



(Semi-)Automatische Kommentarmoderation zur Erhaltung Konstruktiver Diskurse

13

Marco Niemann, Dennis Assenmacher, Jens Brunk,
Dennis M. Riehle, Heike Trautmann und Jörg Becker

1 Einleitung

Das Internet und insbesondere das interaktive Internet – oft auch Web 2.0 genannt – hat die Art und Weise, in welcher Menschen kommunizieren und Informationen austauschen, nachhaltig verändert. Auch den Journalismus tangiert dies auf mehreren Ebenen – unter anderem auch in Bezug auf die Interaktion mit den Leser*innen. Viele Jahre war der Leserbrief das gängige Mittel als Leser*in, die eigene Meinung zu Inhalten eines Printmediums kundzutun (Reich, 2011) inkl. der Verzögerung durch Postumlaufzeiten und die vor Veröffentlichung

M. Niemann (✉) · D. Assenmacher · J. Brunk · H. Trautmann · J. Becker
Institut für Wirtschaftsinformatik, Westfälische Wilhelms-Universität Münster, Münster,
Deutschland
E-Mail: marco.niemann@ercis.uni-muenster.de

D. Assenmacher
E-Mail: dennis.assenmacher@wi.uni-muenster.de

J. Brunk
E-Mail: j_brun17@uni-muenster.de

H. Trautmann
E-Mail: trautmann@wi.uni-muenster.de

J. Becker
E-Mail: joerg.becker@ercis.uni-muenster.de

D. M. Riehle
Institut für Wirtschafts- und Verwaltungsinformatik, Universität Koblenz-Landau, Koblenz,
Deutschland
E-Mail: riehle@uni-koblenz.de

notwendigen Prüfungen. Heutzutage haben Leser*innen die Möglichkeit, über Kommentarspalten und „Diskussionsforen“ mit den Journalist*innen und auch anderen Leser*innen in Kontakt zu treten (Kolhatkar & Taboada, 2017; Pöyhtäri, 2014; Riehle et al., 2020) – mit wesentlich verkürzter Antwortzeit und einem stark erhöhten Grad an Interaktivität (Vogel, 2017). Soziale Netzwerke sind von dieser Entwicklung ebenso betroffen (Chetty & Alathur, 2018; Del Vigna et al., 2017; Mathew et al., 2018) – aufgrund abweichender rechtlicher Grundlagen und technischer Handhabungsmöglichkeiten befinden sich diese jedoch außerhalb des Betrachtungswinkels dieses Beitrags (da hier die Optionen zur (semi-) automatisierten Bekämpfung auch stark von den verfügbaren Schnittstellen der Plattformbetreiber*innen abhängig sind).

Die mit den aufkommenden Diskussionsforen und Kommentarspalten¹ verknüpften Hoffnungen waren gerade zu Beginn hoch: Man erhoffte sich einen positiven Beitrag zum politischen Diskurs und auch eine Förderung der Demokratie im Allgemeinen (Papacharissi, 2004). Aspekte wie eine größere Anonymität und höhere Dezentralität (und damit verbunden eine vergrößerte Basis an Diskutanten) wurden als mögliche Vorteile für offenere und intensivere Debatten aufgefasst (Papacharissi, 2004). Diese Meinung ist auch unter Journalist*innen verbreitet (Gardiner et al., 2016) und findet in der Wissenschaft nach wie vor Rückhalt (Lewis et al., 2014). Über dieses hehre Ziel hinaus sind Kommentare für Medienunternehmen eine Möglichkeit, Besucher*innen ihrer Webpräsenzen länger zu binden und so letztlich die Höhe der generierten Werbeeinnahmen zu steigern (Pöyhtäri, 2014).

Trotz aller positiven Aspekte und assoziierten Hoffnungen waren bereits früh auch mögliche Risiken bekannt – u. a. das Aufkommen von unangemessenen Kommentaren, insbesondere jenen mit explizitem Hass (Lewis et al., 2014; Papacharissi, 2004). Auch mögliche Konsequenzen wurden bereits identifiziert, da in den frühen 2000ern bereits klar war, dass derartige Verhaltensmuster rechtlich problematisch sind (Fišer et al., 2017; Karadeniz, 2009; Pöyhtäri, 2014) und ggfs. andere Nutzer*innen verschrecken (Papacharissi, 2004). Dies bedeutet sukzessive nicht nur mögliche Strafzahlungen, sondern auch Rückgänge von Werbeeinnahmen (Nobata et al., 2016) sowie erhöhte Ausgaben für eine Moderation der Kommentare, um eine gewisse Grundqualität sicherzustellen (Pöyhtäri, 2014). Ein Blick in die Statistiken zeigt, dass sich viele dieser Befürchtungen

¹ Autor*innen wie u. a. Papacharissi (2004) betonen, dass sich Kommentarspalten und Diskussionsforen dahin gehend unterscheiden, dass letztere i. d. R. themenspezifisch ausgestaltet und erstere offener ausgelegt sind. Im Rahmen dieses Beitrages verwenden wir beide Begriffe weitestgehend synonym und beziehen uns explizit auf die Kommentierungsmöglichkeiten unterhalb von Zeitungs- oder Magazinartikeln.

über die Jahre materialisiert haben: So gibt es – je nach Quelle – einen Anteil problematischer Kommentare, der von hohen 2 % bis zu extremen 80 % reicht (Cheng, 2007; Gardiner et al., 2016). Diese immensen rechtlichen und finanziellen Risiken haben bspw. in den vergangenen Jahren 50 % und mehr der (deutschen) Zeitungen dazu gedrängt, Kommentarfunktionalitäten abzuschaffen (Siegert, 2016; Vogel, 2017). Wenn auch mit abweichenden Anteilen, war (und ist) dieser Trend weltweit zu beobachten (Huang, 2016).

Trotz der enormen Dimension des Problems sind viele Journalist*innen nach wie vor überzeugt, dass ein wichtiger Teil ihres Berufsbildes – und auch des Journalismus im Allgemeinen – darin besteht, den öffentlichen Diskurs zu fördern (Heinonen, 2011; Pöyhtäri, 2014). Um dies mittel- und langfristig ermöglichen zu können, müssen die Journalist*innen und Medienschaffenden bei der Prüfung der Kommentare mental entlastet und auch die finanziellen Limitationen (bspw. durch zusätzliche Moderator*innen) reduziert werden (die Problematik des Aufwandes und der Finanzierung wird u. a. auch in Reich, 2011 erläutert). Hieran arbeiten seit ca. einem Jahrzehnt Praktiker*innen (bspw. Nobata et al., 2016 für Yahoo; Wulczyn et al., 2017 für Google; Rajamanickam et al., 2020 für Facebook) und (unabhängige) Wissenschaftler*innen (Waseem & Hovy, 2016; Salminen et al., 2020; Köffer et al., 2018; Niemann, 2019; Jorgensen et al., 2020), welche versuchen, Methoden des maschinellen Lernens und des Natural Language Processing zur automatischen Bewertung/Moderation von Kommentaren zu verwenden. Auch wenn solche Systeme bereits partiell im Einsatz sind, sind unabhängige Studien zu ihrer Qualität noch rar, denn die notwendigen Daten und Modelle sind oftmals nicht frei verfügbar (Dansby et al., 2020). Allerdings geben sich selbst große Unternehmen wie Facebook zurückhaltend und verwenden Algorithmen primär unterstützend (Rosen, 2019) – die Hauptlast tragen nach wie vor Moderator*innen und Dienstleister*innen (Newton, 2019; Rosen, 2019; zu Details über die Moderationspraktiken großer Plattform-Unternehmen s. Mündges Beitrag in diesem Band).

Ziel dieses Buchbeitrages ist es, zunächst einen groben Überblick über den Forschungsstand der (teil-)automatisierten Erkennung von problematischen Kommentaren zu geben. Im Anschluss wird im Rahmen eines Überblicks über zwei zentrale Schwierigkeiten der Domäne berichtet: die Definition von problematischer Sprache und die Annotation von Kommentaren mit entsprechenden Labels sowie die Erstellung von ML-Modellen, welche die Erkennung problematischer Kommentare unterstützen können. Abschließend geben wir einen Überblick über das Forschungsprojekt MODERAT!, welches sich dieser Probleme mittels eines innovativen Geschäftsmodells und einer Analytics-as-a-Service-Plattform

annimmt und speziell auf die (teil-)automatische Erkennung problematischer deutscher Kommentare ausgerichtet ist.

2 Forschungsstand

Die Forschung im Bereich der automatisierten Erkennung von Hassrede und mit ihr artverwandter Formen ist noch eine recht junge Disziplin. Während das allgemeine Problem des Online-Hasses bereits Ende der 1990er-Jahre wissenschaftlich beschrieben wurde (Phillips, 1996), wurde dem Problem bis ca. 2010/2011 keine größere Bedeutung beigemessen (Gardiner et al., 2016). In der Zwischenzeit hatte sich im Internet der Wandel vom eher statischen Web 1.0 zum interaktiven Web 2.0 vollzogen.

Erste maschinelle Verfahren zur Detektion von Hassrede – bzw. von „online harassment“ – wurden ab ca. 2009 entwickelt und machten sich den Fortschritt in der Forschung um maschinelles Lernen zunutze (Yin et al., 2009). Für ihre Arbeit nutzten die Autor*innen aus heutiger Sicht noch vergleichbar simple Verfahren zur Verarbeitung textueller Daten: Kommentare wurden in n -Grams zerlegt (vgl. Abb. 1), um Sentiment-Werte (im Text detektierbare Emotionswerte) und Kontextvariablen (bspw. Nutzer*innennamen, Zeitstempel,...) angereichert und durch eine binäre Support-Vector-Machine (SVM) klassifiziert (vgl. Abb. 2). Forschungsarbeiten wie die von Sood et al. (2012) sowie Bretschneider et al. (2014) formulierten bereits früh ein kritisches Problem der noch jungen Domäne: Bedeutete für Yin et al. (2009) „harassment“ noch die beabsichtigte Belästigung anderer inkl. leicht abweichender Formen, wie der Verbreitung von Hass und dem gezielten Mobbing anderer, wählen Bretschneider et al. (2014) schon die leicht abgewandelte Definition von elektronischen Nachrichten, die ihre Opfer psychologisch verletzen. Andere Autor*innen, für die Sood et al. (2012) hier nur beispielhaft genannt werden, führten weitere Konzepte wie „Beleidigung“ und „Profanität“ in die wissenschaftliche Debatte ein. Was zunächst trivial anmuten mag, ist ein bis heute fortbestehendes Problem in dem Forschungsbereich, da die verschiedenen Definitionen eine Vergleichbarkeit existierender Arbeiten (und damit auch existierender Lösungen) extrem erschweren (Brunk et al., 2019; Ganz, 2019; Niemann et al., 2020). Die vorab genannten Kategorien sind hierbei nur eine kleine Teilmenge des gesamten Definitionsraumes (Brunk et al., 2019; Niemann et al., 2020).

Eng verbunden mit der Frage nach der adäquaten Klassifikation von Hassrede ist die Erstellung von Datensätzen, welche zur Erstellung von maschinellen Modellen genutzt werden können. Hintergrund ist die nach wie vor dominierende

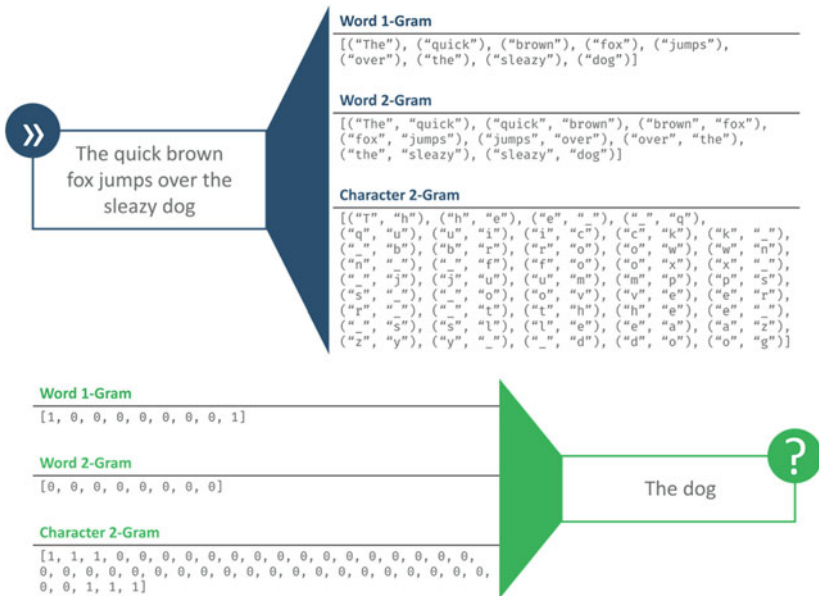


Abb. 1 Exemplarische Darstellung versch. n -gram Zerlegungen eines Satzes

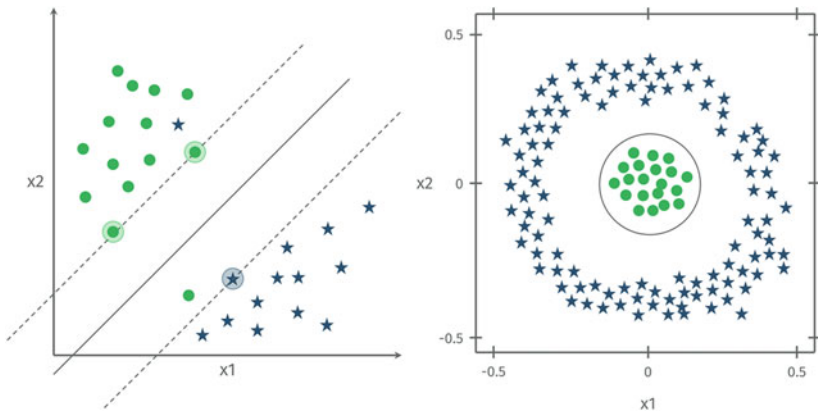


Abb. 2 Beispielhafte Abbildungen der Klassifikation durch eine Support Vector Machine (SVM)

Stellung überwachter Lernverfahren bei der Klassifikation von Hasskommentaren – sprich von Verfahren, welche klar kategorisierte Lerndaten benötigen, um neue Datenpunkte zu bewerten (Bretschneider et al., 2014). Werden unterschiedliche Kategorisierungen verwendet, sind natürlich auch die Datensätze (und mit ihnen erstellte ML-Modelle) nicht kombinierbar und vergleichbar. Nicht nur die Systematik zur Kategorisierung stellt die Schöpfer*innen und Nutzer*innen entsprechender Datensätze vor Herausforderungen, sondern auch die Art und Weise der Annotation der entsprechenden Kategorien an die Kommentare (Brunk et al., 2019). Denn i. d. R. annotieren die Verfasser*innen von Kommentaren diese nicht selbstständig, sondern eine dritte Partei. Klassischerweise sind dies entweder Moderator*innen der entsprechenden Webplattform (Nobata et al., 2016), sogenannte Crowdworke (Howe, 2006; Nobata et al., 2016), oder auch die den Datensatz erstellenden Forscher*innen (Ross et al., 2016). Jede dieser Methoden hat verschiedene Vor- und Nachteile, etwa hinsichtlich des Grades an Übereinstimmung zwischen verschiedenen Personen bzw. an Kontinuität pro Person oder hinsichtlich der Nähe zum kommentierenden Publikum. Es gibt auch hier diverse Prüfmechanismen, bspw. Übereinstimmung verschiedener Bewerter*innen, um die Qualität der Daten bewerten und vergleichen zu können.

Darüber hinaus zeigen aktuelle Literatur-Reviews, dass die verfügbaren Datensätze noch mit weiteren Problemen bzw. Einschränkungen belastet sind (Fortuna & Nunes, 2018). Eine dieser Einschränkungen liegt in den Quellen der typischerweise verwendeten Daten. So nutzen die meisten Datensatzersteller*innen bis zum heutigen Tage Daten aus sozialen Netzwerken – und hier primär dem Kurznachrichtendienst Twitter. Begründet ist dies oftmals mit der einfachen Zugänglichkeit der Daten, da bspw. Tweets über die integrierte Twitter API abgefragt werden können. Problematisch ist, dass Twitter als Kurznachrichtendienst seine Mitteilungen auf 280 (früher 140) Zeichen beschränkt. Dies verändert nicht nur die verwendete Sprache (bspw. höhere Frequenz an Abkürzungen), sondern erzeugt eine künstlich beschränkte Textlängenverteilung, die längere und elaboriertere Kommentare nicht abdeckt. Somit ist die Erstellung von Modellen für Texte anderer Medien stark erschwert. Hinzu kommt, dass viele der Datensätze nur eine einmalige Nutzung erfahren bzw. die Verwendung von Datensätzen durch Dritte nicht der Regelfall ist (Fortuna & Nunes, 2018). Während dies vielfach schlichtweg der Ausdruck divergierender Klassifikationen ist, fehlen dennoch für zahlreiche Datensätze erneute Überprüfungen der Datenqualität für die Erstellung maschineller Modelle.

Auch für das Herzstück der Forschung zur (semi-)automatischen Erkennung von Hassrede – der Entwicklung entsprechender ML-Modelle – gibt es mittlerweile eine umfangreiche Anzahl an Forschungsergebnissen (Brunk et al., 2019;

Fortuna & Nunes, 2018; Niemann, 2019). Äquivalent zu allen anderen ML-Domänen gilt auch für die Erkennung von Hassrede, dass es keinen One Size Fits All-Algorithmus gibt (Wolpert & Macready, 1997). Somit zeigt sich bei einem Blick in die letzten 20 Jahre Forschung in diesem Bereich ein hochgradig diverses Bild bezüglich der verwendeten Algorithmen (Brunk et al., 2019; Fortuna & Nunes, 2018). Der Großteil der frühen Publikationen, aber auch der aktuellen Manuskripte, nutzt primär klassische Klassifikationsalgorithmen. Dazu gehören u. a. Verfahren wie die logistische Regression (Badjatiya et al., 2017; Davidson et al., 2017; Niemann, 2019; Wulczyn et al., 2017), Entscheidungsbäume (und Random Forest) (Burnap & Williams, 2015; Dinakar et al., 2011; Niemann, 2019) und SVMs (Badjatiya et al., 2017; Lee & Yoon & Jung, 2018; Mathur et al., 2018; Mehdad & Tetreault, 2016; Warner & Hirschberg, 2012). Wie in fast allen Bereichen, die sich Methoden maschinellen Lernens bedienen, sind auch in der Erkennung problematischer Sprache neuronale Netzwerke zunehmend populär. Dies reicht von generischen Ansätzen basierend auf RNNs (Recurrent Neural Networks) (Mehdad & Tetreault, 2016; Pavlopoulos et al., 2017; Serrà et al., 2017), CNNs (Convolutional Neural Networks) (Mathur et al., 2018; Pavlopoulos et al., 2017; Švec et al., 2018) und LSTMs (Long Short-Term Memory) (Aken et al., 2018; Kolhatkar & Taboada, 2017) bis zu aktuelleren Transformer-basierten Vorgehen (Bagueño & Mendoza, 2020; Pavlopoulos et al., 2019; Risch et al., 2019; Salminen et al., 2020), welche rein auf die Arbeit mit natürlicher Sprache ausgerichtet sind und in aktuellen Studien vielversprechende Ergebnisse generieren. Auch alternative Konzepte wie bspw. AutoML als Ansatz zur Reduktion des menschlichen Konfigurationsbias finden sukzessive ihren Weg in die Domäne (Jorgensen & Choi, 2019; Jorgensen et al., 2020). Neben der allgemeinen Dynamik der ML-Domäne steht auch das Fehlen einheitlicher Metriken zum Vergleich der Performanz (respektive die konsistente Nutzung existierender Metriken) der weiteren Entwicklung im Wege, da Forschungsergebnisse so oft nur sehr eingeschränkt vergleichbar sind (Brunk et al., 2019).

Darüber hinaus gibt es noch eine Vielzahl weiterer Aspekte, welche bisher noch wenig befohrt worden sind. Dazu zählen u. a. die Einbindung maschineller Modelle in tatsächlich nutzbare Plattformen inkl. eines entsprechenden Geschäftsmodells (Brunk et al., 2019; Riehle et al., 2020), die Sicherstellung algorithmischer Transparenz bzw. der Erklärbarkeit der Entscheidungen der ML-Modelle sowie die Akzeptanz durch die Moderator*innen und die Nutzer*innen der Kommentarspalten (Brunk et al., 2019).

3 Kommentar-Annotation

Wie bereits im Überblick zum Forschungsstand angesprochen, ist ein zentrales Problem bei der Erkennung und (semi-)automatisierten Bekämpfung von problematischer Sprache die angemessene Annotation der Kommentare (für eine Verwendung durch ML-Modelle). Komplikationen ergeben sich hierbei u. a. sowohl durch die Vielzahl verwendeter Konzepte in der Literatur als auch durch oftmals abweichende Definitionen für die verschiedenen Konzepte (Jurgens et al., 2019; Niemann et al., 2020). Gängige betrachtete Konzepte umfassen u. a. *Hassrede* (Niemann, 2019; Nobata et al., 2016; Warner & Hirschberg, 2012), *Beleidigungen* (Niemann, 2019; Sood et al., 2012), *anstößige Sprache* (Y. Chen et al., 2012; Davidson et al., 2017), *Drohungen* (Anzovino et al., 2018; Parliamentary Assembly, 2007), *Mobbing*² (Chatzakou et al., 2017a, b), *Profanität* (Sood et al., 2012) und viele weitere. Viele dieser Konzepte überlappen sich je nach Definition mit anderen, während auch für einzelne Konzepte die Definitionen in den verschiedenen Quellen variieren. Sprachliche Eigenheiten wie Sarkasmus und Ironie (Fortuna & Nunes, 2018; Founta et al., 2018; Geiger et al., 2020; Lachenicht, 1980) sowie unterschiedliche Sichtweisen pro Person kommen bei der Definition und Anwendung erschwerend hinzu (Chetty & Alathur, 2018; Geiger et al., 2020; Mondal et al., 2017). Die daraus resultierende Unklarheit ist für die Erstellung maschineller Modelle besonders problematisch. Denn wenn die Moderator*innen oder Crowdworker, welche Lerndaten bereitstellen, nicht akkurat annotieren – was ohne klare Abgrenzung hochgradig schwierig ist (Fišer et al., 2017) – können auch daraus resultierende ML-Modelle keine qualitativ hochwertigen Ergebnisse liefern, denn die menschliche Performanz ist hier der Referenzwert (Ross et al., 2016; Warner & Hirschberg, 2012).

Ein Blick auf die Domäne und die existierenden Definitionen zeigt, dass ein gewisses Maß an Variation insbesondere auch vor dem Hintergrund divergierender rechtlicher Vorgaben unabdingbar ist (Ullmann & Tomalin, 2020). Um zukünftigen Definitionen und Definitionsansätzen dennoch zusätzlich Struktur zu geben, schlagen Niemann et al. (2020) einen konfigurierbaren, mehrperspektivischen Ansatz vor, der Wissenschaftler*innen und Praktiker*innen bei der Erstellung fundierter und klarer Label unterstützt (siehe Abb. 3). Erste Quelle für Informationen sind aufgrund des linguistischen Fokus der Aufgabe Wörterbücher der entsprechenden Zielsprache. Hintergrund für dieses Vorgehen ist, dass für

² Das Mobbing wird an dieser Stelle nicht weiter betrachtet, da es sich strukturell von den anderen aufgezählten Konzepten unterscheidet. Der primäre Unterschied ergibt sich hierbei daraus, dass Mobbing einen wiederholten Angriff auf eine Person darstellt. Somit kann eine einzelne Nachricht ohne weiteren Kontext nicht als Mobbing klassifiziert werden.

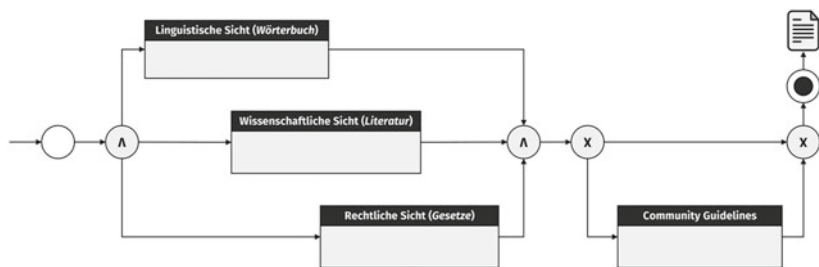


Abb. 3 Konfigurierbarer, multi-perspektivischer Ansatz zur Erstellung von Definitionen für Problematische Sprache (Niemann et al., 2020)

bestimmte Wörter und Konzepte dort bereits gängige Definitionen hinterlegt sind oder bspw. auf artverwandte oder synonyme Begriffe und Konzepte verwiesen wird, welche ebenfalls berücksichtigt werden müssen. Da viele Arten der problematischen Sprache einen justiziablen Hintergrund haben (Brown, 2017; Fišer et al., 2017; Ullmann & Tomalin, 2020), dienen die nationalen und supranationalen Gesetze (insb. für Staaten, die in supranationalen Konstrukten wie bspw. der EU integriert sind, von Bedeutung) als zweiter wichtiger Ansatzpunkt, um eine Definition problematischer Konzepte zu erarbeiten. Abschließend wird auch der aktuelle Forschungsstand berücksichtigt, da bei einem rechtlich komplexen und auch darüber hinaus vielschichtigen Sachverhalt wie problematischer Sprache das Recycling vorhandener Ergebnisse den Erkenntnisprozess beschleunigen kann. Darüber hinaus bieten die wissenschaftlichen Qualitätssicherungsverfahren (u. a. Peer Review für gängige Veröffentlichungstypen) eine zusätzliche Absicherung, dass die generierten Label auf fundierten Erkenntnissen beruhen.

Exemplarisch wurde das Verfahren auf den Fall Deutschland angewendet und zunächst ein theoretisch fundiertes Schema hergeleitet. Hierzu wurden entsprechende einsprachige Wörterbücher, das deutsche Strafgesetzbuch, EU Verordnungen und die gängige wissenschaftliche Literatur herangezogen (Niemann et al., 2020). Auch die Sicht eines großen deutschen Zeitungshauses wurde gemäß der Empfehlungen von Jurgens et al. (2019) und Niemann et al. (2020) prototypisch gepflegt. Hierzu wurde das theoretisch abgeleitete Schema intensiv mit mehreren hauptberuflichen Moderator*innen diskutiert, geschärft und um plattformspezifische Kategorien erweitert. Das vollständige und das theoriegetriebene Konzept sind in Abb. 4 einmal gegenübergestellt:

Während das theoriegetriebene Modell in diesem Fall von anderen Wissenschaftler*innen und Praktiker*innen im deutschsprachigen Raum aufgegriffen

	Label	Erklärung
Theorie- Getrieben	Sexismus	Attacken auf Personen basierend auf ihre(m/r) Geschlecht(sidentität), oft mit einem Fokus auf Frauen
	Rassismus	Attacken auf Personen basierend auf ihrer Herkunft, Ethnie, Nationalität oder Religion; typ. gedacht Hass zu erzeugen
	Drohung	Ankündigung der Verletzung der körperlichen Unversehrtheit des Opfers
	Beleidigung	Verunglimpfende oder verächtliche Äußerungen (in der Regel ohne weitere Angaben)
	Profane Sprache	Verwendung einer sexuell expliziten und unangemessenen Sprache
Organisations- spezifisch	Meta / Organisatorisch	Kommentare, welche sich bspw. auf die Sperrung voriger Kommentare beziehen; kein Themenbezug; diese Kommentare sind i.d.R. nicht missbräuchlich oder problematisch per se, tragen aber oftmals dazu bei, dass sich Konversationen emotional aufladen ohne einen inhaltlichen Mehrwert für die aktive Diskussion beizutragen
	Werbung	Explizites Bewerben von Produkten oder Dienstleistungen; nicht grundsätzlich missbräuchlich problematisch, Beobachtungen zeigen jedoch, dass solche Kommentare zu einer Verschlechterung der Diskussionsqualität beitragen können

Abb. 4 Vollständiges Label-Schema erstellt nach dem Verfahren von Niemann et al. (2020)

werden könnte, ist bei dem vollständigen Konzept stets zu bedenken, dass dies auch die Spezifika einer einzelnen Unternehmensinstanz abbildet. Der Ansatz bietet somit gegenüber anderen bestehenden Vorgehensmodellen den Vorteil, dass eine reproduzierbare Methode verwendet werden kann, um Label-Schemata zu erstellen (Niemann et al., 2020). Die so erzeugten Label-Schemata können sowohl auf einer national-allgemeinen Ebene gehalten werden, um Vergleichbarkeit zu ermöglichen, erlauben es aber auch, die Unterschiede einzelner Anwender*innen dediziert zu berücksichtigen. Solche Unterschiede sind aus rein wissenschaftlicher Sicht wenig wünschenswert, in der Praxis aber de facto unabdingbar (Niemann et al., 2020).

Abschließend bleibt zu erwähnen, dass im Rahmen von Pilotstudien sowohl mit den entsprechenden Labeln als auch in den aktuell in Entwicklung befindlichen Softwarelösungen strikt der sogenannte Multi-Label-Ansatz verfolgt wird. Die u. a. im Label-Erstellungsprozess angefallene Arbeit mit Veröffentlichungen und Praktiker*innen zeigt deutlich, dass, wie u. a. von Jurgens (2013) postuliert, Kommentare oftmals nicht eindeutig Kategorien zugewiesen werden können. Das Gestatten multipler Label pro Kommentar soll helfen, die Annotationsqualität zu erhöhen (kein willkürliches, maschinell schlecht trennbares Einordnen in Kategorien) (Founta et al., 2018; Niemann, 2019), aber auch sinnvolles Feedback im Nutzungsprozess ermöglichen (maschinelle Modelle können mit feingranulareren Labeln mehr Feinheiten aufdecken).

4 Maschinelles Lernen

Die Identifikation von Hassrede in sozialen Medien ist aus Sicht des Data Scientist auf den ersten Blick vorerst ein klassisches Problem, das mit überwachten

Lernverfahren zur Klassifikation voll automatisiert gelöst werden kann. Erst eine detaillierte Betrachtung der Datensituation sowie der Anforderungen an ein potenzielles Lösungsmodell lässt auf eine komplexere Ausgangssituation zurück schließen. Insbesondere Aufgaben, die in den Bereich des Natural Language Processing fallen, benötigen eine nicht zu unterschätzende Anzahl an annotierten Lerndaten, also Daten (in diesem Fall Nutzer*innenkommentare), denen eindeutig zugeordnet ist, ob sie in den Bereich der Hassrede fallen oder nicht. Hier ergeben sich mehrere Problematiken. Einerseits ist die Erhebung einer großen Datenbasis insbesondere für kleinere Medienhäuser oft kein realistisches Szenario, einfach weil die Anzahl der Kommentare pro Tag nicht ausreicht. Andererseits sind vorhandene Datensätze üblicherweise hochgradig ungleichmäßig verteilt, d. h. eine oder mehrere potenzielle Zuordnungsklassen sind über- oder unterproportional häufig im Datensatz vertreten. Im vorliegenden Fall sind Kommentare natürlich im Allgemeinen unkritisch, enthalten also keine Hassrede. Dementsprechend ist das Verhältnis zwischen Hassrede und einem unproblematischen Kommentar nicht ausgeglichen.

Ein weiterer problematischer Aspekt, wie bereits mehrmals in diesem Beitrag erwähnt, ist die Subjektivität der Annotation. Während es in den klassischen Problemen des maschinellen Lernens immer eine eindeutige Klassifikation von Datenpunkten gibt, ist dies in diesem Kontext nicht der Fall und unterscheidet sich nicht nur von Medienhaus zu Medienhaus, sondern teilweise bereits von Mensch zu Mensch. Dementsprechend ist auch eine Vermischung von Datensätzen unterschiedlicher Quellen nicht trivial umzusetzen und kann zur Verzerrung der gelernten Modelle führen.

Üblicherweise ist die Ausgabe eines Klassifikationsmodelles die Entscheidung über die Zugehörigkeit einer Klasse. Diese Zugehörigkeit wird über eine sog. Konfidenz beschrieben (oftmals eine Wahrscheinlichkeit über die Zugehörigkeit). Die finale Entscheidung der Zuordnung erfolgt dann durch die Festlegung einer Entscheidungsgrenze. Erreicht oder übertrifft die vom Modell zugewiesene Konfidenz diese Grenze, wird der Kommentar der jeweiligen Klasse zugeordnet. Fraglich ist nun, wie diese Grenze festgelegt werden muss oder soll und wo eine Zensur von Inhalten stattfindet. Dies ist insbesondere in Deutschland ein sensibles Thema in der Gesellschaft und erfordert nicht nur eine informationstechnische, sondern auch sozialwissenschaftliche Betrachtung des Sachverhalts.

Insgesamt lässt sich feststellen, dass es sich bei der Identifizierung von Hassrede um eine hochgradig komplexe Aufgabe handelt, die nicht vollautomatisch mit Mitteln des maschinellen Lernens gelöst werden kann. Wir glauben an eine teilautomatisierte Lösung des Problems. Maschinelle Lernverfahren können Community Manager*innen und Moderator*innen bei ihrer Entscheidung unterstützen

und den Aufwand der Moderation erheblich reduzieren. Der Mensch muss aber immer Teil des gesamten Prozesses bleiben und ggfs. manuell eingreifen. Im Folgenden wird beschrieben, welche maschinellen Verfahren unterstützend bei der Identifikation von Hasskommentaren eingesetzt werden können.

Die Erstellung umfassender sowie performanterer textueller Klassifikationsmodelle im Kontext des Natural Language Processing ist zeitintensiv und erfordert, wie bereits erwähnt, eine große Zahl von Datenpunkten. Ein weitverbreiteter und vielversprechender Ansatz zur Reduktion des verbundenen Aufwandes zum Trainieren der Modelle ist das sog. „Transfer Learning“, also die Nutzung von bereits vortrainierten Modellen für eine neue Aufgabe (in diesem Fall die Detektion von Hassrede) (Ruderet al., 2019). Üblicherweise wird in diesem Zusammenhang eine Repräsentation der natürlichen Sprache, basierend auf großen Textkorpora, beispielsweise Wikipedia, erzeugt. Dabei wird für jedes Wort oder jeden Satz eine feststehende Vektordarstellung berechnet. Der entstandene Vektorraum zeichnet sich dadurch aus, dass sich Dokumente in ähnlichen Kontexten näher zueinander befinden als unzusammenhängende Dokumente (Bojanowski et al., 2016; Mikolov et al., 2013; Pennington et al., 2014). Basierend auf diesen universellen Repräsentationen wird anschließend das eigentliche Problem – der sog. Downstream Task – gelöst. Dies kann auf unterschiedliche Weise geschehen. Einerseits kann ein von Grund auf neues Klassifikations- oder Regressionsmodell mit den Modell-Repräsentationen der vorliegenden Daten trainiert werden, oder aber das vortrainierte Modell wird selber mit dem gegebenen Datensatz aktualisiert. In den letzten Jahren wurden einige neue und innovative Ansätze der wissenschaftlichen Gemeinschaft vorgestellt. Diese unterscheiden sich sowohl in der Art der Repräsentation und der Modellarchitektur als auch der zugrunde liegenden Sprache (Brown et al., 2020; Devlin et al., 2019; Eisenschlos et al., 2019; Howard & Ruder, 2018; Radford et al., 2019). Es wird hervorgehoben, dass keines der Modelle in allen Aufgabenbereichen dominiert und daher individuell, je nach Problem, der geeignetste Ansatz identifiziert werden muss. Zudem zeichnet sich die Forschungsdomäne zurzeit durch eine hohe Anzahl an neuen Veröffentlichungen aus. Neue Methoden und Ansätze werden in kurzen zeitlichen Abständen publiziert und erzielen immer bessere Ergebnisse. Dementsprechend ist für die Konzeption einer semi-automatisierten Moderationsplattform wichtig, dass Modelle schnell und leicht ausgetauscht werden können, um eine zeitgerechte Leistung zu ermöglichen. Abschließend sei noch einmal deutlich hervorgehoben, dass die Qualität der Modelle nicht nur von der Qualität der Algorithmen und Modellarchitekturen abhängt, sondern auch essenziell von der Qualität der gegebenen Annotationen. Falsch annotierte Daten führen zu einer Verzerrung

der Grundgesamtheit, ergo zu einer falschen Verteilungsannahme. Dementsprechend ist nicht nur das maschinelle Modell, sondern auch die zugrunde liegende Datenbasis ein Indikator für eine zuverlässige Erkennung von Hasskommentaren.

5 Geschäftsmodell

Um die oben geschilderten Herausforderungen zu adressieren und Plattformbetreiber*innen Werkzeuge an die Hand zu geben, um das Kommentaraufkommen mit deutlich geringerem Aufwand moderieren zu können, hat sich das Institut für Wirtschaftsinformatik der Universität Münster mit der Rheinischen Post Mediengruppe zusammengeschlossen. Gemeinsam bearbeiten beide das EU- und EFRE.NRW-geförderte Forschungsprojekt „Reduzierung des Moderationsaufwandes von Nutzer- Kommentaren mithilfe von Automatisierung durch textanalytische Methoden“ (MODERAT!)³. Das Projekt MODERAT! entwickelt durch einen interaktiven und interdisziplinären Ansatz Software-Werkzeuge und eine praxisorientierte Web-Plattform.

Abb. 5 zeigt das konzeptionelle Geschäftsmodell, welches der Umsetzung der prototypischen Web-Plattform zugrunde liegt – der Einfachheit halber umgesetzt im bekannten Schema des Business-Model-Canvas (BMC) von Osterwalder und Pigneur (2010). Kernpunkt des Geschäftsmodells ist, dass die Bewertung von Kund*innenkommentaren als Analytics-as-a-Service (AaaS) über API-Endpunkte angeboten wird. Dies basiert auf der Feststellung, dass viele bestehende Systeme und Forschungsansätze – auch größerer Unternehmen und Konzerne – oftmals nur Teile der oben genannten Herausforderungen adressieren (Brunk et al., 2019). Durch die Bündelung von Ressourcen (u. a. personeller, finanzieller, aber auch datentechnischer Natur) soll die AaaS-Plattform ggü. Individuallösungen einen kompetitiven Vorteil erhalten (Brunk et al., 2019; Naous et al., 2017). Neben einer reinen API-Anbindung soll der AaaS zusätzlich als Web-Plattform in den Moderationsworkflow der Plattformbetreiber*innen eingebunden werden können, sodass die Moderator*innen diese als Self-Service Plattform nutzen können. Das Hauptnutzer*innenversprechen ist ein geringerer Moderationsaufwand (d. h. niedrigere Kosten) und gleichzeitig ein besserer Austausch mit den Nutzer*innen des Diskussionsbereichs. Eine detaillierte Beschreibung des Geschäftsmodells sowie die damit einhergehenden Herausforderungen können in der Publikation des Modells nachgelesen werden (Brunk et al., 2019).

³ <https://www.moderat.nrw/>










Schlüssel-partner  <ul style="list-style-type: none">› Dienstleister für IT-Infrastruktur, z.B. Server Hardware inkl. CPU und GPU› Hosting-Dienstleister, z.B. Versorgung mit Bandbreite und Netzwerk› Kooperationspartner der Medienindustrie, die Kommentardaten zur Verfügung stellen	Schlüsselaktivitäten  <ul style="list-style-type: none">› Forschung und Entwicklung› Modellentwicklung und -training› API-Bereitstellung› Bewertung neuer Daten Schlüsselressourcen  <ul style="list-style-type: none">› Open Source Framework des maschinellen Lernens, z.B. Keras oder Tensorflow› Bewertete Datensätze	Nutzenversprechen  <ul style="list-style-type: none">› Moderationsprozess folgt klarer Struktur und ist gut dokumentiert› Manueller Aufwand wird reduziert, da Kommentare vorbewertet und optional auch vorgefiltert werden› stärkerer Austausch mit Besuchern, da Kommentarsektionen nicht geschlossen sind› Hassfreie Kommentarsektionen attraktiver für Werbetreibende	Kundenbeziehungen  <ul style="list-style-type: none">› Plattform wird als Self-Service den Moderatoren angeboten› Modelle werden kundenspezifisch angepasst und trainiert Marketingkanäle  <ul style="list-style-type: none">› digitale Kommunikation und Datenaustausch via API› Datenaustausch via Push- oder Pull-Prinzip	Kunden-segmente  <ul style="list-style-type: none">› verschiedene Kundensegmente› Kleine Kunden, nur einzelne Bewertungsanfragen je nach Bedarf, pay-by-use› Mittelgroße Kunden, integrierter Moderationsworkflow via API, Paketangebot je nach Anzahl von Bewertungen› Großkunden, unlimitierte Bewertungsanfragen, Festpreis
Kostenstruktur  <ul style="list-style-type: none">› Miete von Hardware und Hosting› Forschungs- und Entwicklungskosten› Aufwand für Beratung und Modellanpassungen für Kunden		Einnahmequellen  <ul style="list-style-type: none">› Einnahmen durch Abonnement-Pläne› Einnahmen durch Beratung und Modellanpassung		

Abb. 5 Geschäftsmodell auf Basis des Business Model Canvas (vgl. Brunk et al., 2019)

6 Analytics-as-a-Service-Plattform

Während das zuvor genannte Geschäftsmodell ein abstraktes Konzept darstellt, befasst sich dieses Kapitel mit einer konkreten Instanziierung dieser Geschäftsmodellidee.⁴ Diese wird aktuell im Rahmen des Forschungsprojektes MODERAT! entwickelt. Um das geschilderte Problem der (semi-)automatischen Hasserkennung unter der Maßgabe beschränkter Ressourcen aufseiten der Medienhäuser zu lösen, benötigt es einer ausgeklügelten IT-Architektur.

Zur Umsetzung entsprechender Architekturen bedient man sich i. d. R. einer sogenannten Mikroservice-Architektur (siehe Abb. 6). Anstelle einer großen, komplexen, allumfassenden Software wird eine Vielzahl kleinerer, spezialisierter Komponenten entworfen, welche dann verknüpft zur Bewältigung der Aufgabe genutzt werden. Dies birgt eine Reihe inhärenter Vorteile, u. a. Erweiterbarkeit, Konfigurierbarkeit und Skalierbarkeit. Insbesondere der letzte Aspekt ist für eine AaaS-Plattform mit einer Vielzahl von potenziellen Kund*innen von zentraler

⁴ Entsprechend sei angemerkt, dass die exakte Ausführung einer AaaS-Plattform abweichen kann und die vorgestellte Lösung nur einen möglichen Lösungsweg darstellt.

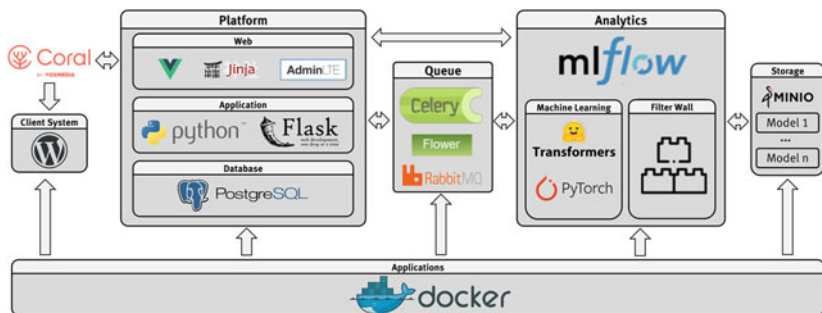


Abb. 6 AaaS-Architektur der MODERAT! Comment.AI Plattform. (Angelehnt an Riehle et al., 2020)

Bedeutung: Wird eine Zahl von n Kund*innen angebunden, die über typischerweise ein $m > 1$ an Kommentarsystemen verfügen⁵, ergeben sich schnell eine Vielzahl von anzubindenden Systemen und ein enormer kontinuierlicher Datenstrom. Zur Umsetzung nutzen wir u. a. die Container-Virtualisierung docker, welche sich zu einem der De-Facto-Standards in diesem Bereich entwickelt hat. Darüber hinaus kann docker u. a. mit Orchestrations-Tools verknüpft werden, um angepasst an die jeweilige Lastsituation den AaaS-Service passend skalierend zu können. Eine gängige Option in diesem Bereich ist Kubernetes. Dieser Lösungsansatz erlaubt es, die Comment. AI-Plattform schnell und unkompliziert an verschiedene Lastsituationen anzupassen und somit kontinuierlich einen optimalen Trade-Off zwischen Performance und verbrauchten Ressourcen sicherzustellen.

Beim Aufrufen der Comment.AI-Lösung wird der/die Journalist*in und/oder Community-Manager*in auf ein modernes, möglichst schlicht gehaltenes Moderationsdashboard geleitet und sieht damit wenig von der unterliegenden Servicestruktur (siehe Abb. 7). Dieses präsentiert die zu moderierenden Kommentare inkl. der Ergebnisse der analytischen Bewertung chronologisch mit diversen Filteroptionen. Um eine möglichst vertraute Benutzer*innenoberfläche bereitzustellen, wird das moderne Javascript Frontend Vue.js zusammen mit dem Bootstrap-Template AdminLTE verwendet. Als Web-Framework wird das Python-basierte Micro-Framework Flask verwendet. Neben der guten Erweiterbarkeit und

⁵ Selbst bei 20–30 Medienhäusern mit jeweils bis zu 5 Systemen (bspw. eigene Kommentarspalte, Forum, Facebook,...) ergeben sich so schnell 100–150 paralleler Anbindungen.

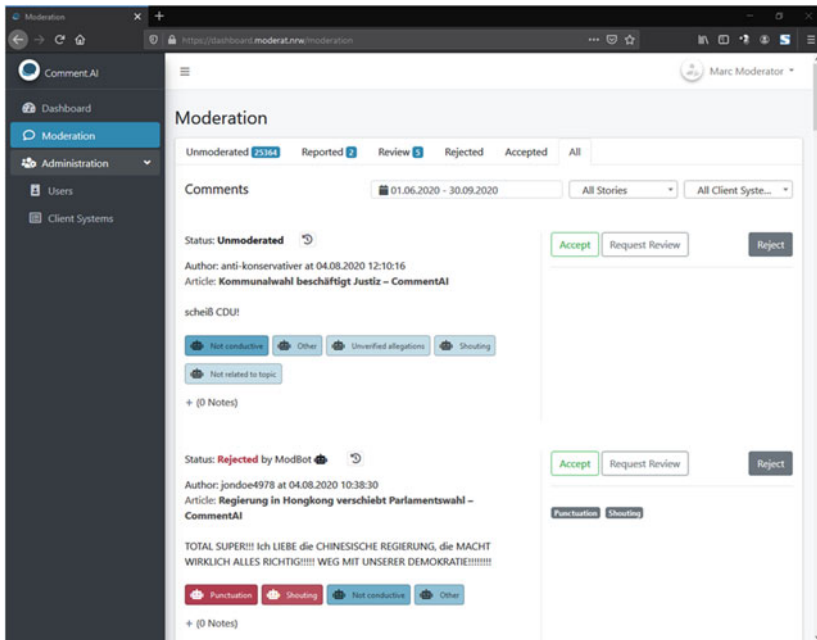


Abb. 7 Moderations-Dashboard der MODERAT! Comment.AI Plattform

der Vielzahl an bereits verfügbaren Erweiterungen war insbesondere die Einfachheit (Grinberg, 2018) und der Aufbau auf Python entscheidend für die Auswahl. Da auch die unten vorgestellte Analytics-Komponente überwiegend auf Python aufbaut (da es im Bereich Machine Learning de facto die Standard-Programmiersprache darstellt), erhöht dies die Einheitlichkeit und Wartbarkeit der Gesamtplattform (und damit die Verfügbarkeit bei gleichzeitig niedrigen Kosten). Für das Persistieren der Daten wird mit PostgreSQL eine der Standard-SQL-Datenbanken verwendet.

Im Bereich der Analytik bedient sich der Comment.AI-Prototyp zweierlei Verfahren: Zum einen werden statische Verfahren verwendet, welche basierend auf einfachen Regelwerken, bspw. Telefonnummern, Namen oder auch abweichende Sprachen (wie Englisch, Russisch, Arabisch,...), detektieren. Dies ist ohne die Verwendung maschinellen Lernens möglich und spart somit Ressourcen bei Erstellung und Verwendung. Der zweite und zentrale analytische Teil sind die Machine Learning-Verfahren. Hier werden aktuell primär sogenannte

Transformer-basierte Verfahren, wie Google's BERT verwendet, welche zur Zeit des Schreibens dieses Beitrags als „State of the Art“-Verfahren gelten (Devlin et al., 2019). Als Backend wird PyTorch verwendet. Da Sprachen typischerweise „lebendige“ Konstrukte sind und da die avisierte AaaS-Plattform eine Vielzahl verschiedener Kund*innen mit verschiedenen Modell-Anforderungen berücksichtigen können soll, wird zusätzlich eine Lösung zur Verwaltung der erzeugten Modelle benötigt. Für diese Aufgabe wird MLflow verwendet, welches sowohl als zentrales Repository aller erstellten Modelle dienen wird als auch als Versionsverwaltung, um bspw. auf Änderungen im Sprachgebrauch reagieren zu können, ohne die Nachvollziehbarkeit voriger Entscheidungen zu gefährden (Zaharia et al., 2018; A. Chen et al., 2020). Die tatsächliche Speicherung der Modelle kann auf einem beliebigen Cloud- oder on-premise-Speicher erfolgen; in der vorgestellten Instanziierung wird hierfür MinIO on-premise genutzt. Von der oben beschriebenen Komplexität sind jedoch die Hauptnutzer*innen von Comment. AI – die Journalist*innen und Community Manager*innen – vollständig entkoppelt, da diese im Hintergrund durch den Plattform-Betreiber/die Plattformbetreiberin betreut wird.

Die Verbindung zwischen der eigentlichen Moderationsplattform und dem Analytics-Backend wird über eine sogenannte Task-Queue hergestellt – in diesem Falle Celery mit dem Task-Broker RabbitMQ. Auch hier ist der Hintergrund primär die Skalierbarkeit der Plattform: Je nach Art der gerade zu kommentierenden Artikel kann die Zahl der eingehenden Kommentare stark variieren. Entsprechend wäre ein sequenzielles Abarbeiten über die Plattform nicht möglich. Über das Zuführen der Kommentare in eine Queue wird der Kommentar in eine Art Warteschlange eingereiht und kann von einer flexibel skalierbaren Anzahl an sogenannten Workern aufgegriffen und bearbeitet werden. Dies erlaubt es aus Sicht der Moderationsplattform, die Kommentarbewertung als „Black-Box“ zu betrachten, in welcher die Anzahl der Worker flexibel (und automatisch) den jeweiligen Gegebenheiten angepasst werden kann.

7 Ausblick

Der Diskurs und öffentliche Debatten sind insbesondere in Demokratien ein hohes Gut. Das Internet als dezentrales Medium erlaubt einer stetig wachsenden Zahl an Menschen, an diesem Diskurs zu partizipieren. Dies erhöht zwar die Vielfalt und Interaktivität entsprechender Debatten – allerdings steigt parallel auch die Zahl und Frequenz problematischer, wenn nicht gar hasserfüllter Beiträge.

Journalist*innen und auch Medien bekennen sich zu ihrer Aufgabe und Verantwortung, die öffentliche Meinungsbildung zu informieren, zu begleiten und zu moderieren – stehen jedoch auch vor enormen Herausforderungen, dies rechtlich und finanziell zu bewerkstelligen. An dieser Stelle setzen Praxis und Wissenschaft vermehrt an, um (teil-)automatisierte Lösungen zur Reduzierung des Problems zu erstellen.

Am Ende des Tages wird der Hass vermutlich weder aus der Gesellschaft noch aus dem Internet noch aus den dortigen Kommentarspalten verschwinden. Dennoch ist es wichtig sicherzustellen, dass vereinzelte, laute und problematische Stimmen nicht den öffentlichen Diskurs verstummen lassen können. Weder die Informatik noch die Wirtschaftsinformatik können die unterliegenden gesellschaftlichen Probleme lösen. Sie können jedoch den Journalist*innen und Moderator*innen Werkzeuge an die Hand geben, welche diesen helfen, auch in Zukunft den öffentlichen Diskurs zu informieren und zu begleiten. Allerdings ist auch davon auszugehen, dass dieses Problem angesichts der Dynamik von Sprache, Technologie und Kontext⁶ nur in einem dauerhaft durchzuführenden Anpassungsprozess angegangen werden kann und einzelne Projekte nur einen temporären Beitrag liefern können.⁷

Vor diesem Hintergrund gibt dieser Buchbeitrag einen kurzen Überblick über zentrale Probleme, denen die Arbeit an diesem Themenkomplex gegenübersteht. Insbesondere die Kategorisierung von Kommentaren ist kein primär technisches Problem – hat jedoch ebenso wie die Auswahl und Konfiguration der entsprechenden Lernalgorithmen einen erheblichen Einfluss auf die Qualität der gewonnenen Analysen und Entscheidungsunterstützung. Die Forschungsergebnisse der letzten 10 Jahre zeigen einen positiven Trend, was die Erkennungsrate und auch den Umgang mit sprachlichen Stilmitteln (*Ironie, Abkürzungen, bewusste Wortverfälschungen*) angeht. Nichtsdestotrotz gibt es noch kein dominierendes Design oder abschließende Erkenntnisse. Darüber hinaus soll dieser Beitrag auch dafür

⁶ Erste wissenschaftliche Erkenntnisse, wie auch die praktischen Erfahrungen der Community Manager zeigen, dass die Sprache nicht nur natürlichen Änderungen unterliegt, sondern oftmals bewusst nach Wegen gesucht wird, maschinelle Moderation auszutricksen (Rojas-Galeano, 2017). Die bewusste, für Menschen einfach nachzuvollziehende, aber technisch schwer zu detektierende Änderung von Elementen ist auch aus anderen Bereichen wie der Bilderkennung bekannt (Tabacof & Valle, 2016).

⁷ Zusätzliche Probleme ergeben sich in all jenen Situationen, in denen neben rein textuellen Beiträgen auch Bild Darstellungen, Ton- und/oder Videosequenzen beigefügt werden können, welche zusätzlich problematisches Material beinhalten können (Gagliardone et al., 2015; Gomez et al., 2020; Mathew et al., 2019).

sensibilisieren, dass eine maschinelle Unterstützung zwar möglich, aber aufgrund der inhärenten Komplexität auch schwierig umzusetzen, zu warten und auch zu bezahlen ist – insbesondere für das Gros der kleineren und mittelständischen Medienhäuser. Ein möglicher Ausweg ist die Zuwendung zur Plattform-Ökonomie und zur Ausgestaltung einer Analytics- as-a-Service-Lösung wie im Beitrag vorgeschlagen. Dies ermöglicht die Bündelung knapper und teurer personeller und technischer Ressourcen, um qualitativ hochwertige ML-Modelle bereitstellen zu können, ohne die Notwendigkeit prohibitiv teurer Einzelprojekte.

Allerdings wird es auch über den aktuellen Forschungs- und Umsetzungsstand hinaus in absehbarer Zukunft Handlungsbedarf geben: In der EU (High-Level Expert Group on Artificial Intelligence, 2019), aber auch weltweit (Garfinkel et al., 2017; Laaksonen et al., 2020) setzt sich zunehmend die Erkenntnis durch, dass zur Sicherstellung von Vertrauenswürdigkeit und Akzeptanz von ML-Systemen die algorithmische Transparenz erhöht werden muss. Allerdings sind gerade die viel verwendeten neuronalen Netzwerke sogenannte „Black-Box“-Modelle, welche zwar i. d. R. gute Erkennungsraten liefern, deren Entscheidungsfindungsprozesse jedoch nicht einfach nachgehalten werden können. Somit besteht an dieser Stelle noch Forschungs- und Umsetzungsbedarf, um die erzielten Erkenntnisse sowohl Community Manager*innen und Journalist*innen als auch den Kommentator*innen ggü. besser aufbereiten zu können.

Einhergehend mit einer erhöhten technischen Transparenz muss zudem sichergestellt werden, dass auch aufseiten der (kommentierenden) Bevölkerung Sensibilisierungs- und Aufklärungsarbeit geleistet wird. Denn Unkenntnis führt schnell zu Unbehagen und Ablehnung entsprechender Ansätze.⁸ Damit droht den Erstumsetzer*innen ein Ausbleiben der Kommentator*innen, während sich die problematischen Kommentare verschieben – ggfs. zusätzlich angeheizt durch den algorithmischen Eingriff. Die Öffentlichkeitsarbeit im Rahmen von Forschungsprojekten wie dem No-Hate-Projekt oder auch MODERAT! sind zwar ein erster Ansatz; dennoch bedarf es entsprechender Aufklärungsarbeit auch auf der Bildungsebene sowie in den Massenmedien, um das Gros der Bevölkerung angemessen zu erreichen.

⁸ Das studentische Projekt [hateminig.de](https://www.hateminig.de), welches primär einen Forschungsüberblick über die Erkennung von Hass im Jahr 2016 geben sollte, war auch Ziel solcher Kritik. Ein Beispiel ist bspw. der Tweet des Users @citoyen_lauris, welcher in den vorgestellten Projektergebnissen Ansätze zur Gesinnungskontrolle von Onlinekommentator*innen verortete (siehe https://twitter.com/citoyen_lauris/status/765928352188334081).

Förderung

Die Forschung, die zu diesen Ergebnissen führte, wurde vom Land Nordrhein-Westfalen und vom Europäischen Fonds für regionale Entwicklung gefördert (EFRE.NRW 2014–2020), Projekt: MODERAT! bzw. gegen MODERAT!-Logo austauschen (moderat_logo.eps) (No. CM-2-2-036a).

Literatur

- Aken, B. van, Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Hrsg.), *Proceedings of the second workshop on abusive language online* (S. 33–42). ALW2. Association for Computational Linguistics.
- Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In M. Silberstein, F. Atigui, E. Kornysheva, E. Métais, & F. Meziane (Hrsg.), *Proceedings of the 23rd international conference on applications of natural language to information systems* (S. 57–64). NLDB 2018. Springer.
- Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on world wide web companion* (S. 759–760). WWW'17 Companion. International World Wide Web Conferences Steering Committee.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching word vectors with subword information*. arXiv: [1607.04606](https://arxiv.org/abs/1607.04606) [cs.CL].
- Bretschneider, U., Wöhner, T., & Peters, R. (2014). Detecting online harassment in social networks. In M. D. Myers & D. W. Straub (Hrsg.), *Proceedings of the international conference on information systems – Building a better world through information systems* (S. 1–14). ICIS 2014. Association for Information Systems.
- Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, 36(4), 419–468.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). *Language models are few-shot learners*. arXiv: [2005.14165](https://arxiv.org/abs/2005.14165) [cs.CL].
- Brunk, J., Niemann, M., & Riehle, D. M. (2019). Can analytics as a service save the online discussion culture? – The case of comment moderation in the media industry. In *Proceedings of the 21st IEEE conference on business informatics* (S. 472–481). CBI 2019. IEEE.
- Bugueño, M., & Mendoza, M. (2020). Learning to detect online harassment on Twitter with the transformer. In P. Cellier & K. Driessens (Hrsg.), *Proceedings of the international workshops of ECML PKDD 2019* (S. 298–306). ECML PKDD 2019. Springer.
- Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 223–242.

- Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017a). Hate is not binary: Studying abusive behavior of #GamerGate on Twitter. In *Proceedings of the 28th ACM conference on hypertext and social media* (S. 65–74). HT'17. ACM.
- Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017b). Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on web science conference* (S. 13–22). WebSci'17. ACM.
- Chen, A., Chow, A., Davidson, A., Dcunha, A., Ghodsi, A., Hong, S. A., Konwinski, A., et al. (2020). Developments in MLflow: A system to accelerate the machine learning lifecycle. In *Proceedings of the 4th workshop on data management for end-to-end machine learning* (S. 1–4). DEEM 2020. ACM. <https://doi.org/10.1145/3399579.3399867>.
- Chen, Y., Zhou, Y., Zhu S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE international conference on social computing and 2012 ASE/IEEE international conference on privacy, security, risk and trust* (S. 71–80). SOCIALCOM-PASSAT'12. IEEE.
- Cheng, J. (2007). *Report: 80 percent of blogs contain offensive content*. <https://arstechnica.com/information-technology/2007/04/report-80-percent-of-blogs-contain-offensive-content/>.
- Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40, 108–118.
- Dansby, R., Fang, H., Ma, H., Moghbel, C., Ozerem, U., Peng, X., Stoyanov, V., Wang, S., Yang, F., & Zhang, K. (2020). *AI advances to better detect hate speech*. <https://ai.faciebook.com/blog/ai-advances-to-better-detect-hate-speech/>.
- Davidson, T., Warmusley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international conference on web and social media* (S. 512–515). ICWSM-2017. AAAI.
- Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In A. Armando, R. Baldoni, & R. Focardi (Hrsg.), *Proceedings of the first Italian conference on cybersecurity* (S. 86–95). ITASEC17. CEUR-WS.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Hrsg.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (S. 4171–4186). NAACL-HLT 2019. ACL.
- Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *The social mobile web, papers from the 2011 ICWSM workshop* (S. 11–17). ICWSM 2011. Association for the Advancement of Artificial Intelligence.
- Eisenschlos, J., Ruder, S., Czaplá, P., Kadras, M., Gugger, S., & Howard, J. (2019). MultiFiT: Efficient multilingual language model fine-tuning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (S. 5702–5707). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1572>.
- Fišer, D., Erjavec T., & Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In Z. Waseem, W. Hui

- Kyong Chung, D. Hovy, & J. Tetreault (Hrsg.), *Proceedings of the first workshop on abusive language online* (S. 46–51). ALW1. Association for Computational Linguistics.
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), 1–30.
- Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the twelfth international AAAI conference on web and social media* (S. 491–500). ICWSM 2018. AAAI.
- Gagliardone, I., Gal, D., Alves T., & Martinez, G. (2015). *Countering online hate speech* (73. Aufl.). UNESCO.
- Ganz, K. (2019). Hate Speech im Internet. In J. Dorer, B. Geiger, B. Hipfl, & V. Ratković (Hrsg.), *Handbuch Medien und Geschlecht: Perspektiven und Befunde der feministischen Kommunikations- und Medienforschung* (S. 1–10). Springer VS.
- Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter D., & Ulmanu, M. (2016). *The dark side of guardian comments*. <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments>.
- Garfinkel, S., Matthews, J., Shapiro, S. S., & Smith, J. M. (2017). Toward algorithmic transparency and accountability. *Communications of the ACM*, 60(9), 5. ISSN: 15577317. <https://doi.org/10.1145/3125780>.
- Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang R., & Huang, J. (2020). Garbage in, garbage out? In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (S. 325–336). FAT*’20. ACM. <https://doi.org/10.1145/3351095.3372862>.
- Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (S. 459–1467). WACV 2020. IEEE. ISBN: 978-1-7281-6553-0. <https://doi.org/10.1109/WACV45572.2020.9093414>.
- Grinberg, M. (2018). *Flask web development: developing web applications with python* (2. Aufl., S. 314). O’Reilly Media, Inc. ISBN: 978-1-491-99173-2.
- Heinonen, A. (2011). The journalist’s relationship with users. In J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, Z. Reich, & M. Vujnovic (Hrsg.), *Participatory journalism: Guarding open gates at online newspapers* (1. Aufl., S. 34–55). Wiley-Blackwell. <https://doi.org/10.1002/9781444340747.ch3>.
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI. Technischer Bericht*. European Commission.
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (S. 328–339). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1031>.
- Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*.
- Huang, C. L. (2016). *The 2016 global report on online commenting, chapter 1: The problem with comments*. <https://blog.wan-ifra.org/2016/10/17/the-2016-global-report-on-online-commenting-chapter-1-the-problem-with-comments>.
- Jorgensen, M. & Choi, M. (2019). Abusive language detection using auto-machine learning for multiple languages. *Veritas: Villanova Research Journal*, 1, 3–4.

- Jorgensen, M., Choi, M., Niemann, M., Brunk, J., & Becker, J. (2020). Multi-class detection of abusive language using automated machine learning. In *Proceedings of the 15th international conference on wirtschafts-informatik* (S. 1763–1775). WI 2020.
- Jurgens, D. (2013). Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of NAACL-HLT 2013* (S. 556–562). NAACL-HLT 2013. ACL.
- Jurgens, D., Chandrasekharan, E., & Hemphill, L. (2019). A just and comprehensive strategy for using NIP to address online abuse. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (S. 3658–3666). ACL 2019. ACL.
- Karadeniz, B. (2009). *Blinder Rassismus auf Websites von Tageszeitungen*. <https://blog.netplanet.org/2009/11/08/blinder-rassismus-auf-websites-von-tageszeitungen/>.
- Köffer, S., Riehle, D. M., Höhenberger, S., & Becker, J. (2018). Discussing the value of automatic hate speech detection in online debates. *Proceedings der Multikonferenz Wirtschaftsinformatik Lüneburg, Deutschland* (S. 83–94).
- Kolhatkar, V., & Taboada, M. (2017). Constructive language in news comments. In Z. Waseem, W. Hui Kyong Chung, D. Hovy, & J. Tetreault (Hrsg.), *Proceedings of the first workshop on abusive language online* (S. 11–17). ALW1. Association for Computational Linguistics.
- Laaksonen, S. M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020). The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data*, 3, 1–16. <https://doi.org/10.3389/fdata.2020.00003>.
- Lachenicht, L. G. (1980). Aggravating language a study of abusive and insulting language. *Paper in Linguistics*, 13(4), 607–687.
- Lee, Y., Yoon, S., & Jung, K. (2018). Comparative studies of detecting abusive language on Twitter. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Hrsg.), *Proceedings of the second workshop on abusive language online* (S. 101–106). ALW2. Association for Computational Linguistics.
- Lewis, S. C., Holton A. E., & Coddington, M. (2014). Reciprocal journalism: A concept of mutual exchange between journalists and audiences. *Journalism Practice*, 8(2), 229–241. <https://doi.org/10.1080/17512786.2013.859840>.
- Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science* (S. 173–182). WebSci'19. ACM. ISBN: 9781450362023. <https://doi.org/10.1145/3292522.3326034>.
- Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhanian, P., Kalyan Maity, S., Goyal, P., & Mukherjee, A. (2018). Thou shalt not hate: Countering online hate speech. In *Proceedings of the thirteenth international AAAI conference on web and social media* (S. 369–380). ICWSM 2019. AAAI.
- Mathur, P., Sawhney, R., Ayyar, M., & Shah, R. R. (2018). Did you offend me? Classification of offensive tweets in Hinglish language. In *Proceedings of the second workshop on abusive language online* (S. 138–148). ALW2. Association for Computational Linguistics.
- Mehdad, Y., & Tetreault, Y. (2016). Do characters abuse more than words? In R. Fernandes, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, & A. Gainer (Hrsg.), *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue* (S. 299–303). SIGDIAL 2016. Association for Computational Linguistics.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M.

- Welling, Z. Ghahramani, & K. Q. Weinberger (Hrsg.), *Advances in neural information processing systems* 26 (S. 3111–3119). Curran Associates, Inc. <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.
- Mondal, M., Araújo Silva, L., & Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media* (S. 85–94). HT 2017. ACM.
- Naous, D., Schwarz, J., & Legner, C. (2017). Analytics as a service: Cloud computing and the transformation of business analytics business models and ecosystems. In *Proceedings of the 25th European conference on information systems* (S. 487–501). ECIS 2017. AIS.
- Newton, C. (2019). *The trauma floor: The secret lives of Facebook moderators*. In *America*. <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>.
- Niemann, M. (2019). Abusiveness is non-binary: Five shades of gray in german online news-comments. In *Proceedings of the 21st IEEE conference on business informatics* (S. 11–20). CBI 2019. IEEE.
- Niemann, M., Riehle, D. M., Brunk, J., & Becker, J. (2020). What is abusive language? Integrating different views on abusive language for machine learning. In *Proceedings of the 1st multidisciplinary international symposium on disinformation in open online media* (S. 59–73). MISDOOM 2019. Springer.
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (S. 145–153). WWW'16. ACM.
- Osterwalder, A., & Pigneur, Y. (2010). *Business model generation* (281). Wiley.
- Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2), 259–283. <https://doi.org/10.1177/1461444804041444>.
- Parliamentary Assembly. (2007). *Recommendation 1805 (2007): Blasphemy, religious insults and hate speech against persons on grounds of their religion*. <http://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=17569%7B%5C&%7Dlang=en>.
- Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). Deep learning for user comment moderation. In Z. Waseem, W. Hui Kyong Chung, D. Hovy, & J. Tetreault (Hrsg.), *Proceedings of the first workshop on abusive language online* (S. 25–35). ALW1. Association for Computational Linguistics.
- Pavlopoulos, J., Thain, N., Dixon, L., & Androutsopoulos, I. (2019). ConvAI at SemEval-2019 Task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th international workshop on semantic evaluation* (S. 571–576). SemEval 2019. ACL.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (S. 1532–1543). <http://www.aclweb.org/anthology/D14-1162>.
- Phillips, D. J. (1996). Defending the boundaries: Identifying and countering threats in a usenet newsgroup. *The Information Society*, 12(1), 39–62.
- Pöyhtäri, R. (2014). Limits of hate speech and freedom of speech on moderated news websites in Finland, Sweden, the Netherlands and the UK. *ANNALES Ser. hist. Sociol*, 24(3), 513–524.

- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.
- Rajamanickam, S., Mishra, P., Yannakoudakis, H., & Shutova, E. (2020). Joint modelling of emotion and abusive language detection. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Hrsg.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (S. 4270–4279). ACL 2020. ACL. <https://doi.org/10.18653/v1/2020.acl-main.394>.
- Reich, Z. (2011). User comments. In J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, Z. Reich, & M. Vujnovic (Hrsg.), *Participatory journalism: Guarding open gates at online newspapers* (S. 96–117). Wiley-Blackwell. <https://doi.org/10.1002/9781444340747.ch6>
- Riehle, D. M., Niemann, M., Brunk, J., Assenmacher, D., Trautmann, H., & Becker, J. (2020). Building an integrated comment moderation system – Towards a semi-automatic moderation tool. In *Proceedings of the HCI international 2020*, Kopenhagen, Dänemark.
- Risch, J., Stoll, A., Ziegele, M., & Krestel, R. (2019). hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model. In *Proceedings of the 15th conference on natural language processing* (S. 403–408). KONVENS 2019.
- Rojas-Galeano, S. (2017). On obstructing obscenity obfuscation. *ACM Transactions on the Web*, 11(2), 1–24. <https://doi.org/10.1145/3032963>.
- Rosen, G. (2019). *Community standards enforcement report, November 2019 Edition*. <https://about.fb.com/news/2019/11/community-standards-enforcement-report-nov-2019/>.
- Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Hrsg.), *Proceedings of the 3rd workshop on natural language processing for computer-mediated communication* (S. 6–9). NLP4CMC III. Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum.
- Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials* (S. 15–18). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-5004>.
- Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S., Almerikhi, H., & Jansen, B. H. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1), 1–34.
- Serrà, J., Leontiadis, I., Spathis, D., Stringhini, G., & Blackburn, J. (2017). Class-based prediction errors to categorize text with out-of-vocabulary words. In Z. Waseem, W. Hui Kyong Chung, D. Hovy, & J. Tetreault (Hrsg.), *Proceedings of the first workshop on abusive language online* (S. 36–40). ALW1. Association for Computational Linguistics.
- Siegert, S. (2016). *Nahezu jede zweite Zeitungsredaktion schränkt Online-Kommentare ein*. <http://www.journalist.de/aktuelles/meldungen/journalist-umfrage-nahezu-jede-2-zeitungsredaktion-schraenkt-onlinekommentare-ein.html>.
- Sood, S. O., Churchill, E. F., & Antin, J. (2012). Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2), 270–285.
- Švec, A., Pikuliak, M., Šimko, M., & Bieliková, M. (2018). Improving moderation of online discussions via interpretable neural models. In D. Fišer, R. Huang, V. Prabhakaran, R.

- Voigt, Z. Waseem, & J. Wernimont (Hrsg.), *Proceedings of the second workshop on abusive language online* (S. 60–65). ALW2. Association for Computational Linguistics.
- Tabacof, P., & Valle, E. (2016). Exploring the space of adversarial images. In *Proceedings of the 2016 international joint conference on neural networks* (S. 426–433). IJCNN 2016. IEEE. <https://doi.org/10.1109/IJCNN.2016.7727230>.
- Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, 22(1), 69–80.
- Vogel, A. (2017). #wortgewalt(ig): Leser_innen- und Nutzer_innen-Kommentare in Medien-öffentlichkeiten. Technischer Bericht. Friedrich-Ebert-Stiftung.
- Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In S. Owsley Sood, M. Nagarajan, & M. Gamon (Hrsg.), *Proceedings of the second workshop on language in social media* (S. 19–26). LSM'12. Association for Computational Linguistics.
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop* (S. 88–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-2013>.
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.
- Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina. In *Proceedings of the 26th international conference on world wide web* (S. 1391–1399). WWW'17. ACM.
- 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* (S. 1–7). CAW2.0.
- Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., & u. a. (2018). Accelerating the machine learning lifecycle with MLflow. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 41(4), 39–45.