



universität
wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Hate Speech und Counter Speech auf TikTok

Formen und Dynamiken von Hate Speech und Counter Speech auf TikTok

verfasst von | submitted by

Liz Freimann BA BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt |
Degree programme code as it appears on the
student record sheet:

UA 066 905

Studienrichtung lt. Studienblatt | Degree pro-
gramme as it appears on the student record
sheet:

Masterstudium Soziologie

Betreut von | Supervisor:

Univ.-Prof. Mag. Dr. Ulrike Zartler-Griesl Privatdoz.

Danksagung

Die vorliegende Masterarbeit wäre ohne die Unterstützung folgender Personen nicht möglich gewesen, weshalb ich ihnen allen an dieser Stelle meinen aufrichtigen Dank aussprechen möchte.

Ein besonderer Dank gilt meiner Betreuerin Dr. Ulrike Zartler für die umsichtige Betreuung, ihr Verständnis und die gute Zusammenarbeit in dem Forschungsprojekt „Young People Against Online Hate“. Ebenso danke ich Dr. Christiane Atzmüller für die kollegiale und wertvolle Zusammenarbeit in diesem Projekt sowie für ihre Unterstützung bei der Konzeptualisierung der Masterarbeit und während des Schreibprozesses. Ich danke für die Möglichkeit, meine Masterarbeit im Rahmen dieses Projekts verfassen zu können und für das entgegengebrachte Vertrauen sowie für die fortlaufende Unterstützung.

Mein Dank gilt ebenso allen Personen, die an dem Forschungsprojekt beteiligt waren. Besonders danken möchte ich den Jugendlichen, die mir im Rahmen des Projektes Einblicke in ihre digitalen Lebenswelten gegeben haben und mit denen ich offene, ehrliche und äußerst lehrreiche Gespräche führen durfte. Ebenso danke ich den Fördergeber*innen des WWTF, der Universität Wien, der FH St. Pölten und Saferinternet.at, die dieses Projekt ermöglicht und damit diesem wichtigen Thema zu mehr Sichtbarkeit in der Wissenschaft verholfen haben.

Meinen Eltern und meiner Familie danke ich von Herzen für ihre umfassende Unterstützung, ihre Geduld und ihr Verständnis. Ein großer Dank gilt auch meinen Freund*innen, die mich aus der Nähe und aus der Ferne begleitet haben und deren Freund*innenschaft mir sehr viel bedeutet.

Ein ganz besonderer Dank gilt meinem Partner Christoph, der mich in allem unterstützt und mich in den letzten Wochen und Monaten in vielerlei Hinsicht getragen hat. Ich danke für die vielen Gespräche, Umarmungen, Mahlzeiten und für das Korrekturlesen der vorliegenden Arbeit. Tausend Dank für alles!

Inhaltsverzeichnis

1. EINLEITUNG UND FRAGESTELLUNGEN	4
2. THEORETISCHER RAHMEN	6
2.1. EINE SOZIOLOGIE DES STREITS NACH DENNAOUI UND WITTE	7
2.2. ZWISCHENMENSCHLICHE KONFLIKTE IM DIGITALEN RAUM	10
2.3. PLATTFORMEN ALS KURATIERTE RÄUME	12
2.4. BESONDERHEITEN VON TIKTOK	15
3. FORSCHUNGSSTAND	16
3.1. KARTOGRAFIE EINES FORSCHUNGSFELDES	16
3.2. FORSCHUNGSSTAND ZU HATE SPEECH	19
3.2.1. HATE SPEECH DEFINITIONEN	20
3.2.2. HATE SPEECH AUF TIKTOK	24
3.2.3. FORMEN VON HATE SPEECH	27
3.3. FORSCHUNGSSTAND ZU COUNTER SPEECH	29
3.3.1. COUNTER SPEECH DEFINITIONEN	29
3.3.2. COUNTER SPEECH AUF TIKTOK	32
3.3.3. FORMEN VON COUNTER SPEECH	33
3.4. FORSCHUNGSSTAND ZU INTERAKTIONEN VON HATE SPEECH UND COUNTER SPEECH	36
4. DATENGRUNDLAGE UND METHODISCHES VORGEHEN	37
4.1. ERSTELLUNG UND AUFBEREITUNG DES DATENSATZES	37
4.2. DATENANALYSE	40
5. ERGEBNISSE	45
5.1. HATE SPEECH	46
5.2. COUNTER SPEECH	66
5.3. INTERAKTIONEN VON HATE SPEECH UND COUNTER SPEECH	82
7. LIMITATIONEN UND AUSBLICK	88
8. LITERATURVERZEICHNIS	89
9. ABSTRACT	99
9.1. DEUTSCHES ABSTRACT	99
9.2. ENGLISCHES ABSTRACT	100

Abbildungsverzeichnis

ABBILDUNG 1: SCREENSHOTS DER RICHTLINIEN VON TIKTOK IM UMGANG MIT HASSREDE UND CYBERMOBBING (TIKTOK SICHERHEITZENTRUM 2025A; 2025B)	22
ABBILDUNG 2: TAXONOMIE ZUR HATE SPEECH KLASSIFIZIERUNG VON FILLIES ET AL. (2025, S. 7)	27
ABBILDUNG 3: ABLAUF EINER STRUKTURIERENDES QUALITATIVEN INHALTSANALYSE IN SIEBEN PHASEN NACH KUCKARTZ UND RÄDIKER (2024, S.132)	40
ABBILDUNG 4: EIGENE DARSTELLUNG DES PROZESSES DER KATEGORISIERUNG NACH KUCKATZ UND RÄDIKER (2024)	42
ABBILDUNG 5: EIGENE DARSTELLUNG DES ENTWICKELTEN KATEGORIENSYSTEMS FÜR HATE SPEECH	45
ABBILDUNG 6: EIGENE DARSTELLUNG DES ENTWICKELTEN KATEGORIENSYSTEMS FÜR COUNTER SPEECH	65

1. Einleitung und Fragestellungen

Mit der zunehmenden Digitalisierung sozialer Interaktionen verlagern sich auch Konflikte verstärkt in digitale Räume, in denen Menschen mit unterschiedlichen Hintergründen, Meinungen und Weltbildern aufeinandertreffen. In diesen Kontexten sind hasserfüllte Äußerungen und digitale Gewalt zu einem weitreichenden Problem geworden. Zwischen September 2023 und August 2024 verzeichnete die österreichische Beratungsstelle gegen Hass im Netz fast zweitausend Meldungen von Hass im Netz und über zweitausend individuelle Beratungen für Betroffene und Zeug*innen digitaler Gewalt (ZARA, 2024). Die Folgen von Hass im Netz sind weitreichend und zeigen sich sowohl auf individueller Ebene in Form von Angst, Stress, Scham, Selbstzensur oder Rückzug aus digitalen Räumen, als auch auf gesellschaftlicher Ebene, wo der Hass zu einer Radikalisierung der Öffentlichkeit, sozialer Spaltung und einer Schwächung demokratischer Prozesse führt (Khaleghipour et al., 2025). Marginalisierte Gruppen sowie Personen des öffentlichen Lebens sind dabei besonders gefährdet.

Jugendliche verbringen in Österreich durchschnittlich dreieinhalb Stunden pro Tag im Internet und auf sozialen Medien und weisen dadurch ein erhöhtes Risiko auf, hasserfüllten und negativen Inhalten ausgesetzt zu werden (Dienlin et al., 2025; Korol & Blaya, 2025). Studien zeigen, dass die bestehenden Jugendschutzkonzepte von Plattformen Jugendliche nicht ausreichend von schädlichen Inhalten schützen (Eltaher et al., 2025).

Vor diesem Hintergrund gewinnt Counter Speech als kommunikative Praxis zur Entkräftung hasserfüllter Inhalte an Bedeutung. Im Unterschied zum Löschen von Inhalten, das überwiegend in der Verantwortung der Plattformbetreibenden liegt, stellt Counter Speech eine Form zivilgesellschaftlichen Handelns dar, die grundsätzlich allen Personen offensteht, die Hass im Internet beobachten. Atzmüller et al. (2022) sprechen in diesem Zusammenhang auch von einer digitalen Zivilcourage oder einer „Zivilcourage 2.0“. Aufgrund zunehmend fragmentierter digitaler Öffentlichkeiten (Abidin, 2021; Gerbaudo, 2024) agieren Jugendliche häufig in Online-Umgebungen, zu denen Erwachsene keinen oder nur begrenzt Zugang haben. Dadurch gewinnt das Konzept Counter Speech auch für Jugendliche an Bedeutung.

Diese Überlegungen bilden den Ausgangspunkt für das Forschungsprojekt „Young People Against Online Hate“, das sich unter der Leitung von Dr. Ulrike Zartler mit computergestützten

Methoden zur Unterstützung jugendlicher Gegenrede beschäftigt. Das Projekt vereint Expertise aus Sozialwissenschaften und Informatik und bindet Jugendliche als Expert*innen ihrer digitalen Lebenswelten aktiv ein. Aufgrund der Bedeutung von TikTok für junge Nutzer*innen und des vergleichsweise hohen Anteils jugendlicher Nutzer*innen wurde TikTok als Grundlage der Untersuchungen gewählt (Universität Wien, o.D; University of Applied Sciences St. Pölten, o.D.) Die vorliegende Arbeit basiert auf Daten, die im Rahmen dieses Projektes erhoben wurden und versteht sich als Beitrag zu dem Forschungsvorhaben des Projektes.

Hate Speech und Counter Speech sind auf TikTok bislang wenig erforscht (Tontodimamma et al., 2021; Chung et al. 2024). Die meisten Erkenntnisse beruhen auf Daten von X (vormals Twitter), Facebook und YouTube. Da Plattformen jedoch jeweils eigene Strukturen und Dynamiken aufweisen (Gillespie, 2018) und Jugendliche spezifische Formen von Counter Speech anwenden (Atzmüller et al., 2022) ist eine plattformspezifische Analyse erforderlich. Die Interaktion zwischen Hate Speech und Counter Speech wird im Rahmen dieser Arbeit als eine Form zwischenmenschlicher Konfliktaustragung verstanden, im Zuge derer gesellschaftliche Normen zum Ausdruck kommen und verhandelt werden. Obwohl zwischenmenschliche Konflikte in der soziologischen Konflikttheorie lange Zeit vernachlässigt wurden, hat bereits Georg Simmel, einer der Gründerväter der Soziologie, dem Streit als eine der lebhaftesten Wechselwirkungen unter Menschen besondere Bedeutung für die Ausgestaltung von Gesellschaften zugeschrieben.

Vor diesem Hintergrund ergeben sich folgende Forschungsfragen, die dieser Arbeit zugrunde liegen und eine Annäherung an die Formen und die Interaktionsdynamiken von Hate Speech und Counter Speech auf TikTok ermöglichen sollen:

- (1) Welche Formen von Hate Speech und Counter Speech lassen sich auf TikTok identifizieren?
- (2) Welche Dynamiken zeigen sich in der Interaktion zwischen Hate Speech und Counter Speech auf TikTok?
- (3) Wie prägen plattformspezifische Merkmale von TikTok die Ausgestaltung von Hate Speech und Counter Speech?

Die Forschungsfragen werden mithilfe der qualitativen Inhaltsanalyse nach Kuckartz und Rädiker (2024) bearbeitet. Der theoretische Hintergrund der Arbeit stützt sich auf das Streitkonzept nach Dennaoui und Witte (2022), auf das Konzept der fragmentierten Öffentlichkeiten nach Abidin (2021) und Gerbaudo (2024) sowie auf das Verständnis digitaler Räume als kuratierte Orte der Konfliktaustragung, wie es Gillespie (2018) und Suler (2004) formulieren. Nach der Darstellung des theoretischen Rahmens und des Forschungsstandes sowie der Beschreibung der angewandten Methodik folgt die Präsentation der Ergebnisse, die anschließend interpretiert und im Hinblick auf die Limitationen des Forschungsvorhabens diskutiert werden.

Ziel dieser Arbeit ist es, ein vertieftes Verständnis für die Wirkung plattformspezifischer Bedingungen auf Hate Speech und Counter Speech sowie deren Interaktion zu gewinnen und ein Kategoriensystem für die Formen beider Phänomene auf TikTok zu entwickeln. Die vorliegende Arbeit versteht sich als Beitrag zum Forschungsprojekt „Young People Against Online Hate“ und unterstützt dessen Ziel, Jugendliche beim Verfassen von Gegenrede zu stärken.

2. Theoretischer Rahmen

Digitale Plattformen wie TikTok sind zum Schauplatz gesellschaftlicher Auseinandersetzungen geworden, wo Menschen mit unterschiedlichen Werten, Weltbildern und Zugehörigkeiten aufeinandertreffen. Hate Speech und Counter Speech erscheinen dort nicht nur als sprachliche Phänomene, sondern auch als Formen zwischenmenschlicher Konfliktaustragung, in denen Normen, Deutungsmacht und Anerkennung verhandelt werden. Zur theoretischen Einordnung dieser Dynamiken wird im Folgenden der konfliktsoziologische Ansatz von Dennaoui und Witte (2022) herangezogen, der zwischenmenschliche Auseinandersetzungen, auch „Streit“ genannt, als zentrales Element des menschlichen Miteinanders versteht, in dem gesellschaftliche Regeln und Normen ausgehandelt werden. Dieser Zugang erlaubt es, Hate Speech und Counter Speech als unterschiedliche Modi der Konfliktaustragung zu verstehen und ihre jeweiligen Ausprägungen genauer zu analysieren. Anschließend werden die Besonderheiten digitaler Räume mit einem Fokus auf TikTok herausgearbeitet und ihr Einfluss auf die Formen und die Dynamiken von Hate Speech und Counter Speech thematisiert.

2.1. Eine Soziologie des Streits nach Dennaoui und Witte

Dennaoui und Witte (2022) weisen darauf hin, dass Konflikte in der soziologischen Theorie zwar viel Aufmerksamkeit erhalten, der Streit als Form zwischenmenschlicher Konfliktinteraktion jedoch weitgehend unbeachtet bleibt. Sie schreiben: „Der hohen Konfliktaffinität insbesondere der soziologischen Theorie steht eine weitreichende Streitblindheit gegenüber“ (S. 16). Dabei hat bereits Georg Simmel, einer der Gründerväter der Soziologie, 1908 auf den Streit als relevantes Forschungsfeld für die Soziologie hingewiesen:

„Dass der Kampf soziologische Bedeutung hat, indem er Interessengemeinschaften, Vereinheitlichungen, Organisationen verursacht oder modifiziert, ist prinzipiell nie bestritten. Dagegen muss der gewöhnlichen Anschauung die Frage paradox vorkommen, ob nicht der Kampf selbst schon, ohne Rücksicht auf seine Folge- oder Begleiterscheinungen, eine Vergesellschaftungsform ist. [...] Wenn jede Wechselwirkung unter Menschen eine Vergesellschaftung ist, so muss der Kampf, der doch eine der lebhaftesten Wechselwirkungen ist, der in der Beschränkung auf ein einzelnes Element logisch unmöglich ist, durchaus als Vergesellschaftung gelten“ (Simmel, 1923, S. 186).

Simmel widmet dem „Streit“ in seinem Werk *Soziologie: Untersuchungen über die Formen der Vergesellschaftung* ein eigenes Kapitel und untersucht dessen positive Wirkungen auf das soziale Gefüge. Streit wirkt für ihn nicht nur trennend, sondern kann auch den Zusammenhalt von Gruppen stärken, Spannungen regulieren, Beziehungen stiften und zur Selbstvergewisserung beitragen.

Trotz dieses Befunds fanden Simmels Überlegungen zum Streit in der weiteren soziologischen Theorieentwicklung kaum Beachtung. Die Konflikttheorie der 1960er und 1970er Jahre knüpft zwar an ihn an (Joas, 2013), verlagert jedoch mit ihrer neuen Begrifflichkeit auch den Fokus. Statt interpersonelle „Wechselwirkungen unter Menschen“, bzw. menschliche Interaktionen zu analysieren, richtet sie – in der Tradition von Marx und Engels – ihre Aufmerksamkeit vor allem auf gesamtgesellschaftliche Konfliktlinien, staatliche Akteur*innen und strukturelle Widersprüche (Joas 2013; Dennaoui & Witte, 2022). Damit bietet die Konflikttheorie nur begrenzte Anknüpfungspunkte für eine Soziologie des Streits, die Formen der Konfliktaustragung im zwischenmenschlichen Bereich untersucht.

Ausgehend von Simmel verstehen Dennoui und Witte (2022) Streit als eigenständige Form konflikthafter Interaktion mit spezifischen Merkmalen, die ihn von anderen Konfliktarten wie Klassen- oder Militärkonflikten unterscheidet. Sie definieren Streit als „eine auf einer spezifischen Differenz beruhende, sprachlich verfasste Praxis von zwei oder mehr Akteur*innen in Interaktionskontexten“ (S. 25). Streit ist dabei sozial strukturiert und wird symbolisch kommuniziert, über Rituale reproduziert, durch Normen stabilisiert und schließlich in organisationalen Zusammenhängen verdichtet. Auf der Grundlage dieser Aspekte unterscheiden Dennoui und Witte (2022) bei der Analyse von konflikthafter Interaktionen zwischen vier verschiedenen Dimensionen:

- (1) Symbolische Dimension
- (2) Normative Dimension
- (3) Dimensionen der Interaktionen und Rituale
- (4) Organisationsdimension

In der symbolischen Dimension (1) macht die Analyse des Streits deutlich, dass Sprache nicht nur Medium der Streitaustragung ist, sondern auch selbst zum Streitobjekt werden kann, da sie Selbstbilder formt, Zugehörigkeiten schafft und Grenzen zieht. In diesem Kontext spielt die Durchsetzung eigener Benennungen eine wichtige Rolle und auch die Bedeutung einer bewussten Verweigerung eines Streits. Bei Online Hate und Counter Speech steht Sprache im Zentrum, da im digitalen Raum nonverbale Kommunikationsformen auf der Ebene der Kommentare entfallen (Suler, 2004) und Konflikte daher vollständig in schriftlicher Form ausgetragen werden. Gleichzeitig wird die Sprache vielfach selbst zum Streitobjekt, da verletzend und ausgrenzend Aussagen (Matsuda et al., 1993) in Counter Speech Reaktionen unmittelbar angefochten werden.

Gesellschaftliche Normen (2) beeinflussen maßgeblich, wie Konflikte ausgetragen werden und werden zugleich häufig selbst zum Gegenstand des Streits. Streit schafft mitunter neue Normen und verändert bestehende. Hate und Counter Speech werden im Zusammenhang der vorliegenden Arbeit als Normen- und Wertekonflikte verstanden, die auf zwischenmenschlicher Ebene ausgetragen und in denen unterschiedliche normative Ansprüche sichtbar gemacht und verhandelt werden.

Dennaoui und Witte verstehen Streit als eine rituelle kommunikative Praxis (3), die eigenen Phasen und Übergangsregeln folgt und über vielfältige Techniken und Strategien für Beginn, Fortführung und Beendigung verfügt. Diese Rituale manifestieren sich meist sowohl auf verbaler, als auch auf non-verbaler, bzw. körperlicher Ebene, wobei Letztere im Kontext von Kommentaren in sozialen Medien entfällt. Im Mittelpunkt dieser Arbeit steht die Analyse dieser Abläufe, mit dem Ziel, verschiedene Formen von Hate Speech und Counter Speech zu identifizieren und zu charakterisieren.

Die vierte Dimension (4) betont, dass Streit stets auf organisatorische und institutionelle Voraussetzungen und Rahmenbedingungen angewiesen ist. Selbst ohne formale Strukturen ist ein Mindestmaß an sozialer und räumlich-zeitlicher Koordination erforderlich, damit ein Streit stattfinden kann. Auch Konflikte in den sozialen Medien setzen voraus, dass den beteiligten Nutzer*innen der entsprechende Beitrag angezeigt wird und sie über die technischen Mittel verfügen, darauf zu reagieren. Plattformen sind keine neutralen Orte der Konfliktaustragung, sondern prägen mit ihren eigenen Strukturen und Logiken die Dynamik eines Konflikts.

Auf der Grundlage ihrer Definition und den vier Dimensionen des Streits unterscheiden Dennaoui und Witte verschiedene idealtypische Streitformen. Da diese jedoch kaum näher ausgeführt werden, lassen sie sich nur bedingt weiterverwenden. Im Folgenden werden daher lediglich die Kriterien festgehalten, auf denen ihre Unterscheidungen beruhen. Als Erstes nennen sie den Rationalitätsgrad eines Streits und damit verbunden die anvisierten Zwecke, die eingesetzten Mittel und die zugrundeliegenden Handlungsorientierungen der Beteiligten. Dieses Kriterium setzt bei den Weltbildern, Beweggründen und Emotionen der beteiligten Personen an und ist überwiegend intrapersonell angelegt, spiegelt sich jedoch auch in den gewählten Streitmitteln und dem Konfliktverlauf. Da sich auf die intrapersonelle Ebene auf Grundlage von Online-Komentaren nur begrenzt zugreifen lässt, können im Rahmen dieser Arbeit lediglich die eingesetzten und damit beobachtbaren Streitmittel analysiert werden. Als zweites Kriterium führen sie das Streitobjekt an. Dieses kann – je nach Abstraktionsgrad – materielle oder immaterielle Güter, Interessen, Macht, Prinzipien oder Werte betreffen. Auf einer Metaebene verorten sie zudem Streitformen, die allein der Machtdemonstration dienen oder sich um unbeteiligte Dritte drehen. Das Streitobjekt prägt dabei oft die Wahl der Mittel, wird jedoch nicht immer explizit gemacht. Ein drittes Kriterium betrifft den Grad an Offenheit

oder Geschlossenheit eines Konfliktverlaufs. Das daraus entstehende Kontinuum reicht von eskalierten, entgrenzten und totalen Streitformen bis hin zu verhinderten und erstickten Streittypen. Ein Streit endet oder wird unmöglich, wenn die andere Person als streitunwürdig betrachtet oder nicht mehr als Person anerkannt wird. Wo sich Hate und Counter Speech Interaktionen auf TikTok hier einordnen lassen, ist Gegenstand der vorliegenden Untersuchung.

Moderne Gesellschaften sind durch einen Ausdifferenzierungsprozess in verschiedene Handlungssphären mit jeweils eigenen Regeln und Strukturen geprägt, was sich auch auf das Streiten auswirkt. Dennaoui und Witte (2022) gehen daher nicht von einer einheitlichen Streitkultur aus, sondern von einer Vielzahl unterschiedlicher Streitkulturen, die je nach Sphäre eigene Formen, Orte, Zeiten und Logiken hervorbringen. Im Rahmen dieser Arbeit werden soziale Medien nicht nur als Orte des Streits, sondern auch als eigene Handlungssphäre verstanden, deren Aufbau die Dynamik und die Ausprägung von Hate Speech und Counter Speech beeinflusst. Da jede Plattform über spezifische Strukturen und Logiken verfügt (Matassi & Boczkowski, 2023), werden soziale Medien nicht als homogener Raum betrachtet, sondern der Fokus wird gezielt auf TikTok gerichtet. Bevor jedoch auf dessen Besonderheiten eingegangen wird, folgt eine Betrachtung der Spezifika digitaler Interaktionen im Allgemeinen.

2.2. Zwischenmenschliche Konflikte im digitalen Raum

Seyfert (2024) beschreibt die Gegenwart als „Zeitalter der Digitalisierung 2.0“ (S. 25), das sich durch eine verallgemeinerte Teilhabe und die Präsenz von Algorithmen in sämtlichen Lebensbereichen von der frühen Digitalisierung der 1990er- und 2000er-Jahre unterscheidet (Roberge & Seyfert, 2017; Seyfert, 2024). Produktion und Konsum digitaler Inhalte gehen dabei zunehmend ineinander über, sodass jede Person beides zugleich tun kann. Die Rede von einer „Demokratisierung“ digitaler Kommunikation weist Seyfert jedoch zurück, da diese auf privatwirtschaftlich organisierten Plattformen beruht (Seyfert, 2024; Zuboff 2018). Die Digitalisierung verändert soziale Interaktionen und damit auch die Austragung zwischenmenschlicher Konflikte grundlegend, was klassische, auf Präsenz ausgerichtete soziologische Theorien vor neue Herausforderungen stellt. Seyfert (2024) spricht in diesem Zusammenhang von einer umfassenden digitalen Transformation, in der digitale Beziehungen

zunehmend dominieren. Sie ersetzen Präsenzbeziehungen jedoch nicht, sondern erweitern sie.

Ein zentrales Merkmal digitaler Kommunikation ist der von Suler (2004) beschriebene Online-Enthemmungseffekt, der erklärt, warum Menschen sich im Internet häufig offener und aggressiver verhalten als im persönlichen Kontakt. Suler (2004) nennt sechs sich gegenseitig verstärkende Faktoren: Anonymität, Unsichtbarkeit, Asynchronität, solipsistische Introjektion, dissoziative Imagination und die Minimierung von Autorität. Anonymität beschreibt die vermeintliche Möglichkeit, das eigene Handeln im Internet von der realen Identität zu lösen, was Verantwortung mindert und enthemmtes Verhalten begünstigt. Unsichtbarkeit verweist auf das Fehlen nonverbaler Rückmeldungen, wodurch die Wirkung des Gesagten für die sprechende Person nicht unmittelbar erfahrbar ist. Ohne dieses zentrale Korrektiv werden Offenheit und Aggressivität gleichermaßen erleichtert. Asynchrone Kommunikation führt dazu, dass Reaktionen zeitlich verzögert erfolgen, was provokante oder persönliche Äußerungen wahrscheinlicher macht. Die solipsistische Introjektion bezeichnet die Tendenz, fehlende Informationen wie Stimme oder Mimik des Gegenübers zu imaginieren und zu projizieren, sodass digitale Interaktionen den Charakter innerer, selbstoffenbarender Dialoge annehmen können. Dissoziative Imagination kann bewirken, dass die digitale Welt wie ein Spiel oder ein fiktives Universum erlebt wird, in dem keine Verantwortung für Handlungen in der physischen Realität empfunden wird. Anonymität und Unsichtbarkeit tragen außerdem zur Abschwächung von Status und Autorität bei und können hierarchische Strukturen schwächen, was Tabubrüche und Regelüberschreitungen begünstigt. Suler (2004) betont jedoch, dass sich im digitalen Raum nicht das „wahre Selbst“ zeigt, sondern situativ bestimmte Persönlichkeits- und Verhaltensmuster hervortreten, die ebenso kontextabhängig sind wie jede andere Form sozialen Handelns.

Soziale Medien erweitern die raumzeitlichen Grenzen sozialer Interaktion, ohne sie vollständig zu entkoppeln. Digitalisierte Interaktionen ereignen sich parallel an den physischen Orten der Beteiligten und auf der jeweiligen Plattform, die eigene Bedingungen setzt. Der Fokus der vorliegenden Arbeit liegt auf zwischenmenschlichen Konflikten, die unter plattformspezifischen Bedingungen in Form von Hate Speech und Counter Speech auftreten. Welche Bedingungen hierbei relevant sind, wird im folgenden Kapitel näher erläutert.

2.3. Plattformen als kuratierte Räume

Plattformen sind keine neutralen Räume, sondern von privatwirtschaftlichen Akteur*innen gestaltete Umgebungen, in denen Algorithmen und Plattformstrukturen Inhalte kuratieren und Sichtbarkeit steuern (Gillespie, 2018; Seyfert, 2024). Matassi und Boczkowski (2023) kritisieren, dass Unterschiede zwischen Plattformen selten systematisch untersucht werden und betonen, dass soziale Medien keine homogene Einheit bilden. Plattformspezifische Unterschiede entstehen durch unternehmerische Entscheidungen, nationale Gesetzgebungen und die Zusammensetzung der Nutzer*innen. Unternehmensentscheidungen betreffen etwa das grafische und technologische Design einer Plattform, den Einsatz bestimmter Algorithmen sowie die Umsetzung von Regulierungs- und Moderationspraktiken (Gillespie, 2018; Matassi & Boczkowski, 2023; Matamoros-Fernández, 2017; Munn, 2020). Nutzer*innen nehmen diese Unterschiede wahr und passen ihre Kommunikations- und Ausdrucksweisen entsprechend an, was sich direkt auf soziale Interaktionen und die Dynamik zwischenmenschlicher Konflikte in digitalen Räumen auswirkt.

Gillespie (2018) beschreibt zwei zentrale Formen der Plattformregulierung: Moderation und Informationsmanagement. Zur Moderation gehören meldebasierte Verfahren sowie die manuelle oder automatisierte Erkennung und Entfernung problematischer Inhalte. Automatisierte Systeme arbeiten häufig mit Wortfiltern, deren Einsatz nach außen hin nicht transparent gemacht wird. Das Informationsmanagement wiederum betrifft die Frage nach der Sichtbarkeit von Inhalten und ist eng mit den eingesetzten Empfehlungsalgorithmen verknüpft.

Im Zusammenhang mit der Frage nach Sichtbarkeit, zeigt Adriana Matamoros-Fernández (2017) mit ihrem Konzept des „platformed racism“, wie Rassismus durch Entscheidungen von Plattformunternehmen auf zweifache Weise begünstigt wird. Einerseits führt die kommerzielle Ausrichtung der Unternehmen dazu, dass wirtschaftliche Interessen oft auf Kosten vulnerabler Gruppen und Minderheiten durchgesetzt werden. Andererseits verstärken Empfehlungsalgorithmen rassistische Dynamiken, indem sie empörungsgetriebene und antagonistische Inhalte bevorzugen und sie mit erhöhter Reichweite und Aufmerksamkeit belohnen. Die Algorithmen der Social Media Plattformen basieren meist auf Empfehlungssystemen, unterscheiden sich jedoch in den berücksichtigten Kriterien, in ihrer

Reichweite und in ihrer Performanz. Der TikTok-Algorithmus ermöglicht es beispielsweise, dass Beiträge vergleichsweise schnell große Reichweiten erzielen – unabhängig vom eigenen Bekanntenkreis (Leander & Burriss, 2020).

Viele Plattformunternehmen sind international tätig und bieten ihre Dienste in vielen verschiedenen Ländern an. Sie entscheiden, wie viele Moderator*innen sie insgesamt und für einzelne Sprachräume und Regionen einsetzen. Diese Entscheidungen beeinflussen unmittelbar die Qualität der Moderation, werden aber selten vollständig transparent gemacht (TikTok Transparenzzentrum, 2025). Moderator*innen mit regionalsprachlicher Expertise sind wichtig, da kulturelle und soziale Unterschiede zwischen den Nutzer*innen bestehen (Matassi & Boczkowski, 2023). Länderspezifische Unterschiede werden auch im Design der jeweiligen Plattformen sichtbar. Auf Facebook kann das eigene Profil aktuell etwa nur in bestimmten Ländern „verriegelt“ und damit Außenstehenden schwerer zugänglich gemacht werden.³ Die Gründe hierfür bleiben intransparent, könnten jedoch mit nationalen gesetzlichen Vorgaben oder wirtschaftlichen Interessen zusammenhängen.

Grafische Designs, algorithmische Systeme und Moderationspraktiken werden von den Plattformunternehmen kontinuierlich verändert und an neue Bedingungen angepasst. Dieser fortlaufende Wandel entsteht durch Wettbewerb, technische Innovationen und gesellschaftliche Entwicklungen und zeigt sich etwa in der zunehmenden Automatisierung der Moderation. TikTok gibt an, dass im ersten Quartal 2025 mehr als 87 Prozent der gelöschten Videos von automatisierten Systemen entfernt wurden (TikTok Transparenzzentrum, 2025; Peterka, 2025). Seit Ende Mai 2025 ist es zudem erstmals möglich, mit Bildern auf TikTok zu kommentieren (Mandyportius, 2025). Gleichzeitig wurde das Zeichenlimit für Kommentare auf TikTok von 150 auf fast 3000 erhöht (Conectere, 2025). Da offizielle Stellungnahmen hierzu fehlen, muss an dieser Stelle auf die Berichte von Nutzer*innen zurückgegriffen werden. Diese Anpassungen wirken sich direkt darauf aus, wie Hate Speech und Counter Speech auf TikTok ausgestaltet sind. Für die Forschung bedeutet dies, zu berücksichtigen, dass unterschiedliche Beiträge unter verschiedenen Bedingungen entstehen können. Es bedeutet auch, dass jede empirische Studie eine Momentaufnahme darstellt, die soziale Dynamiken zu einem bestimmten Zeitpunkt unter spezifischen Bedingungen abbildet. Frühere Rahmenbedingungen lassen sich dabei häufig nur schwer rekonstruieren und nachvollziehen.

Wie bereits erwähnt, setzen Plattformen zunehmend automatisierte Systeme ein, die problematische Inhalte vor und nach der Veröffentlichung unter anderem mittels Wortfiltern erkennen sollen. Herausgefilterte Beiträge können freigegeben, gelöscht, in ihrer Sichtbarkeit eingeschränkt oder verdeckt werden, sodass nur die Verfasser*innen selbst diese weiter sehen können (Gillespie, 2018). 2022 wurde bekannt, dass TikTok in Deutschland und Österreich Kommentare mit Begriffen wie „schwul“, „queer“, „homophob“ oder „Auschwitz“ automatisch ausblendete, bzw. verdeckte, ohne die Verfasser*innen zu informieren (Eckert et al., 2022; Diwisch, 2022). Diese Praxis, auch „Shadow Banning“ genannt (Savolainen, 2022), begründete TikTok mit dem häufigen Zusammenhang dieser Begriffe mit hasserfüllter Rede. Kritiker*innen wiesen jedoch darauf hin, dass mit einem solchen Vorgehen Meinungsfreiheit und demokratische Debatten erheblich eingeschränkt werden (Eckert et al., 2022). Der Fall verdeutlicht das Spannungsfeld zwischen dem Schutz vor Hate Speech und der Sicherung freier Meinungsäußerung. Einige der genannten Begriffe wurden zwar inzwischen wieder zugelassen, doch TikTok bleibt im Umgang mit solchen Filtern weitgehend intransparent. Neben den von TikTok eingesetzten Wortfiltern können auch Nutzer*innen solche Filter für Kommentare unter ihren Inhalten einrichten. Obwohl diese Listen nicht öffentlich einsehbar sind, haben Nutzer*innen ein Bewusstsein für ihre Existenz und ihre Wirkungsweisen entwickelt und ihren Sprachgebrauch entsprechend angepasst. Daraus hat sich der sogenannte „Algospeak“ entwickelt – eine Praxis, bei der Wörter abgeändert oder neu erfunden werden, um algorithmische Beschränkungen zu umgehen (Steen et al., 2023).

Es wurde deutlich, dass soziale Medien keine neutralen Räume sind, sondern aktiv kuratierte Umgebungen, in denen verschiedene Faktoren bestimmen, was sichtbar wird und wie Kommunikation abläuft. Diese Rahmenbedingungen haben unmittelbare Auswirkungen darauf, wie Hate Speech und Counter Speech entstehen und sich entwickeln und welche Formen sie annehmen. Da jede Plattform eigene Strukturen und Funktionslogiken besitzt, liegt der Schwerpunkt in weiterer Folge auf TikTok, da die im Rahmen dieser Arbeit untersuchten Daten von dieser spezifischen Plattform stammen.

2.4. Besonderheiten von TikTok

Ergänzend zu den bereits genannten Formen der Plattformregulierung weist TikTok zusätzliche Besonderheiten auf, durch welche sie sich von anderen sozialen Medien unterscheidet. TikTok ist eine vergleichsweise junge Plattform, zu der bislang nur wenige sozialwissenschaftliche Studien und Konzepte vorliegen. Die App wurde 2016 vom chinesischen Unternehmen ByteDance gegründet und im August 2018 international verfügbar gemacht. Während der Corona-Pandemie stiegen die Zugriffszahlen stark an (Deutschlandfunk Nova, 2020; Saferinternet.at, 2025). Obwohl in den letzten Jahren auch mehr ältere Nutzer*innen hinzukamen (Ng & Indran, 2022), bleibt TikTok vor allem bei jüngeren Menschen sehr beliebt. In Österreich nutzen 2025 rund 72 Prozent der Jugendlichen unter 17 Jahren die Plattform regelmäßig (Saferinternet.at, 2025). Eine detaillierte Altersverteilung seiner Nutzer*innen veröffentlichte TikTok bislang nicht.

Weimann und Masri (2023) betonen, dass sich TikTok in zwei zentralen Punkten von anderen sozialen Medien unterscheidet. Erstens verfügt TikTok als jüngste der großen Plattformen über weniger Erfahrung im Schutz ihrer Nutzer*innen vor problematischen Inhalten. Gleichzeitig bietet dieser Umstand die Möglichkeit, Nutzer*innen gleich zu Beginn an strenge Moderationsmaßnahmen zu gewöhnen. Zweitens besteht ein großer Teil der TikTok-Nutzer*innen aus sehr jungen Menschen, die besonders anfällig für manipulative und schädliche Inhalte sein können. Der Erfolg der Plattform bei jungen Menschen hängt eng mit ihrem Design zusammen. Kurze Videos, zahlreiche Effekte und kreative Bearbeitungsmöglichkeiten sprechen besonders Kinder und Jugendliche an, die tendenziell eine kürzere Aufmerksamkeitsspanne haben, und bieten ihnen Raum für Ausdruck und Anerkennung (Weimann & Masri, 2024). Für das Jahr 2025 gibt es Hinweise darauf, dass Jugendliche TikTok hauptsächlich zur Unterhaltung nutzen und weniger zur Informationssuche oder Selbstdarstellung. Die Ergebnisse zeigen zudem, dass weniger als ein Viertel der befragten Jugendlichen regelmäßig Kommentare auf TikTok verfasst und dass TikTok-Inhalte von Jugendlichen tendenziell eher gelikt als kommentiert werden (Deda- Bröchin et al., 2025).

Gerbaudo (2024) beschreibt TikTok als Teil einer zweiten Generation sozialer Medien, die sich durch eine neue Form der algorithmischen Inhaltsvermittlung unterscheidet. An die Stelle vernetzter Öffentlichkeiten, die auf persönlichen Beziehungen beruhen, treten algorithmisch

gebildete Gruppen von Nutzer*innen mit ähnlichen Interessen und Verhaltensmustern, die sich meist nicht persönlich kennen. Diese Bündelung fördert zwar die Interaktion mit den Plattforminhalten, führt jedoch zugleich zu einer Entpersonalisierung der Onlineerfahrung, zu Intransparenz hinsichtlich der Strukturen digitaler, öffentlicher Räume und zu einer weiteren subkulturellen Aufsplitterung der digitalen Öffentlichkeit. Abidin (2021) hat zur Bezeichnung dieser Dynamiken das Konzept der „fragmentierten Öffentlichkeit“ eingeführt. Gerbaudo (2024) sieht das Phänomen fragmentierter Öffentlichkeiten besonders ausgeprägt auf TikTok – stärker als auf anderen Plattformen. Welche Formen und Dynamiken Hate Speech und Counter Speech vor dem Hintergrund der kuratierenden Eingriffe der Plattformbetreibenden konkret annehmen, zeigt das folgende Kapitel zum Forschungsstand.

3. Forschungsstand

Die Phänomene Hate Speech und Counter Speech werden in der Forschung meist getrennt voneinander untersucht, auch wenn es einige Studien gibt, die beide berücksichtigen. Besonders in der Forschung zu Gegenrede wird Hate Speech häufig mitgedacht, während umgekehrt seltener auf Gegenrede eingegangen wird – meist lediglich als hoffnungsvoller Ausblick. In der vorliegenden Arbeit wird zunächst ein allgemeiner Überblick über das Forschungsfeld zu Hate Speech und Counter Speech gegeben, bevor die jeweiligen Forschungsstände getrennt voneinander dargestellt werden. Dabei ist zu berücksichtigen, dass Counter Speech stets in Bezug auf Hate Speech zu verstehen ist. Hervorgehoben werden dabei Studien, die TikTok-Kommentare als Untersuchungsgrundlage nutzen oder sich explizit auf TikTok beziehen.

3.1. Kartografie eines Forschungsfeldes

Eine Annäherung an die Phänomene Hate Speech und Counter Speech erfordert zunächst einen Blick auf zentrale Befunde zum Forschungsfeld und aktuellen Forschungsstand. In den vergangenen Jahren hat die Zahl wissenschaftlicher Publikationen zu beiden Themen deutlich zugenommen. Tontodimamma et al. (2021) zeigen einen stetigen Anstieg der Veröffentlichungen zu Hate Speech seit 1992, mit einer stärkeren Zunahme ab 2010. Während 2010 noch rund fünfzig Arbeiten zum Thema Hass im Internet erschienen, waren es ab 2018 jährlich über dreihundert (Tontodimamma et al., 2021; Jiménez-Díaz & Del Rey, 2025). Auch

im verwandten Bereich Cyberbullying ist seit 2012 ein deutlicher Zuwachs an wissenschaftlichen Publikationen zu verzeichnen (Leschka & Mirbabaie, 2025). Seit 2019 ist auch das wissenschaftliche Interesse an Counter Speech nachweisbar gestiegen (Chung et al., 2024).

Insgesamt ist das Interesse an Hate Speech und Counter Speech in digitalen Räumen seit 2019 weiter gestiegen. Dieser Anstieg lässt sich laut Matassi und Boczkowski (2023) und Alsagheer et al. (2022) vor allem auf eine Reihe von Terroranschlägen in den USA, Europa und Neuseeland zurückführen, die von Täter*innen verübt wurden, die sich zumindest teilweise online radikalisiert hatten. Besonders der Anschlag auf zwei Moscheen im neuseeländischen Christchurch, bei dem mehr als fünfzig Menschen getötet wurden, führte zu einer breiten öffentlichen und politischen Debatte. Der Attentäter übertrug die Tat live auf Facebook, wo der Stream während der Ausführung rund zweihundert Mal angesehen, jedoch erst zwölf Minuten nach Ende der Übertragung gemeldet wurde. Bis zu seiner Entfernung wurde das Video etwa viertausend Mal aufgerufen und Kopien zirkulieren bis heute im Internet (Matassi & Boczkowski, 2023; Pichler, 2019). Der Vorfall führte zu intensiver öffentlicher Kritik, rechtlichen Auseinandersetzungen und neuen Gesetzesinitiativen in mehreren Staaten (Pichler, 2019; Mittelstaedt, 2019).

Angesichts der gestiegenen Zahl an Publikationen betonen Jiménez-Díaz und Del Rey (2025) die Notwendigkeit, die in den letzten Jahren rapide gewachsenen Erkenntnisse zu Hass im Internet systematisch zu ordnen und in bestehende Wissensbestände zu integrieren. Sie weisen darauf hin, dass das Wissen über Online-Hass derzeit stark fragmentiert ist, was sich auf die Vielzahl unterschiedlicher Begrifflichkeiten, Erhebungskontexte und disziplinärer Zugänge zurückführen lässt. Diese Fragmentierung betrifft sowohl den Forschungsbereich Hate Speech als auch den Bereich Counter Speech. Eine nähere Auseinandersetzung mit den jeweiligen Begrifflichkeiten folgt in den nächsten Kapiteln.

Neben der bislang unzureichenden Systematisierung des Forschungsfeldes machen Matamoros-Fernández und Farkas (2021) auf die geringe Diversität in der Erforschung von Online Hate Speech aufmerksam. Sie kritisieren vor allem die mangelnde Vielfalt in den ausgewählten geografischen Kontexten, Sprachen, Plattformen und methodischen Zugängen

zur Untersuchung der genannten Phänomene. Der Großteil der Untersuchungen zu Hate und Counter Speech stammt aus dem englischsprachigen Raum – vor allem aus Nordamerika – gefolgt von europäischen Ländern wie Deutschland, Schweden, Spanien, Italien und den Niederlanden (Matamoros-Fernández & Farkas, 2021; Tontodimamma et al., 2021; Chung et al., 2024). Diese einseitige Ausrichtung ist relevant, da politische, gesellschaftliche und sprachliche Rahmenbedingungen die Erscheinungsformen von Hate und Counter Speech maßgeblich prägen.

Die meisten Datensätze stammen von Twitter (seit 2023 „X“), gefolgt von Facebook und Youtube, während andere Plattformen bislang kaum systematisch untersucht wurden. Die Dominanz von Twitter-Daten lässt sich vor allem auf den bis zur Übernahme durch Elon Musk vergleichsweise einfachen Zugang zur Plattform-API zurückführen (Matamoros-Fernández & Farkas, 2021; Chung et al., 2024). Mehrere Wissenschaftler*innen, darunter Matamoros-Fernández und Farkas (2021) sowie Chung et al. (2024) heben das Forschungsdefizit zu Hate Speech und Counter Speech auf TikTok hervor und fordern eine intensivere wissenschaftliche Auseinandersetzung mit dieser Plattform.

In methodischer Hinsicht werden Hate Speech und Counter Speech aus unterschiedlichen wissenschaftlichen Perspektiven untersucht, vor allem aus der Computer- und Sozialwissenschaft. In den letzten Jahren ist der Ruf nach einer engeren interdisziplinären Zusammenarbeit zur Erforschung beider Phänomene lauter geworden (Chung et al., 2024). Vor diesem Hintergrund entstand das Forschungsprojekt „Young People Against Online Hate“, an dem Sozialwissenschaftler*innen der Universität Wien und Computerwissenschaftler*innen der Fachhochschule St. Pölten gemeinsam arbeiteten. Computerwissenschaftliche Arbeiten verfolgen überwiegend quantitative Ansätze und nutzen automatisierte Verfahren – häufig mit dem Ziel, Hate Speech automatisiert zu erkennen und zu lokalisieren. Sozialwissenschaftliche Arbeiten setzen sowohl quantitative als auch qualitative Methoden ein und legen ein besonderes Augenmerk auf inhaltliche Deutungen (Chung et al., 2024). Unter den qualitativen Zugängen dominieren inhaltsanalytische Verfahren, während ethnografische Studien und Interviews bislang selten sind (Matamoros-Fernández & Farkas, 2021). Die vorliegende Arbeit schließt an die qualitative

Forschungstradition an und fokussiert sich auf die inhaltliche Analyse und Interpretation der erhobenen Daten.

Insgesamt lässt sich in der Forschung zu Hate Speech und Counter Speech ein klarer Trend zur inhaltlichen Ausdifferenzierung erkennen, wobei die Datengrundlagen und methodischen Zugänge weiterhin eine gewisse Homogenität aufweisen, die auch die Forschungsergebnisse beeinflusst (Tontodimamma et al., 2021; Chung et al., 2024). Angesichts der Vielzahl an Studien und Perspektiven wird das Forschungsfeld zunehmend komplex und schwer zu überblicken. Chung et al. (2024) schlagen eine strukturierende Systematisierung der Forschungsergebnisse zu Counter Speech vor, die auf der leitenden Frage basiert, „wer welche Art von Kommentar an wen und aus welchem Grund richtet“ (S.33). Um die Dimensionen des Ortes und der Folgen erweitert, ergibt sich daraus ein Gliederungsprinzip, das in dieser Arbeit sowohl für die Darstellung der Forschungsergebnisse zu Counter Speech als auch zu Hate Speech herangezogen wird: Wer (1) verfasst was (2) an wen gerichtet (3) an welchem Ort (4) aus welchem Grund (5) und mit welchen Folgen (6)? Die Erkenntnisse werden entsprechend in die Bereiche Sprecher*innen (1), Empfänger*innen (2), Formen und Definitionen von Hate Speech und Counter Speech (3), Plattform und Beitrag (4), Motivation und Ziele (5) sowie Folgen (6) unterteilt. Aufgrund der Spezifika der empirischen Grundlagen der vorliegenden Arbeit liegt der Fokus in den folgenden Kapiteln auf den Erkenntnissen zu den Formen von Hate Speech und Counter Speech (3) und ihren Folgen (6). Erkenntnisse bezüglich anderer Dimensionen werden lediglich punktuell eingebracht.

3.2. Forschungsstand zu Hate Speech

Der Themenbereich Hate Speech ist seit etwa fünfunddreißig Jahren Gegenstand der Forschung (Tontodimamma et al., 2021). Der Diskurs ist jedoch stark fragmentiert, weil unterschiedliche Begriffe zur Bezeichnung des gleichen Phänomens oder Teilaspekte dessen verwendet werden. In der wissenschaftlichen Literatur finden sich entsprechend Begriffe wie „cyber hate“, „online hate speech“, „online hate“, „online hating“ oder „online hate content/material“. Diese Begriffe werden größtenteils synonym gebraucht, erschweren aber eine Zentralisierung der Erkenntnisse (Jiménez-Díaz & Del Rey, 2025). Die Vielfalt der Begriffe verweist auf das bislang wenig strukturierte Forschungsfeld und auf die Schwierigkeiten, das Phänomen präzise zu definieren. Im Folgenden werden zunächst aktuelle Definitionsdebatten

zu Hate Speech dargestellt, bevor jene Definition eingeführt wird, die in dieser Arbeit Anwendung findet. Danach werden zentrale Formen von Hate Speech mit einem Fokus auf Ergebnisse zu TikTok erläutert.

3.2.1. Hate Speech Definitionen

Rund um die begriffliche Bestimmung von Hate Speech hat sich, ausgehend von den Arbeiten von Sellars (2016), eine eigenständige Forschungstradition entwickelt, die sich mit der Frage befasst, welche Phänomene unter diesem Begriff gefasst werden können oder sollten. Die Wahl der jeweiligen Definition kann einen erheblichen Einfluss auf die Forschungsergebnisse haben (Madukwe et al., 2020). Besonders in den Computerwissenschaften besteht ein starkes Interesse an einer einheitlichen Begriffsbestimmung, da bereits geringfügige Abweichungen Auswirkungen auf die automatisierte Erfassung von Hate Speech haben können (Madukwe et al., 2020).

In den letzten Jahren wurden mehrere Datensammlungen angelegt, die unterschiedliche Definitionen von Hate Speech zusammentragen, einander gegenüberstellen und systematisch vergleichen (Fortuna et al., 2021; Khurana et al., 2022; Pachinger et al., 2023; Korre et al., 2025). Diese Arbeiten verfolgen das Ziel, Forschenden eine solide Grundlage an Definitionen bereitzustellen, das Forschungsfeld übersichtlicher zu gestalten und Erkenntnisse aus dem systematischen Vergleich unterschiedlicher Begriffsverwendungen zu gewinnen. Korre et al. (2025) haben insgesamt 493 verschiedene Definitionen von Hate Speech aus unterschiedlichen Sprach- und Kulturräumen zusammengetragen, von denen jedoch nur ein Teil aus wissenschaftlichen Publikationen stammt. Ergänzend wurden auch Definitionen aus Gesetzestexten, Wörterbüchern und von Online-Plattformen berücksichtigt. Die Definitionen unterscheiden sich in ihrer Begriffsweite und insbesondere darin, welche Opfergruppen sie einbeziehen (Khurana et al., 2022). Sellars (2016) hebt hervor, dass die Herabwürdigung bestimmter Gruppen oder einer Person als Mitglied einer dieser Gruppen vermutlich das zentrale Merkmal ist, das Hate Speech von anderen Formen verletzender Kommunikation wie Mobbing, Drohungen oder Beleidigungen unterscheidet. Pachinger et al. (2023) grenzen Hate Speech von verwandten Konzepten verletzender Sprache wie „Abusive Speech“, „Offensive Speech“, „Toxic Speech“ und „Uncivil Speech“ ab, die jeweils eine eigene Definitionstradition aufweisen. Als Gründe für die Vielzahl unterschiedlicher Definitionen nennen Forschende

einerseits subjektive Interpretationen (Fortuna et al., 2021) und andererseits kulturelle sowie sprachliche Unterschiede (Korre et al., 2025). Weit gefasste Definitionen verzichten häufig auf eine genaue Eingrenzung betroffener Gruppen, während eng gefasste Ansätze Hate Speech auf spezifische Diskriminierungsformen wie Rassismus oder Sexismus beschränken (Khurana et al., 2022).

Entgegen dem Wunsch nach einer einheitlichen Definition von Hate Speech (Jiménez-Díaz & Del Rey, 2025; Seemann et al., 2023) plädieren Korre et al. (2025), Khurana et al. (2022), Pachinger et al. (2023) und Fortuna et al. (2023) dafür, dass Forschende aus bestehenden Sammlungen jene Definition auswählen, die am besten zu ihrem jeweiligen Forschungsziel passt. Auf diese Weise soll eine weitere Ausdehnung des Definitionsfeldes vermieden und eine nachhaltige begriffliche Verdichtung unterstützt werden.

Die vorliegende Arbeit stützt sich – ebenso wie das Forschungsprojekt „Young People Against Online Hate“ – auf das Konzept der digitalen Gewalt von Atzmüller et al. (2022), das bereits in Vorgängerprojekten verwendet wurde. Nach Atzmüller et al. (2022) umfasst digitale Gewalt eine breite Palette hasserfüllter und herabsetzender Verhaltensweisen, die darauf abzielen, andere Personen zu erniedrigen oder zu verletzen. Dazu zählen Hassbotschaften, Cybermobbing, Online-Stalking, extremistische Ansichten, Verschwörungserzählungen, Beleidigungen, Falschinformationen, öffentliche Bloßstellungen, Belästigungen, ungewollte Konfrontationen mit schockierenden, pornografischen, diskriminierenden oder illegalen Inhalten, sowie unautorisierte oder manipulierte Veröffentlichungen persönlicher Bilder und Videos, missbräuchliche Verwendungen visueller Inhalte, Falschdarstellungen persönlicher Informationen und digital übermittelte physische Gewalt- oder Tötungsandrohungen.

Der Begriff Hate Speech wird in dieser Arbeit entsprechend weit gefasst und nicht auf bestimmte Opfergruppen beschränkt. Auch abwertende Äußerungen gegenüber Einzelpersonen, die keinem Kollektiv zugeordnet werden können, werden in diesem Verständnis von Hate Speech berücksichtigt. Im Zentrum steht das herabwürdigende Moment der sozialen Interaktion. Die hier verwendete Definition weist deutliche Parallelen zur Begriffsverwendung des weit gefassten Konzeptes respektlosen Verhaltens, auch „Incivility“ genannt, auf (Pachinger et al., 2023).

Sozialwissenschaftliche Definitionen unterscheiden sich mitunter deutlich von juristischen Zugängen. In den vergangenen Jahren wurde verstärkt versucht, Hate Speech rechtlich zu erfassen und zu sanktionieren. Dabei zeigen sich zwischen den einzelnen Ländern sowohl Gemeinsamkeiten als auch Unterschiede, die den jeweiligen gesellschaftlichen Umgang mit dem Phänomen widerspiegeln (Sellars, 2016; Siegel, 2020). In Österreich können entsprechende Postings unter verschiedene Straftatbestände fallen, darunter Verhetzung, Wiederbetätigung, Cybermobbing, üble Nachrede, Ehrenbeleidigung, Verleumdung, Kreditschädigung oder Belästigung fallen (oesterreich.gv.at, 2025). Das juristische Konzept der Verhetzung fokussiert sich dabei auf bestimmte, besonders vulnerable und geschützte Gruppen und entspricht damit enger gefassten Definition von Hate Speech. Zugleich existiert eine Reihe weiterer rechtlicher Bestimmungen, die einzelne Aspekte des Konzeptes digitaler Gewalt nach Atzmüller et al. (2022) abdeckt. Gesetzliche Regelungen unterliegen einem stetigen Wandel und werden fortlaufend an gesellschaftliche Entwicklungen angepasst (Parlament Österreich, 2020).

Auch Plattformen veröffentlichen eigene Definitionen von Hate Speech und Vorstellungen von Gemeinschaftsrichtlinien, die ein respektvolles Miteinander sichern sollen. Diese Richtlinien sind jedoch häufig bewusst vage gehalten, um den Plattformbetreiber*innen größtmögliche Handlungsspielräume im Umgang mit potenziell problematischen Inhalten zu sichern (Sellars, 2016; Gillespie, 2018). In den vergangenen Jahren hat der Druck staatlicher Akteur*innen auf Plattformen zugenommen, mit dem Ziel, die Regulierung öffentlicher Kommunikation im Sinne des Gemeinwohls zu stärken. Eine wesentliche Neuerung in diesem Zusammenhang stellt der „Digital Services Act“ dar, der 2022 vom Europäischen Parlament und dem Europäischen Rat verabschiedet wurde und 2024 in Kraft getreten ist. Eine Studie von Dubois und Reepschlager (2024) zeigt, dass Plattformen wie Facebook, Twitter und Reddit ihre Definitionen von Hass und Belästigung zwischen 2005 und 2020 kontinuierlich verfeinert und komplexere Konzepte von Verantwortung entwickelt haben. Gleichzeitig verweisen die Autor*innen auf anhaltende Defizite in der Transparenz, die sowohl die kontinuierliche Anpassung dieser Definitionen als auch die Kommunikation über konkrete Löschentscheidungen betreffen.

Im Folgenden sind die aktuellen, am 18. Juli 2025 geltenden Definitionen von „Hassrede“ und „Mobbing“ auf TikTok angeführt. Welche Regelversionen genau zum Zeitpunkt der Entstehung

der in dieser Arbeit analysierten Kommentare gültig waren, lässt sich nur schwer nachvollziehen. Die aktuellen Definitionen geben jedoch einen Einblick in den normativen Anspruch TikToks, den Diskurs auf der Plattform zu regulieren und zeigen, dass TikTok – aufgeteilt in die Bereiche Hassrede und Cybermobbing – ein relativ breites Verständnis von hasserfüllten Inhalten hat. Unter dem Begriff Cybermobbing wird dabei sogar ausdrücklich die Verantwortung von Zeug*innen angesprochen.

AUSZÜGE ZU DEN ANGABEN VON TIKTOK ZUM THEMA HASSREDE	AUSZÜGE ZU DEN ANGABEN VON TIKTOK ZUM THEMA CYBERMOBBING
Was ist Hassrede? Hassreden und hasserfülltes Verhalten umfassen das Angreifen, Bedrohen, Entmenslichen oder Herabwürdigen einer Person oder Gruppe aufgrund ihrer geschützten Eigenschaften. Geschützte Eigenschaften sind persönliche Merkmale, mit denen du entweder geboren wurdest, die unveränderlich sind oder im Rahmen derer dir schwerer psychischer Schaden zugefügt werden würde, wenn du gezwungen wärest, sie zu ändern oder wegen ihnen angegriffen würdest. Dazu gehören ethnische Zugehörigkeit, nationale Herkunft, Religion, Stamm, Kaste, sexuelle Orientierung, Geschlecht, Geschlechtsidentität, schwere Krankheit, Behinderung und Einwanderungsstatus. Herabwürdigende Äußerungen, die sich nicht auf die geschützten Eigenschaften einer Person oder Gruppe beziehen, gelten als Mobbing oder Belästigung. Hassreden und hasserfülltes Verhalten erkennen Hassreden und hasserfülltes Verhalten haben keinen Platz auf TikTok. Wir erlauben keine Hassreden sowie hasserfüllten Ideologien und empfehlen keine Inhalte, die negative Stereotypen über eine Person oder Gruppe mit einer geschützten Eigenschaft enthalten. Wir verwenden eine Kombination aus Technologie und menschlichen Moderator*innen, um Konten und Inhalte zu finden sowie zu entfernen, die gegen unsere Richtlinien verstoßen. Hassrede kann durch jede Form des Ausdrucks vermittelt werden, darunter Bilder, Text, Audio, Cartoons, Memes, Objekte, Gesten und Symbole, von denen einige nicht immer offensichtlich sind. Hasserefüllte Personen und Gruppen zielen mit einer Reihe von Taktiken auf ihre Opfer ab, darunter etwa: Direkte negative Angriffe • Beleidigende Angriffe, bei denen eine Person oder Gruppe und ihre Identität beschimpft werden, wobei gewalttätige, entmenslichende oder erniedrigende Sprache verwendet wird. • Kommentieren von Inhalten einer Community oder eines Creators*iner Creatorin mit hasserfüllten Bildern oder herabwürdigender Sprache. • Verwendung einer hasserfüllten Bezeichnung in Verbindung mit einer geschützten Eigenschaft, um einen Creator*ine Creatorin oder seine*ihre Community anzugreifen. Indirekte negative Angriffe • Witze und Memes die Stereotypen verbreiten und Diskriminierung sowie Intoleranz normalisieren. • Auf Hass basierende Verschwörungstheorien, die Misstrauen oder Abneigung gegen eine Person oder Gruppe aufgrund ihrer geschützten Eigenschaften verbreiten. • Verschlüsselte Botschaften, mit denen Hassgruppen untereinander kommunizieren, ohne nach außen hasserfüllt zu wirken, z. B. Codewörter, Symbole oder Audiotrends.	Was ist Mobbing? Mobbing ist in der Regel ein gezieltes Verhalten, das körperlichen, sozialen und/oder psychologischen Schaden verursachen soll. Das Verhalten kann online oder offline von einer Person oder einer Gruppe ausgeübt werden, die ihre Macht oder ihre vermeintliche Macht über eine andere Person oder eine Gruppe von Personen missbraucht, die sich nicht in der Lage fühlen, dies zu verhindern. Wir sind uns bewusst, dass Mobbing Personen extrem belasten kann, und wir tolerieren es auf unserer Plattform nicht. TikTok ist eine globale Community, die von Kreativität und Ausdruck lebt. Es ist wichtig, dass sich alle sicher fühlen, um sich ohne Angst vor Mobbing oder Belästigung auf der Plattform auszudrücken. Wie kann man Mobbing und Belästigung auf der Plattform erkennen? Häufige Formen von Mobbing sind verbales Mobbing, körperliches Mobbing und sexuelle Belästigung. Wir tolerieren keine Sprache oder Verhaltensweisen, die Personen auf der Plattform <i>belästigen</i>, <i>demütigen</i>, <i>bedrohen</i> oder <i>doxxen</i> (persönliche Informationen über jemanden online in böser Absicht veröffentlichen). Erfahre mehr über häufige Formen von Mobbing. Schon gewusst? • Mobbing betrifft diejenigen, die mobben, diejenigen, die gemobbt werden, und sogar diejenigen, die Mobbing sehen. • Umstehende, also Personen, die Mobbing sehen, können eine wichtige Rolle dabei spielen, Mobbing zu stoppen. • Es ist nicht ungewöhnlich, dass Menschen von Personen gemobbt werden, die sie kennen (z. B. Klassenkamerad*innen, Freund*innen, Kolleg*innen). • Mobbing kann in nahezu jedem sozialen Umfeld vorkommen, zum Beispiel in der Schule, im Internet, auf dem Spielplatz und am Arbeitsplatz. • Personen, die gemobbt werden, können auch andere Personen mobben. Diejenigen, die das tun, schüchtern andere oft ein, um ein Gefühl von Macht oder Kontrolle wiederzuerlangen, das sie in einem anderen Bereich ihres Lebens verloren haben. Kann ich negative Kommentare über Persönlichkeiten des öffentlichen Lebens abgeben? • Wir lassen nicht zu, dass Persönlichkeiten des öffentlichen Lebens schwer attackiert werden, was zu körperlichen Schäden oder zur Förderung von Belästigungen gegen sie führen kann. • Wir halten das Risiko körperlicher oder schwerer psychischer Schäden durch diese Angriffe für größer als den Nutzen ihrer Äußerung und entfernen solche Inhalte.

ABBILDUNG 1: SCREENSHOTS DER RICHTLINIEN VON TIKTOK IM UMGANG MIT HASSREDE UND CYBERMOBBING (TIKTOK SICHERHEITZENTRUM 2025A; 2025B)

Weimann und Masri (2023; 2024) überprüften in zwei nicht repräsentativen, im Abstand eines Jahres durchgeführten Studien mit identischem Forschungsdesign, inwieweit TikTok seine im Jahr 2020 angekündigte, verstärkte Moderation antisemitischer Inhalte tatsächlich umgesetzt hat. Die Ergebnisse zeigen eine deutliche Verschärfung der Problemlage: Die Zahl einschlägiger Videos stieg um 41 Prozent, die Zahl entsprechender Usernamen um 1375 Prozent und die Häufigkeit antisemitischer Kommentare nahm um 912 Prozent zu. Die Autor*innen folgern daraus, dass TikTok entweder nicht in der Lage oder nicht willens ist, die eigenen Nutzungsbedingungen konsequent durchzusetzen. Die Befunde verdeutlichen, dass das Auftreten von Hate Speech auf Plattformen einem dynamischen Wandel unterliegt und sich im Zeitverlauf erheblich verändern kann.

3.2.2. Hate Speech auf TikTok

In den Interviews von García-Prieto et al. (2024) berichten die meisten der befragten Influencer*innen, die von Ableism betroffen sind und auf mehreren Plattformen aktiv sind, dass sie auf TikTok deutlich mehr Hassnachrichten erhalten als anderswo. Als Gründe nennen sie ihre größere Reichweite auf TikTok sowie die algorithmische Ausspielung ihrer Videos an ihnen nicht folgende Nutzer*innen, die ihnen häufiger weniger wohlgesonnen sind. Zudem führen sie die erhöhte Anzahl an Hasskommentaren auf das jüngere Publikum und damit verbunden ein geringes Wissen über Disabilities zurück. Die meisten Angriffe kämen von anonymen Profilen, entweder öffentlich in den Kommentaren oder über Privatnachrichten.

Im Gegensatz dazu zeichnet die Analyse von Rodríguez-González et al. (2025) ein deutlich positiveres Bild von TikTok: Während auf Youtube abwertende Kommentare über Mehrgewichtigkeit dominierten, waren die Kommentare auf TikTok von Zuneigung und Anerkennung für die verstorbene Schauspielerin Itziar Castro, die von Fatphobia und Queerfeindlichkeit betroffen war, geprägt. Die Autor*innen sprechen in diesem Zusammenhang von einem „Network of Love“ (S.2).

Was zunächst widersprüchlich wirkt, lässt sich entweder durch verbesserte Moderationstechniken oder durch das Konzept der „fragmentierten Öffentlichkeiten“ (Abidin, 2021) erklären. Durch die algorithmische Vermittlung von Inhalten entstehen zahlreiche Teilöffentlichkeiten mit eigenen Interessen, Werten und Normen. Treffen diese aufeinander,

kommt es häufig zu Konflikten, die sich in Hate und Counter Speech äußern können. An diese Fragmentierung schließt das Phänomen der „dogshows“ an, das laut Cabbuag und Abidin (2025) vor allem unter jungen Nutzer*innen auf TikTok entstanden und inzwischen weit verbreitet ist. Dabei werden Personen durch spöttische Kommentare und Memes öffentlich bloßgestellt und Ziel kollektiven Spotts. Begünstigt wird dieses Verhalten durch die Fragmentierung der Öffentlichkeit, da Widerspruch innerhalb solcher Teilgemeinschaften oft gering bleibt und abwertende Äußerungen als Humor gerechtfertigt werden. Wie Matamoros-Fernández (2023) betont, kann Humor sowohl Machstrukturen herausfordern als auch festigen – letzteres insbesondere dann, wenn er, wie meist bei „dogshows“ nach „unten“ statt nach „oben“ tritt (S.2). Cabbuag und Abidin (2025) sowie Matamoros-Fernández (2023) verweisen auf die weite Verbreitung von „legal-but-harmful“ Humorformen auf TikTok und warnen vor deren Auswirkungen, insbesondere für junge Nutzer*innen. Sie plädieren dafür, solche Formen von Humor ähnlich ernst zu nehmen wie Hassrede. Für TikTok ist dies besonders relevant, da laut Matamoros-Fernández (2023) Humor und Spiel das Zentrum ihrer Plattformkultur bilden.

Eine der wenigen vergleichenden quantitativen Studien, die TikTok berücksichtigt, kommt zu dem Ergebnis, dass auf X und Youtube mehr negative und toxische Kommentare auftreten, während TikTok, Facebook und Twitch ein insgesamt positiveres und kooperatives Kommunikationsklima aufweisen (Vizcaíno-Verdú & Contreras-Pulido, 2024). Für die Durchführung der Studie wurden vortrainierte, automatisierte Systeme eingesetzt, die auf alle erhobenen Daten der verschiedenen Plattformen angewendet wurden. Mögliche plattformspezifische Formen von Hass wurden dabei nicht berücksichtigt, sodass eine Verzerrung der Ergebnisse möglich ist.

TikTok zeichnet sich durch algorithmisch fragmentierte Öffentlichkeiten, eine Entpersonalisierung der Onlineerfahrung, intransparente Moderationspraktiken wie Wortfilter und die weite Verbreitung rechtlich zulässiger, aber schädlicher Formen von Humor sowie eine besonders junge Nutzer*innenschaft aus. Diese Merkmale prägen die Interaktionen auf der Plattform und beeinflussen auch die Erscheinungsformen von Hassrede. Bei Jugendlichen, die einen großen Anteil der Nutzer*innen auf TikTok ausmachen, legen Studien nahe, dass sie zwar häufig mit Hate Speech konfrontiert sind, jedoch meist als

Beobachtende und seltener als Täter*innen oder Betroffene (Kansok-Dusche et al., 2022). Zwischen schulischem Mobbing und Online-Hass besteht dabei ein Kontinuum, das einen fließenden Übergang zwischen analogem und digitalem aggressivem Verhalten schafft (Kansok-Dusche et al., 2022; Korol & Blaya, 2025). Online-Hass, der von Jugendlichen ausgeht, wird durch geringe Selbstkontrolle, Impulsivität, Sensationslust oder Rachemotive begünstigt, während Empathie und familiäre Nähe das Risiko verringern. Insgesamt entsteht Online-Hass im Jugendkontext aus einem komplexen Zusammenspiel individueller Dispositionen, sozialer Zugehörigkeiten und plattformspezifischer Dynamiken (Jiménez-Díaz & Del Rey, 2025).

Was die Forschung zu nicht-jugendlichen Personen angeht, die Hass in sozialen Medien verbreiten, so fokussiert sich diese vor allem auf organisierte Hassgruppen und deren strategisches Vorgehen zur Einschüchterung von Gegner*innen und Rekrutierung neuer Anhänger*innen (Siegel, 2020). Über individuelle Hater*innen ist hingegen wenig bekannt – wohl auch, weil sie zwar sichtbar, aber für die Forschung schwer zugänglich sind. Studien legen nahe, dass Hater*innen überwiegend männlich, unterschiedlichen Alters und sehr aktiv auf Plattformen sind (Siegel, 2020). Viele sind gut vernetzt, nutzen abwertende Inhalte um Reichweite, Aufmerksamkeit und Selbstbestätigung zu erhalten und bewegen sich in Online-Gemeinschaften, in denen hasserfüllte Diskurse normalisiert sind. Ob ihre Beiträge tatsächliche Überzeugungen widerspiegeln oder vor allem performativ motiviert sind, bleibt unklar. Die Intention, die hinter einem hasserfüllten Beitrag liegt, lässt sich nur schwer erfassen (Benesch et al., 2016). Psychologische Ansätze versuchen die Merkmale von Verfasser*innen hasserfüllter Inhalte zu greifen und liefern einige Erklärungsansätze, auf die an dieser Stelle nicht näher eingegangen werden kann (Sorokowski et al., 2020; P.Y.K.L. et al., 2024). Die Datengrundlage dieser Arbeit erlaubt keine tiefgehende Analyse soziodemografischer Faktoren, da lediglich die selbstgewählten Namen der Nutzer*innen vorliegen.

3.2.3. Formen von Hate Speech

Wie in vielen computerwissenschaftlichen Studien, klassifizieren Vizcaíno-Verdú und Contreras-Pulido (2024) Kommentare mithilfe von Natural Language Processing (NLP) automatisiert als „uncivil“. Dabei wird in der Regel auf bereits trainierte Modelle zurückgegriffen, ohne die zugrunde liegenden Annotationsschemata und Codebooks zu diskutieren. Diese sind jedoch entscheidend, da sie – basierend auf den verwendeten Definitionen – bestimmen, welche Formen von Hass in einer Analyse erfasst werden. Seemann et al. (2023) zeigen in ihrem Vergleich verschiedener Datensätze zu verletzender Sprache, dass die Annotationsschemata teilweise erheblich variieren. Ein umfassender Vergleich von Annotationsschemata zum Thema Hate Speech steht bislang noch aus.

Im Gegensatz zu anderen Studien, legen Fillies et al. (2025) die Details ihrer verwendeten Taxonomie offen. Ihr Datensatz ist der erste öffentlich verfügbare, annotierte Korpus deutschsprachiger TikTok-Kommentare. Datensätze mit TikTok-Kommentaren existieren insgesamt erst seit kurzer Zeit – einer der frühesten, allerdings englischsprachigen, stammt von Cools et al. (2024). Der Datensatz von Fillies et al. (2025) entstand mit Fokus auf Rechtsextremismus und wurde auf der Grundlage eines gemeinsam mit Expert*innen aus der Deradikalisierungsarbeit und Extremismusprävention entwickelten Annotationsschemas erstellt, das 32 Kategorien umfasst. Eine feste Definition von Hate Speech wird bewusst vermieden, stattdessen beziehen sich die Autor*innen auf gemeinsame Merkmale von Hate Speech Definitionen: Hate Speech umfasst sprachliche und bildbasierte Handlungen, sowie nicht-faktische Äußerungen, die darauf abzielen, Personen oder Gruppen aufgrund ihrer sozialen Zugehörigkeit oder politischen Haltung zu diskreditieren, verletzen oder einschüchtern. Sie beruht häufig auf stereotypisierenden Zuschreibungen – etwa rassistischer oder islamfeindlicher Art – und kann absichtlich oder unabsichtlich erfolgen. In digitalen Räumen äußert sie sich als Form sprachlicher Gewalt und als Ausdruck von Normverletzung oder „Uncivility“. Es folgt eine Abbildung ihrer Taxonomie, die unterschiedliche Formen von Hate Speech sichtbar macht.

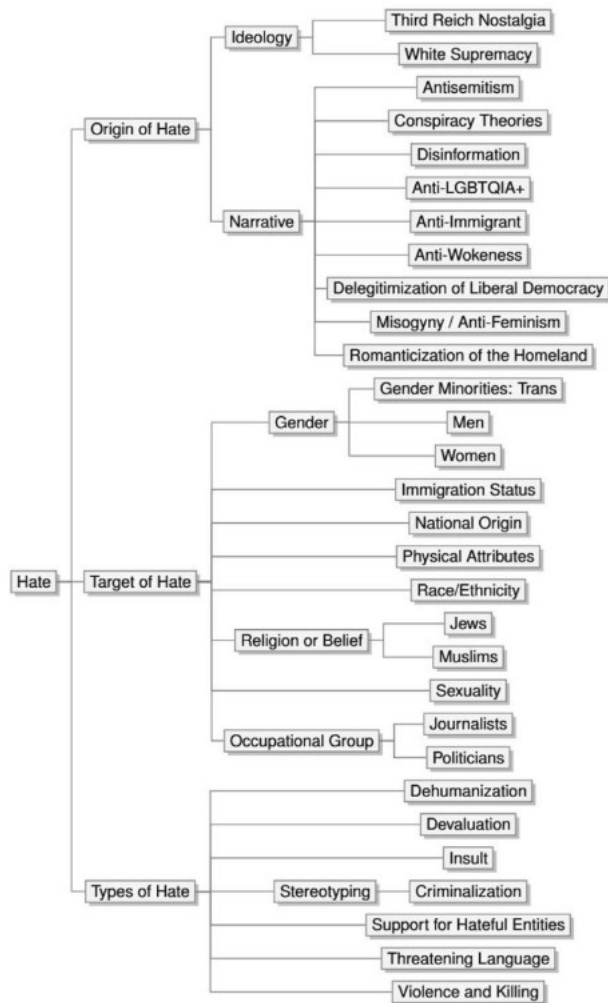


ABBILDUNG 2: TAXONOMIE ZUR HATE SPEECH KLASSIFIZIERUNG VON FILLIES ET AL. (2025, S. 7)

Mit einem ethnografischen Zugang wählt Lorenz (2025) eine ganz andere Herangehensweise an das Thema Hate Speech auf TikTok. Lorenz verbleibt bei der Darstellung der Ergebnisse allerdings auf der deskriptiven Ebene und richtet zudem den Fokus auf eine sehr spezifische Form hasserfüllter Inhalte, nämlich auf vergangenheitsbezogene Hassrede und die Relativierung von NS-Verbrechen.

3.3. Forschungsstand zu Counter Speech

Nachdem Counter Speech in wissenschaftlichen Studien zu Hate Speech einige Zeit lang vor allem als positiver und meinungsfreiheitsschonender Ausblick und Gegenentwurf zum Löschen von Kommentaren betrachtet wurde (Alsagheer et al., 2022), hat das Forschungsinteresse an dem Phänomen seit 2019 deutlich zugenommen. Der Korpus an Studien zum Thema Counter Speech bleibt jedoch deutlich kleiner als jener zu Hate Speech, umfasst aber ebenso wie jener zu Hate Speech Arbeiten aus unterschiedlichen wissenschaftlichen Disziplinen (Chung et al., 2024). Im Zentrum steht dabei meist die Frage nach der Wirksamkeit und den Erfolgsaussichten von Gegenrede (Buerger & Wright, 2021; Siegel, 2020). Wie im Forschungsfeld zu Hate Speech ist auch hier ein Prozess der Ausdifferenzierung zu beobachten, der sich auch in der Beschäftigung mit unterschiedlichen Definitionen und Formen von Counter Speech zeigt. Ein bislang wenig beleuchteter Aspekt betrifft die plattformspezifischen Besonderheiten von Counter Speech. Im Folgenden wird – mit einem Fokus auf TikTok - der aktuelle Forschungsstand zu den Definitionen und Formen von Counter Speech dargestellt.

3.3.1. Counter Speech Definitionen

Die Definitionsvielfalt hinsichtlich Counter Speech ist deutlich weniger ausgeprägt und erforscht als dies beim Thema Hate Speech der Fall ist. Dennoch gibt es auch hier einige Differenzen und konzeptuelle Abweichungen.

Die Arbeit von Benesch¹ et al. (2016) gilt als zentraler Bezugspunkt in der wissenschaftlichen Auseinandersetzung mit Counter Speech. Darin wird Counter Speech definiert als „a response that takes issue with hateful, harmful, or extremist content“ (S.14). Die Definition von Gegenrede hängt folglich eng mit dem Verständnis jener Äußerungen zusammen, auf die sie reagiert. Benesch et al. (2016) vermeiden den Begriff „Hate Speech“, da die Abgrenzung von allgemein beleidigender und verletzender Rede zu schwer, bzw. zu ungenau ist. Der Begriff „Hate Speech“ ist ihnen auch zu stark gerichtet und impliziert für sie ein Gefühl und eine

¹ Susan Benesch hat 2010 das „Dangerous Speech Project“ gegründet, das sich u.a. mit der Erforschung „Counter Speech“ beschäftigt. Durch Arbeiten vom „Dangerous Speech Project“ wurden wichtige Impulse zur Erforschung von Gegenrede gesetzt. Link zum Projekt: <https://www.dangerousspeech.org/dangerous-speech>

Intention des Hasses bei dem*r Sprecher*in, das in der Realität jedoch schwer nachweisbar ist. Benesch et al. (2016) bevorzugen den ihrer Ansicht nach präziseren Ausdruck „hasserfüllte Rede“.

Benesch et al. (2016) unterscheiden weiters zwischen „Counter Speech“ und „Counter Narrative“: Ersteres bezeichnet ausschließlich direkte Reaktionen auf problematische Äußerungen, während der Begriff „Counter Narratives“ allgemeinere Ausdrucksformen umfasst, die andere Narrative oder Weltbilder in Frage stellen. Das kann im Rahmen von Bildungsarbeit und öffentlicher Informationsvermittlung passieren, aber auch im Rahmen von Propaganda. Damit greifen die Autor*innen einen Aspekt auf, der in verschiedenen Definitionsansätzen mitunter unterschiedlich behandelt wird. Es handelt sich um die Frage, ob Äußerungen, die nicht unmittelbar auf einen Hasskommentar reagieren, aber in einem hasserfüllten oder potenziell feindlichen Diskurskontext stehen, als Counter Speech gelten können (Chung et al., 2024).

Mitglieder der Facebookgruppe #ichbinhier² werden beispielsweise dazu animiert, nicht direkt auf Hasskommentare zu reagieren, sondern eigene, sachliche Kommentare unter Facebook-Artikeln zu verfassen, um Solidarität mit Betroffenen zu zeigen, anstatt sich auf Hasskommentare einzulassen und diesen Reichweite zu geben. Diese Praxis weist eine konzeptuelle Nähe zum Ansatz „Hope Speech“ auf (Palakodety et al., 2020; Chakravarthi, 2022; García-Baena et al., 2023; Snyder et al., 2002).

Palakodety et al. (2020) definieren Hope Speech als Inhalte, die in Zeiten politischer Spannungen zur Deeskalation feindseliger Kommunikation in sozialen Medien beitragen. Chakravarthi (2022) versteht darunter Beiträge, die Unterstützung, Zuversicht und Inspiration vermitteln, während García-Baena et al. (2023) darunter Beiträge begreifen, die Gleichberechtigung, Vielfalt und Inklusion stärken. Chung et al. (2024) fassen diese Ansätze zusammen und betrachten Hope Speech als eine Form von Counter Speech, die einen positiven Austausch fördert, verletzender Sprache entgegenwirkt und Optimismus, Resilienz

² Die Facebookgruppe #ichbinhier ist der deutsche Ableger der schwedischen Initiative #jagärhär [„#ichbinhier“], die 2016 als von einer Privatperson gegründet wurde und in Schweden über 70.000 Mitglieder auf Facebook umfasst, die kollektiv auf hasserfüllte Online-Kommentare reagieren (Buerger, 2021). Nach einer mehrjährigen Pause ist der deutsche Ableger #ichbinhier wieder aktiv und zählt aktuell über 38.000 Mitglieder

sowie Diversität und Inklusion stärkt. Die Frage, ob auch eigenständige positive Kommentare ohne direkten Bezug zu Hassrede unter den Begriff Counter Speech fallen, bleibt in ihrer Metastudie jedoch unberücksichtigt.

Neben der Frage, ob auch eigenständige Kommentare als Counter Speech gelten, unterscheiden sich Definitionen darin, ob Counter Speech grundsätzlich nicht-aggressiv sein muss. Viele Studien schließen aggressives Antwortverhalten aus (Benesch et al., 2016; Chung et al., 2024) oder behandeln es als eigene Kategorie (Atzmüller et al., 2022). Neuere Arbeiten hingegen beziehen „Hostile Counter Speech“ ausdrücklich in den Begriff Counter Speech mit ein und entwickeln spezifische Unterkategorien (Hickey et al., 2024).

Ein weiterer Unterschied betrifft die Sprecher*innenposition. Während viele Studien diesen Aspekt nicht weiter berücksichtigen (Chung et al., 2024), zählen Atzmüller et al. (2022) nur jene Beiträge als Counter Speech, die von Bystander*innen verfasst werden, also von Personen, die nicht selbst unmittelbar Ziel des Hasskommentars sind. Dem gegenüber stehen Studien, die sich explizit mit Antwortverhalten von Betroffenen auseinandersetzen und dieses als eine Art Counter Speech deuten (García-Prieto et al., 2024).

Der vorliegenden Arbeit liegt – wie auch dem Forschungsprojekt „Young People Against Online Hate“ – die Definition von Counter Speech nach Atzmüller et al. (2022) zugrunde. Dabei ist zu berücksichtigen, dass diese Arbeit sich im Gegensatz zu anderen (García-Prieto et al., 2024) ausschließlich auf die Ebene von Kommentaren bezieht. Atzmüller et al. (2022) definieren Counter Speech als ein breites Spektrum kommunikativer Praktiken, die als Reaktion auf Hate Speech verfasst werden und darauf abzielen, diese zu konterkarieren oder zu entkräften. Counter Speech wird als Ausdruck zivilcouragierten Handelns verstanden, das den Grundstein für eine aktive, demokratische Online-Gesellschaft schafft, die auf humanistischen und prosozialen Werten beruht (Atzmüller et al., 2022, S. 13-14). Aggressive Formen von Gegenrede, auch „Counter Hate“ genannt, werden eingeschlossen und als eigenständige Ausprägung von Counter Speech betrachtet. Auch wenn sie auf den ersten Blick nicht mit humanistischen Idealen vereinbar erscheinen, können sie aus opferzentrierter Perspektive dem Schutz Betroffener dienen.

Eigenständige Kommentare ohne unmittelbar vorangegangene Hassäußerung werden im Rahmen dieser Arbeit nur dann als Counter Speech gewertet, wenn sie explizit auf hasserfüllte Inhalte verweisen, entweder mittels direkter Antwort oder eigenständigem Kommentar. Eine ähnliche Erweiterung zeigt sich in der aktualisierten Definition des Dangerous Speech Project, das seine Fassung von 2016 präzisiert hat und nun explizit auch Reaktionen einbezieht, die auf unterschiedliche Weise auf hasserfüllte Inhalte reagieren (Dangerous Speech Project, o. D.). Die Abgrenzung zwischen Bystander*innen und Betroffenen ist insbesondere bei gruppenbezogener Menschenfeindlichkeit – etwa bei Rassismus, Sexismus oder Queerfeindlichkeit – schwierig, da die Identität der Verfasser*innen meist anonym bleibt. Zudem können Bystander*innen selbst frühere Opfererfahrungen gemacht haben, wodurch die Grenzen weiter verschwimmen. In dieser Arbeit wird zwischen „Counter Speech“ und „Selbstverteidigung“ unterschieden, wenn erkennbar ist, Verfasser*innen unmittelbar zuvor persönlich angegriffen wurden oder ihre eigene Betroffenheit im Kommentar artikulieren. Es kann jedoch nicht ausgeschlossen werden, dass sich unter den als Counter Speech codierten Kommentaren auch Beiträge Betroffener befinden, sie selbst ihre Betroffenheit jedoch nicht explizit benennen. Die als „Selbstverteidigung“ codierten Kommentare werden aufgrund der vielen Parallelen gesondert berücksichtigt und in die Analyse von Gegenrede miteinbezogen.

3.3.2. Counter Speech auf TikTok

Die Arbeit von Benesch et al. (2016) basiert auf Daten von Twitter. Chung et al. (2024) weisen darauf hin, dass zum Zeitpunkt ihrer Erhebung keine Studien zu Counter Speech auf TikTok vorlagen. Im selben Jahr, in dem Chung et al. (2024) ihre Metastudie veröffentlichten, publizierten García-Prieto et al. (2024) eine Untersuchung video-basierter Reaktionen von Betroffenen auf Hasskommentare. Dabei nutzen sie eine spezifische TikTok-Funktion, die es Nutzer*innen ermöglicht, auf Kommentare unter ihren Videos mit einem neuen Video zu antworten, das automatisch auf den ursprünglichen Kommentar verweist. Aufgrund des klaren Fokus auf diese plattformeigene Praxis wurde keine zusätzliche Definition von Counter Speech oder verwandten Konzepten vorgenommen.

2025 wurden erste Ergebnisse eines Forschungsprojektes aus Italien und Deutschland vorgestellt, das mithilfe quantitativer Methoden die Verbreitung, Inhalte und Reaktionen auf Hate Speech gegen LGBTQ*-Gemeinschaften auf TikTok untersucht (Condom Tibau et al.,

2025) und inhaltliche Parallelen zum Projekt „Young People Against Online Hate“ aufweist. Neben dem Melden, Löschen, Blockieren und Ignorieren hasserfüllter Inhalte nennen die Autor*innen auch Bildung, Aufklärung und Counter Speech als mögliche Reaktionsformen. Counter Speech wird dabei als potenziell nachhaltigere Strategie diskutiert, die auf argumentative Auseinandersetzung statt auf Zensur setzt. Sie gilt als kollaborative Form der Reaktion, die durch alternative oder gegenläufige Narrative Toleranz stärkt, Vorurteile hinterfragt und kritisches Denken fördert.

Da ein großer Teil der TikTok-Nutzer*innen jugendlich ist, sind Erkenntnisse zum Umgang Jugendlicher mit Online-Hass relevant, auch wenn sie sich nicht explizit auf TikTok beziehen. Atzmüller et al. (2022) verweisen darauf, dass Jugendliche online deutlich seltener bei Übergriffen eingreifen als offline. Gründe dafür sind vor allem die Unsicherheiten in der Einschätzung von Situationen, geringe Empathie und mangelnde Bezüge zu den Opfern, unzureichende normative Wertehaltungen, antizipierte Wirkungslosigkeit sowie fehlende Kompetenzen. Unter Jugendlichen dominieren zudem aggressive und täter*innenbezogene Formen von Counter Speech. Das Bystander*innenverhalten variiert dabei nach Geschlecht, Bildung und Migrationshintergrund: Burschen und Jugendliche mit niedrigerem Bildungsniveau oder Migrationshintergrund greifen häufiger zur aggressiven Gegenrede, während Mädchen mit höherem Bildungsniveau eher moralisierend reagieren.

In den Definitionen von Counter Speech wurden einzelne Ausprägungen häufig bereits implizit angedeutet oder ausdrücklich benannt. Im folgenden Abschnitt wird die Vielfalt dieser Formen sowie der mögliche Einfluss plattformspezifischer Merkmale von TikTok auf die Gestaltung von Gegenrede näher betrachtet.

3.3.3. Formen von Counter Speech

Benesch et al. (2016) entwickeln in ihrem grundlegenden Werk nicht nur eine Definition, sondern auch ein mehrstufiges Modell von Counter Speech. Sie unterscheiden vier Typen von Counter Speech – auch „Vektoren“ genannt, eine Taxonomie mit acht Strategien sowie fünf Formen der Reaktion auf Counter Speech, auf die im Exkurs näher eingegangen wird.

Die vier Vektoren beschreiben die Gerichtetheit von Gegenrede: „One-to-One“, „One-to-Many“, „Many-to-One“ und „Many-to-Many“. Sie erfassen, ob eine einzelne Person ihre Aussage an eine oder mehrere andere Personen richtet oder ob ein Kollektiv von Personen auf eine Einzelperson, bzw. ein anderes Kollektiv reagiert. Ein Beispiel für Letzteres ist die Facebookgruppe #ichbinhier, die in ihren koordinierten Aktionen den gleichnamigen Hashtag #ichbinhier nutzen, um Sichtbarkeit zu erzeugen und ihr Vorgehen abzustimmen (Buerger, 2021).

Auf der Ebene der Taxonomie identifizieren Benesch et al. (2016) acht Strategien und verweisen dabei auf Parallelen zur Überzeugungs- und Kommunikationsforschung. Sie unterscheiden zwischen:

- der Präsentation von Fakten zur Korrektur falscher Aussagen oder Wahrnehmungen
- dem Aufzeigen von Widersprüchen oder Heuchelei
- dem Hinweis auf mögliche online und offline Folgen der Äußerung
- dem Ausdruck von Nähe oder Zugehörigkeit zum*r Verfasser*in des ursprünglichen Kommentars oder zur betroffenen Gruppe
- der Verurteilung der Äußerung als hasserfüllt oder gefährlich
- dem Einsatz visueller Medien
- dem Gebrauch von Humor
- der Wahl eines bestimmten Tonfalls, zum Beispiel eines empathischen

Benesch et al. (2016) betonen, dass in einem Counter Speech Kommentar mehrere Strategien gleichzeitig vorkommen können und sich nicht gegenseitig ausschließen. Während sie von „Strategien“ sprechen, wird in dieser Arbeit der Begriff „Formen“ bevorzugt, da Gegenrede nicht immer bewusst strategisch erfolgt, sondern auch spontan oder im Affekt entstehen kann (Atzmüller et al., 2022).

Das Zeichenlimit für Kommentare wurde inzwischen erhöht (Conectere, 2025), doch Links bleiben weiterhin nicht klickbar und müssen manuell in die Adresszeile kopiert werden – ein umständlicher Prozess, der eventuell dafür verantwortlich ist, dass er nur selten genutzt wird. Auch andere Plattformen zeigen Links zunehmend nur als Text an (Stevendie, 2025) –

vermutlich um zu verhindern, dass Nutzer*innen die jeweilige Plattform verlassen und dadurch ihre Verweildauer sinkt. Einschränkungen bestehen möglicherweise zudem bei der Kategorie „Verurteilung der Äußerung als hasserfüllt oder gefährlich“, da TikTok intransparente Wortfilter einsetzt und Nutzer*innen eigene Filter für Kommentare auf ihrem Account aktivieren können. Diese Wortfilter haben in der Vergangenheit nicht nur beleidigende Begriffe, sondern auch Bezeichnungen wie „homophob“ blockiert, die zur Benennung von Hassrede verwendet werden können (Eckert et al., 2022; Diwisch, 2022).

In den Folgejahren nach ihrer Veröffentlichung griffen zahlreiche Forschende die Taxonomie von Benesch et al. (2016) auf und entwickelten sie weiter. Chung et al. (2024) geben hierzu einen Überblick und halten folgende Counter Speech Formen fest: Einbringen von Fakten, Aufzeigen von Widersprüchen, Hinweis auf Konsequenzen, Zugehörigkeit oder Affiliation, Verurteilung, Gegenfragen, Humor und positiver Ton. Zudem betonen sie, dass das Ziel eines Kommentars dessen Ausprägung beeinflusst. Ihre Metastudie unterscheidet drei Hauptziele: eine Veränderung in der Einstellung und dem Verhalten von Täter*innen, eine Einflussnahme auf die Einstellung und das Verhalten von Bystander*innen sowie die Unterstützung der von Hass betroffenen Personen.

Diese Einteilung entspricht weitgehend der Differenzierung von Atzmüller et al. (2022), die zwischen täter*innen-, opfer- und bystander*innenbezogenen Strategien sowie solchen unterscheiden, die sich an alle richten. Sie identifizieren folgende Formen von Counter Speech:

- Täter*innenbezogene Strategien: Warnung, Abwertung, Aufforderung aufzuhören, Faktencheck, Moralischer Appell, Verteidigung des Opfers – an Täter*innen adressiert, Deeskalation, Ablenkung, Empathie, Offline Lösung, Kontakt zu Täter*innen über Privat-Chat
- Opferbezogene Strategien: Opfer öffentlich in Schutz nehmen, Opfer im Privat-Chat kontaktieren
- Bystander*innenbezogene Strategien: Gegenrede anderer Bystander*innen unterstützen, Mit Bystander*innen in Privat-Chat absprechen
- An alle gerichteten Strategien: Einen Schlussspunkt vorschlagen, Deeskalation
- Zusätzlich: Aggressive Gegenattacken

Bisher gibt es keine Kategorisierung oder tiefgehende Analyse von Counter Speech auf der Kommentarebene von TikTok. Vorliegende Studien zu Counter Speech auf TikTok befassen sich entweder mit anderen Formaten (García-Prieto et al., 2024), bleiben rein deskriptiv (Lorenz, 2025) oder nutzen für die Annotation Modelle, die auf Datensätzen anderer Plattformen trainiert wurden (Condom Tibau et al., 2025). Sowohl Chung et al. (2024) als auch Condom Tibau et al. (2025) weisen angesichts plattformspezifischer Unterschiede auf den akuten Mangel an Forschung zu Counter Speech auf TikTok hin.

3.4. Forschungsstand zu Interaktionen von Hate Speech und Counter Speech

Interaktionen zwischen Hate Speech und Counter Speech werden in der Forschung meist im Hinblick auf die Wirksamkeit von Gegenrede diskutiert (Chung et al., 2024). Dieser Erfolg lässt sich allerdings nur schwer erfassen. Häufig richten Studien ihren Fokus auf das Verhalten von Hater*innen und auf die Frage, ob Gegenrede zu einer messbaren Abnahme von Hate Speech führt. Die Befunde hierzu fallen widersprüchlich aus (Ernst et al., 2017) oder deuten auf eine Zunahme hasserfüllter Beiträge nach einer Counter Speech Intervention hin (Garland et al., 2022).

Als Reaktion auf die widersprüchlichen Ergebnisse kam es zu einem Prozess der Ausdifferenzierung, in dessen Verlauf der Fokus stärker auf einzelne Formen von Gegenrede gelegt wurde (Chung et al., 2024). Hangartner et al. (2021) identifizieren hier die empathiebasierte Gegenrede als besonders vielversprechend, während Bilewicz et al. (2021) eine Verringerung verbaler Aggressionen insbesondere bei den Counter Speech Formen „disapproval“, „abstract norm“ und „empathy“ feststellen. Die Interpretation dieser Befunde ist eng an die verwendeten Begriffsbestimmungen und an die konkrete Ausgestaltung der jeweiligen Counter Speech Formen gebunden.

Die Wirksamkeit von Gegenrede kann jedoch auch auf andere Weise gemessen werden. Condom Tibau et al. (2025) betrachten bereits die bloße Existenz von Counter Speech als positiv, unabhängig von ihrer konkreten Ausprägung. Erfolg kann auch daran festgemacht werden, ob eine Intervention weitere Gegenrede anstößt oder ob sie die negativen Folgen von Hate Speech für Betroffene verringert (Buerger, 2022). Hickey et al. (2024) zeigen wiederum, dass aggressive Formen von Gegenrede effektiv darin sein können, Hater*innen aus

bestimmten digitalen Räumen zu verdrängen. Diese Entwicklung kann Betroffene entlasten, birgt jedoch das Risiko einer zunehmenden Radikalisierung, wenn sich Hater*innen in stärker abgeschottete digitale Umgebungen zurückziehen.

Insgesamt zeigt sich, dass der Verlauf nach einer Counter Speech Intervention auf verschiedenen Ebenen untersucht werden kann. Analysen können sich auf die Entwicklung eines einzelnen Threads oder auf die gesamte Kommentarspalte beziehen. Der Fokus kann auf den Hater*innen liegen oder auf den Mitlesenden, auf anderen Counter Speaker*innen oder auf den Betroffenen selbst. Darüber hinaus kann Forschung entweder Counter Speech als Ganzes untersuchen oder den Blick gezielt auf einzelne Ausprägungen richten. Letztlich hängt jede Bewertung der Wirksamkeit davon ab, wie Hate Speech, Counter Speech und deren einzelne Ausprägungen definiert sind und anhand welcher Kriterien Wirksamkeit bestimmt wird.

4. Datengrundlage und methodisches Vorgehen

Die vorliegende Arbeit basiert auf Daten, die im Rahmen des Forschungsprojektes „Young People Against Online Hate – Computergestützte Strategien zur Unterstützung von Gegenrede in sozialen Medien“ erhoben wurden. Im Fokus der Arbeit stehen Hate Speech und Counter Speech Auseinandersetzungen auf Kommentar-Ebene. Die 3.948 TikTok-Kommentare aus denen der Datensatz besteht, wurden in Zusammenarbeit mit der FH St. Pölten gecrawlt und anschließend mit der qualitativen Inhaltsanalyse nach Kuckartz & Rädiker (2024) ausgewertet. Das folgende Kapitel gibt Einblick in die einzelnen Schritte der Datenerhebung und -auswertung.

4.1. Erstellung und Aufbereitung des Datensatzes

Die Analyse basiert auf 3.948 TikTok-Kommentaren zu sechs öffentlich verfügbaren Videos. Da TikTok keine Schlagwortsuche in Kommentaren erlaubt, erfolgte die Auswahl über eine Videosuche mit Schlagwörtern, die gemeinsam mit Jugendlichen und einem*r Social-Media-Experten*in erarbeitet wurden. Die thematische Fokussierung auf LGBTQIA+ und Germany's

next Topmodel³ (GNTM) ergab sich einerseits aus dem Anspruch, sowohl gruppen- als auch personenbezogene Menschenfeindlichkeit abzubilden und andererseits aus projektinternen Befragungen, in denen Jugendliche diese Themen wiederholt als besonders konfliktbeladen und stark von Hassrede betroffen beschreiben.

Auswahlkriterien für die Videoauswahl:

- Aktueller Veröffentlichungszeitpunkt
- Thematische Relevanz
- Überwiegend deutschsprachige Kommentare
- Interaktion von Hate Speech und Counter Speech Kommentaren

Die Identifikation aktueller, thematisch passender Videos war herausfordernd, da die TikTok-Suchfunktion – im Gegensatz zur personalisierten und algorithmusgetriebenen ForYou-Page⁴ – eher ältere Videos anzeigt, die zu einem früheren Zeitpunkt viral gegangen sind. Alle ausgewählten Videos stammen aus dem Zeitraum zwischen März und Oktober 2023. Zudem lieferte die Schlagwortsuche vielfach unpassende Ergebnisse, da TikTok nur die kurzen Captions⁵ und Accountnamen durchsucht und vermutlich Wortfilter einsetzt. Aufgrund der Jugendschutzbestimmungen⁶ von TikTok wurden kaum Videos von Jugendlichen angezeigt.

Ausgewählt wurden drei Videos zum Themenbereich LGBTQIA+ bzw. Queerfeindlichkeit und drei zum Thema GNTM. Die GNTM-Videos stammen vom Account einer Kandidatin der Staffel 2023, die stark polarisierte und auf Social Media wiederholt Ziel von Online-Hass wurde. Thematisch befassen sie sich mit ihrem Umstyling zu einer Kurzhaarfrisur, einer lobenden

³ Germany's next Topmodel (GNTM) ist eine von Heidi Klum moderierte deutsche Castingshow, in der Kandidat*innen wöchentlich gegeneinander antreten, bis am Ende ein*e Gewinner*in gekürt wird. Das Format erzielt regelmäßig hohe mediale Aufmerksamkeit und ist häufig Gegenstand kontroverser Online-Diskussionen. Immer wieder kommt es im Zusammenhang mit der Sendung zu Fällen von Cybermobbing. Besonders bekannt wurde der Fall Lijana Kaggwas, die 2020 während des Live-Finales aus Protest gegen die ihr widerfahrende digitale Gewalt ausstieg. Seither leistet sie Aufklärungsarbeit zum Thema Cybermobbing (Kaggwa, 2023).

⁴ Die ForYou-Page ist die stark personalisierte Startseite von TikTok, auf der Nutzer*innen automatisch Videos angezeigt werden, die der Algorithmus von TikTok für sie aussucht.

⁵ Bei einer „Caption“ handelt es sich um eine Art Begleittext oder Beschreibung inkl. Hashtags, die Nutzer*innen zu ihrem Video dazu posten. 2022 erhöhte TikTok das Limit von 300 auf 2.200 Zeichen für eine Caption. Im September 2023 wurde das Limit auf 4.000 angehoben.

⁶ Link zu „Einstellungen zu Privatsphäre und Sicherheit für Teenager*innen“:

<https://support.tiktok.com/de/account-and-privacy/account-privacy-settings/privacy-and-safety-settings-for-users-under-age-18>

Bemerkung einer Fotografin und ihrem Rauswurf aus der Sendung. Die Auswahl wurde gemeinsam Jugendlichen im Alter von 14 bis 18 Jahren reflektiert, die bestätigten, vergleichbare Inhalte und Kommentare auch in ihren eigenen Feeds zu sehen. Die LGBTQIA+-Videos zeigen ein Statement für Akzeptanz und Gleichberechtigung, einen Beitrag, der traditionelle Geschlechternormen infrage stellt, sowie ein Video eines bekannten deutschsprachigen Youtubers, der beim CSD in Köln provokativ und abwertend mit queeren Personen interagiert. Fünf der Videos sind nur wenige Sekunden lang, eines dauert rund eine Minute und zwanzig Sekunden. Die Zahl der Kommentare pro Video reicht von 67 bis 1095.

Das Crawling der Kommentare erfolgte am 3. Juni 2024 in Kooperation mit der FH St. Pölten mithilfe eines eigens entwickelten Web-Scraping-Programms, das die Daten als Excel-Datei ausgab. Insgesamt wurden 3.948 Kommentare erfasst, inklusive Text, Likes, Datum, Kommentar-ID, Nickname⁷, Profil-Link samt Username und der Information, ob es sich um einen eigenständigen Kommentar („parent“) oder einen Antwortkommentar („child“) handelt. Die Kommentare sind weder im Datensatz noch auf TikTok chronologisch geordnet. Eigenständige Kommentare scheinen algorithmisch – vermutlich anhand von Likes und Interaktionen – gereiht zu werden, während Antwortkommentare innerhalb eines Threads absteigend nach der Anzahl ihrer Likes sortiert sind. Da TikTok keine genauen Zeitstempel und in der Desktop-Version keine klaren Antwortbezüge anzeigt, lässt sich der Gesprächsverlauf nur stark eingeschränkt rekonstruieren. In der App wird zwar angezeigt, auf welche Nutzer*innen sich ein Kommentar bezieht, nicht jedoch, auf welchen konkreten Kommentar im Thread. Dieser Umstand wird besonders umständlich, wenn mehrere Personen mehrfach in einem Thread kommentieren. Zudem wird unter einem eigenständigen Kommentar lediglich die Anzahl der darunterliegenden Antwortkommentare angezeigt. Um die Antworten einzusehen, müssen sie manuell in Dreiergruppen ausgeklappt werden.

Ein Crawling über die App war technisch nicht möglich. Die Zahl der erfassten Kommentare wich von der zum Zeitpunkt des Crawlings auf TikTok angezeigten Anzahl ab. Die Ursachen

⁷ TikTok unterscheidet zwischen „Nicknames“ (auch „Profilname“ genannt) und „Usernames“ (auch „Benutzername“ genannt). Beide beziehen sich auf das gleiche TikTok-Profil, jedoch dient der Benutzername als eindeutiger Identifikator und nur einmal in exakt dieser Form vorhanden sein darf, während der Profilname nicht eindeutig sein muss und mehrere Profile den gleichen Nickname haben können.

dafür sind unklar, könnten jedoch auf technische Probleme beim Laden oder Crawlen der Daten sowie auf gelöschte oder ausgeblendete Kommentare zurückzuführen sein.

Soziale Medien sind ein vergleichsweise junges Forschungsfeld und bislang gibt es nur wenige etablierte Methoden zur Erhebung, Aufbereitung und Analyse entsprechender Daten (Jones et al., 2024). Eine zentrale Herausforderung besteht in der Verarbeitung großer Datenmengen (McKenna et al., 2017). Es werden Programme wie MAXQDA oder NVivo häufig für die qualitative Datenanalyse empfohlen (Jones et al., 2024; Kuckartz & Rädiker, 2024), wohingegen sich die an der Universität Wien verfügbaren Programme MAXQDA und Atlas.ti für den vorliegenden Datensatz als ungeeignet erwiesen. Die Excel-Datei wurde automatisch als quantitativer Datensatz interpretiert, wodurch eine Auswertung der Diskussionsstrukturen oder Metadaten wie Nicknames und Veröffentlichungszeitpunkte nicht möglich war. Ein Grund für die bislang geringe Forschungsaktivität zu TikTok-Daten (Chung et al., 2024) dürfte in der unzureichenden technischen Unterstützung durch gängige Analyseprogramme liegen.

Aus den genannten Gründen erfolgte die Codierung der erhobenen Kommentare manuell in Excel. Dieses Vorgehen ist zwar aufwendiger und potenziell fehleranfälliger, ermöglicht jedoch eine differenziertere Auswertung. Zur Vertiefung der Analyse wurden weitere Variablen eingeführt, etwa zur Position der Antwortkommentare im Thread sowie zur Kommentarhäufigkeit pro Tag und pro Nutzer*in.

4.2. Datenanalyse

Zur Analyse der 3.948 Kommentare kam die inhaltlich strukturierende qualitative Inhaltsanalyse nach Kuckartz und Rädiker (2024) zum Einsatz. Als meistverwendete Methode zur Analyse von Social-Media-Kommentaren (Alafwan et al., 2023), bietet sich die Inhaltsanalyse an, weil sie theoretisch anschlussfähig ist, mit kleinteiligen Sinneinheiten arbeitet und eine hohe Nähe zum Material gewährleistet (Kuckartz und Rädiker, 2024).

Die inhaltlich strukturierende qualitative Inhaltsanalyse nach Kuckartz und Rädiker (2024) wird in sieben Phasen angewendet, die in folgendem Schaubild abgebildet sind. Ihren Ausgangspunkt nehmen diese sieben Phasen in den Forschungsfragen, die jeden der weiteren Schritte begleiten.

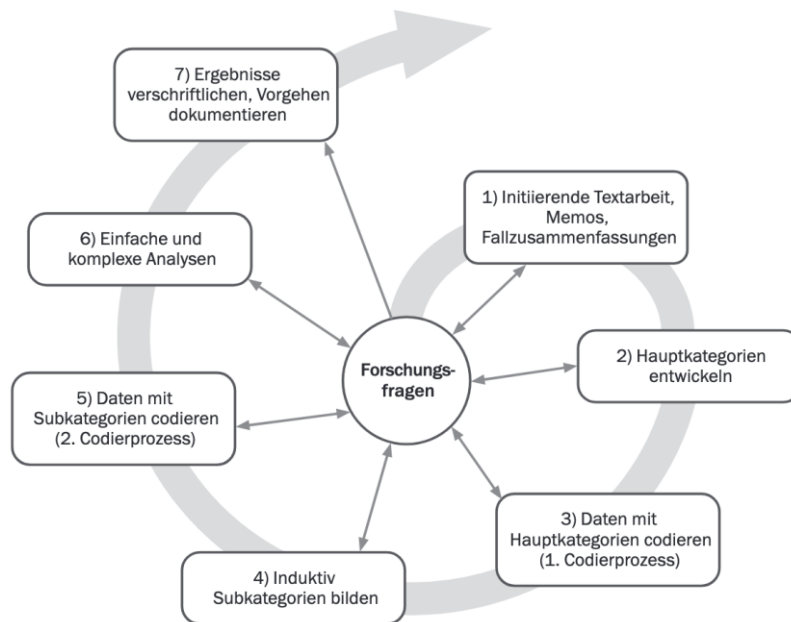


ABBILDUNG 3: ABLAUF EINER STRUKTURIERENDEN QUALITATIVEN INHALTSANALYSE IN SIEBEN PHASEN NACH KUCKARTZ UND RÄDIKER (2024, S.132)

Phase 1: Initiierende Textarbeit, Memos, Fallzusammenfassungen

Die ersten Auswertungsschritte – bestehend aus initiiender Textarbeit, Memos und Fallzusammenfassungen – erfolgten im Rahmen des Forschungsprojektes „Young People Against Online Hate“. Auf Vorarbeiten aufbauend (Atzmüller et al., 2022) wurde ein Analyseraster erstellt, in dem Metadaten, Videoinhalte, gesellschaftlicher Kontext und Charakteristika der Kommentarspalten festgehalten wurden. So konnte ein Überblick über das Material gewonnen werden und erste Ansätze für Codes und Kategorien entwickelt werden (Kuckartz & Rädiker 2024).

Phase 2: Entwicklung von Hauptkategorien

Die inhaltliche Strukturierung der Daten erfolgte in zwei aufeinander aufbauenden Kodierungsphasen. Im ersten Schritt wurden acht Hauptkategorien gebildet, die deduktiv aus dem theoretischen Bezugsrahmen, dem aktuellen Forschungsstand und den Forschungsfragen abgeleitet und im Rahmen eines Probelaufs anhand der Kommentare des ersten Videos ausprobiert wurden:

- **Hate Speech:** alle Formen hasserfüllter Online-Handlungen zur Herabsetzung anderer
- **Counter Speech:** kommunikative Reaktionen auf Hate Speech zu deren Entkräftung oder Gegendarstellung
- **Positive Speech:** Kommentare, die positive Gefühle wie Zustimmung, Dank, Empathie, Bewunderung oder Einsicht ausdrücken und weder Hate Speech noch Counter Speech sind.
- **Negative Speech:** Kommentare mit sachlicher und konstruktiver Kritik, die nicht als Hate Speech oder Counter Speech einzuordnen sind.
- **Affected Speech:** Kommentare, in denen Nutzer*innen persönliche Erfahrungen mit Abwertung oder Diskriminierung teilen. Im Vordergrund steht die eigene Verletzlichkeit, nicht die Reaktion auf Hass.
- **Selfdefensive Speech:** Reaktionen unmittelbar Betroffener auf Hass zur Selbstverteidigung. Nähe zu Counter Speech, aber aus Betroffenenperspektive.
- **Neutral Speech:** Kommentare ohne erkennbare Wertung, wie etwa neutrale Fragen oder Account-Markierungen ohne zusätzliche Information.
- **Unclear Speech:** Kommentare, deren Bedeutung oder Intention unklar bleibt.

Für die Beantwortung der Forschungsfragen sind insbesondere die Hauptkategorien „Hate Speech“, „Counter Speech“, „Positive Speech“ und „Selfdefensive Speech“ zentral.

Phase 3: 1. Codierprozess – Codierung der Daten mit den Hauptkategorien

Im ersten Codierprozess wurden alle 3.948 TikTok-Kommentare als eigene Sinneinheiten einer der acht Hauptkategorien zugeordnet und in der Excel-Datei festgehalten. Einem hermeneutischen Zugang folgend, wurden neben dem Kommentartext auch Video- und Threadkontext berücksichtigt, da sich viele Bedeutungen erst im Zusammenhang erschließen lassen. So kann ein scheinbar positiver Kommentar wie „Gute Besserung“ im Kontext als abwertende Reaktion auf ein Coming-out gemeint sein. Die Kategorien wurden disjunkt angewendet. Der erste Codierdurchlauf diente einer groben Einteilung, während im zweiten, präziseren Prozess Mehrfachzuordnungen berücksichtigt und die Erstcodierungen überprüft wurden.

Phasen 4 und 5: 2. Codierprozess – Bildung von induktiven Subkategorien und Codierung der Daten mit Codes

Die Phasen 4 und 5 wurden angepasst, da viele Kommentare – insbesondere Antwortkommentare – nur im Kontext sinnvoll interpretierbar waren. Im zweiten Codierprozess wurde ein offenes Codiervorgehen angewendet, bei dem Codes direkt am Material entwickelt und anschließend zu übergeordneten Kategorien zusammengeführt wurden. Kategorien stehen dabei auf einer höheren Abstraktionsebene und können mehrere Codes umfassen. Um frühzeitige Verallgemeinerungen zu vermeiden, blieben Codes eng am Text und wurden in Memos ergänzt. Jeder Kommentar wurde einzeln codiert, wobei maximal vier Codes pro Kommentar vergeben wurden. Insgesamt entstanden über 250 Codes, die den acht Hauptkategorien – insbesondere Hate Speech, Counter Speech und Positive Speech – zugeordnet und systematisch zu weiteren Ober- und Subkategorien verdichtet wurden. Die folgende Grafik veranschaulicht den Prozess des Codierens und der Kategorienbildung.

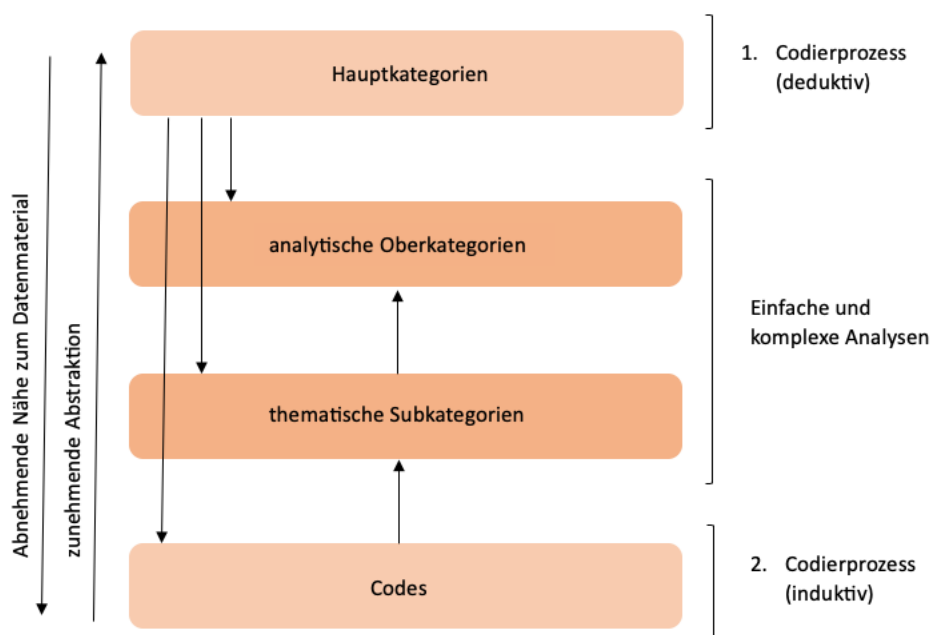


ABBILDUNG 4: EIGENE DARSTELLUNG DES PROZESSES DER KATEGORISIERUNG NACH KUCKARTZ UND RÄDIKER (2024)

Phase 6: Einfache und komplexe Analysen

Die sechste Phase umfasst eine vertiefte Analyse, der im zweiten Kodierungsprozess gewonnenen Codes und der Bildung eines hierarchischen Kategoriensystems zur Bestimmung der Formen von Hate Speech und Counter Speech auf TikTok. Auf Grundlage der Codes wurden thematische und analytische Kategorien entwickelt und systematisch zusammengeführt. Während thematische Kategorien auf einer deskriptiven Ebene operieren, erfassen analytische Kategorien übergeordnete Zusammenhänge und Strukturen. Die Bildung solcher Kategorien ist ein kreativer, erfahrungsbasierter Prozess, der Transparenz und Selbstreflexion im Umgang mit Vorwissen erfordert. Mein eigenes Vorwissen als wissenschaftliche Mitarbeiterin im Projekt „Young People Against Online Hate“, als online Kommentarmoderation bei einer großen deutschsprachigen Tageszeitung und als ausgebildete Counter Speakerin prägte den Analyseprozess und ermöglichte eine zugleich sensible und kritisch-reflektierte Perspektive auf die untersuchten Kommunikationsformen. Kuckartz und Rädiker nennen sieben zentrale Kriterien für die Bildung von Kategorien und Kategoriensystemen, die in der vorliegenden Arbeit berücksichtigt wurden:

- (1) Enge Anbindung an die Forschungsfragen oder relevantes Kontextwissen
- (2) Vollständige Abdeckung aller relevanten Aspekte
- (3) Präzise und trennscharfe Definitionen von Kategorien
- (4) Klare und sprachlich konsistente Formulierungen
- (5) Innere Kohärenz des Kategoriensystems
- (6) Hierarchische Beziehung zwischen Ober- und Subkategorien
- (7) Verständlichkeit und Nachvollziehbarkeit auch für fachfremde Personen

Im Anschluss an die Kategorienbildung wurden die Zusammenhänge zwischen den Subkategorien, insbesondere innerhalb der Hauptkategorien, untersucht und in die finalen Kategoriedefinitionen integriert. Dabei stand das Verhältnis sowie die Abgrenzung der einzelnen Kategorien zueinander im Fokus. Ergänzend ermöglichten vertiefende Einzelfallanalysen einzelner Threads Einblicke in die Dynamiken zwischen Hate Speech und Counter Speech. Die Ergebnisse werden im folgenden Kapitel dargestellt, das zugleich die Phase 7 der Analyse – die Verschriftlichung der Ergebnisse – umfasst.

5. Ergebnisse

In diesem Kapitel werden die Ergebnisse der Inhaltsanalyse vorgestellt. Zunächst werden einige deskriptive Befunde skizziert, bevor das doppelte Kategoriensystem präsentiert wird. Etwa die Hälfte aller Kommentare wurde als Hate Speech und rund 15 % wurde als Counter Speech identifiziert. Letzteres entspricht etwa 600 Kommentaren, von denen rund die Hälfte unter dem Video zum Rauswurf der polarisierenden GNTM-Kandidatin verfasst wurden. Dieser Kommentarbereich weist mit etwa 30% den höchsten Anteil an Gegenrede auf. Insgesamt variiert der Anteil an Counter Speech unter den verschiedenen Videos zwischen 3% und 30%. Ohne das Video mit dem niedrigsten Ergebnis liegt die Spannweite bei 10% bis 30%. Auffällig ist, dass das Video mit dem geringsten Anteil an Gegenrede vergleichsweise hohe Werte für Positive, Affected und Selfdefensive Speech aufweist.

Weniger als ein Prozent aller Kommentare wurden als Affected oder Selfdefensive Speech identifiziert. Nur wenige Kommentare stammten erkennbar von Betroffenen selbst. Das bedeutet jedoch nicht, dass Betroffene nicht mitgelesen oder sich möglicherweise geäußert haben, ohne ihre Betroffenheit offen zu zeigen. Etwa 10% der Kommentare entfallen auf Positive Speech. In drei Threads kam es nach Counter-Speech-Intervention zu Entschuldigungen von Hater*innen. Negative Speech wurde in weniger als einem Prozent der Fälle beobachtet, was darauf hinweist, dass sachliche Kritik deutlich seltener geäußert wurde als hasserfüllte Kommentare.

Etwa 10% aller Kommentare wurden als Neutral Speech und rund 15% als Unclear Speech identifiziert. Neutral Speech umfasst vor allem neutrale Fragen und einfache Accountsmarkierungen ohne inhaltliche Aussage. Der hohe Anteil an Unclear Speech resultiert daraus, dass Antwortkommentare häufig nicht eindeutig zugeordnet werden können, da genaue Zeitstempel und Informationen zum Bezugskommentar fehlen. So kann ein Kommentar wie „same“ je nach Kontext Zustimmung zu Hate oder Counter Speech ausdrücken. Einige Kommentare waren zudem in anderen Sprachen verfasst oder enthielten unverständliche Zeichenfolgen.

Für die beiden Hauptkategorien Hate Speech und Counter Speech wurden jeweils sechs Oberkategorien mit mehreren Unterkategorien herausgearbeitet. Beginnend mit Hate Speech

werden im Folgenden die erarbeiteten Kategorien vorgestellt und anhand exemplarischer Kommentare aus dem Datenmaterial veranschaulicht. Das doppelte Kategoriensystem basiert ausschließlich auf empirischen Befunden und erhebt keinen Anspruch auf theoretische Vollständigkeit. Zur Veranschaulichung wurden schematische Darstellungen beider Systeme erstellt. Anschließend folgen Anmerkungen zu weiteren Hauptkategorien sowie eine Zusammenfassung der Ergebnisse der Einzelfallanalysen.

5.1. Hate Speech

Für die Hauptkategorie „Hate Speech“ wurden insgesamt sechs Oberkategorien identifiziert, die jeweils drei bis maximal neun Unterkategorien umfassen. Es gilt zu beachten, dass ein Kommentar gleichzeitig mehreren Unterkategorien verschiedener Oberkategorien zugeordnet werden kann. Die folgende Abbildung veranschaulicht die Struktur des Kategoriensystems und bietet einen Überblick über dessen innere Gliederung.

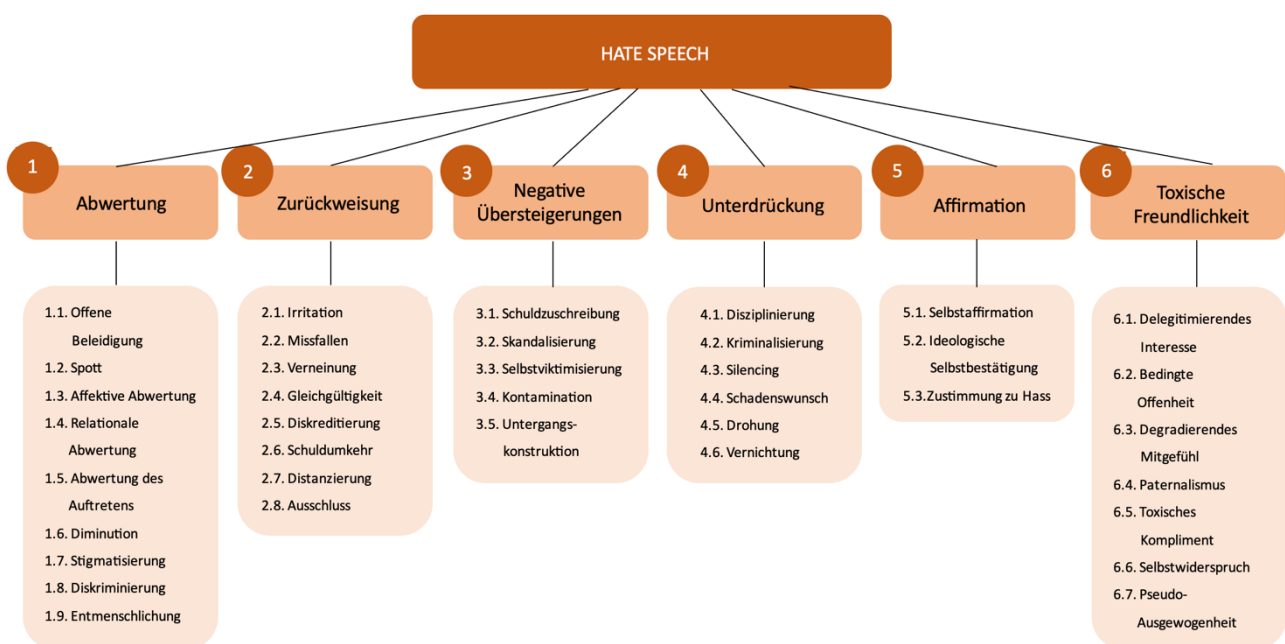


ABBILDUNG 5: EIGENE DARSTELLUNG DES ENTWICKELTEN KATEGORIENSYSTEMS FÜR HATE SPEECH

1.) Abwertung

Diese Oberkategorie umfasst unterschiedliche Formen sprachlicher Herabsetzung, die Personen, Gruppen oder Inhalte in ihrer Bedeutung, Legitimität und gesellschaftlichen Anerkennung untergraben. Sie reicht von Beleidigungen, Spott und affektiven Abwertungen über Pathologisierungen und gruppenspezifische Abwertungen bis zu Entmenschlichungen. Die Vielzahl an Subkategorien zeigt, wie flexibel Sprache zur Stigmatisierung und Abwertung eingesetzt wird. Gemeinsam ist ihnen die Konstruktion und Abwertung von Differenz, durch die soziale Hierarchien fortgeschrieben und stabilisiert werden.

1.1. Offene Beleidigung

Offene Beleidigungen sind direkte Formen sprachlicher Abwertung, die Personen oder Gruppen mit verletzenden oder entwürdigenden Ausdrücken angreifen. Sie verzichten auf rhetorische Tarnung und zielen auf die Herabsetzung von Würde, Status oder Identität ab. Oft dienen sie der Machtdemonstration und Einschüchterung.

Beispiele

- „L“ (V.1, Z.786) [Anmerkung: „L“ steht für „Loser“, bzw. „Verlierer*in“]
 - „npc“ (V1., Z.549) [Anmerkung: „npc“ steht für „non-playing character“ und bezeichnet Personen, die als unwichtig oder bedeutungslos gelten.]
 - „Unsympathisches pack“ (V.5, Z.520)
-

1.2. Spott

Diese Unterkategorie umfasst abwertende Ausdrucksformen, die auf Spott, Ironie, Übertreibung oder Schadenfreude basieren. Durch rhetorische Mittel wie Antiphrasen, Verzerrungen, Bedeutungsumkehrungen und übertriebenen Zuspitzungen werden Personen und Themen ins Lächerliche gezogen und so ihrer Ernsthaftigkeit und Glaubwürdigkeit beraubt.

Beispiele

- „sieht gar nicht mal so gut aus 🔥 🔥 🔥 “ (V.2, Z.51)
 - „Ich identifiziere mich als Katze. Ich brauche endlich ein katzenklo in der schule. Das ist Diskriminierung. Maumaumau“ (V.3, Z.995)
 - „Verdient 🍌 “ (V.6, Z.306) [als Antwort auf Ausscheiden einer Kandidatin bei GNTM]
-

1.3. Affektive Abwertung

Affektive Abwertung nutzt negative Emotionen wie Abscheu, Wut oder Hass als Mittel der Herabsetzung. Anstelle einer argumentativen Auseinandersetzung werden Themen, Personen oder Gruppen emotional aufgeladen, um Ablehnung und Abwertung zu erzeugen.

Beispiele

- „ihhhhhhh“ (V.1, Z. 373)
 - „So viel Tüten gibt’s gar nicht, wie ich kotzen will“ (V.3, Z.952)
 - „wtf 🤢 🤢 🤢 “ (V.5, Z.42)
-

1.4. Relationale Abwertung

Relationale Abwertung erfolgt durch abwertende Vergleiche, die soziale Hierarchien erzeugen und stabilisieren. Inhalte, Personen und Gruppen werden indirekt herabgesetzt, indem sie in eine abwertende Beziehung zu anderen gesetzt werden.

Beispiele

- „Wasser ist Wasser aber warum trinkt ihr nicht von der Toilette“ (V.1, Z.919)
[Anmerkung: als Antwort auf „Love is Love“]
- „Ich bin eig ein loser auf ganzer Linie aber wenn ich das sehe fühle ich mich so überlegen 😏 “ (V.3, Z.364)
- „Gibst mir mehr vibes wie so ne Rossmann Mitarbeiterin:)“ (V.5, Z.491)

1.5. Abwertung des Auftretens

Diese Subkategorie bezeichnet kommunikative Praktiken, die den Fokus vom Inhalt einer Aussage auf das analoge oder digitale Erscheinungsbild des Gegenübers verschieben – etwa auf Schreibweise, Profilgestaltung oder andere Äußerlichkeiten. Diese Verschiebung relativiert die inhaltliche Aussage und entwertet sie über die Herabsetzung der Person oder Gruppe.

Beispiele

- „oke aber warum heulst du“ (V.1, Z.417)
 - „Dieser schnurrbart 😄“ (V.3, Z.913)
 - „Wieso hat sie bei 8 k Follower ein blauen Haken 😄“ (V.5, Z.487)
 - „Google kann dir auch helfen den Unterschied zwischen das und dass zu lernen 😊“ (V.3, Z.459)
-

1.6. Diminution

Diminution bezeichnet eine Form der Abwertung, die Personen oder ihre Aussagen durch Infantilisierung oder Trivialisierung belächelt und dadurch ihre Relevanz mindert. Auf diese Weise wird Betroffenen Reife, Ernsthaftigkeit und Glaubwürdigkeit abgesprochen.

Beispiele

- „das UwU hat gefehlt“ (V.1, Z.592)
 - „Nices Bärtchen brosky“ (V.3, Z.8)
 - „Also deine Leistung ist top, aber dein Sozialverhalten? naja du führst dich manchmal auf wir eine 12 Jährige in der Pupertät“ (V.5, Z.86)
-

1.7. Stigmatisierung

Diese Kategorie umfasst sprachliche Praktiken, in denen Personen oder Gruppen durch die Zuschreibung von Abweichung, Mangel, Verfehlung oder Krankheit abgewertet werden. Durch

Verweise auf gesellschaftliche oder gesundheitliche Normen werden sie als defizitär, irrational oder moralisch verwerflich markiert und sozial herabgestuft.

Beispiele

- „Gute Besserung 🙄💊“ (V.1, Z.136)
- „Das kann doch nicht mehr normal sein...“ (V.3, Z.23)
- „Kopf total kaputt“ (V.3, Z.578)
- „Kann aber auch nicht ganz gesund sein oder. Dafür muss es auch ne Klinik geben..“ (V.3, Z. 623)

1.8. Diskriminierung

Die Kategorie Diskriminierung erfasst sprachliche Abwertungen, die sich auf gruppenspezifische Merkmale wie Geschlecht, sexuelle Orientierung, Herkunft, Religion, Behinderung oder Alter beziehen. Sie äußert sich in sexistischen, rassistischen, antisemitischen, ableistischen oder queerfeindlichen Zuschreibungen, die Personen aufgrund tatsächlicher oder zugeschriebener Zugehörigkeiten herabsetzen und auf spezifischen gesellschaftlichen Stereotypen und Machtverhältnissen beruhen.

Beispiele

- „Mach dir um Deutsche keine sorgen 👍. sind meistens nur die Goldstücke die auf solche menschen los gehen 🗑️.“ (V.2, Z.33)
- „leute in dem kommentar bitte man beleidigt keine behinderten“ (V.3, Z.841)
[Anm.: „behinderten“ wird hier als Schimpfwort gegen queere Menschen verwendet]
- „Das kleine Mädchen ist wieder bockig....“ (V.4. Z. 119)

1.9. Entmenschlichung

Entmenschlichung ist eine extreme Form der Abwertung, bei der Personen ihre Menschlichkeit abgesprochen wird. Dies geschieht durch Gleichsetzungen mit Tieren, Objekten oder durch

das sprachliche Leugnen ihrer Existenz. So wird die betroffene Person nicht nur entwertet, sondern auch in ihrer Identität fundamental delegitimiert.

Beispiele

- „aber ich fühle mich nicht wie eine Person... ich bin ein Pferd 🐎“ (V.1, Z.349)
- „linke zecken“ (V.3, Z. 100)
- „dem ❌ der ❌ das ✅“ (V.3, Z.255)

2.) Zurückweisung

Die Oberkategorie Zurückweisung umfasst verschiedene Ausdrucksformen, in denen positive Inhalte, Identitäten oder Zugehörigkeiten nicht angenommen, abgeschwächt oder abgelehnt werden. Sie wirkt meist subtil und legt den Schwerpunkt eher auf Abwehr und Blockierung als auf Abwertung – auch wenn stets ein abwertendes Moment mitschwingt. Die Zurückweisung operiert über die Subkategorien der Irritation, Missfallen, Verneinung, Gleichgültigkeit, Diskreditierung, Distanzierung und Ausschluss.

2.1. Irritation

Die Subkategorie Irritation umfasst Äußerungen, die positive Inhalte durch Verwunderung, Skepsis oder scheinbar naive Nachfragen infrage stellen. Dabei erfolgt kein argumentativer Widerspruch, sondern eine subtile Unterminierung von Anerkennung und Zugehörigkeit, die positive Diskurse destabilisiert.

Beispiele

- „Ich werd das nie verstehen 🤔“ (V.3, Z. 680)
- „Hä“ (V.3, Z.905)
- „😬😬😬“ (V.5, Z. 601)

2.2. Missfallen

Missfallen bezeichnet eine sprachliche Form der Ablehnung, die sich über negative Urteile oder Einschätzungen äußert. Die Zurückweisung erfolgt über subjektive Meinungen, Geschmacksurteile oder knappe negative Bewertungen, ohne weitere Begründung oder Erläuterung.

Beispiele

- „geht gar nicht. Diese Frisur“ (V.4, Z.107)
 - „Ich finde halt, egal welche Frisur, null wandelbar 🧑“ (V.4, Z.124)
 - „Unsympathisch“ (V.5, Z.238)
-

2.3. Verneinung

Verneinung umfasst sprachliche Praktiken, die positive Aussagen abstreiten oder ins Gegenteil verkehren. Dazu zählen die Ablehnung von Lob und Selbstbehauptungen, die Nicht-Akzeptanz von Vielfalt, die Zurückweisung von Zugehörigkeitsansprüchen sowie das Absprechen von Identitäten oder Zuständen. So werden Anerkennung und Zugehörigkeit verweigert.

Beispiele

- „solltest du nicht sein 🔥 🔥 🔥 “ (V.1, Z.98) [Anm.: bezieht sich auf einen Kommentar, in dem eine Personen ihren Stolz auf ihre lesbische Identität ausdrückt]
- „sowas würd ich niemals sagen“ (V1, Z.483) [Anm.: bezieht sich auf eine Aussage, in der sich für universelle Toleranz ausgesprochen wird]
- „falsche worte 🤔🤔“ (V.1, Z.869) [Anm.: bezieht sich auf eine Aussage, die für universelle Toleranz plädiert]
- „Nein 🙅 “ (V.5, Z.602) [Anm.: bezieht sich auf ein Kompliment]

2.4. Gleichgültigkeit

Die Unterkategorie Gleichgültigkeit erfasst Ausdrucksweisen, die durch minimale Gesprächsbeteiligung, Desinteresse und fehlende Empathie geprägt sind. Sie zeigt sich in der Ignoranz gegenüber Diskriminierung und Leid, spielt die Bedeutung des Gesagten herunter und stabilisiert bestehende Macht- und Ungleichheitsverhältnisse.

Beispiele

- „ok“ (V.1., Z.364)
 - „aha“ (V.1, Z.534)
 - „sie ist weder hübsch noch sympathisch.... aber auch nicht unterhaltsam oder skandalös. einfach langweilig.kann weg, wollen wir nicht sehen.“ (V.4, Z.127)
 - „a dir kann meine Meinung ja egal sein wen juckt das jetzt?“ (V.5, Z.6)
-

2.5. Diskreditierung

Diskreditierung bezieht sich auf sprachliche Praktiken, die andere Personen und ihre Positionen durch den Vorwurf der Lüge, Täuschung oder Einbildung delegitimieren – ohne sich inhaltlich mit deren Aussagen zu befassen.

Beispiele

- „Vom Albaner glaube nicht“ (V1., Z.278)
 - „würdest trotzdem nicht date machen 😏“ (V.1, Z.1029)
 - „Die Kommentare alle gekauft“ (V.5, Z.220)
 - „Lüge“ (V.5, Z.288)
-

2.6. Schuldumkehr

Schuldumkehr beschreibt Aussagen, in denen Verantwortung für abwertende oder diskriminierende Aussagen von sich gewiesen und auf den Betroffenen zugeschoben wird.

Kritik wird heruntergespielt oder Abwertungen als bloße Meinungsfreiheit umgedeutet. So werden Täter-Opfer-Verhältnisse verzerrt und eine ernsthafte Auseinandersetzung mit Kritik verhindert.

Beispiele

- „digga du hasts als erstes gesagt?“ (V.4, Z.66)
 - „Ja ok hate is uncool aber ihr Verhalten erst“ (V.5, Z. 93)
 - „Wer sich in die Öffentlichkeit begibt muss auch mit Kritik umgehen können. 😂“ (V.175)
 - „Selber schuld“ (V.6, Z.698)
-

2.7. Distanzierung

Die Unterkategorie Distanzierung umfasst kommunikative Praktiken, in denen Nutzer*innen den Wunsch äußern, sich von bestimmten Personen, Gruppen oder Inhalten abzugrenzen. Dies zeigt sich durch Ablehnung von Nähe, das Markieren räumlicher oder sozialer Distanz oder die demonstrative Abkehr von gemeinsamen Bezugspunkten.

Beispiele

- „ich wechsele zu YT Shorts wegen den schwulen und so“ (V.1, Z.431)
 - „Ok bei Gay hab ich aufgehört zu schauen will nicht wissen was danach gekommen is“ (V.1, Z.1061)
 - „🏃“ (V.3, Z.222)
 - „Ich will auf dem Mond wohnen 🌕“ (V.3, Z.695)
-

2.8. Ausschluss

Mit der Unterkategorie Ausschluss werden Ausdrucksweisen bezeichnet, die Personen oder Gruppen aktiv aus Gemeinschaften drängen. Diese Praktiken zeigen sich in direkten Zurückweisungen, Aufforderungen zum Rückzug oder der sprachlichen Konstruktion

getrennter Räume. Im Unterschied zur Distanzierung, geht es nicht nur um Abstand und Selbstabgrenzung, sondern um Abweisung und Ausschluss.

Beispiele

- „eig söttemer es eigenes land mache für die“ (V.3, Z. 1084)
- „zuerst mal selber umarmen 😊“ (V.4, Z.50) [Anmerkug: Ausschluss aus Empathiezugeständnis]

3.) Negative Übersteigerungen

Die Oberkategorie Negative Übersteigerungen umfasst Formen hasserfüllter Rede, die gesellschaftliche Entwicklungen, Gruppen oder einzelne Personen dramatisieren, negativ überhöhen oder verzerrt darstellen. Sie arbeitet mit Übertreibung, Alarmismus und falschen Verantwortlichkeiten und reicht von pauschalen Schuldzuschreibungen über Skandalisierung und Selbstviktimisierung bis hin zu Kontaminations- und Untergangsnarrativen. Gemeinsam ist ihnen die übersteigerte Konstruktion von Gefahr, die Betroffene delegitimiert und gesellschaftliche Vielfalt problematisiert.

5.2. Schuldzuschreibung

Die Unterkategorie Schuldzuschreibung bezeichnet eine Praktik, bei der Personen oder Gruppen pauschal für gesellschaftlich negativ bewertete Ereignisse oder Handlungen verantwortlich gemacht werden. Solche Zuschreibungen greifen den eigentlichen Inhalt nicht auf, sondern knüpfen über Scheinkorrelationen an unerwünschte oder verhasste Entwicklungen an. Anders als die Unterkategorie der Schuldumkehr, die Verantwortung für das eigene Handeln abwehrt und verschiebt, konstruiert die Schuldzuschreibung neue, oft willkürliche Verantwortlichkeiten und erklärt ganze Personengruppen kollektiv zu Sündenböcke für Missstände.

Beispiele

- „Ach aber Hauptsache die dürfen Religionen in den Dreck ziehen?!“ (V.1, Z.41)
 - „Mir ist es auch egal aber manchmal macht LGBTQ denn Koran verbrennen und sie erzählen irgendwas von und warum sollten wir sie noch mögen““ (V.1, Z.946)
 - „die verbieten abayas aber sowas nicht ich krieg nh krise“ (V.3, Z.244)
-

3.1. Skandalisierung

Skandalisierung umfasst Ausdrucksweisen, die Inhalte oder Personen übertrieben, dramatisiert oder als unangemessen darstellen. Durch diese Überzeichnung verlieren sie an Ernsthaftigkeit und verlieren ihren Status als legitimer Diskussionsbeitrag.

Beispiele

- „Omg 🐼🐼🐼🐼“ (V.1, Z. 866)
 - „Leuten geht es viel zu gut, dass sie sich Probleme suchen“ (V.3, Z.880)
 - „Du bist so nervig omg“ (V.5, Z. 434)
-

3.2. Selbstviktimsierung

Im Zuge der Selbstviktimsierung inszenieren sich Kommentierende als eingeschränkt, zensiert oder unterdrückt und stellen sich als leidend dar, um die eigenen abwertenden Haltungen zu legitimieren und Kritik abzuwehren. Das behauptete Leid beschreibt keine reale Marginalisierung, sondern soll die eigene Position aufwerten und bestehende Machtverhältnisse verschleiern. Anders als bei der Unterkategorie der kollektiven Zerfallsnarrative steht hier das individuelle Erleben im Vordergrund, während andere Positionen als übergriffig dargestellt werden.

Beispiele

- „ICH KANN DAS ALLES NICHT MEHR 🤖“ (V.1, Z. 521)
- „Möge Gott mich vor solchen Kinder beschützen“ (V.3, Z. 223)

- „Stimmt deswegen zwingt man anderen eine neue „Sprache“ auf usw. brauch man ja jetzt nicht alles aufzählen“ (V.3, Z.634)
- „Mein Therapeut brauch selbst bald einen“ (V.3, Z.860)
- „Ich Spar mir meine Meinung, sonst heulen wieder diese ganzen diversen und andere Versionen von Menschen.“ (V.3, Z. 1043)

3.3. Kontamination

Kontaminationskonstruktionen bezeichnen kommunikative Praktiken, die bestimmte Gruppen, Identitäten oder Ausdrucksformen als verunreinigend oder gefährlich markieren. Oft ist die Rede von „vergifteten“ Orten oder Generationen. Als Vorstufe umfassender Niedergangserzählungen erzeugen sie ein Bild von ansteckender Gefahr, das Vielfalt und Differenz zur Bedrohung macht.

Beispiele

- „Harram“ (V.1, Z.296)
- „Ich hab nicht gesagt das ich sie nicht respektiere ich respektiere es nicht was das mit der Gesellschaft macht“ (V.3, Z.43)
- „Gefahr ! Hiermit werdet ihr gewarnt das es gefährlich sein kann mit soETWAS zu kommunizieren.“ (V.3, Z.1071)

3.4. Untergangskonstruktion

Untergangskonstruktionen rahmen gesellschaftliche Entwicklungen oder die Präsenz bestimmter Personen und Gruppen als Zeichen eines umfassenden Niedergangs. Im Zentrum steht die Vorstellung, dass „das Ganze“ – etwa die Gesellschaft, Nation oder Menschheit – wegen der Existenz bestimmter Personen oder Gruppen zerfällt. Sprachlich kann sich dies in Untergangsmetaphern und Diagnosen kollektiver Verlorenheit zeigen.

Beispiele

- „Wir sind verloren..“ (V.3, Z.25)
 - „Ich hab Angst um meine Kinder... Früher hatte man Angst dass die Kids Kontakt zu "den bösen" hat, heute hat man Angst dass die Kids "zu denen gehören wollen" 🤖“ (V.3, Z.109)
 - „Menschheit stirbt aus...Warrrrum nur ? 😡😡😡😡“ (V.3, Z.662)
 - „es tut weh das anzuschauen das nenne ich armes deutschland und mit dennen sollen wir in den 3. WK na alles klar feuer frei“ (V.3, Z.791)
-

4.) Unterdrückung

Die Oberkategorie Unterdrückung umfasst sprachliche Ausdrucksformen, die darauf ausgerichtet sind Handlungsspielräume, Stimmen, gesellschaftliche Teilhabe oder Existenzen bestimmter Personen oder Gruppen zu begrenzen. Sie zielt auf eine Einschränkung ihrer Möglichkeiten ab und zeigt sich in Disziplinierung, Kriminalisierung, Silencing, ideologischer Dominanzbekundung, Schadenswünschen, Drohungen und Vernichtungsfantasien.

4.1. Disziplinierung

Die Unterkategorie Disziplinierung bezeichnet sprachliche Handlungen, die das Verhalten oder die Identität anderer „korrigieren“ oder normieren wollen. Abweichungen gelten dabei als veränderbar – etwa durch äußeren Druck, Selbstdisziplin oder eine Rückkehr zu vorherrschenden Normen. Besonders problematisch ist dies bei Phänomenen wie Sexualität oder Religion, die dadurch als Phase oder Fehlentwicklung dargestellt werden.

Beispiele

- „Kehre zum glauben zurück es ist nicht zu spät“ (V.1., Z.86)
- „Be straight“ (V.1, Z.829)
- „Komm raus aus deiner phase...“ (V.3, Z.84)

- „Alter die alle sollen IRGENDWAS sinnvolles machen man“ (V.3, Z.799)
- „Dann benimm dich vielleicht auch mal so und zick nicht wegen allem rum.“ (V.5, Z.236)

4.2. Kriminalisierung

Kriminalisierung bezeichnet Äußerungen, die Personen oder Gruppen als verbrecherisch darstellen und ihnen damit Legitimität und Zugehörigkeit absprechen. Die Spannweite reicht von unterschwelligen Assoziationen bis hin zu Forderungen nach Strafe oder Haft. Anders als Kontamination, die Gruppen als moralisch „verderbend“ oder gefährlich rahmt, stellt Kriminalisierung sie explizit als ein strafrechtliches Problem dar.

Beispiele

- „Ja die Welt ist ja so schön bunt 😊 😊 😊 ist ja schon fast alles erlaubt.“ (V.2, Z.61)
- „Eigene Sekte dieser Pronomen Verein“ (V.3, Z.805)
- „bitte kann jemand dieses etwas ins gefängnis stecken? 🗑️ 🗑️“ (V.3, Z.1046)

4.3. Silencing

Silencing umfasst sprachliche Praktiken, die Betroffene direkt zum Schweigen bringen oder ihnen das Recht auf Mitsprache absprechen sollen. Im Zentrum steht die Unterdrückung des Sprechens selbst – etwa durch Lächerlichmachung oder Aufforderungen zu schweigen. So wird das Sprechen von Betroffenen als illegitim, überflüssig oder nicht diskussionswürdig markiert.

Beispiele

- „Sei leise und geh zu deiner mom“ (V.1, Z. 46)
- „ 😬 “ (V.1, Z. 91)
- „hf“ (V.1, Z.768) [Anm.: Abkürzung für „halt die fresse“]
- „Schweigen ist Gold“ (V.5, Z.547)

- „Digga keiner hat mit dir geredet mach du mal ein Spaziergang bei Facebook“ (V.6, Z.174)
-

4.4. Schadenswunsch

Schadenswünsche sind Kommentare, die anderen aktiv Scheitern, Ausschluss oder Benachteiligung wünschen. Im Unterschied zur Schadensfreude beziehen sich Schadenswünsche auf zukünftige Ereignisse und sprechen Betroffenen die Möglichkeit, ihr Potenzial und ihre Identität zu entfalten ab.

Beispiele

- „Ich hoffe sie fliegt sie nervt mega“ (V.5, Z.215)
 - „Ist sie schon draußen?“ (V.5, Z.290)
 - „Schade das sie noch nd draußen ist“ (V.5, Z.356)
-

4.5. Drohung

Unter Drohung fallen Äußerungen, in denen negative Konsequenzen oder Sanktionen angekündigt werden, um Druck auf Betroffene auszuüben. Gewalt kann angedeutet werden, dient hier jedoch in erster Linie der Einschüchterung, Disziplinierung und Machtdemonstration.

Beispiele

- „Ich will aber nicht nett zu dir sein“ (V.1, Z.664)
 - „@TB ich glaube sogar hannes würde den letzten hobs nehmen“ (V.3, Z. 877)
-

4.6. Vernichtung

Die Unterkategorie Vernichtung umfasst Äußerungen, die die vollständige Eliminierung von Personen oder Gruppen fordern oder als wünschenswert darstellen. Die Forderungen oder

Wünsche äußern sich in Anspielungen auf Massenvernichtung, Suizidaufforderungen und in Rückgriffen auf historische Vernichtungspraktiken.

Beispiele

- „Gdu“ (V.1, Z.83) [Anmerkung: Abkürzung für „geh dich umbringen“]
- „der Islam wird siegen 😊“ (V.1, Z.247)
- „Früher wurde sowas anders erledigt“ (V.3, Z.817)
- „Eng und Duschkopf wäre besser.“ (V.3, Z.848) [Anm.: Anspielung auf den Holocaust]
- „Franz, get the Flammenwerfer! We have to defend our Lebensraum!“ (V.3, Z.962)
- „Mit so ner 88 Flak da rein zu halten kommt schon böse“ (V.3, Z.971) [Anm.: hier handelt es sich um ein bestimmtes Waffenmodell, das vor allem während des Zweiten Weltkrieges von den Nationalsozialisten eingesetzt wurde]

5.) Affirmation

Die Oberkategorie Affirmation umfasst Ausdrucksformen, die hasserfüllte Positionen durch Zustimmung, Bestärkung oder positive Bezugnahme stützen. Sie wirken nicht über direkte Abwertung, sondern über die Aufwertung eigener problematischer Weltbilder oder die Zustimmung zu bestehendem Hass. So werden hasserfüllte Äußerungen normalisiert und als sozial anschlussfähig dargestellt.

5.1. Selbstaffirmation

Mit Selbstaffirmation sind Beiträge gemeint, in denen Kommentierende sich demonstrativ erhöhen und dadurch andere implizit oder explizit abwerten. Durch diese Selbsterhebung entsteht eine hierarchische Differenz, die da Gegenüber herabsetzen und die eigene Position aufwerten soll.

Beispiele

- „ich fühl mich auf einmal wie der schlaueste Mensch“ (V.3, Z.461)

- „Ich studiere dann mal psychologie und rette die heutige Menschheit“ (V.3, Z.1011)

5.2. Ideologische Selbstbestätigung

Ideologische Selbstbestätigung umfasst Äußerungen, die diskriminierende, exkludierende oder biologistische Positionen positiv darstellen und als selbstverständlich oder überlegen präsentieren. Sie erfolgt häufig über politische, religiöse oder weltanschauliche Bezüge sowie über Verweise auf Autoritäten, Parteien oder symbolische Zeichen. Im Zentrum steht die Machtdemonstration, die argumentative Auseinandersetzung verdrängt und hegemoniale Normen sowie Ausschlusslogiken weiter festigt.

Beispiele

- „Jo wollte nur sagen das der Regenbogenfür den himmel ist nicht für dein geschlecht und es gibt eh nur 2“ (V.1, Z.441)
- „I am a Mann I will become a solider wenn I will grow op #sigma 🏆“ (V.1, Z.481)
- “@🌈 @Alissa bin homophobic 🏳️ 🏳️ 🏳️” (V.1, Z.739)
- „  “ (V.2, Z.51) [Anm.: Farben der Reichsflagge]
- „embrace masculinity, we need andrew tate more then ever“ (V.2, Z. 58)
- „  “ (V.3, Z.812) [Anm.: blaue Herzen stehen für die AfD]
- „Gott sei Dank bald AFD 😊“ (V.3, Z.964)

5.3. Zustimmung zu Hass

Zustimmung zu Hass zeigt sich in Kommentaren, die andere hasserfüllte Kommentare aufgreifen und bestärken. Das geschieht durch bejahende Kommentare, Wiederholungen, spöttische Anschlusskommunikation oder zustimmende Emojis. Solche Reaktionen normalisieren und verstärken Hate Speech, indem sie hasserfüllte Positionen als gemeinschaftlich geteilt und sozial anschlussfähig erscheinen lassen.

Beispiele

- „same“ (V.1, Z. 793)
 - „👉👉“ (V.3, Z.366)
 - „Bestes Kommentar seit es diese Diskussion gibt 😂😂😂“ (V.3, Z.515)
 - „w“ (V.3, Z.626) [Anmerkung: „w“ steht für „win“]
 - „bester mann“ (V.3, Z.646)
 - „endlich mal eine die es so sieht wie ich“ (V.5, Z.148)
-

6.) Toxische Freundlichkeit

Die Oberkategorie Toxische Freundlichkeit bündelt sprachliche Praktiken, die höflich, mitfühlend oder neugierig wirken, aber dennoch Abwertung transportieren und verschleiern. Delegitimierendes Interesse, bedingte Offenheit, degradierendes Mitgefühl, Paternalismus und ähnliche Formen tragen dazu bei, hasserfüllte Botschaften sozial anschlussfähig wirken zu lassen.

6.1. Delegitimierendes Interesse

Unter delegitimierendes Interesse fallen Nachfragen, die zunächst neugierig wirken, jedoch darauf hinauslaufen, Identitäten, Zugehörigkeiten oder Verhaltensweisen infrage zu stellen und zu problematisieren. Dabei wird ein Eindruck von Offenheit erzeugt, während tatsächlich Zweifel, Skepsis oder Spott vermittelt und Ablehnung verschleiert wird.

Beispiele

- „Aber was haben die menschen dann unten?“ (V.3, Z.1035)
- was heißt alle ? Hund Auch ? (V.3, Z.879)
- „Wie pflanzen sich die vor wenn die geschlechtslos sind“ (V.3, Z.950)
- „Arbeitest du?“ (V.3, Z.63)

6.2. Bedingte Offenheit

Zur Unterkategorie der bedingten Offenheit gehören Kommentare, die zunächst Zustimmung oder Toleranz signalisieren, dann jedoch in Abwertung und Abgrenzung umschlagen. Es entsteht der Eindruck von Offenheit, die jedoch nur unter den Bedingungen der Sprechenden gilt.

Beispiele

- „was hab ich falsch gemacht ? ich liebe diese Person aber nicht ihre sünde 🥰“ (V.1, Z.10)
 - „Ich habe nichts dagegen wenn die Leute sich als was anderes fühlen. Das ist so ein kranker Trend geworden das man das ist das es einfach nur noch“ (V.3, Z.42)
 - „Hä verstehe ich nicht 😂 also schwul bin und lesbisch okay ...aber was ist der Rest ?“ (V.3, Z.1032)
-

6.3. Degradierendes Mitgefühl

Degradierendes Mitgefühl umfasst Kommentare, die scheinbar empathisch klingen, aber sarkastisch oder abwertend gemeint sind. Statt echter Anteilnahme wird Mitleid genutzt, um Betroffene lächerlich zu machen und ihre Identität als bedauernswert darzustellen.

Beispiele

- „Die Leute die gay dind tun mir leid ich stehe immer noch auf Frauen 🤔“ (V.1, Z.477)
- „Ich glaube; dass diese Welt mittlerweile zu viel für einige geworden ist und man versucht sich mithilfe von neuen Identiäten etwas mehr Kontrolle *“ (V.3, Z.656)
- „so traurig was bei der Jugend im Kopf abgeht“ (V.3, Z.838)

6.4. Paternalismus

Paternalismus zeigt sich in Äußerungen, die zwar wohlwollend klingen, aber als Belehrungen aus einer Position angenommener Überlegenheit formuliert sind. Betroffene werden bevormundet und in ihrer Selbstbestimmung herabgesetzt.

Beispiele

- „möge Jesus dir helfen“ (V.1, Z.8)
-

6.5. Toxisches Kompliment

Toxische Komplimente wirken zunächst positiv, enthalten jedoch implizite Beleidigungen und Einschränkungen. Die scheinbare Anerkennung kippt so ins Gegenteil und erzeugt eine subtile Herabsetzung.

Beispiele

- „Warum nicht besser kurz als lang und ungepflegt“ (V.3, Z.46)
 - „Hauptsache du fühlst dich wohl 🧡“ (V.4, 105)
 - „Hübsch bist du ja auch jeder zählt der charakter auch“ (V.5, Z. 441)
-

6.6. Selbstwiderspruch

Die Unterkategorie Selbstwiderspruch zeigt sich in Kommentaren, die Abwertung zunächst ausdrücklich verneinen, sie anschließend jedoch trotzdem äußern. Durch diese Vorwegnahme soll der abwertende Gehalt verschleiert und Kritik an der eigenen Aussage abgewehrt werden.

Beispiele

- „Boh no hate aber Anya nh warum bist du so zickig“ (V.4, Z.77)
- „Heyy,ich will dich nicht haben sondern es nur erklären !💗 Die anderen Leute die dich haben mögen dich nicht weil du dich halt anders verhältst als di““ (V.5, Z.468)

6.7. Pseudo-Ausgewogenheit

Zur Pseudo-Ausgewogenheit gehören diskursive Praktiken, die Kritik oberflächlich auf alle Beteiligten verteilen und dadurch Täter*innen entlasten. Der Eindruck von Neutralität verschiebt bzw. relativiert die Schuld und Verantwortung für hasserfüllte Aussagen.

Beispiele

- „Digga einfach alle unangenehm in dem Video inklusive dem moderator“ (V.3, Z.952)

5.2. Counter Speech

Im Folgenden werden die im Datensatz identifizierten Formen von Counter Speech vorgestellt, die zu sechs Oberkategorien gebündelt wurden. Wie das Kategoriensystem für Hate Speech, erhebt auch dieses keinen Anspruch auf theoretische Vollständigkeit, da es ausschließlich empirisch beobachtete Formen von Counter Speech umfasst. Die folgende Abbildung veranschaulicht die Gliederung des Kategoriensystems für Counter Speech.

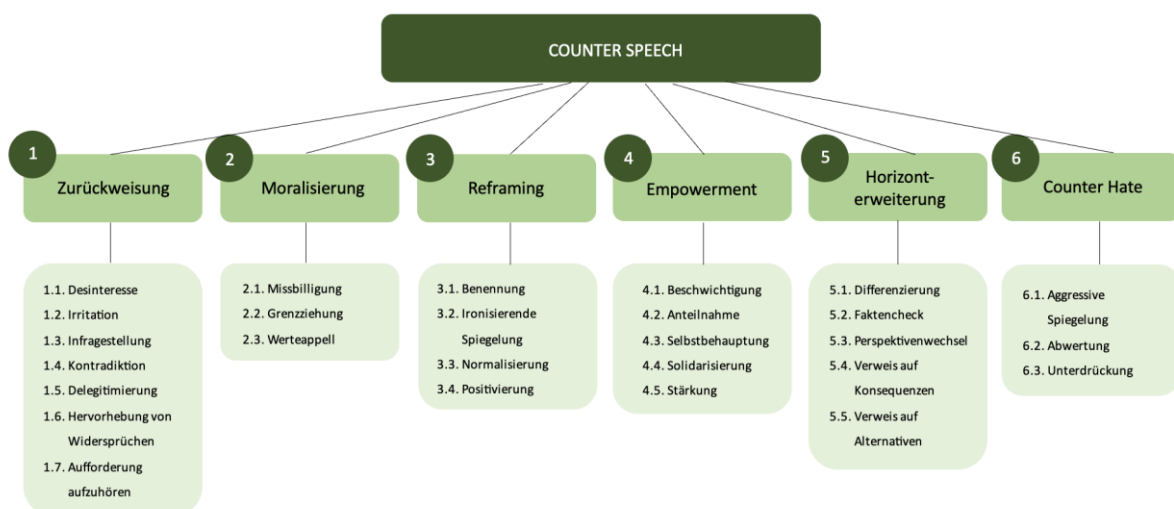


ABBILDUNG 6: EIGENE DARSTELLUNG DES ENTWICKELTEN KATEGORIENSYSTEMS FÜR COUNTER SPEECH

1.) Zurückweisung

Die Oberkategorie Zurückweisung umfasst Formen der Gegenrede, die hasserfüllte Aussagen ohne ausführliche Begründungen oder ergänzende Informationen ablehnen. Sie bedient sich Ausdrucksweisen wie Desinteresse, Irritation, Infragestellung, Kontradiktion, Delegitimierung, der Hervorhebung von Widersprüchen sowie der Aufforderung aufzuhören. Anders als bei anderen Oberkategorien von Counter Speech erfolgt die Gegenrede hier ohne Einbringung zusätzlichen Wissens oder moralischer Maßstäbe.

1.1. Desinteresse

Desinteresse beschreibt eine Art der Gegenrede, in der mangelndes Interesse oder Gleichgültigkeit gegenüber dem vorangegangenen Kommentar und seinem hasserfüllten Inhalt zum Ausdruck gebracht wird. Der Hasskommentar wird dabei nicht inhaltlich kommentiert, sondern durch eine Abwertung seiner Relevanz zurückgewiesen.

Beispiele:

- „Wie unhöflich Hä, auf sowas muss man sich einstellen wenn man im Internet unterwegs ist 🙄“ (V.3, Z.91)
 - „und uns ist deine Meinung egal ?“ (V.5, Z.5)
-

1.2. Irritation

Irritation bringt Unverständnis, Verwirrung oder Fassungslosigkeit über einen hasserfüllten Beitrag zum Ausdruck und stellt diesen dadurch indirekt infrage. Empörte oder irritierte Reaktionen destabilisieren und delegitimieren die hasserfüllten Inhalte und markieren sie als unangemessen. Im Vergleich zur Unterkategorie der Infragestellung steht hier der spontane Charakter der Reaktion im Vordergrund.

Beispiele:

- „Hä ?“ (V.3, Z.620)
 - „wieso schreibt man so etwas? Ich werde es nie verstehen.“ (V.5, Z.107)
-

1.3. Infragestellung

Die Infragestellung nutzt gezielte Rückfragen, um Inhalt, Logik oder Legitimität eines hasserfüllten Beitrags zu hinterfragen und herauszufordern. Kritische, konfrontative oder rhetorische Rückfragen sollen zeigen, dass die Aussage bei genauerem Hinsehen nicht belastbar und unbegründet ist. So kann nicht nur die Aussage selbst, sondern auch ihr ideologischer oder moralischer Hintergrund problematisiert werden.

Beispiele:

- „Ich kenne das System aber was hat bitte lgbtq damit zu tun?“ (V.3, Z.74)
 - „warum interessiert es euch überhaupt was andere so sind oder worauf sie stehen ?
“ (V.3, Z.944)
 - „Ich sehe trotzdem keinen Grund sie jetzt zu hassen?“ (V.6, Z.360)
-

1.4. Kontradiktion

Kontradiktion bezeichnet eine Form der Gegenrede, die hasserfüllte Aussagen durch eine gezielte Verneinung oder das Behaupten des Gegenteils zurückweist. Die Kommentare sind meist knapp formuliert und verzichten sowohl auf ausführliche Begründungen als auch auf positive Neubewertungen.

Beispiele:

- „DOCH“ (V.1, Z.597)
- „Ich bezweifle das es mal so sein wird“ (V.3, Z.97)
- „Ne daran ist nix witzig“ (V.3, Z.537)

1.5. Delegitimierung

Delegitimierung richtet sich nicht gegen den Inhalt einer hasserfüllten Aussage, sondern gegen die Legitimität oder die Kompetenz der sprechenden Person. Durch Verweise auf fehlendes Wissen oder unzureichende Reflexion wird ihr Geltungsanspruch geschwächt und ihre Position im Diskurs als unqualifiziert dargestellt.

Beispiele:

- „Hast wohl die Folge nicht geguckt“ (V.6, Z.76)
 - „Es ging in der folge um hate kommentare und leider haben viele leute nichts mitgenommen 🤪“ (V.6, Z.279)
 - „Warum sind hir alle so gemein zu dir keiner kennt sie von euch persönlich“ (V.6, Z.198)
-

1.6. Hervorhebung von Widersprüchen

Durch das Hervorheben von Widersprüchen werden inhaltliche Unstimmigkeiten, logische Fehler oder widersprüchliche Argumentationsmuster in einem Hasskommentar sichtbar gemacht. Durch das Aufzeigen logischer Inkonsistenzen oder Selbstwidersprüche soll die hasserfüllte Aussage entkräftet werden. Dabei wird auf die Prämissen des ursprünglichen Kommentars zurückgegriffen, ohne zusätzliches Hintergrundwissen einzubringen.

Beispiele:

- „Kommunismus funktioniert nicht wenn eine marginalisierte Gruppe ausgegrenzt wird, dass widerspricht der eigentlichen Philosophie“ (V.2, Z.13)
- „Naja, du willst dass man dich unterstützt, ünterstützt aber andere nicht...macht keinen sinn“ (V.3, Z.287)
- „Ja selbst wenn macht es keinen Sinn. woher soll die Person die Meinung aller Menschen wissen. Ich freue mich nicht! Also Lüge“ (V.6, Z.396)

1.7. Aufforderung aufzuhören

Die Aufforderung aufzuhören ist eine handlungsorientierte Form der Gegenrede, die Hasskommentare direkt adressiert und Verfasser*innen zur Beendigung ihres Verhaltens auffordert. Solchen Aufforderungen liegen häufig moralische oder soziale Normen zugrunde, die jedoch nicht unbedingt explizit benannt werden.

Beispiele:

- „also lass sie doch leben wie sie will junge“ (V.1, Z.28)
- „was hast du davon sowas zu kommentieren? einfach lassen“ (V.6, Z.309)
- „HÖRT DOCH ENDLICH AUF ZU HATEN 🙄“ (V.6, Z.374)

2.) Moralisierung

Unter Moralisierung werden Formen der Gegenrede zusammengefasst, die hasserfüllte Inhalte anhand sozialer oder moralischer Maßstäbe bewerten oder verurteilen. Das moralische Bewusstsein der Adressierten wird angesprochen und gesellschaftlich geteilte Werte und Normen werden aktiviert. Dabei wird sowohl auf bestehende Regeln als auch auf wünschenswerte Haltungen wie Bildung, Selbstreflexion oder Zusammenhalt verwiesen. Moralisierung kann sanktionierend wirken, aber auch Orientierung bieten und Werte stärken.

2.1. Missbilligung

Missbilligung beschreibt eine Form der Gegenrede, in der hasserfüllte Inhalte nicht nur verneint, sondern ausdrücklich negativ bewertet und verurteilt werden. Im Zentrum steht die Kennzeichnung des Beitrags als unangemessen oder verwerflich, häufig verbunden mit einem impliziten Verweis auf verletzte moralische oder gesellschaftliche Maßstäbe. Anders als moralisierende Appelle, die allgemeine Prinzipien hervorheben, reagiert Missbilligung unmittelbar und sanktionierend. Sie kann sich dabei sowohl auf den einzelnen Kommentar als auch auf das zugrunde liegende Weltbild beziehen.

Beispiele:

- „Find ich echt nicht cool Bruder“ (V.3, Z.324)
 - „Wie freundlich“ (V.4, Z.73)
 - „Trauriger Charakter von dir“ (V.6, Z.207)
-

2.2. Grenzziehung

Die Unterkategorie der Grenzziehung bezeichnet eine Form der Gegenrede, die in Reaktion auf hasserfüllte Inhalte eine klare Überschreitung sozialer, moralischer oder rechtlicher Grenzen markiert. In den Daten wurden unter anderem Beleidigungen, bestimmte situative Kontexte und Sprecher*innenposition sowie Falschinformationen als grenzverletzend erachtet.

Beispiele:

- „Es gibt auch Grenzen. Die Meinungsfreiheit gilt nicht für Beleidigungen und dein Kommentar war offensichtlich beleidigend gemeint“ (V.3, Z.28)
 - „Wenn es dir egal ist, dann red' nicht über irgend ein Thema und verbreite falsch Informationen. Und wie genau kommst du darauf, dass du ein Baum bist? 🐼“ (V.3, Z.438)
 - „Ich weiß ja nicht aber ich finde der geht zu weit“ (V.3, Z.728)
 - „Ok freu dich halt aber das gehört verdammt noch mal nicht unter IHR Video“ (V.6, Z.491)
-

2.3. Werteappell

Mit der Unterkategorie Werteappell ist eine Form der moralisierenden Gegenrede gemeint, die hasserfüllte Aussagen mit einem Verweis auf grundlegende moralische und soziale Werte kontert. Durch das Erinnern an Prinzipien wie Gleichheit, Respekt, Toleranz, Mitgefühl, Zurückhaltung und individuelle Freiheit wird die Verantwortung des Gegenübers betont und ein reflektiertes Miteinander eingefordert. Dazu gehören auch Selbstreflexion und

Weiterbildung als Teil individueller Entwicklung und moralischer Bewusstseinsbildung. Werteappelle können durch konkrete Beispiele oder Bildungsangebote gestützt werden. Werteappelle markieren Hass als unvereinbar mit gemeinsamen Normen und zeigen zugleich positive Alternativen für den sozialen Umgang auf.

Beispiele:

- „hättest du was ist dein problem? ob man das äußere von jemandem schön findet ist subjektiv, du kannst deine Meinung haben aber behalt sie für dich“ (V.3, Z.30)
 - „Lass sie doch.Habt ihr nichts von Liana gelernt?“ (V.5, Z.528)
 - „aber das rechtfertigt noch lange nicht so einen kommentar zu schreiben... als hätte niemand gestern aufgepasst bei dem Workshop über hate im Internet“ (V.6, Z.34)
 - „Jeder verdient es, mit Respekt behandelt zu werden. Jeder.“ (V.6, Z.169)
-

3.) Reframing

Die Oberkategorie Reframing umfasst Formen der Gegenrede, die hasserfüllte Inhalte nicht direkt verneinen oder verurteilen, sondern sie sprachlich und semantisch umdeuten und in einen neuen Bedeutungsrahmen überführen. Durch Benennung, ironisierende Spiegelung, Normalisierung und Positivierung werden problematische Inhalte in einen positiveren Kontext gesetzt, wodurch der ursprüngliche Hass an Wirkung verliert und der Diskurs in eine konstruktivere Richtung verschoben wird.

3.1. Benennung

Benennung umfasst Reaktionen, die Hass und seine verschiedenen Erscheinungsformen explizit benennen und begrifflich einordnen. Dadurch wird der problematische Charakter der Aussage sichtbar und in einen neuen Bedeutungsrahmen gestellt. Im Vordergrund steht weniger die moralische Verurteilung als die analytische Klassifizierung der Aussage.

Beispiele:

- „belästigend, diskriminierend und homophob. Er hat sich in dem Video schlechter verhalten als die ganzen Leute, die er befragt.“ (V.3, Z.734)
 - „„Niemand mochte die“ ist hate smh 🧑 “ (V.6, Z.181)
-

3.2. Ironisierende Spiegelung

Die ironisierende Spiegelung greift Elemente des hasserfüllten Beitrags auf und deutet sie mit Humor, Sarkasmus oder Übertreibung so um, dass sie entschärft werden und ins Lächerliche kippt. Durch die humoristische Rahmung wird der ursprüngliche Sinn spielerisch unterlaufen und der Hass in eine absurde oder komische Perspektive überführt.

Beispiele:

- „Möge dir Jesus helfen Nächstenliebe zu praktizieren 🙏 “ (V.1, Z.9)
 - „Ich mag mediamarkt auch“ (V.2, Z.52) [Anm.: als Reaktion auf ein Posting mit den Farben der Reichsflagge]
 - „Du hast recht unsere Gesellschaft ist schlimm die können nicht mal leute für deren Geschlecht respektieren“ (V.3, Z.41)
 - „scheiße wegen dir bin ich jetzt normal 🤡 🤡 🤡 #entemofizierung for life!! Danke das du mir geholfen hast 🍷 🍷 🍷 “ (V.3, Z.85)
-

3.3. Normalisierung

Normalisierung ist eine Form der Gegenrede, die hasserfüllte Inhalte entschärft, indem die angegriffenen Verhaltensweisen, Identitäten oder Erscheinungsbilder als gewöhnlich und unproblematisch beschrieben werden. Durch die Betonung von Selbstverständlichkeit verschiebt sich der Diskurs von Ablehnung hin zu Gelassenheit und Akzeptanz.

Beispiele:

- „Lgbtq ist was normales und gibt es überall lol“ (V.3, Z.20)
 - „wenn Gott alles erschaffen hat, dann auch homosexualitat. Also red nicht über Gott wenn du ihm widersprichst.“ (V.3, Z.639)
 - „Viele Frauen Anfang 20 sehen mit Kurzhaarschnitt eben noch jünger aus. Probleme damit?“ (V.5, Z.417)
-

3.4. Positivierung

Unter die Subkategorie Positivierung fallen Gegenreden, die Hass durch positive Deutungen und wertschätzende Perspektiven ersetzen. Auf negative Zuschreibungen wird dabei oft nicht eingegangen, während der Diskurs in Richtung Anerkennung und Respekt verschoben wird.

Beispiele:

- „Weil sie ein guter Mensch sein“ (V.1, Z.927)
 - „Zeiten ändern sich , zum Glück“ (V.3, Z.291)
 - „Ich mag sie 😊 😊 “ (V.5, Z.31)
 - „Schlimm jemanden zu hassen, nur weil er nicht ist wie die anderen. Es ist doch gut herauszustechen.“ (V.6, Z.871)
-

4.) Empowerment

Die Oberkategorie Empowerment fasst überwiegend opferzentrierte Formen der Gegenrede zusammen, die Betroffene und andere Counter Speaker*innen stärken, unterstützen und in ihrer Handlungsfähigkeit festigen sollen. Durch Beschwichtigung, Anteilnahme, Selbstbehauptung, Solidarisierung und Stärkung sollen Betroffene emotional entlastet und ihre Resilienz gefördert werden.

4.1. Beschwichtigung

Beschwichtigung ist eine Form der Gegenrede, die Betroffene ermutigt, Hasskommentaren keine Bedeutung beizumessen. Mit beruhigenden oder relativierenden Formulierungen wird geraten, den Hass zu ignorieren und emotional auf Abstand zu gehen. Auf diese Weise soll seine Wirkung abgeschwächt werden, ohne in eine direkte Auseinandersetzung zu treten. Mitunter werden dabei sogar vermeintlich positive Effekte der Hasskommentare betont, wie etwa mehr Sichtbarkeit durch höhere Zugriffszahlen.

Beispiele:

- „Denken die sind nur neidisch und sie unterstützen dich ja dann auch kling zwar voll komisch aber sie schreiben ja dann einen kommentar und manche“ (V.1, Z. 224)
„folgen dir ja auch und somit unterstützen sie dich also merk dir ändere dich für niemanden“ (V.1, Z.223) [Anm.: Beitrag wurde auf mehrere Kommentare aufgeteilt]
 - „Ich finde dich sehr sympathisch bitte lass den ganzen hate nicht zu nah an dich ran!!! Warst und bleibst mein Favorit“ (V.6, Z.413)
 - „ich wünsche dir alles gute hör nicht auf die hater“ (V.6, Z.585)
-

4.2. Anteilnahme

Anteilnahme bezeichnet eine Form der Gegenrede, die Mitgefühl für von Hass betroffene Personen ausdrückt und ihr Leid sichtbar macht. Sie bringt Menschlichkeit und Verletzlichkeit in den Diskurs, kann jedoch zugleich den Opferstatus der Betroffenen stärken und ihr Gefühl von Selbstwirksamkeit einschränken.

Beispiele:

- „Du arme wegen den Kommentaren unter dir:(“ (V.1, Z.191)
- „Das tut so weh zu wissen das sie das alles an den Kopf kriegt. Ich hoffe sie kann den hate Ausblenden:(“ (V.6, Z.403)

4.3. Selbstbehauptung

Die Unterkategorie der Selbstbehauptung verweist auf eine Art der Gegenrede, bei der Betroffene abwertende Zustimmungen durch eigene Deutungen und selbstbewusste Positionierungen zurückweisen. Fremdzuschreibungen werden mit persönlichen Erfahrungen entkräftet, wodurch ein Bedeutungsrahmen entsteht, der ihre Souveränität und Deutungshoheit stärkt.

Beispiele:

- „Ich lebe ja nicht um deinem Ideal zu entsprechen sondern mir selbst zu gefallen plus mein aussehen hat nichts mit meiner zukunft zu tun“ (V.3, Z.29)
 - „Ich 16 arbeite seit dem ich 15 bin“ (V.3, Z.86)
 - „Ne warum sollte ich leise sein 😊“ (V.3, Z.538)
-

4.4. Solidarisierung

Solidarisierung bezeichnet eine Form der Gegenrede, bei der sich Kommentierende klar auf die Seite der von Hass betroffenen Personen oder anderer Counter Speaker*innen stellen. Sie äußert sich in ausdrücklicher Zustimmung oder in Form von Verständnis, Nähe und Parteinahme. Dadurch entsteht ein Moment gemeinsamer Haltung gegen Hass, in dem einzelne Stimmen gestärkt und solidarische Unterstützung sichtbar wird.

Beispiele:

- „true“ (V.3, Z.32)
- „sehe ich auch so war schon ziemlich aufdringlich und respektlos“ (V.3, Z.488)
- „Danke dass es jemand sagt“ (V.3, Z.490)
- „Junge lass sie doch, ich hätte mich genau so verhalten wenn mich niemand mögen würde. Ihr müsst sie ja nd mögen her zu sagen endlich ist sie raus ist“ (V.6, Z.863)

4.5. Stärkung

Stärkung umfasst Gegenreden, die Betroffene ermutigen und ihre Handlungsmacht stärken. Im Vordergrund steht die Förderung einer selbstbewussten und widerständigen Haltung, die Betroffene ermutigt, Hass nicht nur auszuhalten, sondern ihm auch aktiv und selbstbestimmt entgegenzutreten.

Beispiele:

- „Das tut mir leid aber du schaffst das“ (V.1, Z. 262)
- „bleib stark bin grosser Fan“ (V.4, Z.43)
- „hört auf zu urteilen über Menschen die ihr nicht kenn und an Anya, stay strong und lass dich nicht unterkriegen“ (V.5, Z.457)

5.) Horizontenerweiterung

Die Oberkategorie Horizontenerweiterung bündelt Formen der Gegenrede, die hasserfüllten Aussagen mit Wissen, Anregungen zum Perspektivenwechsel und dem Aufzeigen möglicher Handlungsalternativen begegnen. Durch die Vermittlung von Informationen und das Einbeziehen unterschiedlicher Sichtweisen soll Einsicht gefördert und Verständnis erweitert werden. Im Vergleich zu anderen Oberkategorien, wie Zurückweisung und Moralisierung stützen sich diese Formen der Gegenrede stärker auf ein spezifisches Hintergrund- und Kontextwissen der Counter Speaker*innen und zielen darauf ab, den Erkenntnishorizont von Hater*innen und Mitlesenden zu erweitern.

5.1. Differenzierung

Differenzierung stellt eine Form der Gegenrede dar, bei der vereinfachende oder verzerrte Aussagen durch präzisierende Unterscheidungen korrigiert werden. Indem Verallgemeinerungen aufgelöst und Nuancen betont werden, wird die argumentative Grundlage hasserfüllter Beiträge geschwächt und der Diskurs rationalisiert.

Beispiele:

- „verschmutztes wasser beeinflusst dich wenn du es trinkst, ein schwuler der eif. nur existiert und dir nix tut beeinflusst dich nd...dont even 🙊💀“ (V.1, Z.563)
 - „man kann nicht sagen ich bin heute nun Hund und morgen ein Elefant,da muss man sich richtige Gedanken drüber machen.iss halt kein fun Thema“ (V.3, Z.250)
 - „Es ist ein unterschied ob ich mich selbst fett nenne oder andere mir das sagen 😬“ (V.4, Z.8)
-

5.2. Faktencheck

Der Faktencheck bezeichnet eine Form der Gegenrede, bei der hasserfüllte und verzerrte Aussagen durch sachliche Informationen, überprüfbare Fakten und kontextualisierende Anmerkungen widerlegt und korrigiert werden. Durch das Einbringen relevanter Kenntnisse werden Inhalte richtiggestellt oder in größere gesellschaftliche, mediale oder strukturelle Zusammenhänge eingeordnet. Auf diese Weise werden Fehlschlüsse sichtbar, Hasskommentaren wird ihre argumentative Grundlage entzogen und der Diskurs auf einer reflektierten, faktenbasierten und sachlichen Ebene geführt.

Beispiele:

- „das ist zwar war das dies in der bibel steht jedoch sagte gott: doch redet nicht zu schlecht mit einander. Les die bibel durch“ (V.1, Z.157)
- „gratuliere du hast entschlüsselt warum die Leute auf Minderheiten rumhacken“ (V.3, Z. 369)
- „Viele sehen Pronomen halt an wie einen "Spitznamen", und den meisten ist dann egal wie sie angesprochen werden, ist halt einem selbst überlassen lol“ (V.3, Z.581)
- „Sie hat klar gestellt dass sie nix gegen die Nachzügler persönlich hatte, sondern es in dem ersten Moment einfach unfair fand. und jetzt?“ (V.5, Z.383)
- „Ich will nicht ihr Verhalten entschuldigen, aber solche Shows schneiden schon gerne Material so zusammen, dass es was zu reden gibt“ (V.6, Z.44)

5.3. Perspektivenwechsel

Der Perspektivwechsel soll Hater*innen dazu anregen, sich in die Lage der Betroffenen zu versetzen und ihre Sichtweise nachzuvollziehen. Durch empathische Annäherungen und hypothetische Szenarien sollen Distanz abgebaut, Verständnis und Selbstreflexion gefördert und der Hasskommentar indirekt entkräftet werden.

Beispiele:

- „imagine es wäre illegal und du würdest solche Kommentare bekommen wenn du mit deinem bf bist“ (V.1, Z. 289)
 - „das kann man auch für sich behalten, du hättest das auch nicht gerne unter deinem video :/“ (V.6, Z.30)
 - „Ich weiß die Meinungen sind gespalten aber sie ist immerhin ein Mensch. Sie ist sehr jung und ist so wie sie ist. Ich finde es sehr schade dass unter“ (V.6, Z.568)
-

5.4. Verweis auf Konsequenzen

Der Verweis auf Konsequenzen macht auf die psychischen, sozialen oder rechtlichen Folgen hasserfüllter Äußerungen aufmerksam – für Betroffene ebenso wie für Verfasser*innen und die Gesellschaft insgesamt. Anderes als der Perspektivenwechsel, der über Einfühlung und Identifikation wirkt, hebt der Verweis auf Konsequenzen die konkrete Tragweite des eigenen Handelns hervor und regt angesichts der möglichen Folgen zur Selbstreflexion und Verhaltensänderung an.

Beispiele:

- „Wen er sagt film mich nicht und du es trotzdem machst, gibt das dann nicht total ärger wenn der das sieht? 🤪“ (V.3, Z.925)
- „Ey diese Kommentare... Manchmal wäre es besser einfach leise zu sein. Worte können weh tun und ihr wollt sicherlich auch nicht solche Worte hören.“ (V.6, Z.638)
- „Ist ja auch okay man kann nicht jeden mögen nur Hate ist nicht okay menschen nehmen doch deswegen ihr Leben...“ (V.6, Z.769)

5.5. Verweis auf Alternativen

Ein Verweis auf Alternativen bietet angesichts hasserfüllter Inhalte andere Handlungswege an und eröffnet konstruktive Entwicklungsmöglichkeiten für Hater*innen. Diese Form der Gegenrede verweist auf positive Alternativen und verschiebt den Diskurs von der Abwertung hin zu Hoffnung, Reflexion und dem Potenzial moralischer Weiterentwicklung.

Beispiele:

- „schreib es über chat wenn es dich so freut. Es ist sehr unhöflich yk“ (V.6, Z.33)
- „Und daher sage ich einfach nur irgendwann wirst du weiter denken und andere Perspektiven bekommen. 😊“ (V.6, Z.688)

6. Counter Hate

Die Oberkategorie Counter Hate umfasst Formen der Gegenrede, die inhaltliche und sprachliche Parallelen zu Hate Speech aufweisen. Sie greifen auf konfrontative, spöttische und abwertende Mittel zurück, um Hater*innen sprachlich zu entmachten und ihre Stellung im Diskurs anzugreifen. Anders als wissens-, empathie- oder werteorientierte Gegenrede spiegelt Counter Hate den Hass und wendet ihn gegen die ursprünglichen Verfasser*innen. Die Unterkategorien aggressive Spiegelung, Abwertung und Unterdrückung zeigen unterschiedliche Ausprägungen dieser reaktiven und oft emotional aufgeladenen Art der Gegenrede.

6.1. Aggressive Spiegelung

Unter aggressiver Spiegelung wird eine Form der Gegenrede verstanden, bei der hasserfüllte Aussagen auf ihre Verfasser*innen zurückgeworfen werden, ohne dass eine inhaltliche Auseinandersetzung mit dem Hasskommentar erfolgt. Die Logik, der Tonfall und die Argumentation des ursprünglichen Beitrags werden aufgegriffen und umgekehrt, sodass die Hater*innen selbst zum Ziel der jeweiligen Ausformung von Hate Speech werden. Die

aggressive Spiegelung kann sowohl subtil, ironisch, als auch direkt konfrontativ eingesetzt werden.

Beispiele:

- „Denke ich mir auch wenn ich deine bio lese“ (V.1, Z. 59)
 - „wenn du selber so ne schlechte Meinung von dir hast hat sie auch jeder andere von Dir!“ (V.3, Z.373)
 - „Beschreibst du dich gerade selbst?“ (V.5, Z.128)
-

6.2. Abwertung

Diese Unterkategorie von Counter Hate umfasst Gegenreden, die sprachlich abwertend agieren und Gemeinsamkeiten mit der gleichnamigen Unterkategorie von Hate Speech aufweist. Sie sind durch sprachliche Herabsetzungen gekennzeichnet, die die Identität und Legitimität des Gegenübers angreifen. In den Daten zeigen sich hierfür Beispiele für Spott, affektive Abwertung, Abwertung des Auftretens, Infantilisierung, relationale Abwertung und Diskriminierung.

Beispiele:

- „ok patrick, geh heulen 😞“ (V.2, Z.8)
- „Wenn er dich zurück in die Türkei kickt lachst du nicht mehr !! 😂😂“ (V.3, Z.589)
- „Wenn jemand keinen Satz ohne Rechtschreibfehler hinkriegt HAHAA“ (V.3, Z.599)
- „Bah alter was ist das für ein Typ“ (V.3, Z.900)
- „naja aber mirella ist deutlich schlechter gewesen und generell meistens einfach schlecht. sie hätte fliegen müssen“ (V.5, Z.3)
- „Kind“ (V.6, Z.40)

6.3. Unterdrückung

Zu dieser Kategorie zählen Formen der Gegenrede, die den Ausprägungen der gleichnamigen Hate Speech Unterkategorie entsprechen und derselben Logik folgen. In den Daten wurden in dieser Hinsicht Silencing und Vernichtungswünsche identifiziert, die das Verstummen oder den Tod der angesprochenen Person fordern oder als wünschenswert markieren.

Beispiele:

- „Einfach mal leise sein danke“ (V.6, Z.81)
 - „hzm du 13 jährige 🧟 🧟 🧟“ (V.6, Z.108)
 - „niemand hält dich davon ab sie [Anm. bezieht sich auf „diese Welt“] zu verlassen ❤️“ (V.3, Z.585)
-

5.3. Interaktionen von Hate Speech und Counter Speech

Für die Einzelfallanalysen wurden rund 85 Interaktionen von Hate Speech und Counter Speech identifiziert, die jeweils mindestens einen Hasskommentar und eine Gegenrede enthielten. Die ausgewählten Interaktionen umfassen zwischen drei und 107 Kommentaren und wurden einzeln betrachtet und analysiert. Es folgt eine deskriptive Zusammenfassung der zentralen Beobachtungen.

Die Analyse der Interaktionen war durch die Struktur der Plattform erheblich eingeschränkt. Nachdem Antwortkommentare auf TikTok ohne Bezüge und zudem nicht chronologisch gereiht werden, sondern gemäß ihrer Anzahl an Likes, war eine genaue Rekonstruktion des Interaktionsverlaufs in vielen Fällen nicht möglich. Eine präzisere Analyse wäre nur unter anderen Bedingungen oder in einer Zusammenarbeit mit den Plattformbetreibenden möglich.

Die meisten Hasskommentare blieben unbeantwortet. Reaktionen auf Hasskommentare umfassen nicht nur Gegenrede, sondern auch Zustimmung und Verstärkung von Hass. Trotz einer hohen Anzahl an Antworten unter bestimmten Ausgangskommentaren erwies sich die Häufigkeit tatsächlicher Interaktionen zwischen Hate Speech und Counter Speech als gering.

Besonders bei Ausgangskommentaren, die selbst Hass enthielten, traten zahlreiche Zustimmungskommentare auf. Unter Hater*innen ist an vielen Stellen eine gegenseitige Unterstützung zu beobachten. Diese zeigt sich nicht nur in Likes und kurzen Bestätigungskommentaren, sondern auch in Angriffen gegen Counter Speaker*innen.

Insgesamt konnten sechs Reaktionsweisen auf Gegenrede identifiziert werden:

- (1) Ignorieren
- (2) Entschuldigen
- (3) Rechtfertigen
- (4) Klären
- (5) Solidarisieren mit anderen Hater*innen
- (6) Eskalieren

Die untersuchten Counter Speech Interventionen blieben selten ohne Reaktion, allerdings stammten sie nicht immer von der Person, die den Ausgangskommentar verfasst hat. Im gesamten Datensatz kamen eine Entschuldigung und ein Versuch einer Klärung vor. In beiden Fällen, waren keine anderen Hater*innen zugegen, die den ursprünglichen Hasskommentar unterstützten. Entschuldigung und Klärung traten unmittelbar nach dem ersten Gegenrede-Kommentar auf und nicht erst nach einem längeren Austausch, wobei letztere insgesamt selten sind. Inhaltlich knüpften beide Counter Speech Kommentare an moralische Werte an und verwiesen auf Widersprüche in der Argumentation des Hasskommentars. Bei Klärungsversuchen kann nicht eindeutig bestimmt werden, ob es sich hier tatsächlich um ein Missverständnis handelt oder ob versucht wird nachträglich zurückzurudern. Deutlich häufiger ließen sich Rechtfertigungsversuche beobachten, bei denen der ursprüngliche Kommentar mit dem Verweis auf Meinungsfreiheit, Ehrlichkeit oder Relativierung bzw. Leugnung des hasserfüllten Gehalts verteidigt wird. Solidarisierungen mit anderen Hater*innen – sofern, sich welche melden – sind sehr häufig. Schließlich kommt es häufig zu Eskalationen, bei denen Counter Speaker*innen beleidigt werden oder mit spöttischen Bemerkungen bedacht werden. Löschungen eigener Hasskommentare in Reaktion auf Counter Speech Interventionen konnten aufgrund der Beschaffenheit der Daten nicht erfasst werden.

Interaktionen fanden nicht nur unter Hasskommentaren statt, sondern auch unter eigenständigen positiven Kommentaren. Hater*innen reagierten häufig auf wohlwollende Kommentare und auf Beiträge, in denen Personen ihre Zugehörigkeit zu einer angegriffenen Personengruppe artikulierten. Während diese Kommentare gezielt angegriffen wurden, erfuhren besonders letztere verstärkt Unterstützung durch Gegenrede – wobei zu erwähnen ist, dass sich die Verfasser*innen dieser Kommentare auch oft selbst verteidigten. Viele dieser Reaktionen waren opferzentriert und entsprachen der Kategorie Empowerment. Es ist allerdings nicht auszuschließen, dass die Unterstützer*innen aus dem privaten Umfeld der angegriffenen Personen stammen.

In einigen Kommentarspalten traten wiederholt dieselben Nutzer*innen an verschiedenen Stellen mit Gegenrede in Erscheinung. Diese besonders aktiven Gegenredner*innen setzten nicht nur insgesamt vermehrt Kommentare ab, sondern führten tendenziell auch längere Interaktionen mit Hater*innen. Zukünftige Forschung könnte sich mit unterschiedlichen Typen von Counter Speaker*innen befassen und untersuchen, ob sie eher Antwortkommentare oder eigenständige Kommentare verfassen, ob sie ihre Kommentare an spezifische Situationen anpassen oder standardisierte Kommentarmuster verwenden. Es wurde beispielsweise auch eine Person identifiziert, die sich sogar unter zwei verschiedenen Videos derselben GNTM-Kandidatin mit Gegenrede engagierte.

Auch besonders aktive Hater*innen ließen sich erkennen. Sie verfassten mehrere Kommentare, sowohl innerhalb einzelner Threads als auch verteilt über verschiedene Beiträge. Auffällig war, dass diese Nutzer*innen oftmals exakt den gleichen oder sich ähnelnde Kommentare in unterschiedlichen Kontexten absetzen.

Das Engagement besonders aktiver Counter Speaker*innen wurde gelegentlich gegen sie verwendet, indem versucht wurde, sie auf Basis ihrer hohen Aktivität zu diskreditieren. Generell wurde häufig aggressiv auf Counter Speech reagiert. In manchen Fällen verwendeten auch Counter Speaker*innen selbst aggressive und diskriminierende Sprache. So wurde beispielsweise in einem Fall versucht, Queerfeindlichkeit mit rassistischer Rhetorik zu bekämpfen.

Nicht abgebildet werden konnten intrapersonale Reflexionsprozesse. Dies betrifft sowohl Fälle ohne sichtbare Rückmeldung auf Counter Speech als auch solche mit erkennbarer Reaktion. Reflexion kann zeitverzögert stattfinden und langfristig wirken. Gleichzeitig besteht auch die Möglichkeit, dass Gegenrede zu Radikalisierungsprozessen beiträgt. Solche Entwicklungen sind jedoch schwer messbar.

6. Interpretation und Diskussion der Ergebnisse

Es zeigt sich, dass einige Formen von Hate Speech und Counter Speech Überschneidungen aufweisen und einzelne Zugänge sogar nahezu deckungsgleich sind. Insbesondere die Zustimmung zu vorangegangenen Kommentaren und die Oberkategorie der Zurückweisung, die in beiden Modellen vorkommt und vergleichbare Unterformen umfasst, sind hiervon betroffen. Dies verweist auf die zentrale Bedeutung des Kontextes für die Bestimmung von Counter Speech und Hate Speech und wirft die Frage auf, ob automatisierte Systeme diese Nähe zwischen manchen Formen von Hate Speech und Counter Speech zuverlässig erkennen und unterscheiden können.

Im Hinblick auf Hate Speech trat Spott mit deutlichem Abstand am häufigsten auf. Dies könnte damit zusammenhängen, dass Humor und humorvolle Kurzvideos das Kernstück der Plattform TikTok bilden und hasserfüllte Inhalte möglicherweise anders bewertet werden, wenn sie inmitten zahlreicher humoristischer Beiträge auftauchen. Die Häufung von Spott erinnert an das von Cabbug und Abidin (2025) beschriebene Phänomen der „dogshows“, bei dem sich große Gruppen versammeln, um Einzelpersonen kollektiv zu verhöhnen. Der plattformspezifische Algorithmus verstärkt diese Dynamiken zusätzlich, da Videos schnell große Reichweiten erzielen und sich innerhalb fragmentierter Öffentlichkeiten rasch ein Mob zusammensetzen kann, der zum gemeinsamen Verspotten zusammenfindet.

Die häufigste Reaktion auf Hate Speech ist Zustimmung zu diesem Hate Speech. Kommentarstränge mit vielen Antworten bestehen überwiegend aus zahlreichen einzelnen Kurzreaktionen verschiedener Nutzer*innen, die dem geäußerten Hass beipflichten. Eine vergleichbare Häufigkeit an Zustimmungskommentaren zu Counter Speech findet sich nicht. Möglicherweise sind potenzielle Counter Speaker*innen in diesen Kommentarbereichen nicht

präsent oder sie sehen aus unterschiedlichen Gründen davon ab, zu reagieren oder bestehende Gegenrede zu unterstützen. Ebenso ist denkbar, dass ihnen das Video nicht angezeigt wurde oder sie die Kommentare nicht öffnen. Da Kommentarbereiche aktiv aufgerufen werden müssen, bleiben sie Personen verborgen, die einfach weiter durch ihren Feed scrollen.

Das Konzept des „platformed racism“ von Matamoros-Fernández (2017) bietet eine mögliche Erklärung für die hohe Zahl an Likes und Unterstützungsbekundungen zu Hasskommentaren im Vergleich zu Gegenrede und positiven Kommentaren. Der Empfehlungsalgorithmus favorisiert empörungsgetrieben und antagonistische Inhalte. Die untersuchten Hasskommentare weisen einen hohen Gehalt an Emotionen wie Empörung, Wut und Ekel und erfüllen somit diese algorithmisch begünstigten Kriterien.

Auch das Plattformdesign beeinflusst diese Dynamiken maßgeblich. Viele Gegenredekommentare werden als Antwortkommentare verfasst und sind damit weniger sichtbar, da sie erst über zusätzliche Klicks geöffnet werden müssen. Eigenständige Kommentare hingegen sind sofort sichtbar, wodurch Gegenrede im Vergleich zu Hasskommentaren an Sichtbarkeit einbüßt. Die damals geltende maximale Zeichenanzahl von 300 pro Kommentar erschwerte zudem ausführliche Gegenrede. Hasskommentare sind oft sehr kurz und benötigen kaum Platz, während Gegenrede mehr Raum verlangt, der hier jedoch nicht zur Verfügung stand. Dies könnte auch erklären, warum Faktenchecks in den Daten selten vorkommen. Die Begrenzung verhindert außerdem das Einfügen längerer Links. Da Links auf TikTok grundsätzlich nicht eingebettet werden und erst manuell kopiert und geöffnet werden müssen, wird die effektive Nutzung externer Quellen zusätzlich erschwert. Vor dem Hintergrund der privatwirtschaftlichen Strukturiertheit von Plattformunternehmen ist anzunehmen, dass es sich hier um eine Maßnahme zur Sicherung der Verweildauer auf der eigenen Plattform handelt. Bei den Gegenredekommentaren der Kategorie Benennung besteht die Möglichkeit, dass sie über intransparente Wortfilter erschwert wird. Die gleichen Wortfilter sind ein Grund warum hasserfüllte Kommentare vielfach sprachliche Umwege suchen und vielfach mit Umkehrungen und Andeutungen arbeiten. Das erschwert die Erkennung – sowohl für automatisierte Systeme, als auch für potenzielle Counter Speaker*innen.

Diese Kontextunsicherheit, die als wichtiger Faktor für unterlassene Counter Speech Interventionen bei Jugendlichen gilt, wird durch die wenig intuitive Strukturierung der Antwortkommentare verstärkt. Dies gilt ebenso für die Entpersonalisierung digitaler Interaktion (Gerbaudo, 2024), wodurch wenig Bezug zu den Betroffenen der wahrgenommenen Hasskommentare besteht (Atzmüller et al., 2022).

Angesichts dieser Strukturen drängt sich die Vermutung auf, dass TikTok kein besonderes Interesse an Diskussionen in Kommentarspalten hat. Die geringe Zahl an Interaktionen zwischen Hate Speech und Counter Speech unterstützt diesen Eindruck. Vieles deutet darauf hin, dass die Plattform kurze Reaktionen bevorzugt, während längere Auseinandersetzungen eher unerwünscht scheinen. Ein solcher Zugang reduziert potenziell den Moderationsaufwand und vereinfacht automatisierte Erkennungsprozesse, was für das Unternehmen wirtschaftlich vorteilhaft ist. TikTok erscheint weniger als Raum für Konfliktaustragung fungieren zu wollen, sondern primär als Ort der Unterhaltung.

Offen bleibt die Frage nach den Konsequenzen, wenn Plattformen Konflikte durch ihr Design und ihre Empfehlungsalgorithmen zwar befördern, jedoch keinen geeigneten Raum für ihre Austragung bereitstellen. Eine angemessene Umgebung für die Austragung zwischenmenschlicher Konflikte müsste auch die Grenzen des Streitbaren berücksichtigen. Dennaoui und Witte (2022) weisen darauf hin, dass Streit spätestens dort endet, wo die andere Person nicht mehr als Mensch anerkannt wird. Diese Form der Aberkennung zeigte sich in einigen der beobachteten Ausprägungen von Hate Speech.

Gleichzeitig gibt es Anzeichen für die Wirksamkeit von Bildungsangeboten. Unter dem letzten Video zum Ausscheiden der GNTM-Kandidatin fand sich der höchste Anteil an Gegenredekommentaren. Viele Kommentare bezogen sich auf die in der Folge thematisierten Inhalte zu Cybermobbing oder auf Lijana Kaggwa, die nach eigenen Erfahrungen mit Cybermobbing öffentlichkeitswirksam aus einer früheren Staffel von GNTM ausstieg und seither mit großer Reichweite über die Dynamiken von Hass im Netz aufklärt.

7. Limitationen und Ausblick

Die Limitationen der vorliegenden Arbeit betreffen in erster Linie die technische Ebene. Mit präziseren Informationen zum exakten Zeitpunkt der Kommentarveröffentlichung, der eindeutigen Zuordnung einzelner Kommentare zu ihren Bezugsbeiträgen sowie einer chronologischen Sortierung wäre eine vertiefte Analyse der Interaktionen möglich gewesen. Auch die Tatsache, dass gängige Auswertungsprogramme wie MAXQDA nicht für die Analyse von Social Media Kommentaren konzipiert sind, erschwerte die Datenauswertung erheblich. Inhaltlich ist zu berücksichtigen, dass es sich um stark reduzierte Informationen handelt, die unter spezifischen digitalen Bedingungen entstehen (Suler, 2004). Sie bilden lediglich einzelne Facetten von Personen ab, ohne jemals die Gesamtheit der dahinterstehenden Individuen zu erfassen oder eine umfassende Auseinandersetzung mit ihnen zu ersetzen.

Soziale Medien stellen ein breites Forschungsfeld dar und zukünftige Studien bieten zahlreiche Möglichkeiten, Fragen der digitalen Interaktion und Konfliktaustragung weiter zu vertiefen. Eine zentrale Erkenntnis dieser Arbeit besteht darin, Social Media Plattformen nicht als homogene Sphäre zu betrachten, sondern ihre jeweilige Spezifik zu berücksichtigen und Forschungsvorhaben entsprechend auszurichten.

8. Literaturverzeichnis

- Abidin, C. (2021). From “Networked Publics” to “Refracted Publics”: A Companion for Researching “Below the Radar” Studies. *Social Media + Society*, 7(1), 1-13.
<https://doi.org/10.1177/2056305120984458>
- Alafwan, B., Siallagan, M. & Putro, U. S. (2023). Comment Analysis on Social Media: A Review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(6).
<https://doi.org/10.4108/eetsis.3843>
- Alsagheer, D., Mansourifar, H. & Shi, W. (2022). Counter Hate Speech in Social Media: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2203.03584>
- Atzmüller, C., Kröncke, L. & Zartler, U. (2022). Cyberheros. Mobilisierung Jugendlicher zu Counter Speech im Umgang mit Online Hass, Cybermobbing und anderen Formen digitaler Gewalt. Endbericht 2022. Institut für Soziologie. Verfügbar unter:
https://zivilcourage.univie.ac.at/fileadmin/user_upload/p_zivilcourage/Cyberheroes_wiss_Endbericht.pdf [letzter Zugriff: 15.09.2025]
- Benesch, S., Ruths, D., Dillon, K. P., Haji, M. S. & Wright, L. (2016). Counterspeech on Twitter: A Field Study. Dangerous Speech Project. Verfügbar unter:
<https://www.dangerousspeech.org/libraries/counterspeech-on-twitter-a-field-study>
[letzter Zugriff: 10.11.2025]
- Bilewicz, M., Tempska, P., Gniewosz, L., Dowgiałło, M., Tanska, M., Urbaniak R. & Wroczynski, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, 47(3), 260-266.
<https://doi.org/10.1002/ab.21948>
- Buerger, C. (2021). #iamhere: Collective counterspeech and the quest to improve to improve online discourse. *Social Media + Society*, 7(4).
<https://doi.org/10.1177/20563051211063843>
- Buerger, C. & Wright, L. (2021). Counter Speech: A Literature Review. Verfügbar auf SSRN:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4066882
- Buerger, C. (2022). Why They Do It: Counterspeech Theories of Change. Verfügbar auf SSRN:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4245211

- Bundesministerium Justiz (o.D.). Digital Services Act. Neue Regeln für Online-Vermittlungsdienste. <https://www.bmi.gv.at/themen/EU-und-Internationales/Digital-Services-Act.html> [letzter Zugriff: 09.11.2025]
- Chakravarthi, B. R. (2022). Multilingual hate speech detection in english and dravidian languages. *International journal of data science and analytics*, 14(4), 389-406. <https://doi.org/10.1007/s41060-022-00341-0>
- Chung, Y., Abercrombie, G., Enock, F., Bright, J. & Rieser, V. (2024). Understanding Counterspeech for Online Harm Mitigation. *Northern European Journal of Language Technology*, 10(1), 30-49. <https://doi.org/10.3384/nejlt.2000-1533.2024.5203>
- Condom Tibau, J. G., Voggenreiter, A., Pavan, E. & Pfeffer, J. (2025). Prevalence, Substance and Responses to Hate Speech Against LGBTQ Communities on TikTok. *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)*, 19(1), 430-442. <https://doi.org/10.1609/icwsml.v19i1.35824>
- Conectere [@conectere]. (2025, 2. Juni). Noch eine neue Funktion hier auf TikTok. Du kannst jetzt richtig lange Texte in den Kommentaren hinterlassen. Anscheinend sind bis zu 2199 Zeichen möglich. #tiktok #kommentare #neuefunktion #zeichen #funktion [Video]. *TikTok*. <https://www.tiktok.com/@conectere/video/7511260564841123094> [letzter Zugriff 07.11.2025]
- Cools, K., de Buy Wenniger, G. M. & Maathuis, C. (2024). Modeling offensive content detection for TikTok. Conference on Digital Platforms and Societal Harms 2024 (DPSH 2024). *arXiv*. <https://doi.org/10.48550/arXiv.2408.16857>
- Dangerous Speech Project (o. D.). What is counterspeech? Please explain the definition. <https://www.dangerousspeech.org/counterspeech> [letzter Zugriff: 11.11.2025]
- Deda-Bröchin, S., Baumann, L., Willems, I., Bernath, J., Waller, G. & Süss, D. (2025). JAMESfocus – Unterhaltung, Information oder Selbstdarstellung: Warum nutzen Jugendliche soziale Netzwerke? Zürcher Hochschule für Angewandte Wissenschaften. https://www.zhaw.ch/storage/hochschule/medien/news/2025/250702_MM_JamesFocus/JAMESfocus_SN_Nutzungsmotive_DE.pdf [letzter Zugriff: 20.11.2025]
- Dennaoui, Y. & Witte, D. (2022). Streit und Kultur. Vorüberlegungen zu einer Soziologie des Streits. In M. Kumlehn & S. Wodianka (Hrsg.), *Kulturen des Streits*.

Deutungsmachtkonflikte zwischen Konsens und Zerwürfnis (1. Aufl., S. 15-40).

Transcript Verlag.

Deutschlandfunk Nova (2020, 15. Mai). Wieso TikTok gerade in der Krise gehypt wird.

<https://www.deutschlandfunknova.de/beitrag/tiktok-wieso-die-app-gerade-in-der-krise-gehypt-wird> [letzter Zugriff: 07.11.2025]

Dienlin, T., Marboe, G. & Plener, P. (2025). Mental Health Days – Studie 2024. Verfügbar unter:

<https://static1.squarespace.com/static/638a6e7a684c250c7777e1e6/t/67938766aeb12e73f050925a/1737721705243/mental+health+days+Studie+2024+250124.pdf>
[letzter Zugriff: 16.11.2025]

Diwisch, C. (2022, 25. März). Umfangreiche Zensur auf TikTok auch in Österreich. *Radio FM4*.

<https://fm4.orf.at/stories/3022945/> [letzter Zugriff: 09.11.2025]

Dubois, E. & Reepschlager, E. (2024). How harassment and hate speech policies have changed over time: Comparing Facebook, Twitter and Reddit (2005-2020). *Policy & Internet*, 16(3), 523-542. <https://doi.org/10.1002/poi3.387>

Eckert, S., Vitlif, O. & Felke, C. [Tagesschau] (2022, 23.03.). Exklusiv: TikTok nutzt in Deutschland Wortfilter [Video]. *YouTube*.

<https://www.youtube.com/watch?v=MaS17H2KqMk> [letzter Zugriff: 09.11.2025]

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G. & Roth, H. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, 10, 1-49.

Eltaher, F., Gajula, R. K., Miralles-Pechuán, L., Crotty, P., Martínez-Otero, J., Thorpe, C. & McKeever, S. (2025). Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms. *arXiv*: <https://arxiv.org/pdf/2505.11160>

Fillies, J., Theisen, E., Hoffmann, M., Jung, R., Jung, E., Fischer, N. & Paschke, A. (2025). A novel German TikTok hate speech dataset: far-right comments against politicians, women, and others. *Discov Data*, 3(4), 1-21. <https://doi.org/10.1007/s44248-025-00020-y>

Fortuna, P., Soler-Company, J. & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing and Management*, 58, 1-17. <https://doi.org/10.1016/j.ipm.2021.102524>

- García-Baena, D., García-Cumbreras, M. A., Jiménez-Zafra, S. M., García-Díaz, J. A. & Valencia-García, R. (2023). Hope Speech detection in Spanish. The LGBTQ case. *Language Resources and Evaluation*, 57(4), 1487-1514. <https://doi.org/10.1007/s10579-023-09638-3>
- García-Prieto, V., Bonilla-del-Río, M., Figuereo-Benítez, J. C. (2024). Disability, hate speech and social media: video replies to haters on TikTok. *Revista Latina de Comunicación Social* 82, 1-21. <https://www.doi.org/10.4185/RLCS-2024-2258>
- Garland, J., Ghazi-Zahedi, K., Young, J., Hébert-Dufresne, L. & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPI Data Science*, 11(3), 1-24. <https://doi.org/10.1140/epjds/s13688-021-00314-6>
- Gerbaudo, P. (2024). TikTok and the algorithmic transformation of social media publics: from social networks to social interest clusters. *New Media & Society*, 0(0), 1-18. <https://doi.org/10.1177/14614448241304106>
- Gillespie, T. (2018). *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Murias Munoz, M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F. & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50), 1-3. <https://doi.org/10.1073/pnas.2116310118>
- Hickey, D., Schmitz, M., Fessler, D. M. T., Smaldino, P. E., Lerman, K., Muric, G., Burghardt, K. (2024). Hostile Counterspeech Drives Users From Hate Subreddits. *arXiv*. <https://arxiv.org/abs/2405.18374>
- #ichbinhier (2025, 10. November). [Facebook-Gruppe] <https://www.facebook.com/groups/718574178311688> [letzter Zugriff: 10.11.2025]
- Jiménez-Díaz, O. & Del Rey, R. (2025). Cyberhate in adolescents and youths: A systematic review of labels and associated factors. *Aggression and Violent Behavior*, 81, 1-11. <https://doi.org/10.1016/j.avb.2024.102023>
- Joas, H. & Knöbl W. (2013). *Sozialtheorie. Zwanzig einführende Vorlesungen. Aktualisierte Ausgabe*. Suhrkamp.

- Jones, C., Williams, K. L., Branch, S., Crow, C., Miller, J., Richardson, W. & Hilton A. A. (2024). Race(ism), Power, Intimidation, & Domestic Terrorism: A Critical Content Analysis of HBCU Bomb Threats on Social Media. *Peabody Journal of Education*, 99(2), 236-240. <https://doi.org/10.1080/0161956X.2024.2331939>
- Kaggwa, L. (2023). *Du verdienst den Tod! Wie Cybermobbing Menschen und die Gesellschaft zerstört und wie wir wieder mehr Respekt ins Netz bringen*. Kompletmedia.
- Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S. & Bilz, L. (2022). A Systematic Review on Hate Speech among Children and Adolescents: Definitions, Prevalence, and Overlap with Related Phenomena. *Trauma, Violence, & Abuse*, 24(4), 2598-2615. <https://doi.org/10.1177/15248380221108070>
- Khaleghipour, M., Koban, K. & Matthes, J. (2025). Listen to me! Target Perceptions of Digital Hate: A Scoping Review of Recent Research. *Trauma, Violence, & Abuse*, 26(5), 1082-1096. <https://doi.org/10.1177/15248380241303725>
- Khurana, U., Vermeulen, I. Nalisnick, E., van Noorloos, M. & Fokkens, A. (2022). Hate Speech Criteria: A Modular Approach to Risk-Specific Hate Speech Definitions. *arXiv*. <https://doi.org/10.48550/arXiv.2206.15455>
- Korol, L. & Blaya, C. (2025). Beyond the screen: Dissecting the nexus of victimization and cyberhate among adolescents through excessive internet use, online interactions with strangers and parental restrictions. *British Journal of Educational Psychology*, 116(2), 432-459. <https://doi.org/10.1111/bjop.12766>
- Korre, K., Muti, A., Ruggeri, F. & Barrón-Cedeño, A. (2025). Untangling Hate Speech Definitions. A Semantic Componential Analysis Across Cultures and Domains. *arXiv*. <https://doi.org/10.48550/arXiv.2411.07417>
- Kuckatz, U. & Rädiker, S. (2024). *Qualitative Inhaltsanalyse. Methoden, Praxis, Umsetzung mit Software und künstlicher Intelligenz* (6. Aufl.). Beltz Juventa.
- Leander, K. & Burriss, S. K. (2020). Critical Literacy for a Posthuman World: When People Read and Become with Machines. *British Journal of Educational Technology*, 51(4), 1262-1276. <https://doi.org/10.1111/bjet.12924>
- Leschka, J. & Mirbabaie, M. (2025). How to analyze cyberbullying on social media. A systematic literature review in information systems. *I-com*, 24(2), 385-405. <https://doi.org/10.1515/icom-2025-0005>

- Lorenz, A. (2025). Von Hashtags und Geschichte: Vergangenheitsbezogene Hate Speech und ihre Multimodalität auf TikTok. *Publizistik* 70, 261-287.
<https://doi.org/10.1007/s11616-025-00893-8>
- Madukwe, K., Gao, X. & Xue, B. (2020). In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 150-161. <https://doi.org/10.18653/v1/P17>
- Mandyporcius [@mandyporcius]. (2025, 30. Mai). Neu auf TikTok: Kommentarfunktion mit Fotos. Echt easy zu verwenden und ein cooles neues Feature. #oliverporcius #fotos #kommentare #funktion #endlichwasneues #freuemichschonaufrückkommentare [Video]. *TikTok*.
<https://www.tiktok.com/@mandyporcius/video/7510138525547810070> [letzter Zugriff 07.11.2025]
- Matamoros-Fernández, A. (2017). Platformed racism. The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and Youtube. *Information Communication & Society*, 20(6), 930-946. <https://doi.org/10.1111/bjet.12924>
- Matamoros-Fernández, A. & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, 22(2), 205-224.
<https://doi.org/10.1177/1527476420982230>
- Matamoros-Fernández, A. (2023). Taking Humour Seriously on TikTok. *Social Media +Society*, 9(3), 1-3. <https://doi.org/10.1177/20563051231157609>
- Matassi, M. & Boczkowski, P. J. (2023). *To Know is to Compare. Studying Social Media across Nations, Media, and Platforms*. The MIT Press.
- Matsuda, M. J., Lawrence III, C. R., Delgado, R. & Crenshaw, K. W. (1993). *Words that Wound. Critical Race Theory, Assaultive Speech, and the First Amendment*. Routledge.
- Maynard, J. L. & Benesch, S. (2016). Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention. *Genocide Studies and Prevention: An International Journal*, 9(3), 69-95. <http://dx.doi.org/10.5038/1911-9933.9.3.1317>
- McKenna, B., Myers, M. D. & Newman, M. (2017). Social media in qualitative research: Challenges and Recommendations. *Information and Organization*, 27(2), 87-99.
<https://doi.org/10.1016/j.infoandorg.2017.03.001>.
- Mittelstaedt, K. (2019, 27. März). Regierung prüft Auflösung der Identitären – rechtlich könnte das schwierig werden. *DerStandard*.

- <https://www.derstandard.at/story/2000100291943/bundesregierung-prueft-aufloesung-der-identitaeren> [letzter Zugriff 07.11.2025]
- Munn, L. (2020). Angry by design: toxic communication and technical architectures. *Humanities & Social Sciences Communications*, 53(7), 1-11.
- Ng, R. & Indran, N. (2022). Not Too Old for TikTok: How Older Adults Are Reframing Aging. *The Gerontologist*, 62(8), 1207-1216. <https://doi.org/10.1093/geront/gnac055>
- Oesterreich.gv.at. (2025, 6. August). Hasspostings und sonstige strafbare Postings. https://www.oesterreich.gv.at/de/themen/onlinesicherheit_internet_und_neue_medien/internet_und_handy_sicher_durch_die_digitale_welt/3/3 [letzter Zugriff: 17.09.2025]
- P.Y.K.L, S., Das, A., Pulabaigari, V. (2024). Racists spreader is narcissistic; sexists is Machiavellian. Influence of Psycho-Sociological Facets in hate-speech diffusion prediction. *Expert systems with applications*, 247, 1-23. <https://doi.org/10.1016/j.eswa.2024.123211>
- Pachinger, O., Hanbury, A., Neidhardt, J. & Planitzer, A. (2023). Toward Disambiguating the Definitions of Abusive, Offensive, Toxic, and Uncivil Comments. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 107-113. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.c3nlp-1.11>
- Palakodety, S., KhudaBukhsh, A. R. & Carbonell, J. G. (2020). Hope Speech Detection: A Computational Analysis of the Voice of Peace. arXiv. <https://arxiv.org/abs/1909.12940>
- Parlament Österreich (2020, 18. November). Hass-Im-Netz-Bekämpfungsgesetz – HiNBG (481 d.B.). <https://www.parlament.gv.at/gegenstand/XXVII/I/481> [letzter Aufruf: 09.11.2025]
- Peterka, H. (2025, 24. Juli) Berliner TikTok-Mitarbeiter streiken gegen Jobabbau dank KI, die sie selbst trainiert habe. *DerStandard*. <https://www.derstandard.at/story/3000000280754/berliner-tiktok-mitarbeiter-streiken-gegen-jobabbau-dank-ki-die-sie-selbst-trainiert-haben> [letzter Zugriff: 07.11.2025]
- Pichler, G. (2019, 15. März). Livestream von Attentat in Christchurch bringt Facebook und Co. in Bedrängnis. *DerStandard*. <https://www.derstandard.at/story/2000099642025/livestream-von-attentat-in-christchurch-bringt-facebook-und-co-in> [letzter Zugriff: 08.11.2025]

- Roberge, J. & Seyfert, R. (2017). Was sind Algorithuskulturen? In R. Seyfert & J. Roberge (Hrsg.), *Algorithuskulturen: Über die rechnerische Konstruktion der Wirklichkeit*, (1. Aufl., S.7–40). Transcript.
- Rodríguez-González, M., Vargas-Veleta, Y. & Marauri-Castillo, I. (2025). The coverage of fatphobia on YouTube and TikTok: the case of Itziar Castro. *ICONO 14. Scientific Journal of Communication and Emerging Technologies*, 23(1): e2219.
<https://doi.org/10.7195/ri14.v23i1.2219>
- Saferinternet.at (2025, 15. Januar). Jugend-Internet-Monitor 2025.
<https://www.saferinternet.at/news-detail/jugend-internet-monitor-2025> [letzter Zugriff: 07.11.2025]
- Savolainen, L. (2022). The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6), 1091-1109.
<https://doi.org/10.1177/01634437221077174>
- Seemann, N., Lee, Y. S., Hölling, J. & Geierhos, M. (2023). The problem of varying annotations to identify abusive language in social media content. *Natural Language Engineering*, 29, 1561-1585. <https://doi.org/10.1017/S1351324923000098>
- Sellars, A. F. (2016). Defining Hate Speech. Berkman Klein Center Research Publication No. 2016-20. Verfügbar auf SSRN:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244
- Seyfert, R. (2024). Die Theorie algorithmischer Sozialität (TaS). *Österreichische Zeitschrift für Soziologie*, 49, 23-46.
- Siegel, A. A. (2020). Online hate speech. In J. Tucker & N. Persily (Hrsg.), *Social Media and Democracy: The State of the Field* (1. Aufl., S. 56-88). Cambridge University Press.
- Simmel, G. [1908] (1923). *Soziologie: Untersuchungen über die Formen der Vergesellschaftung* (3. Aufl.). Verlag Duncker & Humblot.
- Snyder, C. R., Rand, K. L. & Sigmon, D. R. (2002). Hope Theory. A Member of the Positive Psychology Family. In C.R. Snyder & S. J. Lopez (Hrsg.), *Handbook of Positive Psychology*, (1. Aufl., S. 257-276). Oxford University Press.
- Sorokowski, P., Kowal, M., Zdybek, P. & Oleszkiewicz, A. (2020). Are Online Haters Psychopaths? Psychological Predictors of Online Hating Behavior. *Frontiers in Psychology*, 11, 1-5. <https://doi.org/10.3389/fpsyg.2020.00553>

- Steen, E., Yurechko, K. & Klug, D. (2023). You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, 9(3), 1-17. <https://doi.org/10.1177/20563051231194586>
- [stevendie]. (2025, 30. Juni). Why are all links in Facebook Fanpage comments now plain text and unclickable? *Reddit*. [Online Forum].
https://www.reddit.com/r/facebook/comments/1lo2v1e/why_are_all_links_in_facebook_fanpage_comments/ [letzter Zugriff: 11.11.2025]
- Suler, J. (2004). The Online Disinhibition Effect. *Cyberpsychology & Behavior* 7(3), 321-326.
- TikTok Sicherheitszentrum (2025a, 18. Juli): Hass und hasserfülltem Verhalten entgegentreten. <https://www.tiktok.com/safety/de-de/countering-hate> [letzter Zugriff: 26.09.2025]
- TikTok Sicherheitszentrum (2025b, 18. Juli): Mobbing-Prävention.
<https://www.tiktok.com/safety/de-de/bullying-prevention> [letzter Zugriff: 26.09.2025]
- TikTok Transparenzzentrum (2025, 6. Juni). Bericht zur Durchsetzung der Community-Richtlinien. Januar bis März 2025. <https://www.tiktok.com/transparency/de-de/community-guidelines-enforcement-2025-1> [letzter Zugriff: 07.11.2025]
- Tontodimamma, A., Nissi, E., Sarra A., Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, 126, 157-179.
<https://doi.org/10.1007/s11192-020-03737-6>
- Universität Wien (o.D.). Das Projekt Young People Against Online Hate.
<https://zivildcourage.univie.ac.at/laufende-projekte/projekt-young-people-against-online-hate/> [letzter Zugriff: 08.08.2025]
- University of Applied Sciences St. Pölten (o. D.). Counter Speech – Young People Against Online Hate. Computergestützte Strategien zur Unterstützung von Gegenrede in sozialen Medien. <https://research.fhstp.ac.at/projekte/counter-speech-young-people-against-online-hate> [letzter Zugriff: 08.08.2025]
- Vizcaíno-Verdú, A. & Contreras-Pulido, P. (2024). Cross-platform hatred: influencer's strategies for managing affective aversion. *Cuadernos.info*, 60, 68-92.
<https://doi.org/10.7764/cdi.60.85872>
- Weimann, G. & Masri, N. (2023). Research Note: Spreading Hate on TikTok. *Studies in Conflict & Terrorism*, 46(5), 752-765.

Weimann, G. & Masri, N. (2024) Neuer Antisemitismus auf TikTok. In M. Hübscher & von Mering, S. (Hrsg.), *Antisemitismus in den Sozialen Medien*, (1. Aufl., S. 231-252). Verlag Barbara Budrich.

ZARA – Zivilcourage und Anti-Rassismus-Arbeit (2024). 7. #GegenHassImNetz-Bericht.

September 2023 – August 2024. Verfügbar unter:

<https://assets.zara.or.at/media/ghinbericht/7. GegenHassimNetz Bericht 2024.pdf>

[letzter Zugriff: 15.11.2025]

Zuboff, S. (2018). *Das Zeitalter des Überwachungskapitalismus*. Campus.

9. **Abstract**

9.1. Deutsches Abstract

Die vorliegende Arbeit untersucht die Ausprägungen und Dynamiken von Hate Speech und Counter Speech in TikTok-Kommentaren. Auf Grundlage einer inhaltsanalytischen Auswertungsmethodik nach Kuckartz und Rädiker (2024) wird herausgearbeitet, inwiefern plattformspezifische Besonderheiten diese Kommunikationspraktiken prägen. Zu den zentralen Bedingungen zählen der hohe Anteil junger Nutzer*innen, die algorithmische Verstärkung hasserfüllter Inhalte, die Herausbildung fragmentierter Öffentlichkeiten, die Entpersonalisierung der Online-Erfahrung sowie spezifische Moderations- und Sichtbarkeitsmechanismen. Für Hate Speech und Counter Speech wird jeweils ein eigenständiges Kategoriensystem entwickelt, das eine differenzierte Auseinandersetzung mit ihren jeweiligen Formen und Ausprägungen sowie ihrem wechselseitigen Interaktionslogiken ermöglicht. Theoretisch stützt sich die Arbeit auf das Konflikt- und Streitverständnis nach Simmel (1908) und dessen Weiterentwicklung durch Dennaoui und Witte (2022). Die Analyse zeigt, dass plattformspezifische Bedingungen die Art und Weise, wie zwischenmenschliche Konflikte online ausgetragen werden wesentlich beeinflussen und damit die Dynamiken von Hass und Gegenrede auf ihren Plattformen strukturell mitbestimmen. Die Arbeit bietet eine Grundlage für weiterführende Untersuchungen zu der Vielfalt an Konfliktaustragungsformen in algorithmisch geprägten Kommunikationsräumen.

9.2. Englisches Abstract

This thesis examines the manifestations and dynamics of hate speech and counter speech in TikTok comment sections. Drawing on the content-analytic evaluation methodology proposed by Kuckartz and Rädiker (2024), the study explores how platform-specific characteristics shape these communicative practices. Central conditions include the comparatively high proportion of young users, the algorithmic amplification of hateful content, the formation of fragmented publics, the depersonalization of online experiences, and specific mechanisms of moderation and visibility. Separate categorization systems for hate speech and counter speech are developed, each enabling a differentiated examination of their respective forms, patterns, and reciprocal interaction logics. The theoretical framework builds on Simmel's (1908) concept of conflict and its further development by Dennaoui and Witte (2022). The analysis demonstrates that platform-specific conditions substantially influence the ways in which interpersonal conflicts are enacted online and thus structurally shape the dynamics of hate and counter speech within these environments. The study provides a conceptual and empirical foundation for future research on the diversity of conflict practices that emerge within algorithmically conditioned communication spaces.