderungen in der ausgeglichenen Konfiguration am besten, die einen *F1*-Wert von ca. 72% erzielt. Die ausgeglichene Konfiguration erzielt somit deutlich bessere Klassifikationsergebnisse als existierende Verfahren im Bereich der *Online Harassment* Erkennung.

Das vorgestellte Artefakt ist nicht frei von Limitationen. Das Verfahren berücksichtigt keine Gruppen als referenziertes Ziel von *Online Harassment*, da es sich strikt an der Definition von Tokunaga (2010) orientiert. Es ist zudem fraglich, inwieweit sich ein Individuum mit einer Gruppe identifiziert und aufgrund einer Beleidigung gegenüber der Gruppe bei dem Individuum psychischer Schaden eintritt. Eine Erweiterung des Vorverarbeitungsschrittes zur Personenerkennung erlaubt jedoch die Ergänzung dieser Funktionalität. Das verwendete Normalisierungsmodul ist in der Lage, Abkürzungen, Slang und Rechtschreibfehler zu erkennen und die jeweiligen Wörter in ihre kanonische Form zu überführen. Das Modul normalisiert jedoch nicht alle Wörter korrekt. Insbesondere mehrdeutige beziehungsweise kontextabhängige Abkürzungen und Slang stellen das System vor Herausforderungen. Gegenüber existierenden Verfahren erlaubt das eingesetzte *Sequenzmodell* die Verarbeitung von beleidigenden Phrasen, die aus mehreren zusammenhängenden Wörtern bestehen. Das Lexikon enthält häufig verwendete Phrasen im Kontext von *Online Harassment*. Aufgrund der Vielzahl von sprachlichen Varianten sind nicht alle bekannten beleidigenden Phrasen integriert.

[216] Vgl. Bretschneider et al. (2014), S. 12.

Schließlich misst das Verfahren nicht die Intensität von *Online Harassment*, da ein Gewichtungs-Wert im Lexikon fehlt. Diese Funktionalität ist nicht enthalten, da die wahrgenommene Intensität von beleidigenden Wörtern subjektiv ist.[217] Demgegenüber ist ein Intensitäts-Wert für hybride Systeme relevant, um zwischen Fällen zu unterscheiden, die voll- und halbautomatisch verarbeitet werden.

5.2 Detektion von Cyberbullying

5.2.1 Artefakt zur Detektion von Cyberbullying

Das Artefakt zur Detektion von *Cyberbullying* verfolgt das Ziel, wiederholte Beleidigungen durch denselben Autor gegenüber demselben Opfer zu erkennen. Während die Erkennung von Personenbezügen zur Detektion von *Online Harassment* in voneinander isolierten Nachrichten hinreichend ist, erfordert die Detektion von *Cyberbullying* die nachrichtenübergreifende eindeutige Zuordnung der beteiligten Nutzer. Die Autoren von Nachrichten in *Online Communities* sind oft bekannt, da diese Information häufig Teil der Nachrichten-Metadaten ist.[218] Demgegenüber referenzieren die Autoren typischerweise im unstrukturierten Nachrichtentext die Opfer, wodurch ein direktes Auslesen des betreffenden Nutzers nicht möglich ist.[219] Darüber hinaus verwenden Autoren häufig indirekte Personenreferenzen, wie beispielsweise Personalpronomen, die keine direkten Rückschlüsse auf die Identität der referenzierten Person zulassen.[220] Das Software-Artefakt ist, wie in Abbildung 11 dargestellt, um eine Graphen-basierte Datenstruktur und ein Modul zur Identifikation von Nutzern erweitert, um nachrichtenübergreifende Interaktion zwischen identischen Nutzern zu erkennen.

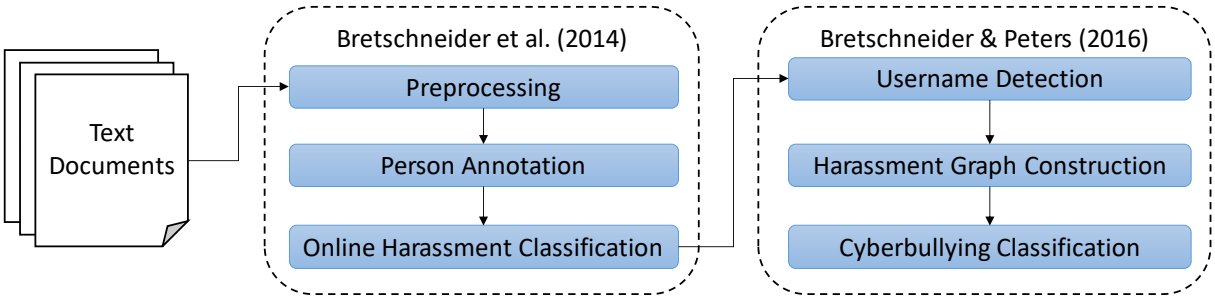

Abbildung 11: Architektur des Artefakts zur Detektion von Cyberbullying[221]

Das "Username Detection" Modul ist für die Identifikation von Nutzern zuständig, die in unstrukturierten Texten referenziert werden. Dazu verwendet es Nutzernamen als eindeutiges

[217] Vgl. Giménez Gualdo et al. (2015), 229f.
[218] Vgl. Bretschneider und Peters (2016), S. 5.
[219] Vgl. Bretschneider und Peters (2016), S. 5.
[220] Vgl. Bretschneider und Peters (2016), S. 5.
[221] Vgl. Bretschneider und Peters (2016), S. 4.

Identifikationsmerkmal, die ein integraler Bestandteil vieler *Online Communities* sind.[222] Das Modul führt die Identifikation mithilfe von vier im Rahmen der Publikation entwickelten Strategien durch.[223] Die Strategien sind im Detail in der entsprechenden Publikation beschrieben. Sie basieren auf den typischen Varianten, andere Nutzer in *Online Communities* zu referenzieren, beispielweise über die Verwendung einer Zitat-Funktion (Quote-Funktion) oder direkt aufeinanderfolgenden wechselseitigen Antworten.[224]

Durch die Identifikation der referenzierten Nutzer in den *Online Harassment* Textpassagen einer Nachricht ist eine nachrichtenübergreifende Analyse und somit eine Re-Identifikation möglich. Das „Harassment Graph Construction" Modul organisiert dazu die detektierten Nachrichten einschließlich dem Autor und dem referenzierten Opfer in einer Graphen-Struktur, dem Harassment Graph.[225] In Abbildung 12 ist exemplarisch ein Ausschnitt eines Harassment Graphen abgebildet. Diese Graphen-basierte Datenstruktur umfasst einen gerichteten Graphen, dessen Knoten Nutzer und dessen Kanten *Online Harassment* Nachrichten repräsentieren.[226] Die gerichteten Kanten repräsentieren den Autor als Quelle und das adressierte Opfer als Ziel sowie die Anzahl der versendeten *Online Harassment* Nachrichten als Kantengewicht.

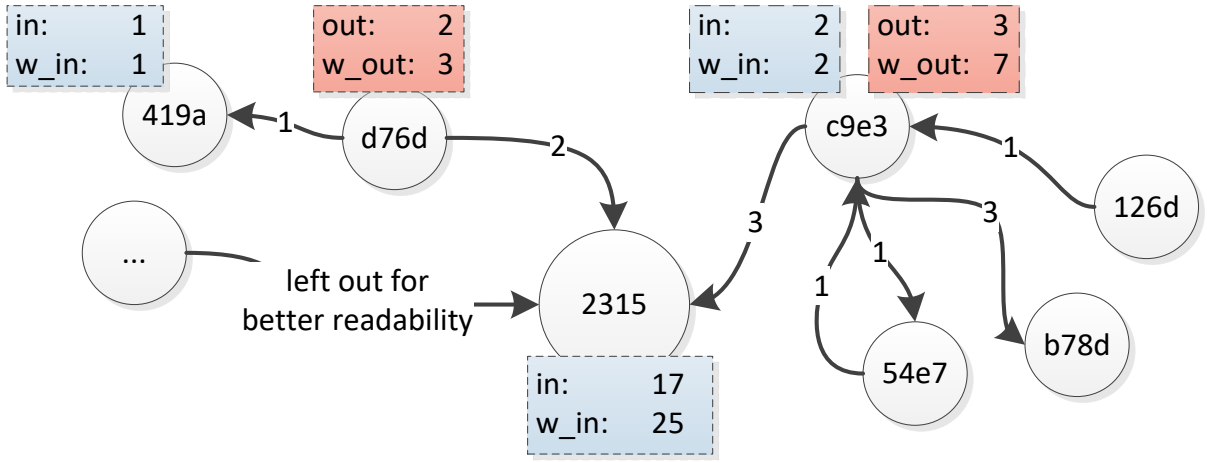

Abbildung 12: Harassment Graph[227]

Im finalen Schritt bestimmt das „Cyberbullying Classification" Modul anhand des Graphen, wie viele *Online Harassment* Nachrichten ein bestimmter Nutzer an ein bestimmtes Ziel gesendet hat.[228] Falls ein Täter wiederholt dasselbe Opfer referenziert, handelt es sich um einen

[222] Vgl. Bretschneider und Peters (2016), S. 4f.
[223] Vgl. Bretschneider und Peters (2016), S. 5.
[224] Vgl. Bretschneider und Peters (2016), S. 5.
[225] Vgl. Bretschneider und Peters (2016), S. 6.
[226] Vgl. Bretschneider und Peters (2016), S. 6.
[227] Vgl. Bretschneider und Peters (2016), S. 6.
[228] Vgl. Bretschneider und Peters (2016), S. 6.

Cyberbullying Fall. Mit dem Harassment Graph sind zusätzlich zu *Cyberbullying* nach Tokunaga (2010) weitere Szenarien klassifizierbar. Darunter fällt beispielsweise die Identifikation von Opfern, die durch mehrere Täter mit mindestens einer *Online Harassment* Nachricht adressiert sind. Weiterhin ist es durch Graphen-basierte Metriken möglich, die Intensität der Fälle zu messen. Diese Metriken erlauben eine quantitative Analyse hinsichtlich der Anzahl der versendeten beziehungsweise empfangenen *Online Harassment* Nachrichten. Der „weighted indegree" (w_in) umfasst beispielsweise die Gesamtzahl der eingehenden *Online Harassment* Nachrichten.[229] In Abbildung 12 ist somit ersichtlich, dass der Nutzer „2315" mit insgesamt 25 *Online Harassment* Nachrichten konfrontiert ist. Der „indegree" (in) und der „outdegree" (out) stellen demgegenüber die Anzahl der Verbindungen zwischen den Knoten unabhängig von den versendeten Nachrichten dar. Ein „indegree" von 17 bedeutet beispielsweise, dass 17 unterschiedliche Nutzer das Opfer mit *Online Harassment* Nachrichten referenzieren.

5.2.2 Evaluation

Die Evaluation des Software-Artefakts findet anhand von zwei Datensätzen aus Foren zu beliebten Online Spielen statt.[230] Eine Annotation aller Online Harassment Fälle einschließlich der referenzierten Opfer findet manuell statt, da existierende Arbeiten diesen Aspekt vernachlässigen.[231] Der resultierende Datensatz ist für die Verwendung in anderen Forschungsarbeiten publiziert.

Die *Cyberbullying* Klassifikation besteht aus drei aufeinander aufbauenden Schritten.[232] Der *Online Harassment* Klassifikator detektiert in einem ersten Schritt *Online Harassment* Nachrichten. In einem zweiten Schritt identifiziert die „Username Detection" die referenzierten Opfer in den zuvor ermittelten *Online Harassment* Nachrichten und fügt sie in den Harassment Graphen ein. Schließlich identifiziert die „*Cyberbullying* Classification" alle *Cyberbullying* Fälle anhand des Graphen. Die detaillierte Evaluation aller Schritte ist in der entsprechenden Publikation enthalten. Im Rahmen dieser Arbeit wird die Evaluation des letzten Klassifikationsschrittes vorgestellt und diskutiert. Fehler, die in den vorgelagerten Stufen aufgetreten sind, sind in der Evaluation des letzten Schrittes einbezogen.[233] Als Evaluations-

[229] Vgl. Bretschneider und Peters (2016), S. 6.
[230] Vgl. Bretschneider und Peters (2016), S. 7.
[231] Vgl. Bretschneider und Peters (2016), S. 7.
[232] Vgl. Bretschneider und Peters (2016), S. 7f.
[233] Vgl. Bretschneider und Peters (2016), S. 7f.

Metriken dienen analog zur Detektion von *Online Harassment* die für Klassifikationsprobleme typischen Metriken *Precision*, *Recall* und *F1*.[234] Die Resultate sind in Tabelle 6 dargestellt.

Datensatz 1			Datensatz 2		
Precision	Recall	F1	Precision	Recall	F1
87,5%	53,85%	66,67%	93,33%	56%	70%

Tabelle 6: Evaluationsergebnisse der Cyberbullying Klassifikation[235]

Im Hinblick auf den dreistufigen Klassifikationsprozess und der Berücksichtigung von Fehlern in vorangegangen Schritten bei der Berechnung der Evaluations-Metriken, sind die erzielten Ergebnisse bezüglich des *F1*-Wertes substantiell hoch. Die erzielten *Precision*-Werte bedeuten eine geringe Rate an falsch positiven Resultaten. Demnach ist nur ein sehr geringer Anteil an Nutzern fälschlicherweise als *Cyberbully* klassifiziert. Demgegenüber weist das System einen moderaten *Recall*-Wert auf, wodurch das Software-Artefakt circa die Hälfte der tatsächlichen *Cyberbullies* erkennt. Dieser Umstand ist hauptsächlich Fehlern aus den vorhergehenden Schritten geschuldet, da das Artefakt entweder die zum *Cyberbullying* Fall zugehörigen *Online Harassment* Nachrichten nicht detektiert oder die Nutzernamen nicht korrekt identifiziert.[236] Als Vergleichsbasis steht nur die Arbeit von Rafiq et al. (2015) zur Verfügung, die auf Basis von Daten der Plattform YouTube und mit impliziten Annahmen zur Identifikation der Opfer einen vergleichbaren *F1*-Wert von 69% erzielen.

5.2.3 Diskussion und Ausblick

Mithilfe des Harassment Graphen ist es möglich, wiederholte Kommunikation zwischen denselben Akteuren zu erfassen. Diese Datenstruktur stellt die Basis für die Identifikation von komplexeren *Online Harassment* Fällen, insbesondere *Cyberbullying*, dar. Ein Großteil der Opfer von *Online Harassment* und *Cyberbullying* reagiert mit Verdrängung und Isolation.[237] Dadurch besteht die Gefahr, dass keine Meldung dieser Fälle erfolgt und somit den Opfern keine Unterstützung bei der Verarbeitung eventueller psychischer Schäden zukommt. Zudem ist es für Administratoren schwierig, *Cyberbullies* zu identifizieren, falls sie *Online Harassment* Nachrichten über einen längeren Zeitraum und in verschiedenen Bereichen der *Online Community* publizieren.

Das vorgestellte Artefakt dient der Unterstützung von Administratoren bei der Detektion dieser Fälle in *Online Communities*. Im Gegensatz zu existierenden Arbeiten steht insbesondere die

[234] Vgl. Sokolova und Lapalme (2009), S. 429f.
[235] Vgl. Bretschneider und Peters (2016), S. 9.
[236] Vgl. Bretschneider und Peters (2016), S. 9.
[237] Vgl. Tokunaga (2010), S. 281; Li (2007), S. 1787.

Identifikation des Opfers im Vordergrund, damit Dritte gegebenenfalls gezielt eingreifen können, um psychische Schäden zu vermeiden oder bei der Verarbeitung zu helfen. Der Harassment Graph dient Administratoren als Werkzeug, um Details von *Online Harassment* und *Cyberbullying* Fällen einzusehen, die ein System zur Klassifikation von *Online Harassment* nicht bietet. Mithilfe des Graphen identifizieren Administratoren *Cyberbullies* einschließlich deren Opfer unabhängig vom Zeitraum und dem Bereich der versendeten Nachrichten. Der Bereich innerhalb einer *Online Community* kann sich beispielsweise über mehrere Diskussionspunkte erstrecken, die unterschiedliche Administratoren verwalten. Dies ist insbesondere für das Phänomen *Cyberstalking* typisch, bei dem ein *Cyberbully* sein Opfer verfolgt und gezielt *Online Harassment* Nachrichten verschickt.[238] Mithilfe der Graphen-basierten Metriken lässt sich zudem die Intensität derartiger Fälle messen. Dadurch lassen sich einerseits problematische Nutzer und andererseits psychisch stark belastete Opfer identifizieren. Das einführend genannte tragische Beispiel des Suizids eines Teenagers verdeutlicht, wie wichtig ein frühzeitiges Eingreifen und der Schutz der Opfer sind.

Die erzielten Klassifikationsergebnisse zeigen, dass das Verfahren präzise *Cyberbullies* detektiert, selbst unter Berücksichtigung von Fehlern bei der vorgelagerten *Online Harassment* Klassifikation. Das Software-Artefakt erzielt *Precision*-Werte ähnlich der Spam Klassifikation[239], wodurch ein vollautomatischer Einsatz möglich ist. Der moderate *Recall*-Wert von über 50% zeigt, dass das Artefakt nicht alle tatsächlichen *Cyberbullies* erkennt. Dies hängt im Wesentlichen mit der Anzahl der versendeten *Online Harassment* Nachrichten zusammen.[240] Je mehr *Online Harassment* Nachrichten ein Täter an ein Opfer sendet, desto höher ist die Wahrscheinlichkeit, dass der Klassifikator den Täter erkennt. Eine Analyse der erkannten *Cyberbullies* in Abhängigkeit der Anzahl der versendeten *Online Harassment* Nachrichten zeigt, dass der Klassifikator die schwerwiegenden Fälle vollständig erkennt.[241] Die Verwendung eines Schwellwertes verbessert den *Precision*-Wert. Der Klassifikator berücksichtigt nur solche *Cyberbullying* Fälle, die sich durch eine über den Schwellwert spezifizierte Mindestanzahl an detektierten *Online Harassment* Nachrichten auszeichnen. Der Klassifikator markiert die verbleibenden Fälle für eine Kontrolle durch Administratoren.

Das Artefakt ist nicht frei von Limitationen. Bisher unterscheidet es nur zwei Rollen: *Cyberbullies* und Opfer. In *Cyberbullying* Fällen existieren darüber hinaus weitere Rollen, wie

[238] Vgl. Aponte und Richards (2013), S. 18:3.
[239] Vgl. Goh und Singh (2015), S. 439.
[240] Vgl. Bretschneider und Peters (2016), S. 9f.
[241] Vgl. Bretschneider und Peters (2016), S. 9f.

z. B. Zuschauer (sogenannte „bystander“)[242] oder Opfer von *Cyberbullying*, die dadurch selbst zu *Cyberbullies* werden (sogenannte „bully-victims“).[243] Der Einfluss dieser Rollen auf die wahrgenommene Intensität von *Cyberbullying* ist bisher wenig erforscht.[244] Der Harassment Graph ist modifizierbar, um „bully-victims“ anhand der Nachrichtenhistorie und der Auswertung der Zeitstempel zu identifizieren. Die Erkennung von „bystandern“ bedarf weiterer Forschung, da diese Nutzer durch die reine Präsenz in *Cyberbullying* Fällen gekennzeichnet sind und diese Art der Partizipation im Harassment Graph nicht vorgesehen ist. Weiterhin erfassen die Graphen-basierten Metriken die Intensität von *Cyberbullying* Fällen ausschließlich anhand von quantitativen Merkmalen. Qualitative Aspekte, wie die Intensität der verwendeten Beleidigungen, finden keine Berücksichtigung, da das Artefakt keine Gewichtung für die Intensität von detektierten *Online Harassment* Nachrichten verwendet. Weiterhin ist die Nachrichtenfrequenz bisher unberücksichtigt. *Online Harassment* Nachrichten, die ein Täter über einen kurzen Zeitraum an das selbe Opfer schickt, richten möglicherweise höheren psychischen Schaden an als solche, die er mit größerem zeitlichen Abstand verschickt.

5.3 Detektion von Directed Hate Speech

5.3.1 Artefakt zur Detektion von Directed Hate Speech

Das in diesem Abschnitt vorgestellte Artefakt verfolgt das Ziel, *Directed Hate Speech* zu detektieren. Die Entwicklung des Artefakts ist getrieben von der Diskussion um nutzergenerierte Inhalte in deutschen sozialen Medien im Rahmen der Flüchtlingskrise. Deshalb fokussiert es im Wesentlichen *Directed Hate Speech* gegenüber Ausländern. Abbildung 13 zeigt die vereinfachte Architektur des resultierenden Artefakts.

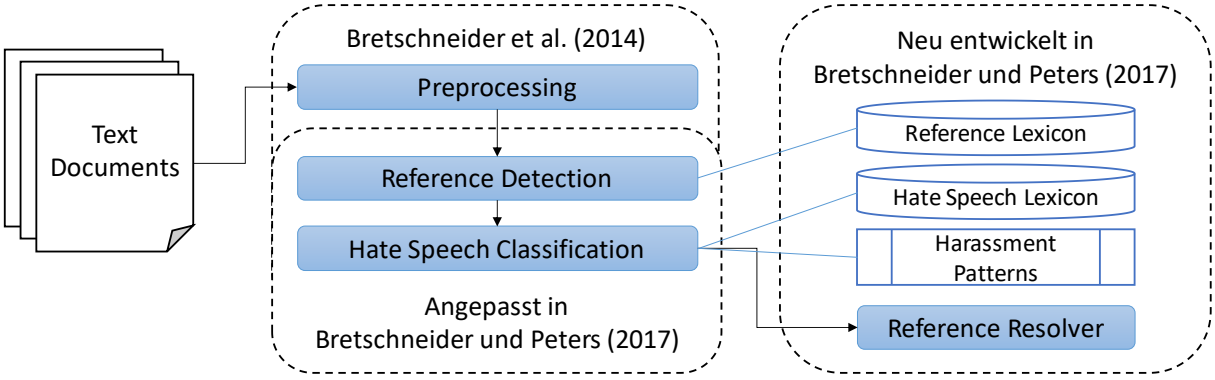

Abbildung 13: Architektur des Artefakts zur Detektion von Directed Hate Speech[245]

[242] Vgl. Xu et al. (2012), S. 657.
[243] Vgl. Giménez Gualdo et al. (2015), 229.
[244] Vgl. Tokunaga (2010), S. 285.
[245] Vgl. Bretschneider und Peters (2017), S. 2217.

Directed Hate Speech gegenüber Ausländern zeichnet sich im Vergleich zu *Online Harassment* und *Cyberbullying* insbesondere dadurch aus, dass Autoren Gruppen als Ziel von beleidigenden, xenophoben oder rassistischen Äußerungen referenzieren.[246] Es findet eine Erweiterung der Personenerkennung aus Artefakt 1 zur „Reference Detection" statt, um Gruppenbezüge und Referenzen zu anderen Entitäten zu erkennen.[247] Darüber hinaus findet eine Anpassung an die deutsche Sprache durch die Einführung eines deutschen „Reference Lexicon" statt. Mithilfe einer Modifikation des Moduls zur Klassifikation von *Online Harassment* lassen sich *Directed Hate Speech* Ausdrücke einschließlich des referenzierten Ziels identifizieren.[248] Das Verfahren basiert auf für die deutsche Sprache entwickelten Patterns in Verbindung mit einem deutschen Lexikon beleidigender Wörter.[249] In Ergänzung dazu unterscheidet es in Anlehnung an Chen et al. (2011) schwere und weniger schwere beleidigende Wörter.[250] Mithilfe dieser Gewichtung erkennt das Verfahren zwei Intensitätswerte von *Directed Hate Speech*.

Der neu konzipierte „Reference Resolver" identifiziert die referenzierten Opfer von detektierten *Directed Hate Speech* Ausdrücken.[251] Im Gegensatz zu Artefakt 2 fokussiert er nicht nur die Identifikation konkreter Individuen, sondern auch die Identifikation von Gruppen. Xenophobe oder rassistische Äußerungen beziehen sich typischerweise auf Volksgruppen und verwenden dazu Wörter, die die Nationalität, die Religion oder besondere Merkmale dieser Volksgruppen umschreiben.[252] Anhand dieser Wörter detektiert der „Reference Resolver" beispielsweise Referenzen zu Ausländern und aggregiert sie in der Klasse „Ausländer und Flüchtlinge".[253] Darüber hinaus sind weitere Klassen definiert, die typischerweise in diesem Kontext Anwendung finden: die Regierung einschließlich aktueller Politiker, die Medien, die Community sowie sonstige Ziele.[254] Die Community umfasst Nutzer, die sich mit der jeweiligen Präsenz in sozialen Medien identifizieren, beispielsweise eine Facebook Seite. Sie ist als Ziel aufgenommen, da sich *Directed Hate Speech* auch gegen diese Nutzer richtet.[255]

[246] Vgl. Bretschneider und Peters (2017), S. 2215.
[247] Vgl. Bretschneider und Peters (2017), S. 2217.
[248] Vgl. Bretschneider und Peters (2017), S. 2217.
[249] Vgl. Bretschneider und Peters (2017), S. 2217.
[250] Vgl. Bretschneider und Peters (2017), S. 2217.
[251] Vgl. Bretschneider und Peters (2017), S. 2218f.
[252] Vgl. Yakushko (2009), S. 44f.
[253] Vgl. Bretschneider und Peters (2017), S. 2217f.
[254] Vgl. Bretschneider und Peters (2017), S. 2216.
[255] Vgl. Bretschneider und Peters (2017), S. 2216.

5.3.2 Evaluation

Die Klassifikation von *Directed Hate Speech* besteht aus zwei Schritten. In einem ersten Schritt identifiziert das Verfahren derartige Ausdrücke.[256] Dieses binäre Klassifikationsproblem wird mit den Metriken *Precision*, *Recall* und *F1* evaluiert.[257] Um die Ergebnisse der Klassifikation zu vergleichen, ist ein Baseline Klassifikator basierend auf einem *Machine Learning* Ansatz implementiert.[258] Tabelle 7 fasst die Ergebnisse der Evaluation der *Directed Hate Speech* Klassifikation zusammen.

	Datensatz 1			Datensatz 2		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	53,57%	76,27%	62,94%	50,65%	71,43%	59,27%
Pattern-basiert	75,26%	61,86%	67,91%	73,89%	53,46%	62,03%

Tabelle 7: Evaluationsergebnisse der Directed Hate Speech Klassifikation[259]

Beide Klassifikatoren erzielen moderate bis gute Resultate hinsichtlich des *F1*-Wertes. Während der Baseline Klassifikator höhere *Recall*-Werte erreicht, erzielt der Pattern-basierte Ansatz deutlich höhere *Precision*-Werte. Dies deckt sich mit den Ergebnissen der zuvor vorgestellten Artefakte im Rahmen dieser Arbeit. Der Baseline Klassifikator scheint falsch positive Ergebnisse aufgrund von *Directed Hate Speech* Nachrichten zu liefern, die zwar beleidigende Wörter enthalten, sich aber nicht gegen Gruppen richten.[260] Demgegenüber grenzt der Pattern-basierte Ansatz diese Fälle besser ab, da er mithilfe des *Sequenzmodells* den Kontext hinsichtlich enthaltener Referenzen untersucht.

In einem zweiten Schritt identifiziert der „Reference Resolver" die referenzierten Ziele der detektierten *Directed Hate Speech* Ausdrücke und ordnet sie in die zuvor definierten Klassen ein.[261] In Tabelle 8 sind die Resultate der Evaluation zusammengefasst. Da mehrere vordefinierte Zielklassen existieren, handelt es sich um ein Mehrklassen-Klassifikationsproblem. Für jede Klasse sind deshalb die Metriken *Precision*, *Recall* und *F1* gesondert ermittelt, wodurch eine feingranulare Evaluation möglich ist.[262] Um die Güte des Verfahrens zur Bestimmung des referenzierten Ziels unabhängig von vorhergehenden Schritten

[256] Vgl. Bretschneider und Peters (2017), S. 2219.
[257] Vgl. Sokolova und Lapalme (2009), S. 429f.
[258] Vgl. Bretschneider und Peters (2017), S. 2219.
[259] Bretschneider und Peters (2017), S. 2219.
[260] Vgl. Bretschneider und Peters (2017), S. 2219f.
[261] Vgl. Bretschneider und Peters (2017), S. 2219f.
[262] Vgl. Sokolova und Lapalme (2009), S. 429f.

zu interpretieren, sind falsch negative Resultate in der *Directed Hate Speech* Erkennung in der Evaluation dieses Schrittes nicht berücksichtigt.[263]

	Datensatz 1			Datensatz 2		
	Precision	Recall	F1	Precision	Recall	F1
Ausländer	**51,79%**	**65,91%**	**58%**	**59,26%**	**33,56%**	**44,44%**
Regierung	**76,32%**	**58%**	**65,91%**	**74,07%**	**51,28%**	**60,61%**
Community	12,5%	20%	15,39%	55,56%	83,33%	66,67%
Medien	81,82%	77,14%	79,41%	80%	100%	88,89%

Tabelle 8: Evaluationsergebnisse der Klassifikation des referenzierten Ziels[264]

Der „Reference Resolver" erzielt in Abhängigkeit der Zielklasse deutlich unterschiedliche Ergebnisse. Er erkennt für die im Rahmen dieser Arbeit zentrale Klasse der „Ausländer" einen moderaten Anteil der Referenzen korrekt. Für die Zielklasse „Regierung" erreicht er ebenfalls moderate Klassifikationsergebnisse. Die manuell annotierte Datenbasis zeigt, dass *Directed Hate Speech* Nachrichten oft sowohl Ausdrücke gegen Ausländer als auch gegen die Regierung enthalten.[265] In diesen Fällen scheint der „Reference Resolver" die referenzierten Ziele nicht korrekt aufzulösen, wodurch Fehlklassifikationen entstehen. Demgegenüber erzielt er sehr gute Ergebnisse für die Klasse „Medien". In dieser Klasse verwenden die Autoren häufig direkte Referenzen, wodurch im Vergleich zu indirekten Referenzen der Auflösungsschritt und damit mögliche Fehler entfallen.[266] Es sind nur wenige *Directed Hate Speech* Nachrichten enthalten, die sich gegen die jeweilige Community richten.[267] Aufgrund der geringen absoluten Anzahl dieser Nachrichten erklären sich die deutlich unterschiedlichen Evaluationsergebnisse, da die Klassifikation einer einzelnen Nachricht große Auswirkung auf die Evaluations-Metrik hat. Eine falsche Ermittlung des referenzierten Ziels von *Directed Hate Speech* gegenüber der Community liegt häufig darin begründet, dass derartige Kommentare nicht eindeutig auf die Community verweisen und oft nur anhand semantischer Zusammenhänge oder anderer Kommentare des jeweiligen Nutzers zuordenbar sind.[268]

5.3.3 Diskussion und Ausblick

Ein aktuelles Projekt der Deutschen Regierung in Kooperation mit Facebook unterstreicht die Bedeutung der Moderation von *Directed Hate Speech*. Diese Kooperation beinhaltet die Gründung einer Task Force bestehend aus Personal von *Online Communities*, Parteien und dem

[263] Vgl. Bretschneider und Peters (2017), S. 2219f.
[264] Bretschneider und Peters (2017), S. 2220.
[265] Vgl. Bretschneider und Peters (2017), S. 2219f.
[266] Vgl. Bretschneider und Peters (2017), S. 2219f.
[267] Vgl. Bretschneider und Peters (2017), S. 2216.
[268] Vgl. Bretschneider und Peters (2017), S. 2216.

deutschen Justizministerium, um insbesondere *Directed Hate Speech* gegenüber Ausländern zu identifizieren und aus Facebook zu entfernen.[269] Durch das hohe Nachrichtenaufkommen in *Online Communities* ist Moderation arbeits- und zeitintensiv.[270] Das im Rahmen der dritten Publikation vorgestellte Artefakt dient als Basis für Systeme, die diese Aufgabe unterstützen.

In Analogie zur Bekämpfung von *Online Harassment* führt der vollautomatische Eingriff in den Publikationsprozess zu einer Unterbindung von *Directed Hate Speech* und somit zu einer Vermeidung der Aufhetzung der Öffentlichkeit. Ein proaktiver Einsatz schützt den Autor und die Plattform vor eventuellen juristischen Konsequenzen. Das Beispiel der Entscheidung des Europäischen Gerichtshofs für Menschenrechte zeigt, dass solche Konsequenzen auch dann für die Plattform entstehen können, wenn die Inhalte nach eingehender Beschwerde bereits durch Moderatoren gelöscht sind.[271] Die Evaluation zeigt, dass das vorgestellte Artefakt mit Einschränkungen für eine vollautomatische Anwendung ohne Beteiligung durch menschliche Kontrollinstanzen geeignet ist. Dabei sind insbesondere der *Precision*-Wert und die Kosten von falsch positiven Resultaten zu betrachten.[272] Ein hoher *Precision*-Wert bedeutet einen geringeren Anteil an falsch positiven Resultaten. Da das Verfahren etwa 70% *Precision* erreicht, fallen 30% falsch positive Resultate an. Diese fälschlich blockierten Nachrichten können die Nutzer frustrieren und sie dazu veranlassen, die *Online Community* zu verlassen.[273] Darüber hinaus steht eine automatische Blockierung von Nachrichten in Konflikt zum Recht auf freie Meinungsäußerung. Es ist daher abzuwägen, ob und gegebenenfalls welche Inhalte zu löschen sind. Im Rahmen dieser Forschungsarbeit wird daher angenommen, dass derartige Festlegungen in den Policies der jeweiligen *Online Community* definiert sind. Die automatische Blockierung lässt sich durch den Intensitätswert feiner abstufen, indem der Klassifikator nur solche Nachrichten automatisch blockiert, die eine hohe Intensität bezüglich des *Directed Hate Speech* Ausdrucks enthalten.[274]

Bei einem halbautomatischen Einsatz markiert das Verfahren *Directed Hate Speech* Ausdrücke, um sie in einem zweiten Schritt einer menschlichen Kontrollinstanz zu präsentieren. Ein wesentlicher Vorteil dieses Ansatzes besteht darin, dass die menschliche Kontrollinstanz individuell entscheiden kann, ob eine Nachricht tatsächlich gegen die Policy verstößt und somit zu löschen ist. Eine Vorauswahl von *Directed Hate Speech* Nachrichten sollte möglichst wenige

[269] Vgl. BBC News (2015).
[270] Vgl. Chen et al. (2011), S. 71; Wise et al. (2006), S. 30.
[271] Vgl. Europäischer Gerichtshof für Menschenrechte (2015); Scott (2015).
[272] Vgl. Miner et al. (2012), S. 889; Witten et al. (2011), S. 163f.
[273] Vgl. Newell et al. (2016), S. 5.
[274] Vgl. Bretschneider und Peters (2017), S. 2220.

falsch positive Resultate enthalten (hohe *Precision*), um den Arbeitsaufwand für die menschliche Kontrollinstanz zu minimieren. Darüber hinaus sollten möglichst viele tatsächliche *Directed Hate Speech* Nachrichten in der Vorauswahl enthalten sein (hoher *Recall*). Die Evaluationsergebnisse zeigen, dass das Artefakt für derartige Anwendungsfälle geeignet ist, da es gute *Precision*-Werte und moderate bis gute *Recall*-Werte erzielt. Im Kontext von *Online Communities* ist neben der Identifikation einzelner *Directed Hate Speech* Nachrichten insbesondere die Identifikation öffentlicher Gruppierungen relevant, die sich durch radikale Ansichten auszeichnen und *Directed Hate Speech* verwenden, um ihre Ideologien zu verbreiten und die Öffentlichkeit aufzuhetzen.[275] Aufgrund der großen Nachrichtenmenge in *Online Communities* ist es hinreichend, einen Teil der *Directed Hate Speech* Nachrichten von derartigen öffentlichen Gruppierungen zu detektieren, um die Gruppierung als Ganzes zu erkennen. Demzufolge sind die Anforderungen an die Klassifikationsgüte gegenüber der Blockierung einzelner Nachrichten geringer.

Das Artefakt weist Limitationen auf. Eine zentrale Limitation besteht in der manuell definierten Menge an Ziel-Klassen. Das Verfahren ordnet detektierte Referenzen zu Individuen oder Gruppen einer der Klassen zu. Zur Hinzunahme weiterer Klassen ist eine Anpassung der „Reference Detection" notwendig. Weiterhin ist die Granularität der Zielklasse vorgegeben. Bisher ist es nicht möglich, die abstrakte Klasse „Ausländer" in feingranulare Klassen, beispielsweise „Ausländer aus Syrien", zu unterteilen. Durch den Einsatz einer Ontologie lassen sich sowohl andere Granularitäten einbeziehen als auch die Klassenbildung automatisieren.[276] Weiterhin ist die Klassifikationsgüte gegenüber den Artefakten zur Detektion von *Online Harassment* und *Cyberbullying* schlechter. Die Resultate sind jedoch nur bedingt vergleichbar, da die ersten beiden Artefakte auf die englische Sprache und das dritte Artefakt auf die deutsche Sprache ausgerichtet sind. Weitere Forschung in Bezug auf Patterns für die deutsche Sprache könnte zu einer Verbesserung der Klassifikationsgüte beitragen. Zudem stehen im Vergleich zur englischen Sprache keine leistungsfähigen Werkzeuge zur Textvorverarbeitung, insbesondere der Normalisierung, bereit. Da die Normalisierung Einfluss auf die Klassifikationsgüte hat[277], führen Verbesserungen in diesen Prozessschritten zur Verbesserung der Klassifikationsergebnisse.

[275] Vgl. Williams und Burnap (2016), S. 213; Glaser et al. (2002), S. 189f.
[276] Vgl. Pustejovsky und Stubbs (2012), S. 76; Russel und Norvig (2010), S. 437f.
[277] Vgl. Sood et al. (2012a), S. 1484.

6 Anwendung der Forschungsergebnisse in anderen Domänen

Während im Kapitel zuvor spezifische Limitationen der jeweiligen Artefakte diskutiert werden, sind in diesem Abschnitt allgemeine Limitationen aufgeführt. Darauf aufbauend wird in Anlehnung an Gregor und Hevner (2013) diskutiert, inwieweit sich die erzielten Forschungsergebnisse generalisieren auf andere Domänen übertragen lassen.[278]

6.1 Limitationen

Eine generelle Herausforderung im Bereich des *Text Mining* ist es, zwischen den Zeilen zu lesen, um die Bedeutung von mehrdeutigen oder versteckten Aussagen zu erkennen.[279] Darunter fallen insbesondere:

- Ironie, Sarkasmus und Zynismus sowie sonstige mehrdeutige Wortspiele[280]
- Umschriebene sexuelle Anspielungen oder Beleidigungen[281]

In Analogie zu Verfahren aus dem Bereich der *Sentiment Analysis*[282] und der Mehrzahl der Verfahren im Bereich der *Hate Speech* Erkennung[283] sind auch die vorgestellten Artefakte im Rahmen dieser Arbeit nicht oder nur mit Einschränkungen in der Lage, derartige Fälle zu verarbeiten. Einzig Dinakar et al. (2012) verwenden eine Wissensbasis in Kombination mit einer Inferenzmaschine, um bestimmte Arten von umschriebenen Beleidigungen zu identifizieren.[284] Zu dieser Art von Beleidigungen zählen sexuell diskriminierende Anspielungen, wie beispielsweise die Aussage „Warum hast du aufgehört Makeup zu tragen?“ gegenüber einer männlichen Person.[285] Diese Ansätze basieren auf einer sorgfältigen Konstruktion der Wissensbasis.[286] Alleine für die Erkennung von sexuellen Anspielungen definieren Dinakar et al. (2012) eine Menge von 200 Fakten.[287]

Die vorgestellten Artefakte kategorisieren nicht die Art der Beleidigung. Während die Artefakte im Rahmen dieser Arbeit die Identifikation der beteiligten Rollen fokussieren, widmen sich andere Arbeiten einer feingranularen inhaltlichen Analyse. Beispielsweise unterscheiden Dinakar et al. (2011) Beleidigungen in Bezug auf die Sexualität, die Rasse oder Kultur und die

[278] Vgl. Gregor und Hevner (2013), S. 351.
[279] Vgl. Miner et al. (2012), S. 996.
[280] Vgl. Miner et al. (2012), S. 996.
[281] Vgl. Miner et al. (2012), S. 996.
[282] Vgl. Feldman (2013), S. 89; Montoyo et al. (2012), S. 678.
[283] Vgl. beispielsweise Delort et al. (2011), S. 27; Spertus (1997), S. 1064.
[284] Vgl. Dinakar et al. (2012), S. 18:12.
[285] Vgl. Dinakar et al. (2012), S. 18:23.
[286] Vgl. Russel und Norvig (2010), S. 307f.
[287] Vgl. Dinakar et al. (2012), S. 18:12.

Intelligenz.[288] Für jede Form der Beleidigung verwenden die Autoren einen eigenen *Machine Learning* Ansatz mit jeweils eigenständigen Trainingsdaten.[289] Delort et al. (2011) wählen eine feinere Einteilung von 11 Klassen[290] in Kombination mit einem zweistufigen Klassifikationsansatz.[291] Die erste Stufe identifiziert *Hate Speech* Nachrichten, die zweite Stufe identifiziert die Kategorie der identifizierten Nachricht.[292] In Anlehnung an dieses Vorgehen lassen sich die vorgestellten Artefakte um diese Stufe erweitern, um die Art der Beleidigung als nachgelagerten Schritt zu bestimmen.

Die vorgestellten Artefakte sind für die englische und deutsche Sprache ausgelegt. Zur Übertragung der Verfahren in andere Sprachen ist die Konstruktion neuer Lexika und *Harassment Patterns* erforderlich. Bisher gibt es kein Rahmenwerk, um diese Anpassung zu systematisieren oder zu teilautomatisieren. Weitere Forschungsarbeit könnte sich der automatisierten Erstellung von *Harassment Patterns* widmen, um eine Adaption für andere Sprachen zu beschleunigen. Um wenige allgemeingültige Patterns zu konstruieren und sprachliche Besonderheiten zu erfassen, ist jedoch eine manuelle Definition hilfreich.

Schließlich sind die Artefakte nicht in der Lage zu entscheiden, ob eine detektierte Beleidigung tatsächlich psychischen Schaden bei einem Opfer auslöst. Dieser Umstand hängt maßgeblich von der Wahrnehmung des Opfers selbst und der jeweiligen Situation ab.[293] In Analogie dazu können die Artefakte nur mit Einschränkungen die wahrgenommene Intensität der Beleidigung messen. Da beide Aspekte von der individuellen Wahrnehmung abhängen, ist eine entsprechende Individualisierung des Klassifikators notwendig, um derartig feingranulare Entscheidungen zu treffen. Die Detektion von *Hate Speech* basiert im Rahmen dieser Arbeit auf der Annahme, dass die typischerweise vorhandenen Policies zum normgerechten Umgang miteinander in *Online Communities* ein Referenzmaß für alle Nutzer darstellen.

6.2 Anwendung in anderen Domänen

Xu et al. (2012) stellen Übertragungspotential in die Domäne der Sozialwissenschaften und der Psychologie heraus. Die Intention besteht dabei in der Datensammlung bezüglich der verschiedenen Formen von *Hate Speech*, um diese aus sozialwissenschaftlichen und psychologischen Gesichtspunkten zu untersuchen.[294] Die Psychologie untersucht insbesondere

[288] Vgl. Dinakar et al. (2011), S. 12.
[289] Vgl. Dinakar et al. (2011), S. 13f.
[290] Vgl. Delort et al. (2011), S. 25.
[291] Vgl. Delort et al. (2011), S. 22ff.
[292] Vgl. Delort et al. (2011), S. 24.
[293] Vgl. Giménez Gualdo et al. (2015), 229f.
[294] Vgl. Xu et al. (2012), S. 657.

Cyberbullying auf Ebene der beteiligten Akteure, wobei Fragen bezüglich des Verhaltens der Opfer im Vordergrund stehen.[295] Ergänzend dazu betrachten die Politik- und Sozialwissenschaften *Hate Speech* auf einer Makro-Ebene.[296] Es wird beispielsweise untersucht, wie sich *Hate Speech* im Zeitverlauf in sozialen Netzwerken entwickelt hat und ob es bestimmte Ereignisse gibt, die zu einer Zunahme oder Abnahme führen.[297] Die vorgestellten Artefakte lassen sich ohne Modifikation für diesen Zweck einsetzen, indem sie derartige Nachrichten nach der Publikation automatisch identifizieren und archivieren. Da der Anteil an *Hate Speech* gegenüber sonstigen Nachrichten typischerweise gering ist[298], können die vorgestellten Software-Artefakte den Arbeitsaufwand der manuellen Datenbeschaffung deutlich reduzieren.

In Kapitel 3.1 ist die *Hate Speech* Detektion als Spezialfall der *Sentiment Analysis* dargestellt. Es bietet sich daher eine Anwendung in dieser Domäne an. Verfahren der *Sentiment Analysis* führen Analysen auf Dokument-Ebene, Satz-Ebene und Aspekt-Ebene durch.[299] Bisher existieren nur wenige Verfahren mit guter Klassifikationsgüte auf der feingranularen Aspekt-Ebene.[300] Die *Harassment Patterns* sind so konzipiert, dass sie auf dieser Ebene arbeiten. Statt der Verbindung zwischen beleidigendem Wort und Personenreferenz, könnten die Software-Artefakte eine Verbindung zwischen stimmungsbehaftetem Wort und einer Entität, wie z. B. einem Produkt, einer Marke oder einem Unternehmen, suchen. Im Rahmen der Marktforschung lassen sich auf diese Weise aus nutzergenerierten Inhalten feingranulare Aussagen bezüglich derartiger Entitäten identifizieren und aggregieren. Hierbei ist zu beachten, dass die Nutzer die Entitäten unterschiedlich referenzieren, beispielsweise durch Synonyme oder abstrakte Begriffe (z. B. „das Unternehmen aus Redmond“, statt „Microsoft“). Ontologien können diese Zusammenhänge abbilden, sodass mit ihrer Hilfe eine Aggregation zu einer Entität (z. B. „Microsoft“) stattfinden kann.[301]

Montoyo et al. (2012) identifizieren in ihrem state-of-the-art Paper zur *Sentiment Analysis* die Analyse von Diskursen, insbesondere in Foren, als offene Forschungsfrage.[302] Bisher noch weitestgehend ungelöste Herausforderungen betreffen unter anderem die Zuordnung von Referenzen zwischen den Nutzern sowie die Berücksichtigung der zeitlichen Aufeinanderfolge

[295] Vgl. Slonje et al. (2013), S. 26.
[296] Vgl. Burnap und Williams (2015), S. 224f.
[297] Vgl. Burnap und Williams (2015), S. 224f.
[298] Vgl. Bretschneider et al. (2014), S. 4; Kontostathis et al. (2013), S. 196; Sood et al. (2012a), S. 1483.
[299] Vgl. Feldman (2013), S. 83.
[300] Vgl. Feldman (2013), S. 88; Tang et al. (2009), S. 10770.
[301] Vgl. Pustejovsky und Stubbs (2012), S. 76; Russel und Norvig (2010), S. 437f.
[302] Vgl. Montoyo et al. (2012), S. 678.

der Nachrichten.[303] Das im Rahmen dieser Arbeit vorgestellte Artefakt zur Detektion von *Cyberbullying* ist in der Lage, Referenzen zwischen Nutzern mithilfe des „Reference Resolvers“ zu erkennen.[304] Mithilfe des Harassment Graphen werden bisher ausschließlich *Online Harassment* Nachrichten einschließlich dem Absender und dem Opfer erfasst.[305] Durch eine Verallgemeinerung des Graphen lassen sich sämtliche Formen von Dialogen mit der zeitlichen Abfolge abbilden. Somit lässt sich mithilfe von Verfahren der *Sentiment Analysis* die Stimmung in den Dialogen erfassen.

[303] Vgl. Montoyo et al. (2012), S. 678.
[304] Vgl. Bretschneider und Peters (2016), S. 5.
[305] Vgl. Bretschneider und Peters (2016), S. 6.

7 Schlussbetrachtung

Hate Speech und die speziell gegen menschliche Ziele gerichteten Formen *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* stellen zunehmend relevante Probleme in *Online Communities* dar.[306] Das Ministerkomitee des Europarats umschreibt *Hate Speech* als Sammelbegriff für jegliche Form der Äußerung mit rassistischen, xenophoben, antisemitischen oder anderen Inhalten basierend auf Hass und Intoleranz.[307] *Directed Hate Speech* zeichnet sich durch Ausdrücke aus, die sich gegen Gruppen von Menschen richten, beispielsweise Menschen mit bestimmter Staats- oder Religionszugehörigkeit.[308] Die ausschließlich gegen Individuen gerichtete Form des *Online Harassment* bezeichnet den Versand einer Nachricht über elektronische Medien mit dem Ziel, bei einem Opfer psychischen Schaden anzurichten.[309] Findet ein wiederholter Versand von *Online Harassment* durch denselben Autor an dasselbe Opfer statt, spricht man von *Cyberbullying.*[310]

In dieser Arbeit wurden drei Artefakte nach dem Forschungsparadigma der *Design Science* konzipiert, implementiert und evaluiert, um *Directed Hate Speech*, *Online Harassment* und *Cyberbullying* in *Online Communities* einschließlich der referenzierten Opfer automatisch zu detektieren. Anhand einer strukturierten Literaturanalyse und den Evaluationsergebnissen im Rahmen dieser Studie wurde gezeigt, dass existierende Ansätze den sprachlichen Kontext von *Hate Speech* unzureichend in die Klassifikation einbeziehen, wodurch die Klassifikationsgüte vergleichsweise moderat ausfällt. Die Präsenz beleidigender Wörter ist nicht hinreichend, um diese Formen präzise zu erkennen, da die Wörter in einem Kontext stehen und gegen Individuen oder Gruppen gerichtet sind.[311] Gegenüber existierenden Artefakten basieren die in dieser Arbeit vorgestellten Ansätze auf einem *Sequenzmodell* zur Modellierung von Texten. Dadurch bleibt der sprachliche Kontext erhalten, um den Text hinsichtlich syntaktischer Verbindungen zwischen beleidigenden Wörtern und Referenzen zu Individuen oder Gruppen zu untersuchen. Diese Verbindungen sind mithilfe von Patterns modelliert, die ein Klassifikator abgleicht, um *Hate Speech* zu detektieren. Weiterhin sind die Artefakte in der Lage, das referenzierte Ziel zu markieren und zu identifizieren. Obwohl in der Literatur bereits auf die Bedeutung der Erkennung der referenzierten Opfer hingewiesen ist[312], ist dieser Aspekt in der Mehrzahl

[306] Vgl. Williams und Burnap (2016), S. 3; Aponte und Richards (2013), S. 18:4; Tokunaga (2010), S. 281; Li (2007), S. 1779f; Campbell (2005), S. 70.
[307] Vgl. Weber (2009), S. 3.
[308] Vgl. Gitari et al. (2015), S. 217.
[309] Vgl. Tokunaga (2010), S. 278.
[310] Vgl. Tokunaga (2010), S. 278.
[311] Vgl. Nahar et al. (2013), S. 51; Sood et al. (2012a), S. 1484f.
[312] Vgl. Cohen et al. (2014), S. 53.

existierender Arbeiten vernachlässigt. Dies ist jedoch insbesondere für die Erkennung von *Cyberbullying* entscheidend, da sich diese Form durch den wiederholten Versand von Nachrichten an dasselbe Opfer auszeichnet. Demnach ist eine Re-Identifikation des referenzierten Opfers über mehrere Nachrichten hinweg notwendig, um eine korrekte Zuordnung zu gewährleisten. Das zweite vorgestellte Artefakt adressiert diese Problematik mit einem Modul zur Identifikation des Opfers. Dieses Modul löst direkte und indirekte Referenzen mithilfe von Nutzernamen auf, die in *Online Communities* typischerweise ein eindeutiges Identifikationsmerkmal darstellen. Das Artefakt erstellt mit der Erkennung des Autors und des referenzierten Opfers einen Interaktionsgraphen (Harassment Graph), auf Basis dessen es *Cyberbullying* Fälle identifiziert. Der Harassment Graph erlaubt die Opfer derartige Fälle zu detektieren, sodass Administratoren gezielt Hilfestellung bei der Verarbeitung möglicher psychischer Schäden leisten können. Schließlich ist die Identifikation der Opfer entscheidend, um *Directed Hate Speech* zu erkennen, die sich gegen menschliche Ziele in Form von Gruppen richtet. Darunter fallen rassistische Äußerungen, die sich typischerweise gegen Nationalitäten, Religionen oder bestimmte Menschengruppen richten.[313] Das dritte Artefakt erkennt diese Art von referenzierten Opfern und aggregiert sie in vordefinierte Klassen, wie beispielsweise die Klasse „Ausländer“. Dadurch ist die gezielte Verarbeitung von *Directed Hate Speech* möglich.

Mithilfe von annotierten Datensätzen und geeigneter Evaluations-Metriken sind die Verfahren evaluiert. Die Datensätze sind in der Forschungscommunity publiziert, um Vergleichbarkeit und Transparenz zu gewährleisten. Die Evaluationsresultate zeigen eine Verbesserung der Klassifikationsgüte gegenüber existierenden Verfahren insbesondere im Bereich der *Online Harassment* und *Cyberbullying* Erkennung. Schließlich wurden die Artefakte und die zugehörigen Evaluationsergebnisse in Hinblick auf die praktische Anwendbarkeit diskutiert. Die Artefakte unterstützen Personal voll- oder halbautomatisch vor dem Hintergrund der typischerweise großen Nachrichtenmenge in *Online Communities* bei der arbeitsintensiven Aufgabe, *Hate Speech* Nachrichten zu moderieren.[314]

Vollautomatische Systeme arbeiten ohne die Beteiligung menschlicher Kontrollinstanzen und skalieren gegenüber halbautomatischen Verfahren besser mit der großen Nachrichtenmenge in *Online Communities*. Mit einem proaktiven Einsatz verhindern sie die Publikation von *Hate Speech*, um psychischen Schaden bei Individuen oder die Gefährdung der Öffentlichkeit gänzlich zu vermeiden. Die Anforderungen an die Klassifikationsgüte derartiger Systeme sind

[313] Vgl. Dinakar et al. (2011), S. 3; Gitari et al. (2015), S. 215; Burnap und Williams (2015), S. 224 f.
[314] Vgl. Chen et al. (2011), S. 71; Wise et al. (2006), S. 30.

hoch, insbesondere bezüglich der Fehlerrate von falsch positiven Resultaten. Sowohl das vorgestellte Artefakt zur Detektion von *Online Harassment* als auch das Artefakt zur Detektion von *Cyberbullying* erfüllen diese Anforderungen, sodass ein vollautomatischer Einsatz möglich ist. Demgegenüber ist der Einsatz von vollautomatischen Systemen in Hinblick auf die Gewährleistung des Rechts auf freie Meinungsäußerung kritisch zu bewerten. Halbautomatische Systeme markieren detektierte *Hate Speech* Nachrichten, um sie in einem zweiten Schritt einem Moderator zu präsentieren. Der Moderator prüft die Nachrichten, um Einzelfallentscheidungen zu treffen. Gegenüber vollautomatischen Systemen skalieren diese Ansätze aufgrund der Beteiligung von Menschen schlechter mit dem Nachrichtenaufkommen in *Online Communities*. Zudem vermeidet dieser Ansatz keine Schäden aufgrund der Publikation von *Hate Speech*, falls die Nachrichten bis zur erfolgten Moderation in der *Online Community* sichtbar sind. Der Einsatz von hybriden IT-Systemen mindert die wechselseitigen Nachteile, indem sie in Abhängigkeit der Intensität von *Hate Speech* ein voll- oder halbautomatisches Vorgehen wählen. Derartige Systeme sind mit einem Schwellwert konfiguriert, sodass eine vollautomatische Verarbeitung nur bei einer hohen Wahrscheinlichkeit einer korrekten Klassifikation stattfindet. Das System markiert alle verbleibenden Fälle, um sie durch einen Moderator zu überprüfen.

Weitere Forschungsarbeit bezüglich der Verbesserung der Patterns trägt zur Erhöhung der Klassifikationsgüte bei. Insbesondere im Rahmen vollautomatischer Systeme können die vorgestellten Ansätze nur mit Einschränkungen mit den Ansätzen aus der häufig vergleichend herangezogenen Spam Erkennung konkurrieren.[315] Im Rahmen der Detektion von *Cyberbullying* unterscheidet das Artefakt nur die Rollen „*Cyberbully*“ und „Opfer“. Durch die Erweiterung des Harassment Graphen lassen sich weitere Rollen identifizieren, wie beispielsweise Opfer, die später selbst wie *Cyberbullies* agieren. Weiterhin erlauben Ontologien im Rahmen der Erkennung von *Directed Hate Speech* eine dynamische Auswahl und Zuordnung zu den Ziel-Klassen.[316] Schließlich besteht weiteres Forschungspotential in der Untersuchung der Übertragbarkeit der vorgestellten Ansätze auf Problemstellungen der *Sentiment Analysis*. Eine Problemstellung, bei der weiterer Forschungsbedarf besteht, ist die Aspekt-orientierte *Sentiment Analysis*, die sich durch die Detektion von Stimmungen beziehungsweise Meinungen bezüglich bestimmter Bezugsobjekte auszeichnet. Da die Patterns ähnliche sprachliche Beziehungen modellieren, erscheint eine Adaption vielversprechend.

[315] Vgl. Dinakar et al. (2012), S. 18:2.
[316] Vgl. Pustejovsky und Stubbs (2012), S. 76; Russel und Norvig (2010), S. 437f.

Literaturverzeichnis

Alby T. (2007): Web 2.0: Konzepte, Anwendungen, Technologien. Hanser, München 2007. ISBN: 978-3-446-40931-6

Alexa (2015): The top 500 sites on the web. URL: www.alexa.com/topsites/global;0. Abruf am 03.03.2015.

Aponte D.F.G., Richards D. (2013): Managing Cyber-bullying in Online Educational Virtual Worlds. In: Proceedings of the 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death, Edinburgh, Melbourne, Australia, S. 18:1–18:9, 2013. doi: 10.1145/2513002.2513006

Arseneault L., Bowes L., Shakoor S. (2010): Bullying victimization in youths and mental health problems: 'Much ado about nothing'?. In: Psychological Medicine, 40(5), S. 717-729, 2010. doi: 10.1017/S0033291709991383

BBC News (2014): Cyberbullying suicide: Italy shocked by Amnesia Ask.fm case. URL: http://www.bbc.com/news/world-europe-26151425. Abruf am: 22.04.2014.

BBC News (2015): Migrant crisis: Facebook backs German anti-racism drive, URL: http://www.bbc.com/news/world-europe-34256960. Abruf am 30.05.2016.

Bernstein M.S., Monroy-Hernández A., Harry D., André P., Panovich K., Vargas G. (2011): 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, S. 50-57, 2011.

Bretschneider U., Wöhner T., Peters R. (2014): Detecting Online Harassment in Social Networks. In: Proceedings of the 35th International Conference on Information Systems (ICIS 2014), Auckland, New Zealand, 2014.

Bretschneider U., Peters R. (2016): DETECTING CYBERBULLYING IN ONLINE COMMUNITIES. In: Proceedings of the 24th European Conference on Information Systems (ECIS 2016), Research Papers, Paper 61, 2016.

Bretschneider U., Peters R. (2017): Detecting Offensive Statements towards Foreigners in Social Media. In: Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS 2017), Hawaii, USA, S. 2213-2222, 2017.

vom Brocke J., Simons A., Niehaves B., Reimer K., Plattfaut R., Cleven A. (2009): RECONSTRUCTING THE GIANT: ON THE IMPORTANCE OF RIGOUR IN

DOCUMENTING THE LITERATURE SEARCH PROCESS. In: Proceedings of the 17th European Conference on Information Systems (ECIS 2009). Research Papers, Paper 161, 2009.

Burnap P., Williams M.L. (2015): Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. In: Policy & Internet, 7(2), S. 223-242, 2015. doi: 10.1002/poi3.85

Campbell, M.A. (2005): Cyber Bullying: An Old Problem in a New Guise?. In: Australian Journal of Guidance and Counselling, 15(1), S. 68-76, 2005.

Cer D.M., de Marneffe M.-C., Jurafsky D., Manning C.D. (2010): Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy. In: Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., Tapias D. (eds.), LREC, 2010. ISBN: 2-9517408-6-7

Chen D., Manning C. (2014): A Fast and Accurate Dependency Parser using Neural Networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Katar, S. 740-750, 2014.

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

Cohen R., Lam D.Y., Agarwal N., Cormier M., Jagdev J., Jin T., Kukreti M., Liu J., Rahim K., Rawat R., Sun W., Wang D., Wexler M. (2014): Using Computer Technology to Address the Problem of Cyberbullying. In: SIGCAS Computers & Society, 44(2), S. 52-61, 2014. doi: 10.1145/2656870.2656876

Conover M.D., Ratkiewicz J., Francisco M., Goncalves B., Flammini A., Menczer F. (2011): Political Polarization on Twitter. In: International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011.

Delort J.-Y., Arunasalam B., Paris C. (2011): Automatic Moderation of Online Discussion Sites. In: International Journal of Electronic Commerce, 15(3), S. 9-30, 2011. doi: 10.2753/JEC1086-4415150302

Dinakar K., Reichart R., Lieberman H. (2011): Modeling the detection of textual cyberbullying. In: AAAI International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain 2011.

Dinakar K., Jones B., Havasi C., Lieberman H., Picard R. (2012): Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. In: ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), S. 18:1-18:30, 2012. doi: 10.1145/2362394.2362400

Domingos P. (2012): A few useful things to know about machine learning. In: Communications of the ACM, 55(10), S. 78-87, 2012. doi: 10.1145/2347736.2347755

Englander E., Muldowney A.M. (2007): Just turn the darn thing off: Understanding cyberbullying. In: Proceedings of the national conference on safe schools and communities, Washington, USA, S. 83-92, 2007.

Esuli A., Sebastiani F. (2006): SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), Paris, France, S. 417-422, 2006.

Europäischer Gerichtshof für Menschenrechte (2015): Case of Delfi AS vs. Estonia. URL: http://hudoc.echr.coe.int/eng?i=001-155105, Abruf am 24.05.2016.

Facebook (2016): Gemeinschaftsstandards. URL: https://www.facebook.com/communitystandards, Abruf am 04.07.2016.

Feldman R. (2013): Techniques and applications for sentiment analysis. In: Communications of the ACM, 56(4), S. 82-89, 2013. doi: 10.1145/2436256.2436274

Giménez Gualdo A.M., Hunter S.C., Durkin K., Arnaiz P., Maquilón J.J. (2015): The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. In: Computers & Education, 82, S. 228-235, 2015. doi: 10.1016/j.compedu.2014.11.013

Glaser J., Dixit J., Green D.P. (2002): Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence?. In: Journal of Social Issues, 58, S. 177-193, 2002.

Godbole N., Srinivasaiah M., Skiena S. (2007): Large Scale Sentiment Analysis for News and Blogs. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), Colorado, USA, S. 219-222, 2007.

Goh K.L., Singh A.K. (2015): Comprehensive Literature Review on Machine Learning Structures for Web Spam Classification. In: Procedia Computer Science, 80, S. 434-441, 2015. doi: 10.1016/j.procs.2015.10.069

Gruzd A., Roy J. (2014): Investigating Political Polarization on Twitter: A Canadian Perspective. In: Policy & Internet, 6(1), S. 28-45, 2014. doi: 10.1002/1944-2866.POI354

Hamel L. (2009): Knowledge Discovery with Support Vector Machines. Wiley-Interscience, New York, NY, USA 2009. ISBN: 978-0-470-37192-3

Hinduja S., Patchin J.W. (2013): Social Influences on Cyberbullying Behaviors Among Middle and High School Students. In: Journal of Youth and Adolescence, 42(5), S. 711-722, 2013. doi: 10.1007/s10964-012-9902-4

Jurafsky D., Martin J.H. (2009): Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, Upper Saddle River, NJ, USA 2009. ISBN: 978-0-13-504196-3

Kay M. (2005): Introduction. In: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.), Oxford University Press, New York, USA 2005. ISBN: 978-0-19-927634-9

Keller H., Sigron M. (2010): State Security v Freedom of Expression: Legitimate Fight against Terrorism or Suppression of Political Opposition?. In: Human Rights Law Review, 10(1), S. 151-168, 2010. doi: 10.1093/hrlr/ngp041

Kim S.W., Douai A. (2012): Google vs. China's 'Great Firewall': Ethical implications for free speech and sovereignty. In: Technology in Society, 34(2), S. 174-181, 2012. doi: 10.1016/j.techsoc.2012.02.002

Kim S-M., Hovy E. (2004): Determining the Sentiment of Opinions. In: Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, Article 1367, 2004. doi: 10.3115/1220355.1220555

King R.A., Racherla P., Bush V.D. (2014): What We Know and Don't Know About Online Word-of-Mouth: A Review and Synthesis of the Literature. In: Journal of Interactive Marketing, 28(3), S. 167-183, 2014. doi: 10.1016/j.intmar.2014.02.001

Kontostathis A., Reynolds K., Garron A., Edwards L. (2013): Detecting Cyberbullying: Query Terms and Techniques. In: Proceedings of the 5th Annual ACM Web Science Conference, Paris, France, S. 195–204, 2013. doi: 10.1145/2464464.2464499

Kraut R.E., Resnick P (2016): Building Successful Online Communities: Evidence-Based Social Design. The MIT Press, Cambridge, USA 2012. ISBN: 978-0262016575

Lenhart A. (2015): Teen, Social Media and Technology Overview 2015. URL: http://www.pewinternet.org/files/2015/04/PI_TeensandTech_Update2015_0409151.pdf, Abruf am 30.06.2016.

Li Q. (2007): New Bottle but Old Wine: A Research of Cyberbullying in Schools. In: Computers in Human Behaviour, 23(4), S. 1777-1791, 2007. doi: 10.1016/j.chb.2005.10.005

de Marneffe M.-C., Manning C.D. (2008): Stanford typed dependencies manual. URL: http://nlp.stanford.edu/software/dependencies_manual.pdf, Abruf am. 26.10.2016.

Miner G., Delen D., Elder J., Fast A., Hill T., Nisbet R.A. (2012): Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, Oxford, UK 2012. ISBN: 978-0-12-386979-1

Montoyo A., Martínez-Barco P., Balahur A. (2012): Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. In: Decision Support Systems, 53(4), S. 675-679, 2012. doi: 10.1016/j.dss.2012.05.022

Murphy K.P. (2012): Machine learning: a probabilistic perspective. MIT press, Cambridge, USA 2012. ISBN: 978-0-262-01802-9

Newell E., Jurgens D., Saleem H.M., Vala H., Sassine J., Armstrong C., Ruths D. (2016): User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In: Tenth International AAAI Conference on Web and Social Media, Köln, Germany, 2016.

Oetheimer M. (2009): Protecting Freedom of Expression: The Challenge of Hate Speech in the European Court of Human Rights Case Law. In: Cardozo Journal of International & Comparative Law, 17(3), S. 427-443, 2009.

Office of the High Commissioner for Human Rights (o. J.): International Covenant on Civil and Political Rights. URL: http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx, Abruf am 08.09.2016.

Pang B., Lee L. (2008): Opinion Mining and sentiment analysis. In: Foundations and Trends in Information Retrieval, 2(1-2), 2008, S. 1-135. doi: 10.1561/1500000011

Patchin J.W., Hinduja S. (2006): Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. In: Youth Violence and Juvenile Justice, 4(2), S. 148-169, 2006. doi: 10.1177/1541204006286288

Patchin J.W., Hinduja S. (2010): Trends in online social networking: adolescent use of MySpace over time. In: New Media & Society, 12(2), S. 197-216, 2010. doi: 10.1177/1461444809341857

Pustejovsky J., Stubbs A. (2012): Natural Language Annotation for Machine Learning. O'Reilly Media, Sebastopol, USA 2012. ISBN: 978-1449306663

Russel S., Norvig P. (2010): Artificial Intelligence: A Modern Approach, 3. Auflage, Prentice Hall, Upper Saddle River, USA 2010. ISBN: 978-0-13-604259-4

Sabella R.A., Patchin J.W., Hinduja S. (2013): Cyberbullying myths and realities. In: Computers in Human Behavior, 29(6), S. 2703-2711, 2013. doi: 10.1016/j.chb.2013.06.040

Scott M. (2015): Estonian News Site Can Be Held Liable for Defamatory Comments, Court Rules. URL: http://www.nytimes.com/2015/06/18/business/media/estonian-news-site-can-be-held-liable-for-defamatory-comments-court-rules.html?_r=2, Abruf am. 21.10.2016.

Slonje R., Smith P.K., Frisén A. (2013): The nature of cyberbullying, and strategies for prevention. In: Computers in Human Behavior, 29(1), S. 26-32, 2013. doi: 10.1016/j.chb.2012.05.024

Sokolova M., Lapalme G. (2009): A systematic analysis of performance measures for classification tasks. In: Information Processing and Management, 45(4), S. 427-437, 2009. doi: 10.1016/j.ipm.2009.03.002

Sood S.O., Antin J., Churchill E.F. (2012a): Profanity Use in Online Communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), Austin, Texas, USA, S. 1481-1490, 2012. doi: 10.1145/2207676.2208610

Sood S.O., Churchill E.F., Antin J. (2012b): Automatic Identification of Personal Insults on Social News Sites. In: Journal of the American Society for Information Science and Technology, 63(2), S. 270-285, 2012. doi: 10.1002/asi.21690

Statista (2014): Age distribution of active social media users worldwide as of 3rd quarter 2014, by platform. URL: http://www.statista.com/statistics/274829/age-distribution-of-active-social-media-users-worldwide-by-platform/, Abruf am: 06.03.2015.

Sunstein C.R. (2002): The Law of Group Polarization. In: Journal of Political Philosophy, 10(2), S. 175-195, 2002. doi: 10.1111/1467-9760.00148

Tang H., Tan S., Cheng X. (2009): A survey on sentiment detection of reviews. In: Expert Systems with Applications, 36(7), S. 10760-10773, 2009. doi: 10.1016/j.eswa.2009.02.063

Tokunaga R.S. (2010): Review: Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. In: Computers in Human Behaviour, 26(3), S. 277-287, 2010. doi: 10.1016/j.chb.2009.11.014

Tsytsarau M., Palpanas T. (2012): Survey on mining subjective data on the web. In: Data Mining and Knowledge Discovery, 24(3), S. 478-514, 2012. doi: 10.1007/s10618-011-0238-6

United Nations (o. J.): The Universal Declaration of Human Rights. URL: http://www.un.org/en/universal-declaration-human-rights/, Abruf am 08.09.2016.

Velikovich L., Blair-Goldensohn S., Hannan K., McDonald R. (2010): The Viability of Web-derived Polarity Lexicons. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, USA, S. 777-785, 2010.

Weber A. (2009): Manual on hate speech, Council of Europe Publishing, Strasbourg Cedex, France 2009. ISBN: 978-92-871-6614-2

Williams M.L., Burnap P. (2016): Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. In: British Journal of Criminology, 56(2), S. 211-238, 2016. doi:10.1093/bjc/azv059

Witten I.H., Frank E., Hall M.A. (2011): Data Mining: Practical Machine Learning Tools and Techniques. 3. Auflage, Morgan Kaufmann Publishers, Burlington, USA 2011. ISBN: 978-0-12-374856-0

Wise K., Hamman B., Thorson K. (2006): Moderation, Response Rate, and Message Interactivity: Features of Online Communities and Their Effects on Intent to Participate. In: Journal of Computer-Mediated Communication, 12(1), S. 24-41, 2006. doi: 10.1111/j.1083-6101.2006.00313.x

Yakushko O. (2009): Understanding the roots and consequences of negative attitudes toward immigrants. In: The Counseling Psychologist, 37(1), S. 33-66, 2009. doi: 10.1177/0011000008316034

Yin D., Xue Z., Hong L., Davison B.D., Kontostathis A., Edwards L. (2009): Detection of harassment on web 2.0. In: Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009, Madrid, Spain, S. 1-7, 2009.

YouTube (2016): Policy and Safety Hub. URL: https://www.youtube.com/yt/policyandsafety/, Abruf am 04.07.2016.

Zhang L., Liu B. (2014): Aspect and Entity Extraction for Opinion Mining. In: Chu W.W. (Hrsg.) Data mining and knowledge discovery for Big Data, Springer, Heidelberg, Germany, S. 1-40, 2014. ISBN: 978-3-642-40836-6

Zhuang L., Jing F., Zhu X.-Y. (2006): Movie Review Mining and Summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06), Arlington, Virginia, USA, S. 43-50, 2006. doi: 10.1145/1183614.1183625

Anhang A: Erläuterungen zu den Co-Autorenschaften

Bretschneider U., Wöhner T., Peters R. (2014): Detecting Online Harassment in Social Networks. In: Proceedings of the 35th International Conference on Information Systems (ICIS 2014), Auckland, New Zealand, 2014.

> Die Publikation wurde von *Uwe Bretschneider* konzipiert. Das in der Publikation vorgestellte Verfahren zur Detektion von Online Harassment wurde durch *Uwe Bretschneider* entwickelt. Die zugehörige Entwicklung eines Moduls zur Personenerkennung, die Implementierung des Systems und die Evaluation wurden ebenfalls von *Uwe Bretschneider* durchgeführt. Die Publikation wurde durch *Prof. Dr. Ralf Peters* wissenschaftlich begleitet. *Dr. Thomas Wöhner* wirkte an der Strukturierung des Beitrags mit. *Prof. Dr. Ralf Peters* und *Dr. Thomas Wöhner* führten eine redaktionelle Aufbereitung des Beitrags durch.

Bretschneider U., Peters R. (2016): DETECTING CYBERBULLYING IN ONLINE COMMUNITIES. In: Proceedings of the 24th European Conference on Information Systems (ECIS 2016), Research Papers, Paper 61, 2016.

> Die Publikation wurde von *Uwe Bretschneider* konzipiert. Das in der Publikation vorgestellte Verfahren zur Detektion von Cyberbullying wurde durch *Uwe Bretschneider* entworfen, implementiert und evaluiert. Die Publikation wurde durch *Prof. Dr. Ralf Peters* wissenschaftlich begleitet und redaktionell aufbereitet.

Bretschneider U., Peters R. (2017): Detecting Offensive Statements towards Foreigners in Social Media. In: Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS), S. 2213-2222, Hawaii, USA, 2017.

> Die Publikation wurde von *Uwe Bretschneider* konzipiert. Das in der Publikation vorgestellte Verfahren zur Detektion von Hate Speech einschließlich dem referenzierten Ziel wurde durch *Uwe Bretschneider* entworfen, implementiert und evaluiert. Die Publikation wurde durch *Prof. Dr. Ralf Peters* wissenschaftlich begleitet und redaktionell aufbereitet.

Anhang B: Publikation: Detecting Online Harassment in Social Networks

Bretschneider U., Wöhner T., Peters R. (2014): Detecting Online Harassment in Social Networks. In: Proceedings of the 35th International Conference on Information Systems (ICIS 2014), Auckland, New Zealand, 2014.

Detecting Online Harassment in Social Networks

Completed Research Paper

Uwe Bretschneider
Martin-Luther-University
Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
uwe.bretschneider@wiwi.uni-halle.de

Thomas Wöhner
Martin-Luther-University
Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
thomas.woehner@wiwi.uni-halle.de

Ralf Peters
Martin-Luther-University
Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
ralf.peters@wiwi.uni-halle.de

Abstract

Online Harassment is the process of sending messages for example in Social Networks to cause psychological harm to a victim. In this paper, we propose a pattern-based approach to detect such messages. Since user generated texts contain noisy language we perform a normalization step first to transform the words into their canonical forms. Additionally, we introduce a person identification module that marks phrases which relate to a person. Our results show that these preprocessing steps increase the classification performance. The pattern-based classifier uses the information provided by the preprocessing steps to detect patterns that connect a person to profane words. This technique achieves a substantial improvement compared to existing approaches. Finally, we discuss the portability of our approach to Social Networks and its possible contribution to tackle the abuse of such applications for the distribution of Online Harassment.

Keywords: Online Harassment, Cyber Bullying

Introduction

Web 2.0 applications enable users to publish content and connect with each other. In particular Social Networks and Social Broadcast Services like Facebook and Twitter enjoy huge popularity. Especially young people use them as a tool to maintain their relations to friends, classmates or fellow students (Easley and Kleinberg 2010). Since these platforms allow an unfiltered and sometimes anonymous exchange of content, new problems arise as well. Such problems are the missing protection of private information within user profiles, the questionable authenticity of users and the possibility of sending or broadcasting spam and offending messages.

Offending communication has already been an issue at schools and colleges in the form of Harassment and Bullying. With the rise of Social Networks the problem has been extended to Online Harassment and Cyber Bullying (Li 2007). Online Harassment is the process of sending messages over electronic media to

cause psychological harm to a victim. If such messages are sent several times by the same person to the same victim the process is called Cyber Bullying (Tokunaga 2010). Online Harassment and Cyber Bullying are growing in relevance and may lead to serious consequences like depression for the victims (Aponte and Richards 2013; Tokunaga 2010; Li 2007; Campbell 2005). Particularly the case of the 14 years old girl Nadia from Italy which committed suicide after being harassed on the Social Network Ask.fm attracted public attention (BBC News 2014). In this work we concentrate on Online Harassment methods since they are part of Cyber Bullying detection.

A victim has only limited options to defend himself. An offending message can be deleted if it is sent directly or if it is posted on the victim's profile. However, this requires the victim to read the message which might already inflict psychological harm. Depending on the reaction time of the victim the message might also be read by others before it is deleted. If the message is broadcasted to several receivers like on Twitter, the victim cannot delete it directly. Some Social Networks implement a reporting function to delete such messages by an administrator and possibly suspend the corresponding account. However, an offender can easily create a new account to bypass such restrictions. Another problem is that a high fraction of the victims isolate themselves and do not report such cases (Li 2007). Furthermore the barrier to send offending messages is lower in Social Networks compared to direct interaction. The victim cannot react in a direct way since the offender is possibly far away or anonymous. Finally, the reach of messages published online extend the reach of direct communication, especially if they are posted on a victims profile or send as broadcast message like on Twitter (Campbell 2005).

As suggested by Aponte and Richards (2013) these problems can be addressed by software systems which are able to block or mark Online Harassment messages. Software systems are more efficient than personnel due to the vast amount of messages in Social Networks. However, there is a lack of effective methods to realize an automatic detection for Social Networks (Kontostathis et al. 2013; Dinakar et al. 2012). Emerging approaches try to adapt methods from the research area of sentiment analysis. Sentiment analysis offers methods to classify texts regarding their contained sentiment into positive or negative (Tsytsaru and Palpanas 2012; Pang and Lee 2008). The detection of Online Harassment can be interpreted as a special problem of sentiment classification since such messages contain negative sentiment. We found, that sentiment analysis methods incorrectly classify a large amount of messages as Harassment (false positives). Since Online Harassment messages express a harmful statement related to a person and these methods are not able to detect relations between sentiments and persons, they are not suitable for Online Harassment detection. Moreover, messages in Social Networks contain noisy language including spelling errors, word variations and slang (Sood et al. 2012). Therefore, the methods for Online Harassment detection must also be robust against noisy language. Even though research on normalization of texts exists, these methods have not yet applied in the context of Online Harassment detection.

In order to address these problems, we extend current research on Online Harassment detection. Our contribution is twofold: Firstly, we introduce a new preprocessing step that indentifies phrases referring to a person. Secondly, we propose a pattern-based approach to identify Online Harassment. It is based on the detection of profane words and their links to recognized persons expressed by typical patterns. A text is interpreted as sequence-based model to match such patterns. We evaluate our method on the basis of a labeled Twitter dataset.

The rest of this paper is organized as follows: Section 2 introduces an overview of related work on Online Harassment detection. In section 3 the proposed method is presented in detail. Furthermore, we describe the development of the datasets which we construct from Twitter data. Section 4 contains the evaluation method. Each module of our proposed method is evaluated separately to distinguish between their effects on the classification results. Section 5 discusses the results and their portability into Social Networks. In section 6 limitations of the proposed approach are described. Finally, section 7 summarizes the results of this work and points out open problems for future research.

Related Work

Since the research field of Online Harassment detection is still emerging, there is only a limited amount of work available. Currently three Online Harassment techniques approaches exist: wordlist-based, machine learning and rule-based approaches.

The first approach is based on wordlists containing known profane words. A document is interpreted as a bag-of-words model which is matched against the wordlist. The document is classified as Online Harassment if a match is found. Since the bag-of-words model treats all words in an isolated manner, these approaches are not able to explicitly model relations between persons and profane words. Furthermore, the classification performance varies considerably depending on the wordlist used (Sood et al. 2012). The work of Kontostathis et al. (2013) reveals that large wordlists result in the detection of a high percentage of Online Harassment messages while smaller wordlists result in less misclassifications.

The second approach is based on machine learning methods. These methods are able to learn classification rules automatically by detecting patterns in Online Harassment messages. They require manually annotated training data to learn such rules. However, due to the sparse amount of Online Harassment messages it can be cumbersome to collect an adequate amount of training data (Kontostathis et al. 2013; Sood et al. 2012). Machine learning approaches achieve slightly better classification performance than wordlist-based approaches (Kontostathis et al. 2013; Sood et al. 2012; Dinakar et al. 2012; Dinakar et al. 2011). However, they also treat input documents as a bag-of-words model sharing the limitations of wordlist-based approaches (Kontostathis et al. 2013).

The third approach is based on rule engines to analyze semantic relations within documents. Wordlist and machine learning techniques rely on explicitly formulated statements in a text. Dinakar et al. (2012) investigate the effect on the performance by incorporating a knowledge database and a rule engine in the classification process. Online Harassment content which is built upon implicit knowledge can be detected by such methods. For example, the sexually discriminatory message sent to a male: “why did you stop wearing makeup?” (Dinakar et al. 2012). Such techniques require thorough construction of knowledge databases. For the problem of detecting sexuality related harassment alone, Dinakar et al. (2012) construct around 200 assertions. These assertions allow the rule engine to infer conclusions whether a given statement is sexual harassment.

In addition to scientific methods, first commercial systems like XRayData[1] have already been introduced to monitor certain Social Network accounts. Parents can use such tools to intervene in potentially harmful conversations. These systems assume that a human will analyze messages marked by the system and draw a final conclusion. However, they are only able to protect a predefined set of certain users. Furthermore, there are no investigations regarding the classification performance of such tools yet.

Contrary to these approaches we propose a method that treats a document as a sequence of words to preserve their order. This allows us to focus on the identification of references between profane words and potential victims. We specify patterns that express typical links between such words to improve the classification performance. In contrast to rule-based approaches, our approach relies on a small set of patterns reducing the effort compared to the maintenance of a knowledge database.

Proposed method

In this section we first introduce the system architecture of our proposed method. The associated modules are described in detail in the subsequent sections including a prototypical implementation using the example of Twitter.

System architecture

The goal of the proposed method is the automatic detection of Online Harassment messages in Social Networks. Two requirements arise from the definition of Online Harassment: Firstly, the method has to identify content that might cause psychological harm. Secondly, the method has to detect links between such content and references to a person. Additionally, according to Sood et al. (2012), the method has to be robust against noisy language. To meet these requirements we introduce the architecture shown in Figure 1 which maps each requirement to a module. The modules are further organized in a three step process consisting of text preprocessing, person identification and classification. This modular architecture allows us to evaluate and exchange each module separately.

[1] http://xraydata.com/

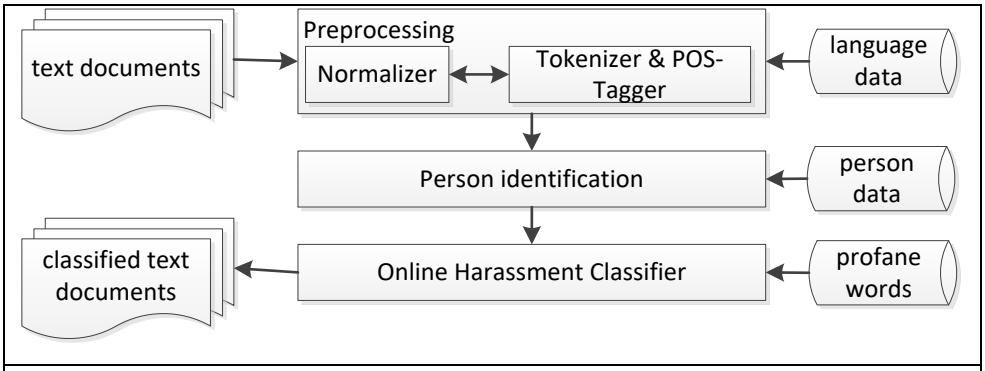

Figure 1. System architecture

In a first step the text documents are preprocessed. A minimal requirement for a wordlist-based classifier comprises a tokenization of the unstructured text into word chunks. In addition these chunks might be annotated by their part-of-speech (POS) tag, for example as noun, verb or adjective. While current work focuses on these preprocessing steps, we investigate an additional step and its effect on the classification performance. We integrate a normalization module that transforms noisy text consisting of spelling mistakes, slang and abbreviations into a canonical form. The canonical form of a word corresponds to its form found in a reference dictionary. Using this module we address the requirement stated by Sood et al. (2012) to handle dynamically changing and noisy language of Web 2.0 applications.

After preprocessing, the tokens are annotated by the person identification module. Existing work covers only partially the requirements stated in the definition of Online Harassment. The pure presence of a profane word is not sufficient to classify a message as Online Harassment. The purpose of the person identification step is to incorporate the requirement of addressing a victim within a document. For this purpose it identifies and marks words or phrases that refer to a person using POS tags in combination with language data and data from the corresponding Social Network, i.e. usernames.

Finally, the Online Harassment classifier uses the information from the preceding steps to solve the binary classification problem. In contrast to existing research, the document is only classified as Online Harassment if a link between the profane word and the word that relates to the victim exists. As stated by Sood et al. (2012) it is necessary to incorporate the context of profane words to achieve better classification results compared to a bag-of-words model. In order to improve the performance of the classification, we propose a pattern-based approach which treats a text document as a sequence of words. The sequence model preserves the order of the words and allows for the analysis of such links.

We use the example of Twitter to implement our proposed approach in a Java program. Twitter is a popular Social Network allowing users to exchange messages directly or broadcast them to several receivers. We first develop an evaluation dataset to evaluate our method which is described in the next section.

Development of the datasets

Since no dataset is available, we collected our own set of Twitter messages (Tweets) from the public stream between 2012-10-20 and 2012-12-30. The labeling process is accomplished by three annotators. A message is classified as Online Harassment if there is a consensus between at least two of the three annotators. Because the annotation of the messages is cumbersome, we classified the Tweets in their chronological order until a substantial amount of Online Harassment messages was found. For the annotation process we exclude non English, spam, empty and Re-Tweets (messages starting with “RT” or being completely enclosed by quotation marks). Re-Tweets are filtered to avoid duplicates and misclassification, because they forward a text written by another author.

The final dataset consists of 220 Online Harassment and 5162 neutral messages and is further denoted as main dataset. The sparse amount of 4.09% Online Harassment messages confirms findings of Kontostathis et al. (2013) and Sood et al. (2012) and thus represents a realistic proportion. A realistic

proportion is important, since neutral messages might contain profane words without expressing Online Harassment. Consequently, a lower amount of such messages might lead to a lower false positive rate of the classifier and thus distorting the performance measurements. However, we developed a second dataset with a similar amount of Online Harassment messages to provide an independent evaluation dataset. Since schools and colleges are our primary domain of interest, we collected the data by filtering the public stream data for tweets containing the words "school", "class", "college" and "campus". We labeled randomly selected tweets until a substantial amount of Online Harassment messages were found. We refer to the resulting data as school dataset which consists of 194 Online Harassment messages and 2599 neutral messages. We provide access to the datasets under the URL http://www.ub-web.de/research/index.html.

Word normalization

The quality of user generated content in terms of correct speech varies within different Web 2.0 applications. While some applications like Wikipedia try to improve the text quality by offering an editing function to community members[2], other applications like Twitter do not allow a correction after a text is published. For the purpose of this work, we focus on the example of Twitter. Twitter limits the length of Tweets to 140 characters[3] which encourages users to use abbreviations and slang. The normalization process transforms such noisy words into their canonical forms. Wordlist and machine learning approaches benefit from this step because noisy profane words cannot be found in a dictionary unless all the noisy variations are stored as well. Explicitly storing all variations of profane words is laborious (Sood et al. 2012). In the same manner the person identification module benefits from normalized forms of personal pronouns as well.

To assess the relevance of a normalization step we investigate the main dataset in more detail. Our investigation is comprised by two steps. In a first step we determine whether a word is in its canonical form or an out-of-vocabulary (OOV) word. Every word is looked up in a reference vocabulary after removing common pre- and suffixes. If no match is found, the word is judged as OOV word (Jufarsky and Martin 2009). This decision is automated and part of the preprocessing step. However, the selection of an appropriate reference vocabulary is required first. Thus, we examine three common vocabularies regarding the resulting percentage of OOV words on our evaluation dataset. The results are summarized in Table 1.

Table 1. Vocabularies		
	Number of OOV words	Fraction of OOV words
Wordnet 3.0	19.915	37.6657%
Hunspell	10.291	19.4636%
Moby project	3.542	6.6991%

Table 1. Vocabularies

Wordnet 3.0 (Fellbaum 1998) is popular among natural language processing applications (Jufarsky and Martin 2009) but is performing poorly compared to the other wordlists. Instead of covering a high fraction of all existing words, the main focus of Wordnet is to provide high quality syntactic and semantic information. The Hunspell[4] wordlist is used in machine translation tasks (Herrmann et al. 2011) and existing work regarding text normalization (Mosquera et al. 2012). The Moby project wordlist[5] comprises several publicly available vocabularies and is designed for applications that incorporate phonetic

[2] http://en.wikipedia.org/wiki/Wikipedia:About

[3] https://support.twitter.com/articles/15367-posting-a-tweet#

[4] http://hunspell.sourceforge.net/

[5] http://www.infochimps.com/collections/moby-project-word-lists

information. Since this wordlist performs best on the given dataset and thus reducing the normalization effort, we employ it as a basis for the investigation.

In a second step we manually count the profane words and words that refer to a person among the marked OOV words. We count 180 profane words and 323 words that relate to a person. Together they comprise 14.2% of the OOV words. Both word types are useful information for the classification of Online Harassment, which would be lost without a normalization step.

We implement a normalization module based on the method described by Mosquera et al. (2012). We selected this approach since it results in better normalization performance than current machine translation approaches (Mosquera et al. 2012). In a first step the module tries to match the OOV word against a slang and abbreviation dictionary which we built from noslang.com. If no match is found, the module tries to normalize the word by computing a simplified phonetic representation with the double metaphone algorithm. We then look for words that share the same phonetic representation within a prebuilt index based on the Moby project and the profane wordlist. To evaluate the module regarding our requirements we repeat the above-mentioned process. After the normalization step we count 28 remaining profane words and 25 remaining words that relate to a person. In addition we find 13 incorrect normalized profane words and 3 incorrect normalized words concerning a person. The normalization module successfully reduces the amount of these words to 1.95%. We further investigate the effect on the Online Harassment classification in the evaluation section.

Person identification

Relations to persons can be stated explicitly by a name or implicitly by personal pronouns like “you”. In case of Twitter, messages can also be broadcasted to several receivers addressed with their usernames. We identified four reference types that are summarized in Table 2 accompanied by an example of their usage in Online Harassment and neutral messages. The person identification module marks sequences of tokens that relate to a person and annotates them with the type of the reference. As a basis we use insights from the research field of Named Entity Recognition which includes the subtask of person identification (Jufarsky and Martin 2009). Named Entity Recognition treats a text as a sequence of tokens and searches for patterns that describe an entity including persons, locations and organizations. However, such tools are not suitable for our purposes. They are restricted to explicit (named) references and they cannot be extended to dynamically incorporate usernames. Thus, we implemented our own module that focuses on persons and can distinguish between the types summarized in Table 2.

Table 2. Person references

	Online Harassment example	Neutral example
Implicit reference by personal pronoun	“Fuck you and your mom..”	“Im a super bitch today #watchout”
Implicit reference from the point of the authors view	“My new psych advisor is such an asshole”	“So my dumbass ended up dropping my phone in the locker room and now it has like 2 dents.”
Explicit reference to a common name	“So gabby eat ass”	“@<anonymziedUser> Eric hahahahaa i fell like such a moron but i actually thought ya got kidnapped or somet”
Explicit reference to a user	“@<anonymziedUser> you asshole!”	“@<anonymziedUser> HAPPY BIRTHDAYYY!!!!!!!! ❤”

Table 2. Person references

Implicit references can be detected by a list of personal pronouns. A disambiguation between references to others and to the author himself is important to avoid false positives. The person identification module marks tokens that contain a form of “I” as self-reference. While a form of “you” is an unambiguously reference to another person, a form of “we” represents a self-reference and a reference to one or more other persons.

An implicit relation from the point of the authors view can be detected by the possessive determiner "my" in combination with a list of nouns that relate to persons. Such relations are often used in the context of schools to refer to a certain teacher. As the example shows, they are not necessarily built upon a strict sequence consisting of "my" and a noun. Adjectives, conjunctions and additional nouns might be included to further specify the person. Thus, we apply an acceptor which is a special form of a finite state machine to detect such sequences with variable length. The acceptor consumes tokens until either the state "isPerson" or "error" is reached. The state "isPerson" is reached if a combination of "my" and a noun that describes a person is found. The acceptor ignores preceding clarifying this noun more in detail as mentioned above. Such an acceptor can be further configured to match a certain type of implicit referenced persons, i.e. teachers or classmates in context of schools.

Explicit references stated by names can be identified by their word type (noun) in combination with a list of common names. We use the one provided by the Moby project vocabulary. Such lists permit to determine if a recognized first name is typically female or male. This information could be used in fine grained rules as described by Dinakar et al. (2012). Twitter offers a comfortable way to detect references to users by a special token called Twitter Mention starting with the symbol "@" and followed by a username. The addressed user will receive the message in the Twitter network together with all the followers of the author. This kind of reference is the most direct way to address a person since Twitter user names are unambiguous.

Online Harassment Classifier

The Online Harassment Classifier solves a binary classification task which separates Online Harassment documents from other documents. We propose a pattern-based approach which incorporates information from the preceding person identification step. To prevent overfitting we deduce just a small set of general applicable patterns from our dataset. However, since these links depend on the type of the profane phrase, we deduce a set of profane types first and specify the patterns that express the link in a second step.

We introduce an extended profane word lexicon first, which includes profane words or phrases, their POS-tags and a profane type. We use the wordlists provided by Noswearing.com (2014), Broadcasting Standards Authority (2013) and Hargrave (2000) as a reference. We deduced four profane types which are summarized in Table 3 by analyzing the evaluation dataset.

Table 3. Profane types	
	Example
Profane noun	"@<anonymizedUser> cunt"
Profane property	"@<anonymizedUser> is dumb… #illuminati"
Profane verb	"@<anonymizedUser> SHUT YOUR FACE RIGHT NOW!!!!!-_-"
Profane imperative	"Your presence is making my life awkward… Die"

Table 3. Profane types

Each type of profane word is used in different ways in a text sequence in general and in particular in the context of Online Harassment. These forms are tightly related to the patterns introduced in the next section. Except the profane imperative, the profane types can be determined by the POS tag.

Profane nouns are used in is-a-relations and exclamations representing the most common use case in the dataset. Due to the noisy language and incomplete sentence structure found in Tweets, such a relation might not be stated correctly in terms of grammar. However, the relation between a person and a noun is implicitly clear even without a form of "to be" as shown in the example. In contrast, a profane property requires a form of "to be" to establish a link between the word and a person.

Profane verbs are used to express actions consisting of a phrase that describes the action and a relation to a victim. Current wordlist-based approaches focus on single words restrained by their underlying bag-of-words model limiting their capability to detect phrases like in the example above. Often a verb is

ambiguous and is only considered harmful in certain phrases. We focus on profane phrases covered by our wordlist to prevent overfitting. However, an extension of the wordlist would improve the percentage of detected Online Harassment messages. Finally, we consider profane imperatives which are tightly related to verbs. They require another kind of links to persons to express a harmful statement. Thus, we introduce them as a separate profanity type. As shown in the example above the reference to the person is included in the preceding sentence statement while the imperative stands separately.

We identified the following patterns without the claim of completeness to express a connection between a profane type and a person by analyzing our dataset:

Table 4. Proposed patterns

Pattern	Description	Example
n-direct reference before	Person reference is at most n (n=3) tokens off, adjective or determiner might be in between	"@<anonymizedUser> Y r U on Fast? U always wrong. plus u r an rude asshole."
n-is (a)	Person reference and a form of "to be" is at most n (n=3) tokens ahead, adjective, adverb or determiner might be in between	"@<anonymizedUser> is dumb… #illuminati"
n-direct reference after	Person reference is at most n (n=2) token behind, preposition might be in between	"fuck you Tim haha"
Subject predicate object	profane word is in between a self-reference and reference to someone else, a form of future tense might be in between	"I fucking hate you"
Unambiguous reference	there is a (potentially distant) person reference, but no neutral or self-references in the whole document	"People get unfollowed, don't take it personal. Chances are you were just a little too dumb, ugly, or complete fucking garbage."
n-locality of reference	there are person references and neutral or self-references, but the person reference is n (n=3) steps closer to the profane word	"@<anonymizedUser> @<anonymizedUser> you're being a whiny racist prick because you didn't get your way and im laughing at you … im bored now, thanks tho."
Separately standing exclamation	the profane word stands separately at the end or right after a sentence	"Your all compulsive liars. Everything Shit thing that happens somehow ends up my fault. and you wonder why I don't wanna be here. Cunts"

Table 4. Proposed patterns

The configuration of the patterns contains several degrees of freedom. Profane phrases do not necessarily link a profane word and a person reference directly within a successive sequence. In the first example above the words "an" and "rude" are enclosed in a profane phrase. Thus, we introduce a distance (denoted with n) to permit certain words that are in between a profane word and a person reference. The specification of the distance and the set of allowed enclosed words specified by POS types is part of the configuration. We computed an optimal configuration for each pattern by evaluating them in an isolated manner. The resulting values are denoted in parentheses in Table 4. Further research is needed to confirm these settings on other datasets and exclude the possibility of overfitting.

The Online Harassment classifier is configured by providing a matrix which assigns each profane type a set of Online Harassment patterns. Whenever a profane word is detected in a sequence, the classifier tries to match at least one of the associated patterns at the corresponding position. If a pattern matches, the sequence is classified as Online Harassment. The composition of the matrix adds more degrees of freedom to the configuration of the proposed method. Thus, there are additional possibilities to configure the

proposed method by selecting which patterns are linked to which profane type. We discuss this configuration and its impact on the classification performance in the next section.

Method and Evaluation

We discussed the proposed pattern-based approach in the previous section. This section is intended to assess the effectiveness of our proposed artifacts.

Method

To evaluate our proposed approach we compare it against a naive wordlist classifier based on a bag-of-words model. This baseline classifier examines a text for the presence of at least one profane word. If such a word is found, the message is classified as Online Harassment. To further investigate the role of person references we extend this approach by our proposed person identification module. The extended naive classifier only judges a message as Online Harassment if at least one profane word and a person reference are found. Additionally, the normalization module is evaluated separately for each classifier.

We evaluate each classifier with the main and school dataset since they do not require training data. As evaluation metrics we compute recall, precision and f1 values as proposed in (Jufarsky and Martin 2009; Hirschmann and Mani 2003). Precision measures the proportion between correctly classified messages and all messages classified as Online Harassment. Recall measures the proportion between correctly classified messages and the number of real Online Harassment messages. The f1 value is the harmonic mean between precision and recall. Accuracy is not considered because of the sparse nature of Online Harassment. The dataset contains 4.09% Online Harassment messages, which means that by simply classifying every message as neutral, an accuracy value of 95.91% can be achieved.

Evaluation

We omit the evaluation results for the class of neutral messages since the proportion of such messages is substantially high. Thus, the evaluation metrics will yield good results without added value for our conclusions. Table 5 shows the results of the evaluation.

Table 5. Evaluation of Online Harassment classifier

	Main dataset			School dataset		
	Precision	Recall	F1	Precision	Recall	F1
Naive	37.47%	64.55%	47.41%	45.85%	65.46%	53.93%
Normalization	35%	70%	46.67%	43.63%	70.62%	53.94%
Naive with person recognition	63.82%	57.73%	60.62%	73.29%	60.83%	66.48%
Normalization	**58.75%**	**64.09%**	**61.3%**	**70.05%**	**67.53%**	**68.77%**
Pattern-based (balanced setting)	77.72%	68.18%	72.64%	80.14%	58.25%	67.46%
Normalization	**73.45%**	**71.82%**	**72.64%**	**79.01%**	**65.98%**	**71.91%**
Pattern-based (precision setting)	94.52%	31.36%	47.1%	91.67%	22.68%	36.36%
Normalization	**94.74%**	**32.73%**	**48.65%**	**91.67%**	**22.68%**	**36.36%**

Table 5. Evaluation of Online Harassment classifier

The normalization module improves the recall values for all examined classifiers while the precision is only reduced marginally. This confirms our preliminary examinations and the findings of Sood et al. (2012) who point out the noisy character of Web 2.0 texts. The classifiers are not able to match profane phrases which are not in their canonical form thus yielding lower recall values. Contrary to the three classifiers listed on top, the normalization step improves the results for the pattern-based classifier (precision setting) both in recall and precision.

The baseline classifier performs poorly on the datasets in terms of the achieved f1 value. However, while the precision value is very low, the baseline classifier achieves moderate recall values. Classifiers yielding substantial recall values are able to detect a large fraction of the Online Harassment messages. However, if the precision is low simultaneously, they also detect a large amount of false positives. The content of the profane wordlist influences this relationship and thus the performance of the classifier. Large wordlists lead to high recall values but many false positives as confirmed by the findings of Kontostathis et al. (2013). Their complete wordlist achieves high recall (78%) but low precision (44%) values. Subsets of this wordlist, which are optimized for their dataset, achieve high precision (84%) but low recall (37%) values. All investigated variations only achieve moderate f1 values between 28% and 57% (Kontostathis et al. 2013). Similar results are found by Sood et al. (2012). They accomplish precision values between 49% and 63% and recall values between 20% and 41%. Machine learning methods achieve slightly better classification results regarding the f1 values, which vary between 47% and 63% on their individual evaluation datasets (Kontostathis et al. 2013; Dinakar et al. 2012; Sood et al. 2011; Yin et al. 2009).

The person identification module improves substantially the precision value of the naive bag-of-word model while decreasing the recall marginally. This effect is more pronounced in the school dataset and might be caused by a larger amount of direct and unambiguous insults. With this naive setting alone, the classifier is able to achieve good results compared to existing work. Thus, the link between a profane word and a person reference is a relevant feature in Online Harassment detection. However, one supporting factor could be the limited length of a twitter message, which makes person references often unambiguous.

The pattern-based classifier requires the person identification module. Configured with the balanced setting, the classifier performs best with respect to the f1 value in both datasets. Similar performance measurements in both datasets indicate that the patterns are not specifically fitted to the main dataset. We achieve an improvement of 15% compared to existing wordlist-based approaches respectively 9% compared to machine learning approaches. The pattern-based approach allows affecting the recall and precision values without modifying the wordlist by adjusting its configuration. It can be configured by associating a subset of the available patterns to the types of profane words. We selected a combination of patterns for a balanced and a precision setting to demonstrate this effect. The precision setting achieves a precision value greater than 90% which makes it suitable for automatically blocking potential Online Harassment messages. However, the examination of the full space of configuration possibilities is computationally very expensive. Each configuration can be measured and captured in a diagram in terms of their resulting recall and precision values. More research is needed to compute all the Pareto efficient combinations in that manner.

Practical applications within Social Networks

Aponte and Richards (2013) analyze different forms of Online Harassment and Cyber Bullying and suggest solutions for practical applications with the objective to prevent psychological harm. Such messages should be blocked or marked. This can be achieved by using a method as described in this work.

Our proposed method can be used to extend Social Networks enabling them to automatically detect and block Online Harassment messages in a proactive manner as stated by Patchin and Hinduja (2006). Such systems are able to analyze messages in real time causing negligible delay within the Social Network. However, since no human control instance verifies whether the message is blocked correctly, false positives can arise. If such messages are blocked the author might be frustrated at the Social Network. Thus, these approaches rely on a classifier with a high precision value. None of the approaches introduced in previous publications is capable of achieving substantial precision values. In contrast, our proposed approach can be configured to achieve precision values greater than 90% with the cost of low recall values around 20% to 30%. Anyhow, despite the low recall values a substantial fraction of Online Harassment messages could be blocked due to the vast amount of messages within Social Networks.

Systems that automatically block Online Harassment messages address the following problems. First, an Online Harassment message causes psychological damage once a victim reads it, regardless of the time it is visible afterwards. Hence, if the message is blocked before, no psychological damage can occur. In addition, Tokunaga (2010) suggests ignoring the author of an Online Harassment message as a coping strategy. Consequently, an immediately blocked message seems as it was ignored by the victim from the

point of the offenders view. Second, Online Harassment messages can be distributed easily within or between Social Networks by spreading them in a viral manner (Li 2007). Thus, it might be difficult for human control instances to cope with the amount of messages spread. However, automated systems can deal with such an amount of messages since they are scalable. Third, Online Harassment messages can be sent anonymously (Li 2007). Even if a victim blocks potentially offending accounts within a Social Network, he can only protect himself against anonymous accounts with a proactive system.

Systems that mark Online Harassment messages rely on a classifier with a high recall value. Our proposed method can be used to add this functionality to Social Networks or to external programs like parental control systems. Such systems reduce the effort for personnel to act as human control instance. They only need to consider marked messages and decide if it really is Online Harassment and whether further actions have to be taken. Thus, it is desired to cover a high fraction of potential Online Harassment messages while keeping the amount of false positives low. The proposed pattern-based classifier achieves high recall and precision values which makes it suitable for such tasks. Additionally, the system can be further improved when it is combined with a blended mechanism. Low precision values can be compensated by human interaction allowing the system to receive feedback regarding the classification and eventually learn from it.

The following problems are addressed by systems that mark Online Harassment messages. First, the vast amount of messages in Social Networks causes substantial effort for personnel to act as human control instance. However, several authors suggest introducing Social Network policies at schools (Sonhera et al. 2012; Li 2007; Patchin and Hinduja 2006; Campbell 2005). Consequently, to ensure these policies the messages among students need to be monitored by personnel (Sonhera et al. 2012). Our proposed approach can support this task by preselecting potential harassing messages. Second, a substantial fraction of Online Harassment victims does not inform their parents or other adults about these incidents (Tokunaga 2010; Li 2007). While some of these victims avoid involving an adult because they want to cope with the situation themselves or fear restrictions regarding their access to Social Networks, others are just overwhelmed by the situation (Tokunaga 2010). Parental control systems allow parents to be aware of these incidents by monitoring the accounts of their children so they can decide if an intervention is necessary. Third, the barrier to communicate in an offending way is lower in Social Networks compared to face-to-face communication (Tokunaga 2010). Furthermore, offenders might not be aware of the harm they cause with their messages or they might overreact caused by a recent event. The system could notify the authors before publishing potential Online Harassment messages. Psychological harm can be prevented, if the author reconsiders the publication.

Besides the applications mentioned above, the investigated methods can be used to improve Cyber Bullying detection. Cyber Bullying is based on Online Harassment messages that are sent repeatedly to the same victim by the same author. The person recognition module can be extended to detect such relations. Even multiple Social Networks could be analyzed to track cyber bullies using different communication channels as proposed by Dadvar and de Jong (2012). Furthermore, retrieving training data for Online Harassment classification is a challenging task due to the sparseness of such messages. Systems with high recall can help to preselect Online Harassment candidates for a subsequent labeling step. The data can also be used within other research fields like psychology as described by Xu et al. (2012).

Finally, aspects regarding the freedom of speech need to be considered when using either of the systems. While monitoring public channels belonging to schools seems adequate, monitoring of private accounts especially without their knowledge might interfere with the right of privacy. Particularly parental control systems require the approval of the account being monitored. Instead of enforcing the integration of such systems, parents could try to achieve a consensus with their children about using such systems as a safety mechanism allowing the parents to intervene if necessary. In addition, the difference between Online Harassment and harsh criticism might be subtle. Systems that block messages containing criticism regarding these entities restrict the freedom of speech. In this case the prevention of psychological harm and the preservation of freedom of speech are conflicting goals. However, it needs to be distinguished between individuals along with children in particular and public figures, companies, governments and other referenced entities. Consequently, the protection of individuals might deserve a more vigorous consideration than the preservation of unrestricted message exchange, especially if a blocked message can be rephrased. Our proposed approach is capable of achieving this goal by performing a fine grained

person identification excluding entities not related to individuals. Furthermore, the person identification can be used to protect only certain accounts like those from children.

Limitations

The proposed approach cannot measure whether a message really causes psychological harm to a person. Additionally, no quantification of the severity of a profane phrase is applied. Further work could assign weights to the profane phrases within the dictionary by common consensus about which phrase is considered more or less harmful. In this work we assume that the severity of the damage cannot be quantified appropriately since it is based on the individual's perception and its current context (i.e. current mood, relation to sender, visibility of the message) (Tokunaga 2010; Patchin and Hinduja 2006).

Since we assume that only messages containing direct references to persons are considered Online Harassment, we cannot correctly classify messages that refer to a group of size n. However, even for human annotators it is hard to decide whether the group is small enough, so the message can be judged as inflicting psychological harm to the individuals.

The normalization module relies on language data containing common slang and abbreviations. Some abbreviations are context dependent and can lead to misclassification, i.e. "af" could mean "as fuck" or "autofocus" in the context of cameras. If several word candidates exist, a context dependent decision is necessary.

Our approach enables the classifier to match profane phrases instead of matching just single words. However, profane phrases consisting of several words can be stated in various ways. A full enumeration of all possible combinations is laborious and results in large lexica. Our lexicon contains only a small number of such phrases and thus could be extended to further improve the percentage of detected Online Harassment messages. Finally, sarcasm is a general problem in text mining and especially in Sentiment Analysis (Tsytsaru and Palpanas 2012; Pang and Lee 2008). Online Harassment can be veiled by sarcastic formulations or metaphors. The proposed method cannot detect such figures of speech.

Conclusion

Online Harassment is the process of sending messages over electronic media to cause psychological harm to a victim (Tokunaga 2010). In this paper, we have presented a pattern-based approach to detect Online Harassment in Social Networks and discussed its practical applications based on the suggestions of Aponte and Richards (2013).

Such systems should be able to block or mark Online Harassment messages. The pattern-based approach is suitable to realize these use cases by adapting its configuration. Due to the vast amount of messages within a Social Network and the sparse nature of Online Harassment messages, a manual classification is laborious. A balanced configuration of our proposed approach is able to mark potential Online Harassment messages. It achieves f1 values of around 72% which exceeds existing wordlist-based and machine learning approaches by 15% respectively 9%. It further helps to reduce the amount of work for a human control instance which can draw a decision afterwards and might initiate further actions. However, since such actions are reactive in their nature, harm still occurs to the victim if he reads the message. A high precision setting can help to prevent such harm by blocking messages that are very likely Online Harassment. Our approach achieves precision values greater than 90% which outperforms existing approaches by 30%. A high precision value reduces the number of false positives and makes the classifier more suitable for practical applications in Social Networks.

Despite the associated low recall value, a large amount of Online Harassment messages can be blocked among the vast total amount of messages within Social Networks. Previous research focuses on classifiers which are based on bag-of-words models. These approaches primarily analyze text documents regarding the presence of profane words. We use a sequence-based model that preserves the order of words in a document. Since Online Harassment targets at a person we further introduce a person identification module which marks words or phrases referring to persons within this sequence. Our proposed pattern-based approach incorporates information of this step to find links between a detected profane phrase and the addressed person. Such links are expressed by typical patterns we deduced from our dataset. This way

we are able to improve substantially the classification performance regarding the combined measure of precision and recall.

Because of the lack of datasets we provide two sets of manually annotated messages of the Social Broadcast Network Twitter. The labeling process is accomplished by three annotators since it is even for humans considerably hard to decide whether a message is classified as Online Harassment. The datasets are available at this URL[6] and can be used to evaluate other approaches. The analysis of the main dataset reveals that the language used within the messages is noisy. Noisy language contains spelling mistakes, abbreviations and slang. Thus, we extend our approach by a normalization module which transforms noisy text into its canonical form. We found that the module improves the performance of any of the investigated classifiers regarding their achieved recall values.

Future work could determine optimal configuration settings for the pattern-based approach to further improve the classification results. Several Pareto efficient combinations regarding the achieved recall and precision values are possible. The trade-off between recall and precision influences the practical applications of the classifier. Moreover, our deduced patterns need to be confirmed by further research. Furthermore, our proposed method can be extended to detect Cyber Bullying. Cyber Bullying detection is user centered and requires the identification of the bully and the victim across several messages. The proposed person identification module already provides a piece of this information. We also excluded Re-Tweets from our investigation. By incorporating these messages the original author and the users that support him by spreading the message could be identified.

References

Aponte, D. F. G. and Richards, D. 2013. "Managing Cyber-bullying in Online Educational Virtual Worlds," in *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, Edinburgh, Melbourne, Australia, pp. 18:1–18:9.

BBC News. 2014. *Cyberbullying suicide: Italy shocked by Amnesia Ask.fm case*. http://www.bbc.com/news/world-europe-26151425. Last accessed 22/04/2014.

Broadcasting Standards Authority. 2013. *What not to swear: The acceptability of words in broadcasting*. http://bsa.govt.nz/images/assets/Research/Acceptability_of_Words_2013_WEB.pdf. Last accessed 22/04/2013.

Campbell, M. A. 2005. "Cyber Bullying: An Old Problem in a New Guise?," *Australian Journal of Guidance and Counselling* (15:1), pp. 68-76.

Dadvar, M. and de Jong, F. 2012. "Cyberbullying Detection: A Step Toward a Safer Internet Yard," in *Proceedings of the 21st International Conference Companion on World Wide Web*, Lyon, France, pp. 121–126.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. 2012. "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2:3), pp. 18:1-18:30.

Dinakar, K., Reichart, R., and Lieberman, H. 2011. "Modeling the Detection of Textual Cyberbullying," in *Proceedings of the International Conference on Weblog and Social Media (Social Mobile Web Workshop)*.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, New York, NY, USA: Cambridge University Press.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database,* Cambridge, MA: MIT Press.

Hargrave, A. M. 2000. *Delete Expletives?: Research Undertaken Jointly by the Advertising Standards Authority, British Broadcasting Corporation, Broadcasting Standards Commission and the Independent Television Commission,* London, UK: Advertising Standards Authority.

Herrmann, T., Mohammed, M., Niehues, J., Waibel, A. 2011. "The Karlsruhe Institute of Technology Translation Systems for the WMT 2011," in *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, Scotland, UK, pp. 379–385.

Hirschman, L., and Mani, I. 2003. "Evaluation," in *The Oxford Handbook of Computational Linguistics*, Ruslan Mitkov (ed.), Oxford University Press, Oxford, pp. 414-429.

[6] http://www.ub-web.de/research/index.html

Jurafsky, D., and Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Upper Saddle River, NJ: Prentice Hall PTR.

Kontostathis, A., Reynolds, K., Garron, A., and Edwards, L. 2013. "Detecting Cyberbullying: Query Terms and Techniques," in *Proceedings of the 5th Annual ACM Web Science Conference*, Paris, France, pp. 195–204.

Li, Q. 2007. "New Bottle but Old Wine: A Research of Cyberbullying in Schools," *Computers in Human Behaviour* (23:4), pp. 1777-1791.

Mosquera, A., Lloret, E., and Moreda, P. 2012. "Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation," in *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey, pp. 9–14.

Noswearing.com. 2014. *Bad Word List & Swear Filter*. http://www.noswearing.com. Last accessed 22/04/2014.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval* (2:1-2), pp. 1-135.

Patchin, J. W., and Hinduja, S. 2006. "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," *Youth Violence and Juvenile Justice* (4:2), pp. 148-169.

Sonhera, N., Kritzinger, E., and Loock, M. 2012. "A proposed cyber threat incident handling framework for schools in South Africa," in *SAICSIT Conf.*, Centurion, South Africa, pp. 374-383.

Sood, S. O., Churchill, E. F., and Antin, J. 2012. "Automatic Identification of Personal Insults on Social News Sites," *Journal of the American Society for Information Science and Technology* (63:2), pp. 270-285.

Tokunaga, R. S. 2010. "Review: Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization," *Computers in Human Behaviour* (26:3), pp. 277-287.

Tsytsarau, M., and Palpanas, T. 2012. "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery* (24:3), pp. 478-514.

Xu, J., Jun, K.-S., Zhu, X., and Bellmore, A. 2012. "Learning from Bullying Traces in Social Media," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montrèal, Canada, pp. 656-666.

Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis, A., and Edwards, L. 2009. "Detection of Harassment on Web 2.0.," in *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, Madrid, Spain.

Anhang C: Publikation: Detecting Cyberbullying in Online Communities

Bretschneider U., Peters R. (2016): DETECTING CYBERBULLYING IN ONLINE COMMUNITIES. In: Proceedings of the 24th European Conference on Information Systems (ECIS 2016), Research Papers, Paper 61, 2016.

DETECTING CYBERBULLYING IN ONLINE COMMUNITIES

Research

Bretschneider, Uwe, Martin-Luther-University Halle-Wittenberg, Halle, Germany, uwe.bretschneider@wiwi.uni-halle.de

Peters, Ralf, Martin-Luther-University Halle-Wittenberg, Halle, Germany, ralf.peters@wiwi.uni-halle.de

Abstract

Online communities are platforms enabling their users to interact over the web. In particular, they are popular among adolescents as a tool to discuss topics of mutual interest. However, offending communication is a growing issue in these online environments. In its basic form, the process of sending messages over electronic media to cause psychological damage to a victim is called online harassment. In a more severe form, cyberbullying is the process of sending offending messages several times to the same victim by the same offender. In this work, we propose an approach to detect cyberbullies and their victims. Identifying and aiding victims received only brief attention in existing work. We introduce a harassment graph to capture multiple message exchanges comprising cyberbullying cases. We show that our approach is able to precisely detect cyberbullies and their victims. Additionally, we propose metrics to measure the severity of online harassment and cyberbullying cases in terms of quantitative aspects. In particular, the metrics allow to identify victims of severe cyberbullying cases and might be used as an early indicator to provide fast and selective aid by administrators. We further propose use cases for our approach in online communities to tackle the problem of cyberbullying.

Keywords: cyberbullying, online harassment, offending communication, online communities

1 Introduction

Online communities are platforms that enable their users to interact over the web. Common forms of online communities are, for example, forums, discussion boards and social networks. They are popular among adolescents as a tool to discuss topics of mutual interest (Lenhart, 2015). More than 90% of teenagers are online on a daily basis including over 20% that are almost constantly online to stay up to date (Lenhart, 2015). Since these platforms allow unrestricted and often anonymous exchange of content, they are vulnerable for abuse, especially in form of offending communication. Offending communication includes online harassment and cyberbullying, which are growing issues in online environments that involve user interaction (Jones, 2013). Online harassment is the process of sending messages over electronic media to cause psychological harm to a victim. If the same person sends such messages several times to the same victim, the process is called cyberbullying (Tokunaga, 2010).

Online harassment and cyberbullying may lead to serious consequences like depression for the victims (Patchin and Hinduja, 2013; Tokunaga, 2010; Li, 2007). In extreme cases, the consequences can be even more severe, especially for adolescents. Particularly, two cases of suicide attracted public attention in 2014. A 14-year-old girl from Italy committed suicide after being offended on the social network Ask.fm by several anonymous users (BBC News, 2014). Another suicide was caused by a Facebook user threatening repeatedly a 17-year-old boy to make him stay away from his former girlfriend (Dailymail, 2014). In both cases the message exchange caused psychological damage to the victims. More seriously, it was not recognized that the victims require aid to endure this kind of damage. As a recent decision from the

European Court of Human Rights shows, the operators of an online community might be legally responsible for psychological damage that is caused by the publication of content (European Court of Human Rights, 2015). Consequently, a lot of online communities like Facebook and YouTube introduced community standards to approach the problem of offending communication. They encourage victims of online harassment and cyberbullying to report such cases. Administrators manually review these reported cases. However, due to the vast amount of messages within online communities this task is cumbersome and time-consuming.

Current research offers approaches to detect online harassment automatically. Most of these methods focus on an isolated analysis of the corresponding messages excluding the message context. However, the detection of cyberbullying requires the analysis of interrelated messages and the identification of the involved roles (Xu et al., 2012) as a bully sends its victim offending messages multiple times. Consequently, an extension is required to adapt existing methods for the detection of online harassment to the problem of cyberbullying detection. Additionally, current methods to detect online harassment predominantly rely on bag-of-words or n-gram models, which have limited capabilities to detect the persons referenced in online harassment messages (Chen et al., 2012; Bretschneider et al., 2014). Therefore, the identification of the involved roles in cyberbullying is even more difficult.

In this work, we extend prior research by concentrating on the detection of cyberbullying and its involved actors, especially the victims. Our contribution is threefold: First, we introduce a graph-based model to structure observed offending communication between users of an online community. We build this graph by analyzing the message context, which includes all messages exchanged between these users. Second, we propose a method to detect cyberbullying in online communities based on the identification of online harassment and the identified actors in the graph model. In contrast to existing approaches, we also focus on the detection of victims. Furthermore, we propose metrics based on the harassment graph to measure the severity of online harassment and cyberbullying cases. Third, we develop two annotated datasets including the referenced victims to evaluate our approach. Additionally, the datasets might be used as a benchmark for further research.

The paper is structured as follows. Section 2 presents an overview of existing research on cyberbullying detection. In section 3 the proposed method is explained in detail. We evaluate our proposed method in section 4. Practical applications and limitations are discussed in section 5 and section 6. Finally, the results of this work are summarized in section 7.

2 Related Work

The detection of cyberbullying consists of the detection of online harassment messages and the involved persons, especially the offender and his victim (Xu et al., 2012). Consequently, the detection of cyberbullying relies above all on approaches to identify online harassment. Since online harassment detection is an emerging research field, there is only a limited amount of work available. Existing work predominantly employs lexicon and machine learning approaches.

Lexicon approaches utilize wordlists containing known profane words to match them against a given text. In their naïve form, they classify a text as online harassment, if it contains offending words. The classification performance varies considerably depending on the wordlist used (Sood et al., 2012; Kontostathis et al., 2013). Kontostathis et al. (2013) observe that large wordlists improve the recall while reducing the precision. In contrast, smaller wordlists containing mainly severe offending words lead to the opposite effect. The performance is considerably improved by searching for user identifiers and offending words in combination (Chen et al., 2012; Bretschneider et al., 2014). Machine learning approaches are able to learn classification rules automatically by analyzing pre-classified training examples. A preprocessing step transforms a given text into an n-dimensional vector containing the features characterizing the text. These features are comprised of the words itself, n-grams, part-of-speech (POS) tags and other characteristics or a combination of these. Machine learning approaches often achieve slightly better classification performance compared to lexicon approaches (Kontostathis et al., 2013;

Sood et al., 2012; Dinakar et al., 2012). However, due to the sparse amount of online harassment messages and the lack of annotated datasets, it can be cumbersome to collect an adequate amount of training data (Kontostathis et al., 2013; Sood et al., 2012).

The above-mentioned approaches use bag-of-words or n-gram models to structure texts. The bag-of-words model disregards the sequence of the words and structures them in an isolated manner within a multiset. The n-gram model retains the sequence of n consecutive words to preserve a small proportion of the context. However, these approaches have limited capabilities to model relations between persons and profane words explicitly as such relations are expressed by a potentially large sequence of words. Consequently, the identification of the referenced victim is more difficult. Chen et al. (2012) overcome this restriction by introducing a finer grained text model based on a dependency graph used by the Stanford natural language processing toolkit. Thereby, relations between words within a sentence are accessible. The authors propose a lexicon approach extended by grammatical rules that leverage these relations to detect online harassment directed to a person. The parser needs to provide a comprehensive dependency graph for the approach to work, which is more difficult, if the text contains slang and abbreviations or has no clean sentence structure. Furthermore, the parser is not able to capture relations between words that are outside of a sentence. The authors remove punctuation marks to bypass this restriction. However, Chen et al. (2012) base their research on a different definition of offensive communication. They classify sentences as offensive that contain at least one strong offensive word or a combination of a weak offensive word and a person identifier. Consequently, the identification of a victim is not necessarily required and thus not the main focus of their work. In contrast, online harassment requires by definition a reference to a victim.

Although Cohen et al. (2014) underline the importance to aid victims of cyberbullying, this aspect is often excluded in automated analyses. There is only limited work available that focuses on the detection of the involved roles in cyberbullying cases. Xu et al. (2012) use a sequence labeling approach to identify different roles involved in online harassment cases. They first identify online harassment messages called bullying traces in social media with a machine learning approach. Bullying traces are isolated messages that express an experience with bullying or cyberbullying. After identifying these messages, role labeling for the author of the message and the mentioned entities within the message is applied. Xu et al. (2012) achieve good classification performance results for the role labeling of the author. Nevertheless, the classification performance for the mentioned persons within the text, especially the victim, is moderate. The role labeling is performed in an isolated manner ignoring multiple references across several messages to the same victim. Furthermore, victims are not uniquely identified, especially if the reference to the victim is expressed implicitly, for example, by using personal pronouns. Dadvar and de Jong (2012) detect online harassment messages by incorporating user metadata and the presence of profane words in a machine learning approach. The authors plan to extend this approach to identify bullies across several social networks. However, the identification of the victims is not yet part of their classification process. Hosseinmardi et al. (2014) examine cyberbullying behavior in the social network ask.fm by analyzing a substantial number of user profiles and messages and deriving an interaction graph. The interaction graph contains the users of the social network as nodes and the likes a user grants to another user as edges. In contrast to other approaches, the graph contains victims of cyberbullying by capturing the number of offending messages a user has received. However, Hosseinmardi et al. (2014) define cyberbullying as the process of sending a message containing at least one negative word to a user profile within this network. Thus, repeated interaction is not considered. Furthermore, they do not distinguish between negative content directed against the user profile and negative content without a certain target. Yet, their analysis reveals that users react differently to negative messages in dependence of their social isolation. Social isolation is measured by the number of likes a user grants to and receives from other users. As a key finding, they discover that users receiving a substantial amount of negative messages without any positive support in form of likes by others, grant less likes in return. As the positive support increases, these users tend to be more active. Hosseinmardi et al. (2014) assume that socially isolated users are more vulnerable to cyberbullying and thus require special attention.

Contrary to existing online harassment and cyberbullying detection approaches, we employ a pattern-based method as described in Bretschneider et al. (2014). This approach treats a text document as a sequence of words to preserve their order. In contrast to existing methods based on bag-of-words models, the sequence model allows to access the context of a detected profane word, which is important to detect the referenced victim of an offending statement. We choose the approach from Bretschneider et al. (2014) as it is able to detect an offending passage and allows us to search for the referenced victim within the context of this passage. In addition, we leverage a unique identifier, the username of the underlying online community, to recognize victims of multiple offending messages in different communication processes.

3 Proposed Method to Detect Cyberbullying

In this section, we describe our method to detect cyberbullying cases. After presenting the system architecture, we describe the tasks of username detection and the construction of the harassment graph in detail.

3.1 System Architecture

Our proposed method is based on the online harassment classification described in Bretschneider et al. (2014). We extend this approach by introducing three additional processing steps to identify cyberbullies and their victims. The resulting system architecture is depicted in figure 1.

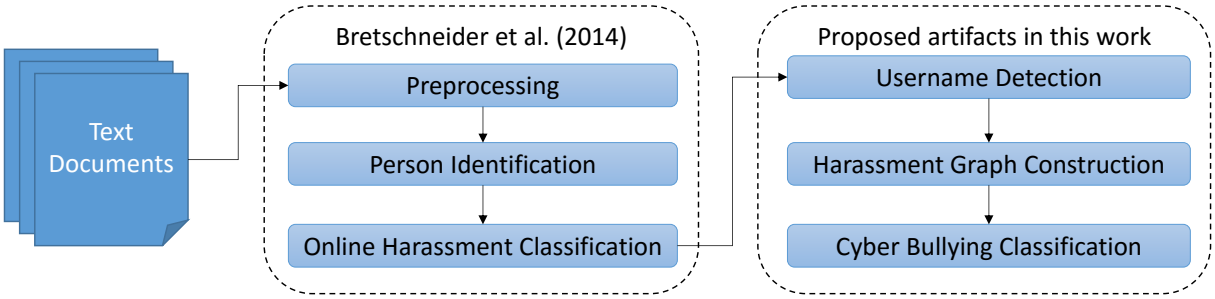

Figure 1. System architecture of the proposed cyberbullying detection approach.

The preprocessing step decomposes unstructured text into its components. These tokens are organized in a sequence to preserve their order and context. Additionally, the preprocessing module annotates these tokens with part-of-speech tags indicating their grammatical type, for example noun, verb or adjective. Furthermore, we utilize a normalization module that corrects spelling mistakes and resolves abbreviations or slang, which are typical for user generated content (Sood et al., 2012). In a next step, the person identification marks tokens that reference persons, i.e. usernames or personal pronouns. Finally, the online harassment classifier searches for relations between these person references and offending words by matching harassment patterns. Bretschneider et al. (2014) introduce seven harassment patterns that are able to match various ways of expressing online harassment within a sequence of words.

Approaches to detect online harassment treat offending messages in an isolated manner. Thus, they are not sufficient for cyberbullying detection as a cyberbully sends several interrelated messages to the same victim. Consequently, we propose three additional steps to detect interrelated online harassment messages that form cyberbullying cases. First, we introduce a module to identify the usernames and their roles involved in a cyberbullying case. We distinguish two roles, the cyberbully and the referenced victim. To correctly assign these roles, we employ the username as a unique identifier. We introduce a module that detects users referenced in the plaintext of a message by leveraging its context. In a second step, we build a directed graph containing the identified users as nodes and their roles indicated by directed edges representing the online harassment messages sent from a user to a victim. This way, we can map victims of offending communication and the corresponding offenders including the number of messages they sent. Finally, the cyberbullying classification step analyzes the resulting harassment

graph. The module classifies users harassing other users multiple times as cyberbullies. In addition, we are able to identify victims that are harassed by multiple users including the amount of online harassment messages they received. Furthermore, we propose metrics to measure the severity of online harassment and cyberbullying cases.

3.2 Username Detection

Usernames are an inherent part of most online communities used as a unique identifier for online personas. We leverage usernames to identify the roles involved in online harassment and cyberbullying cases. The detection of the author of an offending message is trivial as his username is part of the message metadata. In contrast, the detection of referenced victims within the plaintext of a message is more complicated.

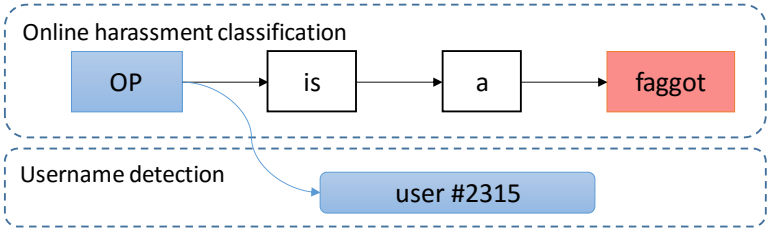

Figure 2. Online harassment classification and username detection.

As shown in figure 2, the pattern-based classifier detects offending passages within the plaintext containing a person reference and an offending word or phrase. Often, these references are stated implicitly, for example, by using personal pronouns. The usage of implicit references is typical for the progress of a discussion comprised of several interrelated messages. However, they are not sufficient as a unique identifier. The username detection resolves implicit references to explicit references expressed by a username. We propose the strategies listed in table 1 to perform this task. To aid this process, we collect usernames of authors from the message metadata distinguishing users occurring in a current discussion. Every implicit reference that is not resolvable with these strategies is disregarded and ignored in the following steps.

Strategy	Example
Reference to original poster	OP is <offending word>
Preceding reference to user	@<user> … you are <offending word>
Quote	[<user> said: …] you are <offending word>
Follow-up message	

Table 1. Username detection strategies.

The first strategy accounts for the abbreviation “OP” or its full form “original poster” that might be used directly in the offending passage as shorthand for the author publishing the first message of a discussion. Figure 2 shows an example from our dataset that contains such an abbreviation. We extend the person identification preprocessing step to detect such special words and recognize them as a person reference. Simultaneously, we are able to improve the online harassment classification performance as these types of references are not considered by Bretschneider et al. (2014). The username detection resolves then the reference by determining the corresponding author from the first message. The following strategy resolves implicit references in offending passages expressed by personal and reflexive pronouns. The corresponding explicit reference might be present in the context of the message. The sequence text model allows us to search for such an explicit reference within the sequence of words. If a username or a reference to the “OP” can be found, the implicit reference is resolved accordingly.

Typically, online communities offer a quote function to allow users to reference statements from others within the progress of the discussion. The quote strategy tries to find the author of a quoted text snippet,

if such a quote is present before the offending passage. The quoted text snippet is matched against the preceding messages until its origin is found to determine the corresponding author. Finally, users might refer to directly preceding messages as an immediate response. Since a discussion contains messages in a chronological order, this temporal relation can be leveraged to identify follow-up messages. We assume, that users that actively participate in a discussion directly answer within the same day. Furthermore, a day is often the smallest unit displayed in the message timestamp. Thus, we resolve the implicit reference to the author of a directly preceding message, if the message is published within this time frame.

3.3 Harassment Graph Construction

The harassment graph is a directed graph containing harassment messages, their authors and the victims addressed in these messages. As an example, a snippet of the harassment graph resulting from the annotation process described in the subsequent chapter is depicted in figure 3.

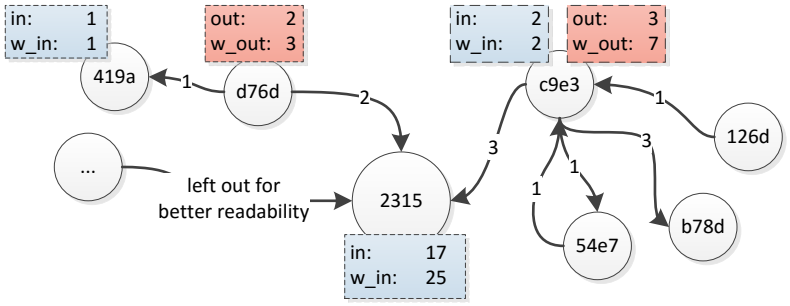

Figure 3. Harassment graph.

Formally, the harassment graph $G = (V, E)$ is comprised of a set of nodes or vertices $V = (1, \ldots, n)$ and a set of directed edges $E = (1, \ldots, k)$. The nodes represent users uniquely identified by their usernames. We anonymize the usernames by computing a hash value displaying the first four characters for better readability in the figure. The directed edges indicate the source and the target of an online harassment message. The edges are represented in an adjacency matrix X of the size $n \times n$ that contains binary variables. Each cell x_{ij} with a value of 1 represents a directed connection between the user i and j. Additionally, the number of online harassment messages sent from a user i to j are represented as weights in the weight matrix W of the size $n \times n$.

The classifier analyzes the harassment graph to identify cyberbullying cases. A cyberbullying case includes one offender sending at least two online harassment messages to a victim (Tokunaga, 2010). The harassment graph contains this information, as offenders are the source nodes of directed edges with a weight value greater than 1. In contrast to existing work, the harassment graph also allows us to focus on the identification of victims threatened by multiple offenders. For example, figure 3 depicts such a case where user "2315" is offended by multiple other nodes. To the best of our knowledge, current literature offers no definition for such cases. Thus, we define the problem of victim classification in analogy to cyberbullying as identifying victims offended by at least two distinct users.

Victim metric	Formula	Cyberbully metric	Formula
Indegree	$in(i) = \sum_{j=1}^{n} x_{ij}$	Outdegree	$out(i) = \sum_{j=1}^{n} x_{ji}$
Weighted indegree	$w\_in(i) = \sum_{j=1}^{n} x_{ij} * w_{ij}$	Weighted outdegree	$w\_out(i) = \sum_{j=1}^{n} x_{ji} * w_{ji}$

Table 2. Metrics to measure online harassment and cyberbullying severity.

In addition to existing research, we propose four metrics based on graph and social network analysis to further quantify the severity of online harassment and cyberbullying cases. The metrics are summarized in table 2. To apply them exclusively to cyberbullying cases, only directed edges with a weight greater than 1 are considered.

We employ the degree prestige metric (Wasserman and Faust, 1994) and the weighted indegree to indicate the psychological strain a victim has to bear. In social network analysis, the degree prestige is used to measure the popularity of a user. However, this interpretation reverses in the scenario of online harassment. For example, in figure 3 user "2315" has an indegree of 17, which means he faces 17 offenders (omitted in the figure for better readability).Thus, large values indicate severe cases including a large number of offenders referring to a single victim. Yet, the number of offenders alone does not cover repeated incidents caused by only a few offenders. Therefore, we employ the weighted indegree to account for such scenarios. This metric combines the number of offenders and the number of online harassment messages sent to a victim. User "2315" has a weighted indegree of 25 and thus received 25 online harassment messages in total.

In a similar way, we use the outdegree and weighted outdegree to quantify the aggressiveness of offending users. The outdegree measures the number of victims a user addresses. For example, user "c9e3" is referring to three victims resulting in an outdegree value of the same amount. A large value is an indicator for aggressive offenders in general as they refer to several victims. This assumption is supported more strongly, if the cases occur in different topics over a large time frame and the proportion of online harassment messages compared to neutral messages of this author is substantially large. However, these aspects are not yet considered and might be subject to further research. In analogy to the weighted indegree, the weighted outdegree considers the number of victims addressed by the offender and the total number of outgoing online harassment messages. If the weighted outdegree is substantially larger than the outdegree, it indicates that the offender focuses on only a few victims. User "c9e3", for example, has an outdegree of 3 and a weighted outdegree of 7. He seems to focus on the users "2315" and "b78d" as he directed six of seven online harassment messages against them.

4 Method and Evaluation

We discussed the proposed system architecture to detect cyberbullying cases in the previous section. This section is intended to assess the effectiveness of the artifacts contained in this architecture.

4.1 Dataset

To evaluate our proposed artifacts, we require an annotated dataset containing online harassment cases including the usernames of the offender and the victim. To the best of our knowledge, there are not yet any reference datasets containing this information. Consequently, we collect two datasets by downloading the general forum of the popular online games World of Warcraft[1] (dataset 1) and League of Legends[2] (dataset 2). We select these forums because of their popularity among adolescents. Adolescents, in particular, are vulnerable against online harassment and cyberbullying (Li, 2007). Furthermore, we evaluate the approach on two different datasets to prevent overfitting.

Since the amount of online harassment messages is typically sparse (Kontostathis et al., 2013; Sood et al., 2012) and the total amount of messages in these forums is substantially large, we preselect topics containing potential online harassment messages by searching for offending words contained in the wordlist from noswearing.com. This way, we selected 20 topics for each dataset. The annotation is performed by three human experts labeling each message in these topics. Online harassment cases are annotated as a tuple of the form: *(offender, victim, message)*. We only include tuples in the final dataset,

[1] http://eu.battle.net/wow/en/forum/872818/

[2] http://na.leagueoflegends.com/board/

if there is a consensus between at least two of the three annotators excluding the remaining tuples. The resulting dataset 1 contains 16975 messages with 137 harassment cases and dataset 2 contains 17354 messages with 207 harassment cases. To measure the inter-annotator agreement, we employ Fleiss' Kappa. As there is a substantial amount of neutral messages that would distort this measurement, we only consider cases judged as online harassment by at least one annotator. As a consequence, the measurement is more realistic. We measure a Fleiss' Kappa value of 0.51 for dataset 1 indicating moderate agreement and for dataset 2 a value of 0.72 indicating substantial agreement. These results emphasize the difficulty to identify offending passages, especially for borderline cases. In dataset 2, we observe more severe statements resulting in larger agreement between the annotators. We anonymized the usernames by employing a hash function on each username for the purpose of the publication. We provide access to the datasets under the URL http://ub-web.de/research/.

4.2 Method

Since the cyberbullying and victim classification process consist of three consecutive steps, we evaluate each step separately. First, we evaluate the online harassment classification. As evaluation metrics we employ precision, recall and f1-measure as recommended in (Sokolova and Lapalme, 2009). Precision measures the ratio of correctly classified instances to all instances classified as online harassment. Recall measures the ratio of correctly classified instances to all instances that really are online harassment. The f1 value is the harmonic mean between precision and recall. Second, we evaluate the username detection. We measure the amount of correctly detected usernames of victims among all detected online harassment cases. Third, we evaluate the detection of cyberbullies and the detection of victims of multiple offenders as described in the previous chapter. We measure precision, recall and f1 for both classification tasks.

To compare our results achieved in the online harassment classification step with the pattern-based approach from Bretschneider et al. (2014), we implement a baseline classifier as described in Chen et al. (2012). However, Chen et al. (2012) base their work on a different definition of offensive communication resulting in limited comparability to our results. Each message containing at least one strong offending word regardless of contained person references is classified as offending. While this is correct in their evaluation, it is not specific enough for our definition of online harassment, which necessarily requires a person reference. Since there is no public implementation of the Lexical Syntactic Feature (LSF) framework available, we followed their descriptions to implement the classifier. Additionally, we apply a support vector machine using the software RapidMiner as this approach performed moderately well in existing work (Kontostathis et al., 2013; Sood et al., 2012; Dinakar et al., 2012). We evaluate different configurations for the SVM (kernels: polynomial, dot, radial, anova and epachnenikov) and present the best result in terms of f1 in the evaluation section. To account for the substantially skewed class distribution, we activate the balance cost option in RapidMiner to adjust the settings accordingly. As machine learning approaches require training data, we split the dataset into training and test data. We follow the suggestions from Witten et al. (2011) applying a 3-fold cross validation for the evaluation. To ensure that the class distribution in each fold represents the class distribution of the whole dataset, we employ stratification.

4.3 Evaluation

The evaluation results for the online harassment classification are listed in table 3. The classification task is a difficult problem as the measurements demonstrate. The moderate overall results are associated with the characteristics of offending language in general and online harassment in particular. While online harassment is sparse in nature, offending language that is not necessarily directed against a person is fairly common in our datasets. Dataset 1 contains 4.17% offending messages marked by the LSF-like classifier from Chen et al. (2012), while only 0.81% of the messages are annotated as online harassment. Capturing the minor difference between offending language and online harassment by machine learning approaches requires adequate features and training examples. However, the imbalanced class ratio

makes the collection of training data more difficult. In contrast to the results from Chen et al. (2012) and Bretschneider et al. (2014), we were not able to reproduce the substantially high f1 values achieved in their respective evaluations. Both classifiers were evaluated on short messages from Twitter (Bretschneider et al., 2014) and YouTube (Chen et al., 2012). The messages in our datasets are substantially longer containing statements with various referenced objects (i.e. companies or game-related characters).

	Dataset 1			Dataset 2		
Classifier	Precision	Recall	F1	Precision	Recall	F1
LSF-like classifier (baseline 1)	10.73%	55.47%	17.99%	13.56%	64.53%	22.41%
Machine learning (baseline 2)	14.35%	67.65%	23.68%	22.78%	55.39%	32.29%
Pattern-based	59.17%	52.21%	55.47%	74.10%	60.29%	66.49%

Table 3. Evaluation of online harassment classification.

Both baseline classifiers, the LSF-like and the machine learning approach (anova kernel), can only achieve moderate results in terms of precision and f1. A low precision value indicates a high amount of false positives resulting in falsely classified cyberbullies and victims in the subsequent steps. Thus, they increase the amount of work for administrators to manually examine the corresponding messages correcting these results. A high recall value indicates that a large amount of the actual offending messages are detected. By definition, cyberbullies write at least two offending messages to the same victim (Tokunaga, 2010) and thus, they can only be identified, if at least two cases are detected correctly. Consequently, it is important to detect a large amount of offending messages to cover preferably the complete amount of the messages sent by an offender. The pattern-based online harassment classifier achieves moderate recall values and precision values making it more suitable for the given task.

	Dataset 1		Dataset 2	
Strategy	Correctly transformed	Incorrectly transformed	Correctly transformed	Incorrectly transformed
Explicit references	1	0	8	0
Reference to original poster	2	0	7	2
Preceding reference to user	5	4	7	1
Quote	44	2	88	0
Follow-up message	7	1	6	2
Total amount	**59**	**7**	**116**	**5**
Relative amount (in %)	**89.39**	**10.61**	**95.87**	**4.13**

Table 4. Evaluation of username detection.

Table 4 contains the evaluation results for the username detection grouped by strategies. As the results show, users rarely use explicit references to refer to other users in an offending statement. Instead, they frequently use quotes. Quotes often can be resolved accurately by parsing the html code or applying text matching techniques. In contrast, follow-up messages and preceding references to other users are prone to errors, especially, if multiple references exist in the context. However, the approach is able to resolve a reasonable amount of indirect references used in the detected online harassment cases. In dataset 1 are 7 and in dataset 2 are 5 incorrectly transformed cases. Thus, the username detection is able to resolve 89.39% (dataset 1) and 95.87% (dataset 2) references correctly. However, there are 6 references in dataset 1 and 2 references in dataset 2 that could not be resolved as no strategy was applicable. By further analyzing these unresolved cases, we observe that most of these references are implicit. Especially in mutual discussions, users tend to omit explicit references as they are clear due to the ongoing discussion and its content. This way, users refer to each other even though their messages might be spread over the

discussion. The approach is not able to analyze such references as they are expressed in a semantic way and thus none of the strategies is applicable. These cases might be addressed in further research.

	Cyberbully classification			Victim classification		
Dataset	Precision	Recall	F1	Precision	Recall	F1
Dataset 1	87.5%	53.85%	66.67%	66.67%	53.33%	59.26%
Dataset 2	93.33%	56%	70%	71.43%	57.69%	63.83%

Table 5. Evaluation of cyberbullying and victim classification.

Finally, table 5 summarizes the classification results for the cyberbully and victim classification. Considering the three consecutive steps performed to classify cyberbullies and victims, the achieved f1 values for cyberbully classification are considerably high and reasonable for victim classification. To correctly identify a cyberbullying case, at least two online harassment messages from the same author to the same correctly resolved victim need to be identified. As the results indicate, we are able to detect a fair amount of the actual cyberbullies and victims. In addition, the precision values indicate a low false positive rate. The low recall values mainly result from errors during the preceding steps. As we are interested in a measurement for the complete process, we consider in the calculation of the evaluation metrics for cyberbullying and victim classification errors during these preceding steps. Consequently, false negatives during online harassment classification and unresolved usernames reduce the recall value of cyberbullying classification.

By further analyzing the correctly identified cases, we observe that we were able to detect the severe cases measured by the metrics introduced in the previous section. This relation between the number of correctly detected cyberbullies and their weighted outdegree, respectively the number of correctly detected victims and their weighted indegree is depicted in figure 3. We choose the weighted out- and indegree to account for the total number of online harassment messages sent and received.

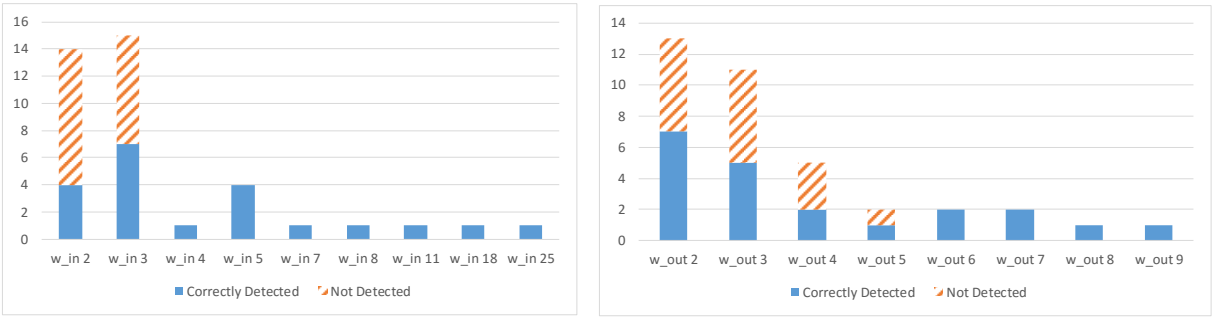

Figure 4. Correctly detected cyberbullies by weighted outdegree (left) and correctly detected victims by weighted indegree (right).

Since there is only a limited amount of work on the problem of cyberbullying classification available, we cannot directly compare our results. Although Xu et al. (2012) present the task of role labeling to classify the role of the author of a message and to identify the roles of the persons mentioned in the plaintext, they base their work on different definitions of these roles. Additionally, the evaluation is performed on a dataset with different characteristics containing short Twitter messages. The author role classification treats each message in an isolated manner deciding if the author is an offender. Thus, this problem is more similar to our online harassment classification task. Their person identification is a sequential tagging task annotating the roles of persons mentioned in a plaintext. Thus, it is comparable to our person identification and username detection task. Xu et al. (2012) achieve results of 53% precision and 42% recall emphasizing the difficulty of this task.

5 Practical Applications

Numerous online communities voluntarily committed themselves to introduce and enforce policies to maintain socially acceptable mutual communication. Such policies enable administrators to intervene in offending communication and punish the corresponding offender. However, due to the vast amount of messages published in online communities this task is labor-intensive. As a consequence, some online communities are not actively moderated and rely on their users to report abuses. However, a lot of victims isolate themselves to cope with online harassment or cyberbullying and thus do not report these cases (Li, 2007). The approach presented in this work can be utilized to aid human control instances and reduce their effort by automatically marking online harassment and cyberbullying cases.

Additionally, the harassment graph reveals further insight that an online harassment detection system alone cannot provide. First, if the system is implemented in the online community, additional information might be integrated in the harassment graph. Online harassment messages and their corresponding origin, i.e. a topic or a discussion, might be stored within the directed edges. This way, an administrator can easily navigate to the relevant section of the online community and evaluate the message context. Furthermore, the administrator can display all online harassment messages sent from or received by a certain node. Second, the introduced metrics help to estimate the severity of an online harassment or cyberbullying case. The offender metrics are an indicator to identify malicious users. The administrator might investigate all the published messages from this user and decide if further actions need to be taken. The victim metrics are an indicator for the severity of the perceived psychological damage. Instead of only punishing the offender, the administrator could aid the victim by asking about his condition. In online communities an offender might create new accounts to bypass message publication restrictions. Thus, protecting the victim might be more effective to avoid ongoing psychological damage. Furthermore, our approach is able to identify victims of severe online harassment cases with substantially high precision. As the case of the 14-year-old girl demonstrates, the identification of such cases is important, especially to intervene in an early stage.

Finally, the approach might be used to provide data for other research disciplines, i.e. social sciences, as proposed by Xu et al. (2012). The system is able to identify cyberbullying cases including the corresponding messages and involved users. The message context might be manually analyzed by experts to identify other roles like cyberbully assistants or bystanders.

6 Limitations

The severity of online harassment cases in terms of quality is not assessed by the proposed approach. Thus, the proposed metrics are entirely based on the quantity of offending messages instead of a combined measurement of quantity and quality. However, to assess the severity of online harassment cases is a challenging task even for expert annotators as the perceived severity varies among individuals (Tokunaga, 2010). Additionally, no temporal relation is considered. Online harassment messages sent in a small time frame to the same victim might cause more psychological damage than messages spanning over a larger period. Yet, the classification performance in terms of the achieved precision value is not sufficient for automated systems blocking malicious users or content. Falsely blocked users or messages due to low precision might cause frustration for the corresponding authors as they received unjust penalty. Precision might be further improved by introducing severity values and thus focusing on severe cases. Such cases might be blocked automatically while less severe cases might be reviewed by human control instances. Furthermore, we are not able to detect other roles involved in a cyberbullying case. Assistants, for example, intensify the severity of such cases as they support the involved bully (Xu et al., 2012). Finally, the detection of irony, sarcasm or longer statements paraphrasing online harassment is still an open problem. Currently, only the approach proposed by Dinakar et al. (2012) is capable of detecting paraphrased sexual harassment that alludes to characteristics of the opposite sex. Thus, the remaining cases need to be examined manually by personnel.

7 Conclusion

Offending communication is a growing issue in online environments that involve user interaction (Jones, 2013). In its basic form, the process of sending messages over electronic media to cause psychological damage to a victim is called online harassment. In a more severe form, the process of sending offending messages several times to the same victim by the same offender is called cyberbullying (Tokunaga, 2010). In this work we propose an approach to detect cyberbullies and their victims in online communities.

In current research, online harassment and cyberbullying cases are examined in an isolated manner detecting offending messages separately without focusing on the addressed victims. However, cyberbullying cases consist of several interrelated messages referring to the same victim. We extend current research to detect cyberbullying cases consisting of multiple message exchanges between the same users. First, we introduce the harassment graph to capture all offending messages as directed edges and the corresponding actors as nodes. We leverage the usernames already available in online communities to uniquely identify the involved actors. In discussions, users typically refer implicitly to each other, for example, by using personal pronouns as other users can derive the addressed user from the context of the discussion. However, implicit references are not sufficient to map them unambiguously to the harassment graph. Thus, we propose strategies to detect the usernames of referenced victims within the plaintext of a message as a second step. In a third step, we examine the resulting graph to classify cyberbullies. In contrast to existing research, we also focus on identifying victims of online harassment caused by several offenders referring to one victim. Identifying and aiding victims received only brief attention in existing work. Finally, we propose metrics to measure the severity of online harassment and cyberbullying cases in terms of quantitative aspects. The results show that our approach is able to detect the most severe cases accurately.

We introduce two labeled datasets to evaluate our approach as there are no reference datasets available that include information about the referenced victims. We provide access to the datasets as a benchmark and for further research[3]. The approach is evaluated in terms of precision, recall and f1-measure. Since there is only limited amount of work on cyberbullying detection available, we cannot directly compare our results to other approaches. We were able to achieve reasonable results for the cyberbullying classification task. Our approach yields 87.5% (dataset 1) and 93.33% (dataset 2) precision and 53.85% (dataset 1) and 56% (dataset 2) recall, which is substantially better than the achieved values of 53% precision and 42% recall from Xu at al. (2012) on their similar sequential tagging task. The results indicate that the presented approach might be used to aid administrators by identifying users involved in online harassment and cyberbullying cases within the vast amount of messages exchanged among the users of online communities. Administrators can easily view the corresponding messages and their context to decide if further actions need to be taken, especially if the victims might need assistance.

Further research might be conducted on improving the classification performance to create systems suitable for fully automated systems that restrict offending users. Currently, the harassment graph focuses on offenders and their victims. Other roles, for example cyberbully assistants, might be in the graph as well to gain further insight in the severity of such a case. As each online harassment message contains a timestamp, the message history is available in the online harassment graph. Thus, a dynamic analysis is possible, for example, to detect users that act as cyberbullies after being cyberbullied themselves. Further research might also extend the metrics in terms of qualitative aspects, for example, the severity of online harassment expressed by each message. At this time, only quantitative aspects, namely the number of offenders and messages are considered.

[3] http://ub-web.de/research/

References

BBC News (2014). "Cyberbullying suicide: Italy shocked by Amnesia Ask.fm case." URL: http://www.bbc.com/news/world-europe-26151425 (visited on 09/14/2015).

Bretschneider, U., Wöhner, T. and Peters, R. (2014). "Detecting Online Harassment in Social Networks - Building a Better World through Information Systems." In: *Proceedings of the International Conference on Information Systems 2014*. Auckland: New Zealand.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012). "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." In: *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT)*. Amsterdam: Netherlands, 71–80.

Cohen, R., Rawat, R., Sun, W., Wang, D., Wexler, M., Lam, D.Y., Agarwal, N., Cormier, M., Jagdev, J., Jin, T., Kukreti, M., Liu, J. and Rahim, K. (2014). "Using computer technology to address the problem of cyberbullying." In: *ACM SIGCAS Computers and Society* 44 (2), 52–61.

Dadvar, M. and de Jong, F. (2012). "Cyberbullying detection." In: *Proceedings of the 21st International Conference on World Wide Web*. Lyon: France, 121–126.

Daily Mail Online (2015). "Teenage boy drowns himself in the sea after being trolled on Facebook by a former friend who was dating his ex-girlfriend." URL: http://www.dailymail.co.uk/news/article-2638053/Teenage-boy-drowns-sea-trolled-Facebook-former-friend-dating-ex-girlfriend.html (visited on 09/14/2015).

Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. (2012). "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying." In: *ACM Transactions on Interactive Intelligent Systems* 2 (3), 1–30.

European Court of Human Rights (2015). "Grand Chamber judgment Delfi AS v. Estonia - liability of Internet news portal for offensive online comments." URL: http://hudoc.echr.coe.int/eng-press?i=003-5110487-6300958 (visited on 09/15/2015).

Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q. and Mishra, S. (2014). "Towards Understanding Cyberbullying Behavior in a Semi-Anonymous Social Network." In: *ASONAM: IEEE Computer Society 2014*. Beijing: China, 244–252.

Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). "Detecting cyberbullying." In: *the 5th Annual ACM Web Science Conference*. Paris: France, 195–204.

Lenhart, A. (2015). "Teen, Social Media and Technology Overview 2015." URL: http://www.pewinternet.org/files/2015/04/PI_TeensandTech_Update2015_0409151.pdf (visited on 09/14/2015).

Li, Q. (2007). "New bottle but old wine: A research of cyberbullying in schools." In: *Computers in Human Behavior* 23 (4), 1777–1791.

Jones, L.M., Mitchell, K.J. and Finkelhor, D. (2013). "Online harassment in context: Trends from three Youth Internet Safety Surveys (2000, 2005, 2010)." In: *Psychology of Violence* 3 (1), 53–69.

Patchin, J.W. and Hinduja, S. (2013). "Cyberbullying among adolescents: implications for empirical research." In: *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine* 53 (4), 431–432.

Sokolova, M. and Lapalme, G. (2009). "A systematic analysis of performance measures for classification tasks." In: *Information Processing and Management* 45 (4), 427–437.

Sood, S.O., Churchill, E.F. and Antin, J. (2012). "Automatic identification of personal insults on social news sites." In: *Journal of the American Society for Information Science and Technology* 63 (2), 270–285.

Tokunaga, R.S. (2010). "Following you home from school: A critical review and synthesis of research on cyberbullying victimization." In: *Computers in Human Behavior* 26 (3), 277–287.

Wasserman, S. and Faust, K. (1994). "Social Network Analysis. Methods and Applications." Cambridge: Cambridge University Press.

Witten, I. H., Frank, E. and Hall, M. A. (2011). "Data Mining: Practical Machine Learning Tools and Techniques." San Francisco: Morgan Kaufmann Publishers.

Xu, J.-M., Jun, K.-S., Zhu, X. and Bellmore, A. (2012). "Learning from Bullying Traces in Social Media." In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg: USA, 656–666.

Anhang D: Publikation: Detecting Offensive Statements towards Foreigners in Social Media

Bretschneider U., Peters R. (2017): Detecting Offensive Statements towards Foreigners in Social Media. In: Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS 2017), Hawaii, USA, S. 2213-2222, 2017.

Detecting Offensive Statements towards Foreigners in Social Media

Uwe Bretschneider
Martin-Luther-University Halle-Wittenberg
uwe.bretschneider@wiwi.uni-halle.de

Ralf Peters
Martin-Luther-University Halle-Wittenberg
ralf.peters@wiwi.uni-halle.de

Abstract

Recently, politicians and media companies identified an increasing number of offensive statements directed against foreigners and refugees in Europe. In Germany, for example, the political group "Pegida" drew international attention by frequently publishing offensive content concerning the religion of Islam. As a consequence, the German government and the social network Facebook cooperate to address this problem by creating a task force to manually detect offensive statements towards refugees and foreigners. In this work, we propose an approach to automatically detect such statements aiding personnel in this labor-intensive task. In contrast to existing work, we assess severity values to offensive statements and identify the referenced targets. This way, we are able to selectively detect hostility towards foreigners. To evaluate our approach, we develop a dataset containing offensive statements including their target. As a result, a substantial amount of offensive statements and a moderate amount of the referenced victims was detected correctly.

1. Introduction

The ongoing civil war in Iraq and Syria and its consequences are dominantly present in the media. The crisis led to the displacement of millions of refugees that are forced to search asylum in other countries. For example, Germany alone expected around one million asylum-seekers in 2015 [1]. These large numbers of refugees arriving in Europe caused controversial political discussions about the feasibility and consequences of their accommodation [2]. In this context, social media is growing in popularity to organize political discussions, exchange opinions and form groups of mutual interest [3,4]. In recent years, social media platforms have recorded a substantial increase in user numbers. Facebook, for example, has over 1 billion daily active users [5]. New content rapidly spreads in social media networks reaching a large amount of users [6] and enabling similar minded people to easily find and connect with each other [7].

Besides people having sympathy for the critical situation of the refugees, there are also people sharing a negative view. In extreme cases, they direct offensive statements towards refugees or foreigners in general expressing their fear and aggression [2]. In Germany, for example, the political group "Pegida" drew international attention by frequently publishing new content containing offensive statements towards foreigners, especially towards followers of Islam [4]. This form of offensive language is often referred to as cyberhate or hate speech, which is a general problem in social media [8,9].

Recently, German politicians recognized hostility towards foreigners in social media as a growing problem since it might facilitate public incitement against foreigners. Moreover, radical groups and political parties might take advantage of the recent situation spreading their ideology and eventually recruiting new supporters [9,10]. Social media platforms intensify this problem by the possibility to anonymously create content rapidly reaching a large number of users [6]. More importantly, content containing one-sided and radical viewpoints might be a problem in political opinion-formation, if users have only restricted access to credible opposing opinions [11,12]. This way, an important concept of democracy is violated: taking informed decisions in the context of competing opinions and ideas [13].

In a current project, the social network Facebook cooperates with the German government to address this problem by introducing an action plan. The plan contains a task force consisting of people from online communities, political parties and the German justice ministry to detect offensive statements towards refugees and foreigners [1]. However, due to the vast amount of messages in social media, the task of detecting hate speech is labor-intensive and time-consuming [3,14]. Additionally, there is only a limited amount of automated approaches that are able to detect hate speech directed against a certain target [15]. These approaches are not effective as hate speech towards foreigners is

URI: http://hdl.handle.net/10125/41423
ISBN: 978-0-9981331-0-2

often paraphrased and complex [9]. As a result, they are not capable of detecting the target of hate speech. This is, however, important to distinguish hate speech without certain targets from hate speech directed towards certain people or groups.

We extend current research on the detection of hate speech by the following contributions. First, we present an approach to detect hate speech towards foreigners in social media including the referenced target. Second, we develop an annotated dataset to assess the performance of our approach as there are no reference datasets yet. We provide access to this dataset as a benchmark for further research. Third, we discuss applications of our approach and strategies to tackle the problem of hate speech towards foreigners and refugees.

The rest of this paper is organized as follows: Section 2 contains the theoretical background of this study including a discussion of freedom of speech versus hate speech in the context of social media and an overview of exisitng work in hate speech detection including its related forms. In section 3, the development of the annotated datasets containing user comments from public Facebook pages is presented. In section 4, the proposed approach is introduced in detail. An evaluation based on the annotated datasets is presented in chapter 5. Section 6 discusses practical applications in social media platforms. Finally, section 7 summarizes the results and points out aspects for further research.

2. Theoretical background

2.1. Freedom of speech versus hate speech in social media

Freedom of expression, especially freedom of speech, is regarded as a fundamental individual right anchored in the Universal Declaration of Human Rights of the United Nations that is ratified by the majority of the countries in the world [16]. In the legally binding instrument of this declaration, the "International Covenant on Civil and Political Rights" (ICCPR), freedom of speech is defined as the right to "receive and impart information and ideas of all kinds" [17].

Article 19 (3) of the ICCPR defines restrictions to freedom of speech as it might conflict with "the rights or reputations of others" or "the protection of national security or public order […], or of public health or morals" [17]. The interpretation of the exceptions stated in article 19 (3) ICCPR as well as their implementation in national law is different from country to country [18]. China, for example, applies a very restrictive interpretation in terms of national security and system critic opinions [19]. In democracies, freedom of speech is regarded as a fundamental right and core concept [19,20]. In the United States, for example, freedom of speech is anchored in the first Amendment [19]. A liberal and self-regulating approach is applied based on the principle that ideas contest each other in a marketplace of competing ideas [13,19].

In this work, we follow the interpretation of freedom of speech from the European Union. In contrast to the United States, the European Union is more restrictive, especially with respect to hate speech [20]. In line with current research [9,14], the Council of Europe's Committee of Ministers notes that no universally accepted definition of hate speech exists [21]. As an orientation for European case law, they state that hate speech "covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance." [21]. Violations concerning the publication of hate speech might lead to legal consequences primarily for the author of offensive content [20]. As a recent decision from the European Court of Human Rights shows, the social media platform might be held responsible as well [22].

As a consequence, a conflict between the protection of the victims, the social media platform and the fundamental right of freedom of speech exists. The primary focus of this work is to propose an approach to detect hate speech and their referenced targets that might be used in different ways to comply with national rights. In section 6 we discuss these ways in form of practical applications and their potential consequences on freedom of speech.

As stated above, freedom of speech is an important element of democracy fostering political discussions of competing opinions and ideas [13]. However, using hate speech in political discourse to prevail extremist viewpoints might deter other users wishing to participate in a civil discussion [3]. Consequently, users expecting civil discussions often favor moderation to restrict uncivilized behavior by removing messages that do not conform with community norms [3]. In the context of social media, moderation is a labor-intensive task that causes financial costs [3,14]. In addition, coping with uncivilized behavior causes emotional costs both for moderators and participants with a civil but potentially opposing opinion. Based on the theory of Hochschild's "emotion work", such users need to perform "deep acting" to adjust their inner emotions to match the expectations on emotions required in a civil discussion [23]. While this theory originates from face-to-face communication [23], other researches apply it to the digital context. Menking and Erickson [24], for example, found that women avoid engaging in the Wikipedia as it requires them to perform "deep acting" to cope with harassment.

Another obstacle for discourse are echo chambers, a phenomenon first described by Key [25] in the political context. In social media networks, they might facilitate homogenous viewpoints by superseding opposing viewpoints [11]. Within such a network, users create mutual connections, for example by friendship or follower relations as well as by forming groups. The content displayed to a user often depends on these relations, for example, Facebook's EdgeRank filters by analyzing such relations [26]. Content published by friends or connected groups is more likely to be displayed than content from other users. More importantly, the resulting content often contains one-sided viewpoints as friends typically share similar interests and opinions [11]. In the context of political opinion-formation, echo chambers might be a problem if users are exposed to homogenous opinions favoring an extreme political viewpoint while having restricted access to credible opposing opinions [11,12].

Detecting and automatically resolving these obstacles characterized by hate speech might help administrators to moderate discussion eventually fostering civil discourse. Furthermore, the problem of echo chambers containing mostly hate speech and homogenous viewpoints might be addressed as stated in section 6.

2.2. Approaches to detect hate speech

Hate speech, cyberhate and offensive language are umbrella terms often used in the context of social media to denote offending content in general [9,14]. Hostility towards foreigners is, in particular, characterized by a referenced victim similar to the related form of online harassment. Tokunaga [27] defines online harassment as the process of sending messages over electronic media to cause psychological harm to a victim [27]. Thus, we consider existing approaches in the research fields of hate speech as well as online harassment detection. As we are interested in applying an approach to exclusively detect hate speech towards foreigners including the referenced target, we discuss their strengths and weaknesses in this regard. The related approaches are subsumed in table 1.

Table 1. Existing approaches

	[14]	[28]	[29]	[30]	[31]	[15]
Hate speech	X	X	-	-	X	-
Online Harassment	-	-	X	X	-	X
Referenced victim	X	-	-	X	-	X
Victim identification	-	-	-	-	-	-

The majority of the publications apply either lexicon [28,31] or machine learning approaches [14,29,30]. Lexicon approaches entirely rely on a lexicon containing offensive words typically used in hate speech. In their basic form, they classify a text as hate speech, if it contains at least one offensive word. A major advantage of these approaches is their simplicity and independence of training data as well as easy adoption in other languages by providing adequate lexica by experts. However, their practical applicability is limited, especially in the context of online harassment detection as they achieve only reasonable to moderate classification performance [28]. As a consequence, they are often used to preselect potential offending messages to perform subsequent analyses [31].

Machine learning approaches, in contrast, rely on training data to automatically learn rules to classify hate speech messages. As these rules are derived from statistical relationships, they require numerical inputs in form of features. These features are derived by experts from characteristics of hate speech messages and include, for example, the presence of offending words defined in a lexicon [14,30] and the presence of words typically referring to persons [30]. Compared to lexicon approaches, the classification performance is only slightly better [28,29,31]. Additionally, the collection of an adequate amount of training data is cumbersome due to the lack of annotated datasets [28,31].

All of the above-mentioned approaches rely on bag-of-words models representing a text as a vector of words. As a consequence of these simple models, the order of the words and thus their context is lost. However, the context of the offending passage is important to detect links between offending words and the targeted victims. These approaches are neither capable of detecting such links nor of detecting the passage with the referenced victim. As a consequence, they only achieve moderate classification results in online harassment classification as this form is characterized by containing a link to a victim [14,15].

Chen et al. [14] introduce a refined machine learning approach to address these shortcomings. They note that strong offensive words often occur in unambiguous hate speech messages while weak offensive words are only considered offensive when they are directed against a person. As a consequence, they apply a lexicon distinguishing between strong and weak offensive words. They compute the dependency graph of a given text to analyze its grammatical relations eventually detecting links between offending words and persons. In contrast to bag-of-words models, the dependency graph is a complex text model representing sentences of a text as sets of grammatical relations [14]. The ability to process such relations is the main advantage of the

underlying text model. However, the model is designed for short texts that are treated as a single sentence to capture their whole context possibly resulting in incorrect grammatical relations [14]. Moreover, the approach requires a dependency parser for each language and dismisses the detected victim references as they are not required for further processing.

Xu et al. [30] apply a sequence label task in addition to a machine learning approach to identify online harassment cases including involved roles. First, the machine learning approach is used to detect online harassment. In a second step, role labeling is applied to assign the author of the message, the victim and additional roles. They achieve reasonable results for the identification of the offender. However, the performance for assigning the other roles mentioned within the text, especially the victim, is moderate [30]. Furthermore, an additional training data set is required to perform the sequence label task [30].

More recently, Bretschneider et al. [15] proposed a pattern-based approach to detect offending passages in text messages including the referenced victim. Instead of a bag-of-words model, they apply a sequence model that preserves the order of the words. In contrast to the dependency graph in [14], the sequence model is not restricted in length and easier to compute [15]. Compared to the other approaches that exclusively detect online harassment, they achieve substantially improved classification results by employing patterns that represent typical ways to link offending passages to persons [15]. Similar to the grammatical relations in [14], these patterns need to be defined by experts.

Even though the approaches presented in [15] and [30] are capable of detecting referenced victims, none of the existing approaches further process them to actually identify the victim. Moreover, while online harassment messages are directed towards a person, xenophobic or racist content is typically directed towards groups of people, nationalities or races. Currently, there is only limited amount of work available that addresses the detection of xenophobic or racist content in social media including the referenced victims. The sheer detection of passages referencing a victim is not sufficient to unambiguously identify the target. Often, the offender refers to people by using indirect references that need to be resolved first [15]. In this work, we extend existing approaches to detect text passages containing hostility towards foreigners and identify the referenced target.

3. Construction of the dataset

We constructed three datasets by accessing publicly available Facebook pages, to evaluate our proposed approach and to acquire training data. We crawled Facebook posts including the comments published in response to them. The two popular Facebook pages "Pegida" (dataset 1) and "Ich bin Patriot, aber kein Nazi" ("I'm a patriot, not a nazi") (dataset 2) were selected as they are known for their critical view regarding foreigners and refugees [4] and thus presumably contain offensive statements. In addition, we select the page "Kriminelle Ausländer raus" ("Criminal foreigners get out") (dataset 3) as a training dataset since it is known for xenophobe and racist comments. We crawled the latest 50 posts including their comments beginning from February 2016. We only included 20 posts for dataset 1 to acquire a comparable amount of comments for dataset 1 and dataset 2. Two human experts annotated the datasets marking offensive statements, their severity and the intended target. To the best of our knowledge, there are not yet any reference datasets containing this information.

Each offending passage is marked and assessed with a severity value. Statements that are perceived by the experts as slightly offensive to offensive are denoted with a severity value of 1 and explicit to substantial offensive statements with a value of 2. The severity value is applied in different evaluation scenarios and practical applications described in the method section and section 6 respectively. Additionally, we leverage this information in the training dataset to derive severity values for the offending words in our lexicon.

We employ Cohen's Kappa to measure the inter-rater agreement for offensive statement annotation. The assessed severity value is used as class label. Since the class distribution between offending and neutral messages is substantially skewed in favor of neutral messages, the resulting kappa value would overestimate the agreement. Consequently, we compute a kappa value only considering offending messages marked by at least one annotator. The results indicate a substantial agreement and are denoted in table 2 along with other descriptive metrics of the datasets.

Table 2. Constructed datasets

Dataset	**1**	**2**	**3**
#comments	2649	2641	546
#cases (severity = 1)	99	112	50
#cases (severity = 2)	137	112	130
Cohens Kappa	0.78	0.68	0.73
Target Foreigner	**24.38%**	**37.95%**	**76.67%**
Target Government	33.88%	33.04%	3.89%
Target Press	17.36%	8.04%	2.22%
Target Community	3.72%	4.91%	6.67%
Target Other	16.12%	14.29%	8.89%
Target Unknown	5.37%	1.79%	1.67%

Furthermore, the annotators identified the referenced target. We focus on offending statements directed towards foreigners and refugees and find

evidence that a substantial amount of these statements is indeed directed towards foreigners, especially in dataset 3. However, the coding process revealed that frequently other related entities are referenced, for example the German government. As a consequence, we derive 6 target groups frequently referenced in the datasets: foreigners and refugees, the government represented by political parties and politicians, the community of the Facebook group, the press and media, other identifiable targets and unknown targets. Unknown targets arise if the human annotators are not able to resolve the reference.

A consensus annotation is computed by merging the annotations from both annotators. Severity values are combined by computing the average and rounding down. For example, a severity value pair of 1 and 2 results in a consensus severity value of 1. For the assessed targets, we only consider targets marked by both annotators. If there is no consensus, we classify the target as unknown. We anonymized the dataset by employing a hash function on each username for the purpose of the publication. We provide access to the datasets under the URL www.ub-web.de/research/.

4. Proposed method

4.1. System architecture

In this work, we propose the system architecture depicted in figure 1. The architecture is based on elements employed without modification as described in [15], which are denoted in the dotted line box.

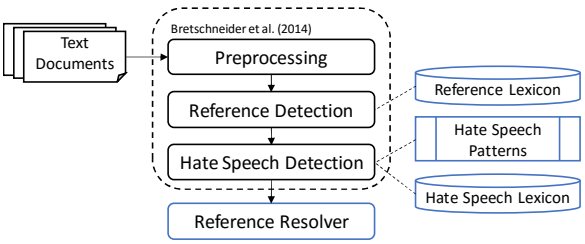

Figure 1. System architecture

Our decision to select this particular approach is primarily justified by the requirement to detect and identify the referenced victims. Only the approaches described in [15] and [30] are capable of accessing the passage including the referenced victim. However, the classification results achieved in [30] are moderate, while the results from [15] are more promising. Compared to the dependency graph in [14], the underlying sequence model in [15] is suitable for longer texts and does not require a dependency parser for the German language.

In a first step of the resulting architecture, the text documents are preprocessed by decomposing the unstructured text into tokens. In addition, these tokens are normalized removing common abbreviations and slang. In contrast to bag-of-words models, a sequence model is applied preserving the order and thus the context of the words. In a second step, the reference identification module marks tokens in the sequence referring to entities of interest for this study. These entities are, for example, foreign nationalities, political groups and the government.

After these preprocessing steps, the hate speech detection module searches for offending words in the sequence. Once such a word is found, the hate speech patterns are applied searching relations between the offending word and a reference to a victim. If a pattern matches, the text is classified as hate speech directed towards a victim. Finally, we identify the victims referenced in these passages by performing a reference resolution. As a consequence of this architecture containing consecutive tasks, the reference resolver can only process cases that are correctly detected in the previous step. Thus, we are interested in detecting a preferably complete amount of offending statements without the cost of too many classification mistakes in the form of false positives. To achieve this goal, we follow the proposals presented in [15] and [14]. In line with Chen et al. [14], we distinguish two forms of offensive statements: severe offending statements not necessarily containing a referenced victim and offending statements directed against a target. While only focusing on the latter has the advantage of a low false positive rate, it also comes with the disadvantage of a lower detection rate [15].

Finally, the original method described in [15] is designed for text documents in English. As our dataset contains text documents in German, we modify the approach accordingly by creating a reference lexicon, a hate speech lexicon and hate speech patterns as described in the subsequent sections. These modifications are required for each language.

4.2. Reference detection

The reference detection is a preprocessing step that marks references to entities of interest that are further processed in subsequent steps. We distinguish between static and dynamic references that are both stored in a dynamic lexicon. Static references are expressed by common words found in appropriate lexica and are further classified into direct and indirect references. Experts need to define this part of the lexicon for each language manually.

Direct references refer among others to nations or religious groups. For example, the sentence “sieht wien

scheiß kosovoalbaner aus" ("looks like a damn Kosovo-Albanian") taken from the dataset contains the direct reference "kosovoalbaner" referring to the ethnical group of Kosovo-Albanians. In contrast, indirect references are often used as a shorthand for direct references or to paraphrase a reference to a victim that is apparent in the context. As an example, the sentence "Dieses Ratten Pack bringt nur unruhen" ("This rat rabble only brings unrest") contains an indirect reference consisting of an article in combination with a word typically referring to a group of people ("pack"). In this case, the reference points at refugees in general and can be resolved by analyzing the corresponding Facebook post, which contains a short story about refugees. We employ the German dictonary "Duden" as a lexical resource to define such static references, especially by using the synonym functionality.

Finally, dynamic references are based on special terms and names that relate to the current political context and characteristics of the social media platform. Usernames, for example, are often unique identifiers in social media platforms to refer to each other. Publicly known names, for example the current German chancellor "Angela Merkel", are often subject of political discussions. Political groups, for example "Pegida" arise and dissolve over time. To account for such dynamic terms, we build a dynamic database by employing expert knowledge. In further work, such information might be derived automatically, for example from knowledge databases like DBpedia.

For each reference we additionally store the corresponding group as defined in the previous section. For example, the chancellor "Angela Merkel" belongs to the government group. In further work, an ontology might be applied instead.

4.3. Offensive statement detection

As our dataset contains text documents in German, we need to modify the approach from Bretschneider et al. [15] accordingly by employing a German offending word lexicon [32] and creating new hate speech patterns tailored for the German language. Our resulting patterns are listed in table 3.

Table 3. Constructed hate speech patterns

Pattern	Example
Reference before	"**Dieses Ratten Pack** bringt nur Unruhen" („This rat rabble only brings unrest")
Is-a-expression	"**Fluechtlinge sind Parasiten**!" („Refugees are parasites")
Reference after	"**Scheiß Pegida**" („Shit Pegida")
Isolated expression	"Achtkantig rausschmeißen**, die Penner**" („Throw these hobos out on their ears")
Compound	"Raus mit dem **Antifapack**" („Out with this anti-facist-rabble")
Explicit sentence	„Eben echte **Arschlöcher**" („Simply real assholes")
Physical violence	„Schwanz **abhacken**" („Cut the dick off")

As described in the data section, we use a separate dataset to develop the patterns to prevent overfitting. In line with [15], we derive general speech patterns expressing several ways to relate offending words to entities. We were able to adapt four of the seven harassment patterns to the German language with minor modifications accounting for possible intermediate tokens between the offending words and the detected reference. As an example, the "is-a-expression" pattern is depicted in figure 2, relating an offending word to an entity reference by a form of "to be".

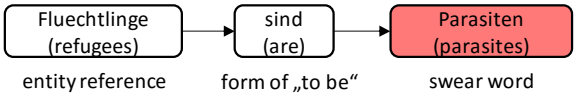

Figure 2. Is-a-expression pattern

In addition, we introduce the compound pattern. The German language allows to use compound words, for example by composing two nouns into a single word. The compound pattern relies on a preprocessing step that splits such compound words into its atomic components. If a combination of offending words and reference is found, the pattern will match. The physical violence pattern searches for combinations of words expressing physical violence towards human beings or parts of the human body. In line with [14], we additionally introduce the explicit sentence pattern. This pattern matches, if a sentence contains severe offensive words typically exclusively referring to persons, regardless of detected references. We distinguish severe offensive words from others by assessing a property to them in our lexicon. As the annotators marked severe offensive statements, we are able to identify the corresponding offensive words by analyzing our training dataset.

4.4. Reference resolver

The reference resolver identifies victims that are addressed in offensive statements detected in the previous step and maps them to one of the groups that are described in section 3. We propose four strategies to resolve such references.

First, if a pattern matches that already contains a direct reference, we directly process this reference and

retrieve the corresponding group from the lexicon. Second, if the reference is indirect, we search for a direct reference in the context of the matching offending passage. If a direct and unambiguous reference can be found, it is resolved accordingly. For the case of detected ambiguous direct references, the closest one is chosen.

Typically, an article or post is the subject of discussion in the context of social media. Users eventually refer to this subject by publishing comments containing indirect references. As a third strategy, we try to analyze the content of the corresponding article searching for direct references using our reference detection module. If such references are found, we resolve them accordingly. Finally, we analyze all comments that are responses to the current article and compute the number of occurrences of direct references ordered by the corresponding group. In this strategy, we assume that comments containing only indirect references typically refer to the same subject that most of the other comments also refer to.

5. Method and evaluation

5.1. Method

We implemented our approach to detect offending statements towards foreigners in two consecutive steps. First, we performed a binary classification task identifying offensive statements. To assess the performance of this task, we computed the evaluation metrics precision (p), recall (r) and f1 as recommended in [33]. Second, we performed a binary classification task assigning each detected offensive statement to the classes offensive (severity = 1) or severely offensive (severity = 2). As the classifier can only process cases that are correctly detected in the first step (true positives), we computed the evaluation metrics without accounting for errors in the first step as we are interested in the performance considering the aspect of practical applicability of the system as discussed in section 6. Finally, we performed a multi class classification task assigning the identified victims to the classes we described earlier in section 3. In particular, we are interested in the performance of the approach to detect offensive statements directed towards foreigners. In analogy to the severity classification, we computed evaluation metrics without considering errors in the previous step. Finally, we implemented a baseline classifier to compare our evaluation results. The baseline classifier consists of a machine learning approach based on a bag-of-words model as described in [14]. We used the software "Rapid Miner" to evaluate different machine learning algorithms. In contrast to [14], we achieved the best results using a naïve bayes classifier without any modifications in Rapid Miner.

5.2. Evaluation

The evaluation results for the offending statement classification are listed in table 4. Both approaches, the baseline classifier and our pattern-based approach, achieved moderate to good results in terms of f1. However, while the baseline classifier achieved higher recall values, the pattern-based approach achieved substantially better precision values.

Table 4. Offending statement classification results (in %)

	Dataset 1			Dataset 2		
	p	r	f1	p	r	f1
Baseline	53.57	76.27	62.94	50.65	71.43	59.27
Pattern-based	75.26	61.86	67.91	73.89	53.46	62.03

The baseline classifier seems to causes false positives by misjudging cases that contain direct or indirect references not belonging to offensive statements. As the approach is based on a bag-of-words model, the context of offensive statements cannot be analyzed directly. In contrast, the pattern-based approach yields less false positives resulting in better precision values. Better precision values reduce the effort for personnel as fewer false positives are detected that need to be corrected in a subsequent step. Additionally, substantial precision values are more suitable for fully automated classification.

Furthermore, the pattern-based approach is able to assess severity values to detected offensive statements. We further investigated the classification performance by distinguishing between the classes offensive (severity 1) and severely offensive (severity 2). The results for each form are denoted in table 5.

Table 5. Severity classification results (in %)

	Dataset 1			Dataset 2		
Severity	p	r	f1	p	r	f1
1	49.51	83.33	62.11	42.17	74.47	53.85
2	69.64	84.78	76.47	70.24	81.94	75.64

While the results for cases with a severity value of 1 are moderate, we were able to achieve good results in terms of f1 value for the detection of severe offending statements. The precision values indicate, that the system is reasonably accurate in detecting such statements and might be used accordingly in practical applications as we will discuss in the next section.

Table 6. Target classification results (in %)

	Dataset 1			Dataset 2		
	p	r	f1	p	r	f1
Foreigner	**51.79**	**65.91**	**58**	**59.26**	**33.56**	**44.44**
Government	**76.32**	**58**	**65.91**	**74.07**	**51.28**	**60.61**
Community	12.5	20	15.39	55.56	83.33	66.67
Press	81.82	77.14	79.41	80	100	88.89

Finally, the evaluation results for the reference identification are subsumed in table 6. The performance measurement for the multi class problem yields contrary results. The results show that a moderate amount of the offensive statements directed towards foreigners was detected correctly, which is the main focus of our study. Frequently, offending statements towards foreigners come along with statements towards the government. In these cases, the classifier seems to misjudge foreigner references for government references and vice versa resulting in moderate overall performance for both of these classes. Additionally, substantial results for press and media class were achieved. These targets are often referenced directly and thus no indirect reference resolution is needed.

6. Practical applications

6.1. Automatic blocking of hate speech

Our approach can be used as a basis for systems that are able to automatically block offending comments. In a proactive manner, the system prevents offending content from its publication. This way, other users are not influenced by the content of the message in a way that facilitates incitement towards foreigners or political parties. Moreover, emotional costs are avoided as they do not have to cope with such content. In contrast to moderators, automated systems are capable of processing a vast amount of messages, which is important in the context of social media platforms as messages can rapidly spread in a viral manner [6]. Furthermore, users with the intention to facilitate incitement might create multiple accounts to bypass suspensions from the social media platform. A proactive system prevents the publication of offending content independent of the account and its message history.

The evaluation revealed that the presented approach is suitable for this kind of practical application with limitations. In automated processing no human control instance that examines the results is involved and thus the cost of false positives need to be considered. A high precision value results in fewer occurrences of false positives and thus reducing these costs. However, precision values around 70 percent result in a fair amount of false positives. Such falsely blocked messages might frustrate users as their message is deleted without proper reason. However, the presented approach allows to assess severity values to indicate the offensiveness of a message. By assuming that substantial offensive content is more likely to violate existing policies or laws, the system can automatically block or delete such messages selectively. As the results in the evaluation section show, the approach can distinguish between offensive and severely offensive statements with substantial precision.

Furthermore, blocking comments is opposed to the right of freedom of speech. Thus, a goal conflict exists between preserving freedom of speech and protecting the victims, authors and the social media platform against potential legal consequences caused by hate speech. It needs to be considered that the decision whether or not a concrete statement from a user violates a certain law is subject to courts and cannot be judged by an automated system.

6.2. Marking comments

The proposed approach can be used in a semi-automated way by automatically marking comments potentially containing offensive statements to present them to a moderator in a subsequent step. Moderators can examine the selected messages and decide, if further actions need to be taken. As a consequence, the effort is reduced compared to manually examining the vast amount of messages in total. Furthermore, communities that are characterized by a substantial amount of published hate speech can be detected as intended by the task force of the German government and Facebook [1]. Marked comments might also be displayed to the author himself before their publication. This way, the author might reconsider the formulation of the message. An offensive comment that is a result of hastily reactions or is, despite its formulation, not intended to be offensive, might be prevented. Finally, community managers might use the system as a third party tool to analyze the comments in response to their published posts. The community manager can then detect problematic comments independently of administrators and eventually remove them.

Considering the moderate to good overall classification performance, the approach is useful for such a task. Compared to automatic blocking or deletion, marking potentially offensive comments shifts the responsibility for the final decision to the human control instance. As a human being is able to take more informed decisions considering multiple aspects on a case-to-case basis, freedom of speech might be preserved more accurately.

6.3. Breaking echo chambers

To tackle the problem of homogenous viewpoints in echo chambers, they first need to be detected, especially those characterized by polarized and homogenous right-wing opinions concerning foreigners or refugees. Our presented approach is suitable for this task, as it is able to detect the referenced victims. If a substantial amount of the offensive statements detected in a community (or in our case Facebook page) is directed towards foreigners, it is likely that the community is characterized by such an echo chamber. The evaluation results reveal, that the foreigner group can be identified precisely and thus, such a detection is possible. Due to the large amount of messages, the chance of detection is improved further.

After the detection of such echo chambers, the beliefs of the users might be challenged by presenting them controversial and well-researched information [11]. The EdgeRank in Facebook, for example, could be adjusted to selectively inject such content. This way, freedom of speech is not violated and each user can decide on his own whether to consider the presented content in its opinion-formation process or not. The presented approach is not able to select appropriate information and selectively inject it into social media. However, prior research addressed this problem in the context of news [34] as well as political discourse in blogs [35]. Such methods might be applied to select appropriate information sources.

7. Conclusion

Recently, offending statements towards refugees and foreigners in social media drew attention to the broader public and are recognized by politicians and social media companies as a growing problem [4,1]. In this work, we proposed, implemented and evaluated an approach to automatically detect offensive statements directed towards foreigners to aid social media platforms in the labor-intensive task of moderation.

We modified the pattern-based approach from Bretschneider et al. [15] to support the German language and to detect and resolve referenced victims, especially foreigners and refugees as well as the government. This step is required as users often refer to their targets indirectly, for example, by paraphrasing or referring to content of the corresponding article or post. Finally, the approach assesses severity values to indicate slightly to offensive statements and severe offensive statements.

To evaluate our approach, we developed an annotated dataset with two human experts providing access to it as a benchmark for further research under the URL www.ub-web.de/research/. The annotations contain offending passages, the referenced victim and a severity value. As evaluation metrics were applied precision, recall and f1-measure. Compared to a machine learning baseline classifier our pattern-based approach yields substantial precision values (75.26% and 73.89%) and moderate overall classification performance in terms of f1 value (67.91% and 62.03%).

We discussed three practical applications: automated blocking and marking of offensive content as well as detecting echo chambers. The achieved precision values allow automated processing of offensive content with limitations as there is a fair amount of remaining false positives. The approach could be used selectively by distinguishing between severely offending content that might be automatically blocked and other offending statements that might be presented to moderators in a semi-automated manner. As we are able to identify the referenced victims, the approach can be used to detect echo chambers containing homogenous xenophobic or racist viewpoints. To aid users kept in such echo chambers, controversial and well-researched information might be presented to them [11]. This way, the existing, potentially polarized, beliefs of social media users are challenged and the political opinion-formation-process is based on more diverse information [13]. Applying such an approach has ethical implications that need to be carefully considered. A major concern is the conflict between preserving freedom of speech and protecting others from hate speech possibly conflicting with their individual rights [17]. Furthermore, if the system is used in an automated manner the responsibility of judging the behavior of users entirely relies on a machine.

Further research is desired on several aspects. First, we did not consider characteristics of the sender of hate speech as the approach can be applied in anonymous contexts. However, such characteristics might improve the classification performance. Second, the approach is not capable of detecting paraphrased offending statements, for example in the form of gender based harassment. To identify such cases, semantic approaches might be applied as an extension. Moreover, to apply the method to different languages, a general framework or guideline could be created to aid this process in a structured way. Finally, the system is not capable of incorporating cross-cultural differences in the perception of offending content. To capture such differences, several configurations containing different hate speech patterns could be analyzed.

10. References

[1] URL: http://www.bbc.com/news/world-europe-34256960, last accessed 05/30/2016.

[2] URL: http://wapo.st/1LMY05q, last accessed 05/30/2016.

[3] K. Wise, B. Hamman, and K. Thorson, "Moderation, Response Rate, and Message Interactivity", *Journal of Computer-Mediated Communication*, vol. 12, no. 1, pp. 24-41, 2006.

[4] URL: http://www.bbc.co.uk/newsbeat/article/30694252/why-are-thousands-of-germans-protesting-and-who-are-pegida, last accessed 05/30/2016.

[5] URL: http://newsroom.fb.com/company-info/, last accessed 06/02/2016.

[6] R.A. King, P. Racherla, and V.D. Bush, "What We Know and Don't Know About Online Word-of-Mouth", *Journal of Interactive Marketing*, vol. 28, no. 3, pp. 167-183, 2014.

[7] E. Gilbert, and K. Karahalios, "Predicting Tie Strength with Social Media", in *SIGCHI Conference on Human Factors in Computing Systems*, Boston, USA, pp. 211-220, 2009.

[8] L.M. Jones, K.J. Mitchell, and D. Finkelhor, "Online harassment in context: Trends from three Youth Internet Safety Surveys (2000, 2005, 2010)", *Psychology of Violence*, vol. 3, no. 1, pp. 53-69, 2013.

[9] M.L. Williams, and P. Burnap, "Cyberhate on Social Media in the aftermath of Woolwich", *British Journal of Criminology*, vol. 56, no. 2, pp. 211-238, 2016.

[10] J. Glaser, J. Dixit, and D.P. Green, "Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence?", *Journal of Social Issues*, vol. 58, pp. 177-193, 2002.

[11] A. Gruzd, and J. Roy, "Investigating Political Polarization on Twitter: A Canadian Perspective", *Policy & Internet*, vol. 6, no. 1, pp. 28-45, 2014.

[12] M.D. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, A. Flammini, and F. Menczer, "Political Polarization on Twitter", in *International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.

[13] C.R. Sunstein, "The Law of Group Polarization", *Journal of Political Philosophy*, vol. 10, no. 2, pp. 175-195, 2002.

[14] Y. Chen, Y. Zhou, Y. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety", in *International Conference on Privacy, Security, Risk and Trust*, Amsterdam, Netherlands, pp. 71-80, 2012.

[15] U. Bretschneider, T. Wöhner, and R. Peters, "Detecting Online Harassment in Social Networks", in *International Conference on Information Systems*, Auckland, New Zealand, 2014.

[16] URL: http://www.un.org/en/universal-declaration-human-rights/, last accessed 22/08/2016.

[17] URL: http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx, last accessed 22/08/2016.

[18] H. Keller, and M. Sigron, "State Security v Freedom of Expression", *Human Rights Law Review*, vol. 10, no. 1, pp. 151-168, 2010.

[19] S. W. Kim, and A. Douai, "Google vs. China's 'Great Firewall'", *Technology in Society*, vol. 34, no. 2, pp. 174-181, 2012.

[20] M. Oetheimer, "Protecting Freedom of Expression", *Cardozo Journal of International & Comparative Law*, vol. 17, no. 3, pp. 427-443, 2009.

[21] A. Weber, "Manual on hate speech", Council of Europe Publishing, Strasbourg Cedex, France, 2009.

[22] URL: http://hudoc.echr.coe.int/eng?i=001-155105, last accessed 24/05/2016.

[23] A. R. Hochschild, "Emotion Work, Feeling Rules, and Social Structure", *American Journal of Sociology*, vol. 85, no. 3, pp. 551-575, 1979.

[24] A. Menking, and I. Erickson, "The Heart Work of Wikipedia: Gendered, Emotional Labor in the World's Largest Online Encyclopedia", in *ACM Conference on Human Factors in Computing Systems*, Seoul, Korea, pp. 207-210, 2015.

[25] V. O. Key, "The Responsible Electorate: Rationality in Presidential Voting", Belknap Press, Cambridge, USA, 1966.

[26] URL: https://www.facebook.com/help/327131014036297/, last accessed 06/02/2016.

[27] R.S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization", *Computers in Human Behavior*, vol. 26, no. 3, pp. 277-287, 2010.

[28] S.O. Sood, E.F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites", *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 270-285, 2012.

[29] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying", *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 2, pp. 1-30, 2012.

[30] J.M. Xu, K.S. Jun, X. Zhu, and A. Bellmore, "Learning from Bullying Traces in Social Media", in *Conference of the NAACL: HLT*, Stroudsburg, USA, pp. 656-666, 2012.

[31] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying", in *the 5th Annual ACM Web Science Conference*, Paris, France, pp. 195-204, 2013.

[33] URL: http://www.hyperhero.com/de/insults.htm, last accessed 06/02/2016.

[34] M. Sokolova, and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information Processing and Management*, vol. 45, no. 4, pp. 427-437, 2009.

[35] S. Park, S. Kang, S. Chung, and J. Song, "NewsCube: delivering multiple aspects of news to mitigate media bias", in *SIGCHI Conference on Human Factors in Computing Systems*, New York, USA, pp. 443-452, 2009.

[36] A. Oh, H. Lee, and Y. Kim, "User Evaluation of a System for Classifying and Displaying Political Viewpoints of Weblogs", in *AAAI International Conference on Weblogs and Social Media*, San Jose, USA, 2009.

Erklärung über verwendete Hilfsmittel

Hiermit erkläre ich, dass ich die Dissertation “Detektion von Directed Hate Speech, Online Harassment und Cyberbullying in Online Communities“ selbstständig angefertigt habe und dass ich alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlehnenden Ausführungen meiner Arbeit besonders gekennzeichnet und die Quellen zitiert habe.

Halle/Saale, den 20.03.2017

Gutachter der schriftlichen Fassung:

- Prof. Dr. Ralf Peters, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für E-Business
- Prof. Dr. Stefan Sackmann, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Betriebliches Informationsmanagement

Prüfungskommission zur Verteidigung der Dissertation:

- Prof. Dr. Ralf Peters, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Electronic Business
- Prof. Dr. Stefan Sackmann, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Betriebliches Informationsmanagement
- Prof. Dr. Taïeb Mellouli, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Wirtschaftsinformatik und Operations Research
- Prof. Dr. Jörg Laitenberger, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Finanzierung und Banken
- Prof. Dr. Martin Klein, Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für Internationale Wirtschaftsbeziehungen

Verteidigungsdatum: 11.07.2017