

# Abschlussbericht DeTox

*August 2022*

## 1. Methodik und Ergebnisse

### a. Kurzbeschreibung des Projekts

‘Hate Speech’ umfasst das weite Spektrum von dem Gebrauch von Schimpfwörtern über Beleidigungen und Diskriminierungen bis hin zu Gewaltandrohungen (Ruppenhofer et al. 2018). Wir verwenden den Begriff ‘Hate Speech’ stellvertretend für die Vielzahl möglicher offensiver Inhalte. Es ist zu beachten, dass der Begriff Hate Speech, Hasskommentar bzw. Hassrede nicht legal definiert ist. Als Grundlage des Projekts diente die Begriffserklärung der zentralen Meldestelle „Hasskommentare“ des Hessen3C: „Postings, Kommentare und Bilder, die Menschen aufgrund ihrer zugeschriebenen oder tatsächlichen Nationalität, ihrer ethnischen Zugehörigkeit, Hautfarbe, Religionszugehörigkeit, Weltanschauung, physischen und/oder psychischen Behinderung oder Beeinträchtigung, ihres Geschlechts, der sexuellen Orientierung und/oder sexuellen Identität, ihrer politischen Haltung, Einstellung und/oder Engagements, ihres äußeren Erscheinungsbildes oder sozialen Status angreifen, entsprechende Äußerungen fördern, rechtfertigen oder dazu anstiften. Hassrede ist demnach durch seine "gruppenbezogene Menschenfeindlichkeit" gekennzeichnet“<sup>1</sup>.

Das Gesamtziel des Projekts war die Detektion, Identifizierung und Bewertung von Hasskommentaren.

Für die wissenschaftliche Forschung haben wir neue Standards für die Klassifikation gesetzt, die u.a. die strafrechtliche Relevanz der Postings beleuchten. Wir haben aktuelle Methoden zur automatischen Klassifikation auf neue Art kombiniert und weiterentwickelt. Außerdem haben wir einen großen, qualitativ hochwertigen annotierten Textkorpus erstellt, unsere Modelle darauf trainiert und der Forschungsgemeinde zur Verfügung gestellt.

Für den Transfer haben wir mit der Meldestelle Hasskommentare kooperiert und so die wissenschaftlichen Erkenntnisse direkt in die Anwendung überführt. Es wurde eine Demo-Version erstellt, um die Klassifikationsmethoden und deren Ergebnisse darzustellen. Ein

---

<sup>1</sup><https://innen.hessen.de/sicherheit/cyber-competence-center-hessen3c/zentrale-meldestelle-hasskommentare>

Extraktionstool erleichtert den Anwender\*innen im H3C direkt die Arbeit. Die Klassifikationsmethoden wurden schließlich in das Extraktionstool integriert.

### **b. Forschungsmethodik**

Die enge Kooperation und ständige Kommunikation zwischen Forschenden und Anwendenden hatte diese Folgen:

- Die Entwicklung von Tools war durch den Bedarf der Anwender direkt gesteuert.
- Aktuelle wissenschaftliche Erkenntnisse flossen direkt in Anwenderwerkzeuge ein.
- Dem Projekt standen realistische und aktuelle Daten zur Verfügung, mit denen die Modelle trainiert werden konnten.
- Das neuartige Annotationsschema für die Klassifikation wurde in enger Zusammenarbeit zwischen Forschung und Praxis entwickelt.
- Die Annotation des Datenkorpus anhand dieses Annotationsschemas wurde wissenschaftlich stringent durchgeführt und evaluiert.

Die Forschung und Entwicklung im Projekt wurde in fünf Arbeitspaketen durchgeführt:

1. Projektkoordination, Dokumentation und Dissemination
2. Datensammlung und Annotation
3. Identifikation durch Klassifikation
4. Bestimmung der Toxizität und Online-Aggression
5. Netzwerkanalyse

### **c. Forschungs- und Projektergebnisse**

Die detaillierten Forschungs- und Projektergebnisse sind in den Deliverables der Arbeitspakete beschrieben, die diesem Abschlussbericht beigelegt sind. In diesem Dokument stehen daher nur Zusammenfassungen.

#### **AP 1: Projektkoordination, Dokumentation und Dissemination**

In diesem Arbeitspaket wurden das Projekt geplant und die Kommunikationsstrukturen aufgebaut (Deliverable 1.1.: Projektplanung und interne Abstimmung). Es wurde die wissenschaftliche Dissemination geplant und durchgeführt (siehe Abschnitt e. Wissenschaftliche Publikationen). Darüber hinaus wurde eine externe Webpräsenz für das Projekt aufgebaut (<https://projects.fzai.h-da.de/detox/>), und es wurden interne Daten- und Kommunikationsplattformen errichtet (Deliverable 1.3: Interne und externe Webpräsenz).

Das Projekt stand unter der Herausforderung, präsenzfrei agieren zu müssen. Diese Herausforderung konnte im Projekt gut gemeistert werden, durch Maßnahmen, die die Kommunikation effizient unterstützen.

Die Projektkoordination wurde durch diese Maßnahmen organisiert:

- Virtuelle Projektmeetings in jeder 2. Woche, mit Protokoll
- Wöchentliche virtuelle Arbeitstreffen
- Tägliche Kommunikation über den Slack-Kanal
- Austausch von Dokumenten über einen Server an der Hochschule Darmstadt
- Austausch von Software über einen Github-Account

## AP 2: Datensammlung und Annotation

Im Rahmen des Projekts sind zwei Datensätze entstanden (Deliverable 2.1: Datensatz von toxischen und nicht toxischen Inhalten):

- Ein Datensatz aus Kommentaren, Tweets und Beiträgen, die dem hessischen Cyber Competence Center (H3C) gemeldet und für das Projekt transkribiert wurden
- Ein für die Shared Task GermEval 2021 von uns zusammengestellter Korpus aus 1,1 Millionen Tweets, die sich auf deutsche Talkshows im Zeitraum 2019 beziehen

Ein Teil der Kommentare in den erstellten Datensätzen wurden zum Training von Modellen zur Klassifikation manuell nach dem erarbeiteten Annotationsschema annotiert (Deliverable 2.2: Annotierter Datensatz). Dazu wurde ein Annotationstool an der Hochschule Mittweida aufgebaut.

Bis zum Ende der Annotationsphase Ende März 2022 wurden 12447 Kommentare von jeweils drei Personen annotiert. Die Qualität der Annotationen wurde nach wissenschaftlich stringenten Methoden evaluiert.

Um eine Erweiterung des Datensatzes über die Projektlaufzeit hinaus zu ermöglichen, wurde ein Feedbacksystem in das Extraktionstool integriert (Deliverable 2.3: Feedback-System). Das Extraktionstool wurde weiterhin durch eine OCR-Texterkennung erweitert, sodass Texte auf Bildern direkt in Text umgewandelt werden können. Damit ist es möglich, Screenshots direkt umzuwandeln, der manuelle Schritt des Abtippens entfällt.

## AP 3: Identifikation durch Klassifikation

Die zentralen Forschungsarbeiten fanden in diesem Arbeitspaket statt. Hier wurden Klassifikationsmethoden evaluiert, entwickelt, neu kombiniert und auf den neuen Datensätzen

trainiert. Die Ergebnisse sind nicht nur im „Deliverable 3.1: Ergebnisse der Klassifikationsexperimente“ dokumentiert, sondern auch in den wissenschaftlichen Publikationen des Projekts.

In der ersten Projektphase haben wir an der Shared Task „GermEval2021 - Toxic, Engaging, & Fact-Claiming Comments“ teilgenommen. Diese Teilnahme hatte mehrere Vorteile: Wir bekamen Zugang zu annotierten Textdaten, wir konnten unsere Experimente im realen Umfeld testen und bekamen Feedback und den direkten Vergleich mit anderen Forschungsgruppen, wir konnten am dazugehörigen Workshop teilnehmen und unsere Klassifikationsexperimente mit anderen Wissenschaftler\*innen aus dem Forschungsgebiet diskutieren und wir konnten unsere Experimente in einem Text im Tagungsband des Workshops publizieren.

Im Projekt wurden verschiedene Modelle miteinander auf neuartige Weise kombiniert:

- Neuronale Netze und SVM für binäre Klassifikationen (z.B. Hate Speech)
- Multi-Label Transformer für multiple Klassen (z.B. Paragraphen und Target)
- Pattern Matching für Klassen mit extrem wenigen Beispielen (z.B. Gefahr)

Dabei wurde das Problem des „Overfittings“, also die potenzielle Überanpassung an Trainingsdaten, die dazu führt, dass neue Daten schlechter klassifiziert werden können, analysiert und vermieden.

Machine Learning Modelle, die wir hier genutzt haben, werden oft auch mit dem Term „Black-Box“ assoziiert, da diese immer weniger transparent für Menschen sind. Umso besser solche Modelle den Kontext „verstehen“, desto komplizierter wird es die Vorhersagen der Modelle nachzuvollziehen. Aus diesem Grund haben wir uns mit der Erklärbarkeit der Modelle beschäftigt, die im Kontext von Hate Speech besonders wichtig ist (Deliverable 3.2: Erklärbare Klassifikation).

#### AP 4: Bestimmung der Toxizität und Online-Aggression

In enger Kooperation zwischen den Projektpartnern und dem H3C wurden Annotationsrichtlinien erstellt, die die Klassen toxischer Inhalte genauer beschreiben und die als Grundlage für die Annotationen dienten (Deliverable 4.1.: Klassen toxischer Inhalte). Nach Abschluss der Annotationen wurde untersucht, wie sich die Klassen in den Daten verteilen (Deliverable 4.2: Statistische Daten zur Verteilung der toxischen Inhalte auf die Klassen).

## AP 5: Netzwerkanalyse

Neben der Detektion und Bewertung von Hasskommentaren ist es notwendig zu verstehen, wie auf Hassbotschaften reagiert wird, denn erst dadurch kann sich Hass in einem Netzwerk weiterverbreiten. Von Bedeutung ist dabei nicht nur der Hasskommentar selbst, sondern zusätzlich spielen Nutzereigenschaften und deren Position im Netzwerk eine Rolle. Dieses war Thema des Arbeitspakets „Netzwerkanalyse“ (Deliverable 5.1: Netzwerkanalyse). Die Verbreitung des Sentiments im Zusammenhang mit toxischen Postings wurde ebenfalls im Arbeitspaket untersucht (Deliverable 5.2: Sentimentanalyse).

**d. Wissenstransfer, durch bspw. Lehrveranstaltungen, Abschlussarbeiten**

Mina Schütz promoviert seit dem 01.01.2021 an der Hochschule Darmstadt zum Thema „Disinformation Detection: A Visual and Explainable Semi-Supervised Transfer Learning Approach“ (Arbeitstitel). Diese Arbeit steht im unmittelbaren Projektzusammenhang.

Jonas Pitz hat seine Masterarbeit zum Thema „Methoden der Klassifikation von Hasskommentaren im Internet“, die im unmittelbaren Projektkontext entstanden ist, im September 2021 fertiggestellt und damit sein Masterstudium abgeschlossen

Sebastian Kraußold hat im Projekt sein Praktikum (im Rahmen des BA-Studiengangs Information Science an der Hochschule Darmstadt) abgeschlossen und eine Nutzerdokumentation für die Tools, die im Projekt entwickelt wurden, geschrieben.

Durch die Einbeziehung von Studierenden als studentische Hilfskräfte können diese ihre im Studium erworbenen Fähigkeiten im Projekt praktisch anwenden und wiederum Erfahrungen, die sie im Projekt erworben haben, im Studium einsetzen.

Die im Projekt entstandenen Datensätze werden in Lehrveranstaltungen der Hochschule Darmstadt und der Hochschule Mittweida verwendet.

**e. Wissenschaftliche Publikationen**

## Textbeiträge

Mina Schütz, Alexander Schindler, Melanie Siegel (2021). Disinformation Detection: An Explainable Transfer Learning Approach. In CODE Conference 2021 – Science Workshop for Ph.D. and masters' theses research proposals.

Schütz, Mina and Demus, Christoph and Pitz, Jonas and Probol, Nadine and Siegel, Melanie and Labudde, Dirk: DeTox at GermEval 2021: Toxic Comment Classification. In Proceedings of GermEval 2021.

Demus, Christoph and Pitz, Jonas and Schütz, Mina and Probol, Nadine and Siegel, Melanie and Labudde, Dirk. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Köhler, J., Shahi, G.K., Struß, J.M., Wiegand, M., Siegel, M., Mandl, T., Schütz, M.: Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection. In: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum. CLEF '2022, Bologna, Italy (2022)

*Noch nicht erschienen, aber angenommen:*

Christoph Demus and Mina Schütz and Jonas Pitz and Nadine Probol and Melanie Siegel and Dirk Labudde. 2022. Hass im Netz – Aggressivität und Toxizität von Hasskommentaren und Postings, Detektion und Analyse. In: Handbuch Cyberkriminologie, hrsg. von Thomas-Gabriel Rüdiger und Petra Saskia Bayerl. Springer.

Christoph Demus, Dirk Labudde, Jonas Pitz, Nadine Probol, Mina Schütz, Melanie Siegel. 2022. Automatische Klassifikation offensiver deutscher Sprache in sozialen Netzwerken. In: Digitale Hate Speech, hrsg. von Sylvia Jaki und Stefan Steiger. Springer.

Fachvorträge:

Mina Schütz, Alexander Schindler, Melanie Siegel (2021). Disinformation Detection: An Explainable Transfer Learning Approach. In *CODE Conference 2021 – Science Workshop for Ph.D. and masters' theses research proposals*.  
[https://www.unibw.de/code/events/code2021\\_content/code-2021-cfp-science-track](https://www.unibw.de/code/events/code2021_content/code-2021-cfp-science-track)

Eingeladener Vortrag auf der Online-Tagung: Interdisziplinäre Perspektiven auf Hate Speech und ihre Erkennung (IPHSE) (8.2.2021) (Melanie Siegel)

Eingeladener Vortrag auf der Konferenz "Qurator 2021 – Conference on Digital Curation Technologies" (10.2.2021) (Melanie Siegel)

Eingeladener Vortrag bei der Digitalstadt Darmstadt (20.4.21) (Melanie Siegel)

Eingeladener Vortrag bei „Hessenmetall Verband der Metall- und Elektro-Unternehmen Hessen e.V.“: "Hass und Hetze im Netz" (15.6.2021) (Melanie Siegel)

Ringvorlesung „Cybercrime: Wie Künstliche Intelligenz uns täuschen kann“ (Dirk Labudde)

Jonas Pitz: Vortrag zum Projekt beim Science Wednesday der Hochschule Darmstadt (17.11.2021).

Melanie Siegel: Vortrag zum Projekt beim Forschungszentrum für Angewandte Informatik (18.11.2021)

Die Teilnahme von Jonas Pitz an der Poster-Session der Deutschen Gesellschaft für Sprachwissenschaften (DGfS) 2022.

Online-Vortrag von M. Siegel für die Gesellschaft für Informatik am 22.3.2022.

Vortrag beim Forensik-Abend der Hochschule Mittweida: Der Einsatz von Sprachtechnologie gegen Online-Hetze

Vortrag beim ECT Lab Online-Seminar "Just digital futures": DeTox – Detection of toxicity and aggression in postings and comments on the internet

Vortrag beim Datenjournalistentreffen der ARD in Frankfurt: Impuls: Was man mit dem Computer aus Texten alles herauslesen kann - und was nicht

#### Sonstige

Zahlreiche Presse-Veröffentlichungen, u.a. WDR, Darmstädter Echo, abitur-und-studium.de, Lauterbacher Anzeiger, Berliner Zeitung, der Frankfurter Neuen Presse und im Deutschlandfunk, Frankfurter Neuen Presse, Osthessen News, Die Zeit, WELT, Süddeutsche Zeitung, heise online, Gießener Allgemeine Zeitung, BILD Frankfurt

## 2. Diskussion und Bewertung

Aufgrund der Corona-Situation mussten die Projektkollaborationen in DeTox fast vollständig online stattfinden. Die Mitarbeiter\*innen im Projekt haben auf professionelle Art Kommunikations- und Kollaborationsstrukturen aufgebaut, die das ermöglicht haben, sodass in den Ergebnissen keine Einschränkungen zu spüren sind.

Im Vergleich zur Planung bestand ein stärkerer Fokus auf dem Aufbau eines qualitativ und quantitativ hochwertigen annotierten Datensatzes deutschsprachiger Posts. Auch der Einsatz von studentischen Hilfskräften zur Annotation war nicht Teil der Planung. Durch die frühe Beteiligung der Projektgruppe an einer Shared Task wurde klar, dass ein solcher Datensatz

für die Klassifikationsmodelle grundlegend ist und dass er in der Forschungslandschaft noch nicht verfügbar war. Wir haben die Aufstellung daher als zentrales Thema gesehen. Der Einsatz von Studierenden zu Annotation ermöglichte uns eine wissenschaftlich stringente Evaluation der Annotationen und erhöhte damit die Qualität. Für die Studierenden selbst war der Einsatz ein interessanter Einblick in die Forschungsarbeiten zur automatischen Klassifikation von toxischen Posts.

Die Forschungsergebnisse aus DeTox orientieren sich am aktuellen Stand der Forschung und gehen darüber hinaus. Die Teilnahme an der GermEval Shared Task gleich zu Beginn des Projekts brachte die Projektmitarbeiter\*innen direkt in Kontakt mit dem aktuellen Stand der Forschung und mit Forscher\*innen im Fachgebiet. Gleichzeitig arbeitete Jonas Pitz an seiner Masterarbeit zum Thema, in der er aktuelle Forschungsergebnisse untersuchte. Mina Schütz arbeitet an ihrer Promotion an der Hochschule Darmstadt, wobei sich Projektergebnisse und Promotionsarbeiten gegenseitig befruchten.

Die enge Kooperation mit dem H3C führte zu Forschungs- und Entwicklungsarbeiten, die sich eng an der Anwendung orientieren.

Die Arbeiten im Projekt wurden in der Öffentlichkeit stark wahrgenommen, zahlreiche Presse-Veröffentlichungen und Interviews haben das Projekt begleitet.

### **3. Ergebnistransfer**

Im Projekt ist zunächst ein Demonstrator entstanden, in dem die Klassifikationsmethoden gezeigt und die Modelle evaluiert werden können.



DeTox - Detektion und Klassifikation von Hatespeech in Kommentaren
Menu ▾

Geben Sie einen Kommentar ein:

...und aus die Maus

Toxizität berechnen

**Toxizität**

5

Wahrscheinlichkeit: Sehr hoch

**Hate Speech**

1

Wahrscheinlichkeit: Sehr hoch

**Extremismus**

0.67

Wahrscheinlichkeit: Uneindeutig

**Gefahr**

0.33

Wahrscheinlichkeit: Gering

**Strafrechtliche Relevanz**

§130 STGB Volksverhetzung

Der Demonstrator gibt außerdem eine Analyse der Daten.



Dieser Demonstrator wurde an den Anwender H3C übergeben.

Im nächsten Schritt wurden die Arbeitsschritte beim Anwender analysiert, um eine maximale Unterstützung mit den entwickelten Werkzeugen zu erreichen. Daraus entstand das Extraktions- und Klassifikationstool, das nun beim H3C eingesetzt wird. Es unterstützt die Arbeit folgendermaßen:

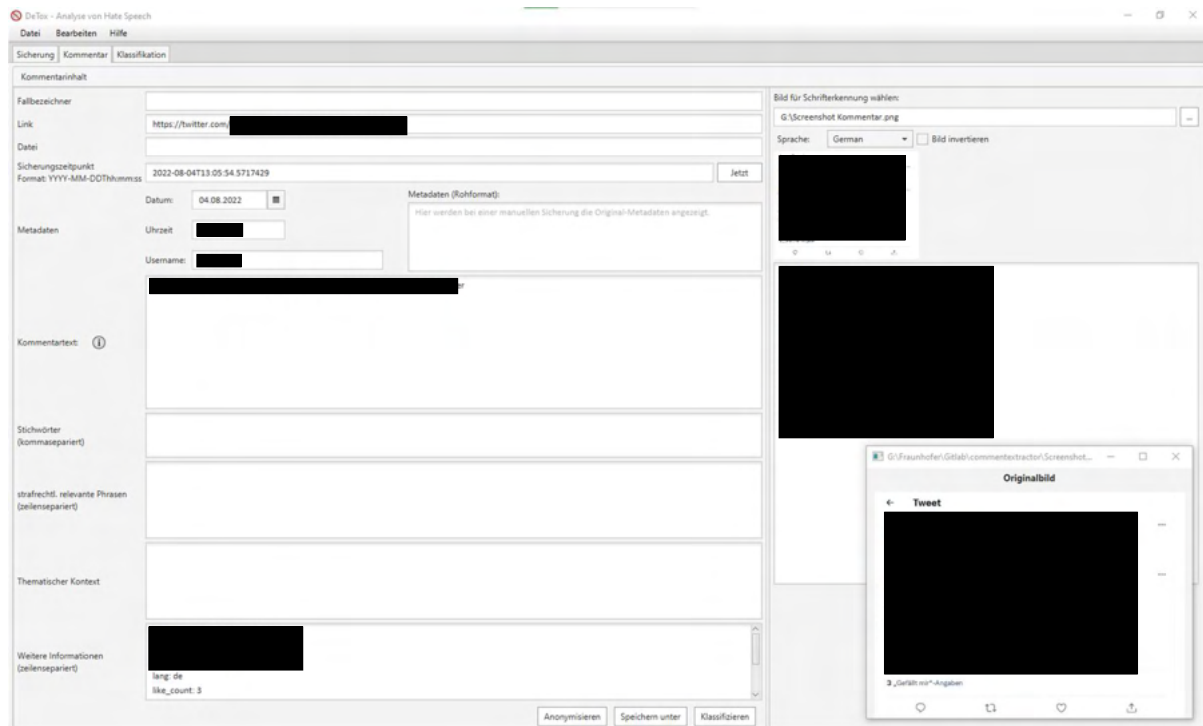
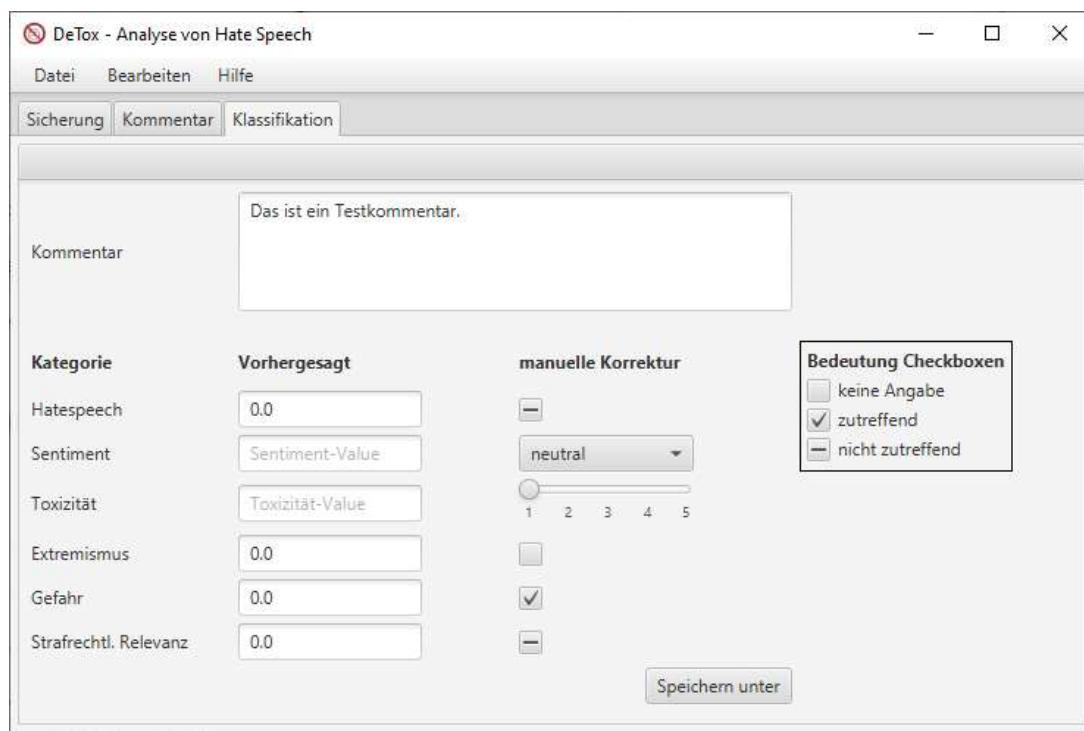
- Die Anwenderin kann einen Link zu einem gemeldeten Post aus Twitter, Youtube oder einer anderen Quelle angeben und bekommt automatisch den Kommentartext und Metadaten in einem Formular angezeigt, das sie dann mit Informationen zur Toxizität und strafrechtlicher Relevanz anreichern kann.
- Wenn eine Meldung per Screenshot erfolgt, kann der Text daraus mit OCR automatisch extrahiert werden.
- Mit einem Klick auf „Klassifizieren“ kann sie den Kommentartext automatisch mit den DeTox-Modellen vorklassifizieren. Diese Klassifikation kann sie manuell korrigieren, falls notwendig und den Post zur weiteren Bearbeitung abspeichern.
- Der Post wird als JSON-Datei zur weiteren Verarbeitung gespeichert.

Damit wurden Werkzeuge entwickelt, die sich direkt am Bedarf der Anwender beim H3C orientieren und diese unmittelbar in ihrer täglichen Arbeit unterstützen.

The screenshot shows the 'DeTox - Analyse von Hate Speech' application window. It has a menu bar with 'Datei', 'Bearbeiten', and 'Hilfe'. Below the menu is a tabbed interface with 'Sicherung', 'Kommentar', and 'Klassifikation'. The 'Kommentar' tab is active, showing an 'Übersicht' section with a text input for 'Link des Kommentars' containing a Twitter URL. A 'Link Prüfen' button is next to it. Below this, a message states: 'Als Quelle wurde TWITTER identifiziert. Kommentar kann automatisch abgerufen werden.' followed by the same URL. The 'Twitter Sicherung' section is expanded, showing a 'Twitter-ID' input field and three checked checkboxes: 'abzurufende Informationen', 'Tweet', 'Public Metrics', and 'Autor (Name, Username)'. A 'Daten abrufen' button is at the bottom of this section. Below the 'Twitter Sicherung' section are two collapsed sections: 'Manuelle Sicherung' and 'YouTube Sicherung'.

## Detektion von Toxizität und Aggressionen in Postings und Kommentaren im Netz

(Hass ist Gift)

## 4. Fazit und Ausblick

Das Projekt DeTox hat zu nachhaltigen Ergebnissen geführt, die über die Projektlaufzeit hinaus hohe Relevanz haben.

Auf der Ebene der Forschung sind da vor allem die annotierten Daten zu nennen, die die Forschung zur deutschsprachigen Klassifikation erheblich beeinflussen werden. Nicht nur Qualität und Quantität der Daten sind dabei neuartig, sondern auch das Annotationsschema, das wesentlich komplexer ist als nur die binäre Klassifikation in Hate Speech oder nicht, die die meisten bisherigen Annotationen verwenden. Die Publikation der Beschreibung unserer Daten bei der WOAH-Konferenz im Juli 2022 führte bereits zu Anfragen nach den Daten. Die Daten werden darüber hinaus in Lehrveranstaltungen an der Hochschule Mittweida und der Hochschule Darmstadt eingesetzt werden. Mina Schütz wird sie für Arbeiten an ihrer Promotion verwenden.

Aber auch die entwickelten Klassifikationsmodelle stehen zur Verwendung und Weiterentwicklung zur Verfügung. Auch diese werden von Mina Schütz verwendet und auch in der Forschungsgruppe von Prof. Siegel weiterentwickelt.

Die Ergebnisse der Arbeiten zur Netzwerkanalyse haben hohe Relevanz für die Forschung in der forensischen Analytik und werden von der Forschungsgruppe von Prof. Labudde aufgegriffen.

Die entwickelten Anwendungen kommen im H3C zum Einsatz und unterstützen dort die tägliche Arbeit.

Die Forschungsgruppe wird versuchen, weitere Forschungsmittel zu akquirieren, um die Arbeiten fortzuführen. Das Potenzial der im Projekt entwickelten Daten ist noch längst nicht ausgeschöpft. Hier einige Ideen:

- Die Netzwerkanalyse im Zusammenhang mit der Analyse der Toxizität sollte dazu führen, dass Strategien zum Umgang mit toxischen Posts entwickelt werden. Welche Strategien sind hilfreich, um Aggression zu stoppen? Was kann man tun, um Personen, die von Hate Speech betroffen sind, effektiv zu unterstützen, sodass es nicht zu Folgen wie bei der Ärztin Kellermayr (<https://www.tagesschau.de/ausland/europa/kellermayr-corona-aerztin-tot-101.html>) kommt?
- Der Aufbau von Medienkompetenz vor allem bei Kindern und Jugendlichen ist ein wesentlicher Faktor, um gegen Desinformation und Hate Speech vorzugehen. Die Ergebnisse des Projekts könnten in ein Portal einfließen, das hilft, diese Medienkompetenz aufzubauen und das in Schulen eingesetzt werden kann.
- Der Zusammenhang zwischen toxischen Beiträgen und Desinformation sollte weiter untersucht werden. Sind es dieselben User, die Hate Speech und Fake News verbreiten? Sind es dieselben Themen? Mina Schütz und Melanie Siegel haben an der

Organisation der Shared Task „CLEF-2022 CheckThat! Lab Task 3 on Fake News Detection“ 2022 mitgewirkt. Jetzt gilt es, die beiden Gebiete miteinander zu verknüpfen.

- In der Kooperation mit dem H3C sollte ein Fokus auf die strafrechtliche Relevanz von Posts gelegt werden. Diese Klassifikation ist in der Forschung noch unbeachtet, in der Anwendung aber äußerst relevant. Die manuelle Verbesserung der automatischen Klassifikationsergebnisse kann in die Modelle einfließen und diese damit effektiv verbessern.

# Deliverable 1.1.: Projektplanung und interne Abstimmung

31. Januar 2021

Das Projekt „DeTox: Detektion von Toxizität und Aggressionen in Postings und Kommentaren im Netz“ ist am 1. Januar 2021 gestartet. Die ersten vier Wochen des Projekts waren von organisatorischen Aufgaben der internen Abstimmung geprägt (Arbeiten in AP 1).

## Zusammenstellung des Teams

Für das Team an der h\_da wurde als studentische Hilfskraft zum 1.1.2021 Nadine Probol eingestellt. Weiterhin wurde Jonas Pitz zum 15.1.2021 ebenfalls als studentische Hilfskraft eingestellt. Er wird in diesem Sommersemester sein Master-Studium der Informationswissenschaft abschließen und steht dann als wissenschaftlicher Mitarbeiter im Projekt zur Verfügung. Mina Schütz wird auf einer 25%-Stelle als wissenschaftliche Mitarbeiterin im Projekt arbeiten. Der Arbeitsvertrag dafür startet am 1. Februar 2021.

Für das Team am SIT wurde Christoph Demus zum 15.1.2021 auf einer 50%-Stelle als wissenschaftlicher Mitarbeiter am Standort Mittweida eingestellt.

## Kick-Off-Workshop

Der Kick-Off-Workshop fand am 13.1.2021 als virtuelles Meeting statt. Teilgenommen haben:

- [REDACTED]
- [REDACTED]
- Prof. Dr. Dirk Labudde (F-SIT)
- Jonas Pitz (h\_da)
- Nadine Probol (h\_da)
- [REDACTED]
- Mina Schütz (h\_da)
- Prof. Dr. Melanie Siegel (h\_da)
- [REDACTED]
- [REDACTED]

**Protokoll Kick-Off (Protokollant: [REDACTED]):**

1. Kurzvorstellung Projekt
2. Beratung der Projektmeilensteine / Vereinbarung Jour Fixe-Termine
  - a. Jour Fixe-Termine alle drei Monate (3, 6, 9, 12, 15, 18)

- b. ggf. zusätzliche JF bei Bedarf
  - c. Einladung zu JF durch VII 4
  - d. Sachstandsbericht/Ergebnisprotokoll durch Zuwendungsempfänger
  - e. Abnahme durch Fachreferat und finale Abnahme durch VII 4
- 3. Berichtspflichten (Formulare, Fristen etc.)
  - a. Bereitstellung eines einheitlichen Formulars pro Bericht durch VII 4 (s. Anhang)
  - b. Berichtspflichten durch Zuwendungsempfänger, nach
    - (1) Kick-Off mit detaillierter Projektplanung (bis 25.01.21),
    - (2) Zwischenbericht in der Mitte der Projektlaufzeit (bis 30.09.21) und
    - (3) Abschlussbericht (bis 31.10.22)
  - c. Fachliche Abnahme durch Fachreferat
  - d. Finale Abnahme durch VII 4
- 4. Zuleistungen HMdIS
  - a. werden in JF zu o.g. Projektmeilensteinen, bzw. nach Bedarf, beraten
- 5. Veröffentlichungen
  - a. ggf. über Beirat Cybersicherheit und Plattform Cybersicherheitsforschung
  - b. zu gegebener Zeit (ca. Projekthalbzeit) werden Veröffentlichungswege beraten und festgelegt

## Kommunikation im Projekt

Es wurde eine projektinterne Austauschplattform für Daten an der Hochschule Darmstadt eingerichtet: [https://\[REDACTED\]](https://[REDACTED])

Als projektinternen Kommunikationskanal nutzen wir Slack [\[REDACTED\]](#)

Regelmäßige Projekttreffen in Form virtueller Webkonferenzen finden alle zwei Wochen statt. Genutzt wird dafür der virtuelle Besprechungsraum

[\[REDACTED\]](#), der auf einem Server der Hochschule Darmstadt installiert ist.

## Öffentlichkeitsarbeit

Es wurde eine Projekt-Webseite aufgestellt: <https://projects.fzai.h-da.de/detox/>

Zwei Vorträge im Rahmen des Projekts sind für Februar geplant:

- Titel: Automatische Klassifikation offensiver Sprache – Erfahrungen aus zwei Shared Tasks  
Vortragende: Prof. Dr. Melanie Siegel (Hochschule Darmstadt)  
Datum/Uhrzeit: 10.2.2021, 14:00 Uhr  
Tagung: Workshop Medien der Qurator Conference 2021
- Titel: Automatische Klassifikation offensiver Sprache – Erfahrungen aus zwei Shared Tasks  
Vortragende: Prof. Dr. Melanie Siegel (Hochschule Darmstadt)  
Datum/Uhrzeit: 8.2.2021, 14:15 Uhr  
Tagung: Interdisziplinäre Perspektiven auf Hate Speech und ihre Erkennung (IPHSE)



# Deliverable 1.3:

## Interne und externe Webpräsenz

26. März 2021

Für das Projekt wurden eine externe und eine interne Webpräsenz eingerichtet.

### Externe Webpräsenz

Das Ziel der externen Webpräsenz ist die Dissemination der Projektergebnisse und die Diskussion mit der Forschungsgemeinde.

#### Zugang

Die Website ist öffentlich erreichbar über <https://projects.fzai.h-da.de/detox/>. Alle Projektteilnehmer haben Zugang zum Backend der Website und können Änderungen und Updates vornehmen.

#### Struktur

Die Homepage besitzt:

- Eine allgemeine Seite
- Eine Seite für News
- Eine Seite mit Informationen über das Projekt
- Eine Seite, auf der die Partner aufgelistet sind
- Eine Seite mit den Ergebnissen
- Ein Impressum
- Eine Seite zum Datenschutz

### Interne Webpräsenz

Die interne Webpräsenz dient der internen Abstimmung im Projekt und unterliegt den Richtlinien zum Datenschutz. Dort werden Zwischenergebnisse und Textkorpora ausgetauscht und interne wissenschaftliche Diskussionen geführt.

#### Synology-Server

Zum Austausch von Daten und Korpora wurde ein Synology-Server eingerichtet, der den Projektteilnehmern über [REDACTED] zugänglich ist. Sie haben dafür Zugangsdaten bekommen.

Der Synology-Server ist mithilfe von Ordnern organisierbar.

**Slack-Workspace**

Für die interne Kommunikation wurde zudem ein Slack-Workspace eingerichtet. Er ist den Teilnehmern des Projekts über [REDACTED] zugänglich. Der Workspace ist in verschiedene Channels unterteilt, um Kommunikation zu verschiedenen Themen zu sortieren und gleichzeitig schnell zu ermöglichen.

**GitLab**

Des Weiteren wurde ein internes GitLab eingerichtet mit einer Projektgruppe. Dieses Git ist erreichbar über [REDACTED]. Die Projektteilnehmer haben alle einen persönlichen Zugang zum GitLab.

In einer readme-Datei wird eine Literaturliste angelegt und zur Verfügung gestellt.

# Deliverable 2.1: Datensatz von toxischen und nicht toxischen Inhalten

14.10.2021

Im Rahmen des Projekts sind zwei Datensätze entstanden:

- 1. Ein Datensatz aus Kommentaren, Tweets und Beiträgen, die dem hessischen Cyber Competence Center (H3C) gemeldet und für das Projekt transkribiert wurden
- 2. Ein für die Shared Task GermEval 2021 von uns zusammengestellter Korpus aus 1,1 Millionen Tweets, die sich auf deutsche Talkshows im Zeitraum 2019 beziehen

Die Datensätze liegen auf dem Synology-Server und werden im weiteren Verlauf des Projekts zum Teil gelabelt und veröffentlicht.

## H3C-Datensatz:

Bisher wurden 1457 Daten vom H3C transkribiert und an uns übergeben. Diese bestehen größtenteils aus Tweets, Facebook-Komentaren, Forenbeiträgen und anderen Online-Komentaren sowie deren Datum und Uhrzeit und ggf. Kontext. Außerdem wurden relevante Phrasen markiert, die auf einen Hasskommentar hindeuten können. Die Daten sind fast ausschließlich toxisch und häufig Hasskommentare.

## GermEval 2021 Twitter-Korpus:

Für die GermEval 2021 Shared Task zur Klassifikation von deutschen Facebook-Komentaren in toxisch oder nicht toxisch wurden von uns ca. 1,1 Millionen Tweets gesammelt. Diese wurden nach Bezug zu verschiedenen deutschen politischen Talk-Shows ausgewählt, um dem von der GermEval zur Verfügung gestellten Datensatz zu ähneln. Dieser Korpus kann zum einen zum Vortrainieren, zum Beispiel von Transformer-Modellen, verwendet werden. Zum anderen können daraus kleinere Datensätze extrahiert und gelabelt

werden. Eine Teilmenge des Twitter-Korpus bildet ein Datensatz mit 5555 Tweets, die Profanität und Hassworte enthalten sowie automatisch als toxisch vorsortiert wurden.

## Deliverable 2.2: Annotierter Datensatz

25.05.2022

Ein Teil der Kommentare in dem erstellten Datensatz (siehe auch Deliverable 2.1) werden zum Training von Modellen zur Klassifikation manuell nach dem erarbeiteten Annotationsschema (Deliverable 4.1) annotiert. Dafür wurden 6 Bachelor-Studierende der Hochschulen Darmstadt und Mittweida eingestellt. Alle studieren im Bachelor in den Fachbereichen Informationswissenschaften oder Digitale Forensik.

Zu Beginn wurden Probeannotationen durchgeführt, damit die Hilfskräfte ein ähnliches Verständnis von Hatespeech und Toxizität im Rahmen dieses Projekts erlangen konnten und eine einheitliche Annotation erreicht wird. In diesem Zusammenhang wurden Kommentare mit stark unterschiedlichen Annotationen der Hilfskräfte ausführlich in der Gruppe diskutiert. Da sich die Diskussion von Kommentaren positiv auf das Ergebnis ausgewirkt hat, wird diese über die gesamte Zeit der Annotation beibehalten (im Rhythmus von ein bis zwei Wochen).

Nach den Probeannotationen wurden die Probanden in zwei Dreiergruppen aufgeteilt, wobei jede Gruppe dann unterschiedliche Kommentare annotiert. Dadurch kann eine große Menge Kommentare annotiert werden und trotzdem wird jeder Kommentar noch von drei Personen annotiert.

Die zu annotierenden Datensätze setzen sich aus Kommentaren unterschiedlicher Quellen zusammen (Abbildung 1). Neben Kommentaren vom Hessen3C wurden Kommentare von Twitter annotiert. Letztere wurden nochmals unterschieden in Twitter-Konversationen (ganze Konversationen werden annotiert) und in Twitter-Kommentare, die mit Hilfe einer Stichwortsuche von Twitter bezogen und vorgefiltert wurden, um einige harmlose Kommentare rauszufiltern.

Bis zum Ende der Annotationsphase Ende März 2022 wurden 12447 Kommentare von jeweils drei Personen annotiert. Davon wurden rund 20% als Hatespeech bzw. toxisch annotiert und ca. 6 % wurden als strafrechtlich relevant gelabelt. Die genauen Zahlen der Häufigkeiten der annotierten Klassen sind in Tabelle 1 und Abbildung 3 dargestellt.

Für die Annotationen wurde das Inter-Annotator-Agreement, d.h. die Übereinstimmung der Annotatoren in ihren Annotationen, bestimmt. Als Maß wird Gwet's-AC1-Koeffizient betrachtet (Abbildung 2). Der Wert liegt zwischen -1 und 1, wobei 1 eine komplette Übereinstimmung der

Annotationen bedeutet. Ab 0.6 ist eine gute Übereinstimmung, ab 0.8 eine sehr gute Übereinstimmung der Annotationen gegeben. Wie im Diagramm erkennbar ist die Übereinstimmung der Annotationen für die meisten Label gut bis sehr gut. Lediglich bei Gefahr liegt Gruppe 1 im Bereich der moderaten Übereinstimmung.

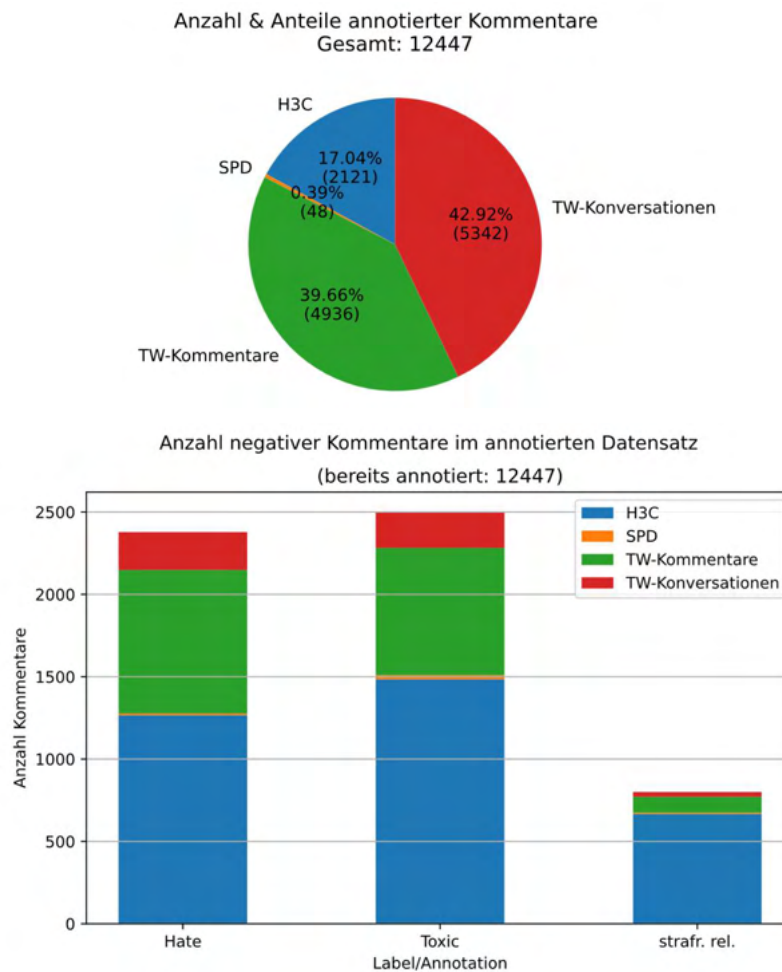


Abbildung 1: Menge und Aufteilung der Kommentare im annotierten Korpus (oben) und Anteile negativer Kommentare (Hatespeech, toxisch oder strafrechtl. relevant) in den verschiedenen Quellen(unten)

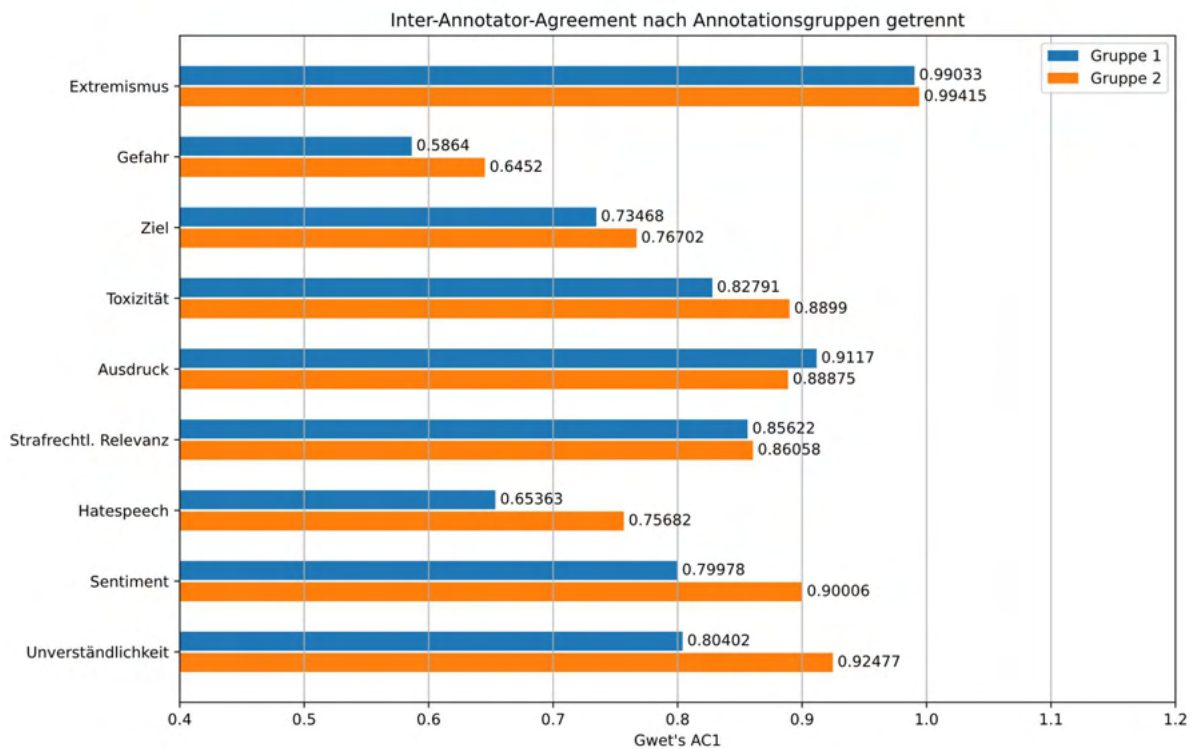


Abbildung 2: Inter-Annotator-Agreement (Güte der Übereinstimmung) der Annotationen. Als Maß wird Gwet's AC1-Koeffizient betrachtet. Interpretation: **0.4 – 0.6: moderate** Übereinstimmung, **0.6 – 0.8: gute** Übereinstimmung, **0.8 – 1.0: sehr gute** Übereinstimmung.

Label	Anzahl	Anteil
Unverständlich	386	3,1 %
Hatespeech	2378	19,11 %
Strafrechtl. Relevanz	800	6,43 %
Extremismus	616	4,95 %
Gefahr	147	1,18 %

Tabelle 1: Häufigkeiten der binären Label im annotierten Datensatz

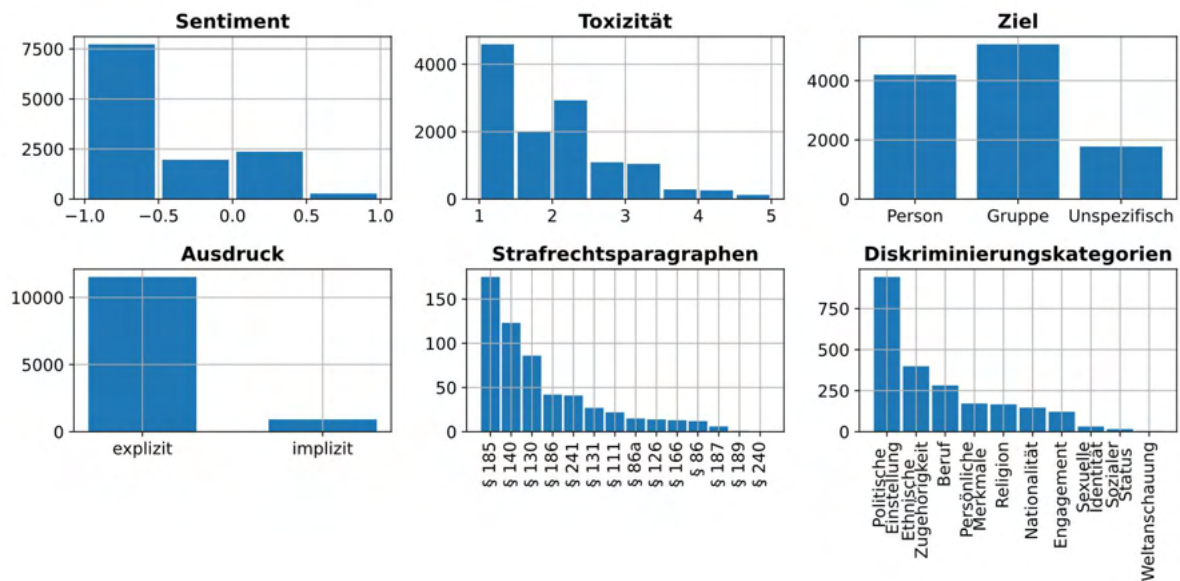


Abbildung 3: Häufigkeiten der nicht-binären Label im annotierten Datensatz



# Deliverable 2.3: Feedback-System

04. Mai 2022

In das Extraktionstool wurde ein Feedback-System eingebaut, welches es ermöglicht, die Ergebnisse der automatischen Klassifikation mit der Einschätzung des Bearbeiters des jeweiligen Kommentars zu vergleichen. Umgesetzt wurde das Feedback-System in einer grafischen Oberfläche im Extraktionstool (Abbildung 1), die die automatischen Klassifikationsergebnisse anzeigt und es dem Bearbeiter ermöglicht, die Einschätzung zu verändern/anzupassen. Sowohl die Ergebnisse der automatischen Klassifikation als auch die Korrektur des Bearbeiters werden im JSON-Format zusammen mit dem Kommentar gespeichert. Das ermöglicht im Nachgang einen umfassenden Vergleich und somit eine Einschätzung der Performance der Klassifikationsmodelle.

The screenshot shows the 'DeTox - Analyse von Hate Speech' application window. It has a menu bar with 'Datei', 'Bearbeiten', and 'Hilfe'. Below the menu is a tab bar with 'Sicherung', 'Kommentar', and 'Klassifikation'. The 'Kommentar' tab is active, showing a text area with 'Das ist ein Testkommentar.' Below this is a table for classification results and manual corrections.

Kategorie	Vorhergesagt	manuelle Korrektur
Hatespeech	0.0	<input type="checkbox"/>
Sentiment	Sentiment-Value	neutral
Toxizität	Toxizität-Value	<input type="range" value="1"/>
Extremismus	0.0	<input type="checkbox"/>
Gefahr	0.0	<input checked="" type="checkbox"/>
Strafrechtl. Relevanz	0.0	<input type="checkbox"/>

Below the table is a 'Speichern unter' button. To the right of the table is a legend titled 'Bedeutung Checkboxen':

- ☐ keine Angabe
- ☒ zutreffend
- ☐ nicht zutreffend

Abbildung 1: Anzeige der Klassifikationsergebnisse eines Testkommentars im Extraktionstool (Für Kategorien, die nicht klassifiziert werden können, erscheinen noch keine Vorhersagewerte). Links wird die Vorhersage der Modelle angezeigt, auf der rechten Seite können Korrekturen der Klassifikation vorgenommen werden. Die Angaben der korrigierten Klassifikation wurden lediglich zufällig gesetzt, damit alle Werte einmal beispielhaft vorhanden sind)

Darüber hinaus können die Daten verwendet werden, um Schwachstellen der Modelle aufzudecken und um Modelle bei Bedarf oder in bestimmten Zeitintervallen neu zu trainieren. Das trägt zur Verbesserung der Klassifikationsergebnisse bei und sorgt für eine stetige Anpassung der Modelle an aktuelle Themen und Sprachgewohnheiten.

# Deliverable 3.1: Ergebnisse der Klassifikationsexperimente

27.06.2022

In diesem Deliverable werde wesentliche Ergebnisse der durchgeführten Klassifikationsexperimente vorgestellt und die Performance der finalen Modelle angegeben.

Die Performance (Klassifikationsgenauigkeit) eines fertigen Modells kann meist nur geschätzt werden. Grund ist der, dass die Modelle auf Trainingsdaten trainiert und auf Testdaten getestet werden. Anhand der Testdaten kann abgeschätzt werden, wie gut ein Modell funktioniert. Um bestmögliche Modelle zu erhalten, wurden für die finalen Modelle jedoch alle vorhandenen Daten zum Training genutzt, d.h. es sind keine Trainingsdaten zur Evaluation übrig. Aus diesem Grund wurden für die Evaluation weitere Modelle trainiert, denen Testdaten beim Training vorenthalten wurden. Das ist ein gängiges Vorgehen.

In der relevanzbasierten Klassifikation werden oft die Performancemaße Precision, Recall und F1-Score verwendet, die hier kurz erklärt werden sollen. Für alle drei Maße gilt, dass sie Werte von 0 bis 1 annehmen können, wobei 0 am schlechtesten und 1 am besten ist.

- **Precision:** Anteil richtig klassifizierter Objekte unter den positiv klassifizierten Objekten (Wie viele positiv klassifizierte Objekte sind tatsächlich positiv?).
- **Recall:** Anteil der positiven Datenpunkte, die das System als solche erkannt hat.
- **F1-Score:** Der F1-Score ist das gewichtete harmonische Mittel aus Precision und Recall. Dieses „belohnt“, wenn sowohl Precision als auch Recall gleichermaßen hoch sind. Sobald einer der Werte deutlich niedriger ist als der andere, sinkt auch der F1-Score beträchtlich.

Precision und Recall beeinflussen sich gegenseitig - wenn ein Wert steigt, verringert sich der andere in der Regel. Bei der Anwendung zur Detektion von Hatespeech ist es insbesondere wichtig, einen hohen Recall zu erreichen. Ein hoher Recall bedeutet hier, dass nur wenige Hasskommentare vom Modell „übersehen“ werden. Dafür müssen Abstriche bei der Precision in Kauf genommen werden, d.h. unter den vom Modell als Hatespeech markierten Kommentaren, können auch einige sein, die keine Hatespeech enthalten.

Klasse	Modell	Prec	Recall	F1-Score	Anmerkungen
<b>Hatespeech</b>	Transformer	0.84	0.78	0.74	Klasse „kein Hate Speech“ wird besser erkannt als „Hate Speech“
<b>Sentiment</b>	Transformer	0.70	0.70	0.69	Klasse „negativ“ wird am besten erkannt, positiv am wenigsten gut.
<b>Toxizität</b>	Multiclass-SVM	0.43	0.55	0.45	Klassen 1-3 werden sicherer erkannt als Klassen 4 und 5
<b>Extremismus</b>	Transformer	0.75	0.75	0.75	Klasse „Extremismus“ wird unter 50% erkannt
<b>Strafrechtliche Relevanz</b>	Transformer	0.73	0.71	0.71	Klasse „strafrechtlich relevant“ wird nur ca. bei 50% erkannt
<b>Gefahr</b>	Kein intelligentes Modell, sondern Pattern Matching, da in der Klasse zu wenige Daten vorhanden sind. Es werden alle im Datensatz vorhandenen Gefahr-Kommentare erkannt, wie gut es generalisiert ist <b>unklar</b> .				

Tabelle 1: Performance (Macro-Average-Werte) der final verwendeten Modelle im Extraktions-/Klassifikationstool.

In den Experimenten hat sich außerdem gezeigt, dass die H3C-Daten, obwohl es mit rund 2000 Kommentaren nur ein relativ kleiner Anteil war, zu einer starken Verbesserung der Modelle geführt haben. Im Vergleich zu Modellen, die ohne die H3C-Daten trainiert wurden, konnte die Performance um 10 – 20 % gesteigert werden. Das unterstreicht die Wichtigkeit „reale“ Daten zur Verfügung zu haben und zeigt, dass diese auch durch eine größere Menge künstlich zusammengestellter Daten (z.B. Sammlung von Kommentaren auf Twitter) nur schwer ersetzt werden können.

Im Folgenden werden die finalen Ergebnisse der Experimente mit den Transformer Modellen dokumentiert. Es wurde ein vortrainiertes Modell herangezogen (gbert-base), dass schon auf generischen Daten trainiert wurde (pre-training). Normalerweise werden dazu z.B. öffentlich zugängliche Daten wie Wikipedia verwendet. Wir haben dieses Modell noch ein weiteres Mal auf generischen Twitter Daten weitertrainiert. Dazu haben wir aus dem nicht-annotierten Tweet-Korpus 1 Millionen Tweets per Zufall gewählt. Das Modell wurde für 5 Epochen, mit einer Batch Size von 32, einer Learning Rate von  $2e-5$  und einer Masked Language Probability

von 15% trainiert. Zum Verfeinern (fine-tuning) können diese Modelle auf den jeweiligen Aufgaben mit annotierten Daten trainiert werden. Dabei haben wir die Modelle jeweils einmal mit H3C Kommentaren und einmal ohne H3C Kommentare trainiert und miteinander die Ergebnisse verglichen. Die Größe der Datensätze unterscheidet sich hierbei leicht, da für jede Klasse die Kommentare gelöscht wurden, bei denen keine finale Klasse im Annotierungsprozess durch z.B. Unverständlichkeit definiert wurde.

### Hate Speech:

- Trainingsset:
  - Ohne H3C: 8.237 Kommentare
  - Mit H3C: 9.992 Kommentare
- Validierungsset:
  - Ohne H3C: 916 Kommentare
  - Mit H3C: 1.111 Kommentare
- Testset:
  - Ohne H3C: 1.018
  - Mit H3C: 1.234

*Ergebnisse auf dem Validierungsset während dem Training für Hate Speech:*

Evaluierungsmetrik	Ohne H3C	Mit H3C
Precision	72,55 %	77,95 %
Recall	70,84 %	74,38 %
F1	70,40 %	74,12 %
Accuracy	90,41 %	87,32 %

*Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – OHNE H3C:*

Klasse	Precision	Recall	F1	Anzahl
<b>Kein Hate Speech (0)</b>	0.93	0.96	0.94	906
<b>Hate Speech (1)</b>	0.52	0.38	0.44	112

*Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – MIT H3C:*

Klasse	Precision	Recall	F1	Anzahl
<b>Kein Hate Speech (0)</b>	0.91	0.95	0.93	996
<b>Hate Speech (1)</b>	0.73	0.60	0.66	238

### Extremismus:

- Trainingsset:
  - Ohne H3C: 8.237 Kommentare
  - Mit H3C: 9.992 Kommentare
- Validierungsset:
  - Ohne H3C: 916 Kommentare
  - Mit H3C: 1.111 Kommentare
- Testset:
  - Ohne H3C: 1.018
  - Mit H3C: 1.234

*Ergebnisse auf dem Validierungsset während dem Training für Extremismus:*

Evaluierungsmetrik	Ohne H3C	Mit H3C
Precision	93.59%	75.44%
Recall	93.97%	75.36%
F1	93.77%	74.81%
Accuracy	99.25%	95.71%

*Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – OHNE H3C:*

Klasse	Precision	Recall	F1	Anzahl
<b>Kein Extremismus (0)</b>	0.99	1.00	1.00	1010
<b>Extremismus (1)</b>	0.00	0.00	0.00	8

*Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – MIT H3C:*

Klasse	Precision	Recall	F1	Anzahl
<b>Kein Extremismus (0)</b>	0.96	0.98	0.97	1172
<b>Extremismus (1)</b>	0.49	0.32	0.39	62

### Sentiment:

- Trainingsset:
  - Ohne H3C: 8.243 Kommentare
  - Mit H3C: 9.998 Kommentare
- Validierungsset:
  - Ohne H3C: 916 Kommentare
  - Mit H3C: 1.111 Kommentare
- Testset:
  - Ohne H3C: 1.018
  - Mit H3C: 1.235

*Ergebnisse auf dem Validierungsset während dem Training für Sentiment:*

Evaluierungsmetrik	Ohne H3C	Mit H3C
Precision	86.73%	69.97%
Recall	86.74%	70.72%
F1	86.50%	68.65%
Accuracy	97.84%	75.96%

Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – OHNE H3C:

Klasse	Precision	Recall	F1	Anzahl
Neutral	0.65	0.67	0.66	413
Positiv	0.73	0.48	0.58	23
Negativ	0.77	0.76	0.77	582

Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – MIT H3C:

Klasse	Precision	Recall	F1	Anzahl
Neutral	0.64	0.70	0.67	439
Positiv	0.50	0.39	0.44	23
Negativ	0.83	0.79	0.81	773

#### Toxizität:

- Trainingsset: 9865 Kommentare (inkl. H3C)
- Testset: 2467 Kommentare (inkl. H3C)

Ergebnisse auf dem Testset nach dem Training für jede Toxizitätsklasse:

Toxizität	Precision	Recall	F1	Anzahl
1 (nicht toxisch)	0.69	0.73	0.71	939
2	0.65	0.49	0.56	1051
3	0.42	0.46	0.44	347
4	0.24	0.51	0.32	108
5 (max. toxisch)	0.14	0.55	0.23	22
Accuracy			0.58	2467
Macro Average	0.43	0.55	0.45	2467
Weighted Average	0.61	0.58	0.59	2467



*Konfusionsmatrix der Klassifikationsergebnisse für die Toxizität (1-5):*

		Vorhergesagte Klasse				
		1	2	3	4	5
Tatsächliche Klasse	1	<b>648</b>	195	35	20	5
	2	277	<b>514</b>	166	72	22
	3	18	74	<b>159</b>	77	19
	4	8	3	16	<b>55</b>	26
	5	0	0	2	8	<b>12</b>

*Interpretation der Konfusionsmatrix:* Anhand der Tabelle kann erkannt werden, welche Klassen (Toxizitätsstufen) besser und welche schlechter klassifiziert werden. Die Spalten stehen für die vom Klassifikator vorhergesagten Klassen, die Zeilen für die tatsächlichen Klassen. Die Werte in der Tabelle geben jeweils eine Anzahl an Kommentaren an. Es kann dann die Verteilung abgelesen werden, mit der Kommentare vorhergesagt wurden. Beispielsweise für die tatsächliche Klasse 1 (Zeile 1) wurden 648 Kommentare korrekt als Klasse 1 klassifiziert, 195 Kommentare wurden in Klasse 2 statt Klasse 1 einsortiert, 35 Kommentare in Klasse 3 statt Klasse 1 usw. Auf der Hauptdiagonalen ist demzufolge die Anzahl korrekt klassifizierter Kommentare zu sehen.

### Strafrechtliche Relevanz:

- Trainingsset:
  - Ohne H3C: 8.237 Kommentare
  - Mit H3C: 9.992 Kommentare
- Validierungsset:
  - Ohne H3C: 916 Kommentare
  - Mit H3C: 1.111 Kommentare
- Testset:
  - Ohne H3C: 1.018
  - Mit H3C: 1.234

*Ergebnisse auf dem Validierungsset während dem Training für Strafrechtliche Relevanz:*

Evaluierungsmetrik	Ohne H3C	Mit H3C
Precision	88.20%	73.45%
Recall	88.20%	71.56%
F1	88.11%	71.22%
Accuracy	98.49%	94.20%

*Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – OHNE H3C:*

Klasse	Precision	Recall	F1	Anzahl
<b>Kein straf. Rel. (0)</b>	0.99	1.00	0.99	1005
<b>Straf. Rel. (1)</b>	0.33	0.08	0.12	13

*Ergebnisse auf dem Testset nach dem Training für die jeweilige Klasse – MIT H3C:*

Klasse	Precision	Recall	F1	Anzahl
<b>Kein straf. Rel. (0)</b>	0.96	0.98	0.97	1154
<b>Straf. Rel. (1)</b>	0.65	0.41	0.50	80

# Deliverable 3.2: Erklärbare Klassifikation

02.07.2022

In diesem Deliverable werde wesentliche Ergebnisse der durchgeführten Experimente zur Erklärbarkeit von Klassifikationsmodellen vorgestellt.

Machine Learning (ML) Modelle werden oft auch mit dem Term „Black-Box“ assoziiert, da diese immer weniger transparent für Menschen sind. Umso besser solche Modelle den Kontext „verstehen“, desto komplizierter wird es die Vorhersagen der Modelle nachzuvollziehen. Im inneren von solchen Modellen werden komplexe mathematische Formeln angewandt, die diese zu einer „Black-Box“ machen. Dementsprechend ist die Forderung in der Wissenschaft – und von Endnutzer\*innen – hoch, einen besseren Einblick in die Entscheidungsstrategien von z.B. Klassifikationsmodellen zu geben. Im Bereich der Verarbeitung von natürlichsprachigen Daten (Natural Language Processing – auch: NLP) gibt es im aktuellen Stand der Forschung verschiedene Methoden, um dieses Problem zu lösen. Die zwei Methoden, die am häufigsten in NLP verwendet werden, haben wir im Folgenden auch auf unsere Modelle angewendet: *SHAP* (Shapley Additive exPlanations) [1] und *LIME* (Local Interpretable Model-Agnostic Explanations) [2]:

- **SHAP:** Bei SHAP wird einem trainierten Modell ein Textinhalt gegeben und bspw. jedem Wort ein anderes Gewicht zugeordnet. Basierend auf den veränderten Vorhersagen wird ermittelt, welche Wörter am meisten Einfluss auf die Entscheidung des Modells haben.
- **LIME:** Bei LIME wird einem trainierten Modell ein Textinhalt gegeben und dabei geschaut auf Basis welcher Eigenschaften („Features“) sich die Vorhersage ändert. Bei Texten werden z.B. Wörter entfernt und dann die Ausgabe des Modells analysiert. Basierend darauf werden dann Visualisierungen erzeugt.

Da ML Modelle, wie z.B. Transformer und neuronale Netze, im Deep Learning Bereich (DL) mittlerweile ganze Sätze verarbeiten, werden natürlich alle Wörter von den beiden Methoden herangezogen. Bei früheren ML Modellen, die oft die Anzahl von wichtigen und seltenen Wörtern benutzt haben, sind die Ergebnisse oft nachvollziehbar, dafür ist aber die Performance (Klassifikationsgenauigkeit) meist niedriger. Dementsprechend leidet durch die Komplexität

von DL Modellen auch die Transparenz. Werden Methoden zur Erklärbarkeit von Klassifikationen verwendet, so wird dies auf ein fertig trainiertes Modell angewandt. So können, neben den Standard-Evaluierungsmetriken (siehe D3.1), die Modelle auch weitergehend untersucht werden. Beide Methoden geben die Wahrscheinlichkeit für die vorhergesagten Klassen aus. So kann bspw. bei Hate Speech, das Modell den Wert 0.75 (75%) für die Klasse „Hate Speech“ und 0.25 (25%) für die Klasse „Kein Hate Speech“ ausgeben. Dies heißt, dass das Modell glaubt es könnte sich zu 75% Wahrscheinlichkeit um Hate Speech handeln. Der Standard-Wert zur Ermittlung zu welcher Klasse ein Text zugewiesen wird liegt normalerweise bei der Grenze von 0,5. Wenn es über 0,5 ist, ist es in dem Beispiel automatisch die Klasse Hate Speech. Bei finalen Klassifikationsausgaben geht diese Information oft verloren: denn, wenn das Modell nur zu 0.53 sagt, dass es zur Klasse Hate Speech gehört, ist die Vorhersage sehr ungenau und in diesem Fall sogar fast „geraten“. Des Weiteren geben die Methoden zur Erklärbarkeit Wörter an, die dem Modell geholfen haben, sich entweder für die eine oder andere Klasse zu entscheiden. Letztlich, können auch die Texte visualisiert werden, mit farblichen Highlights, um zu zeigen an welcher Stelle diese Wörter im Text vorkamen.

Hier ist es nur problematisch, dass eben Transformer Modelle, wie sie bei uns im Einsatz sind, den ganzen Satz als Input zum Trainieren nehmen – das heißt, dass auch Wörter wie „der“, „die“, „das“, „und“, „oder“ etc. Einfluss auf das finale Ergebnis haben. Solche Wörter werden in NLP auch Stoppwörter genannt und geben wenig Aufschluss darüber, wieso ein Text der einen oder anderen Klasse zugehörig ist. Es gibt auch Studien, die besagen, dass z.B. die Anzahl von Personalpronomen oder andere Arten von Wörtern, wie bestimmte Verben, Einfluss auf die Entscheidung haben können und sich Texte so unterscheiden. Allerdings müsste dies dann in Visualisierungen unterstützen zu den Ausgaben der Methoden zur Erklärbarkeit angezeigt werden. Dies herauszufinden ist allerdings sehr ressourcen-aufwendig und oft nicht in kurzer Zeit machbar.

## Literatur

- [1] Lundberg, Scott M.; Lee, Su-In. (2017). A unified approach to interpreting model predictions. <http://arxiv.org/abs/1705.07874>.
- [2] Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos (2016). „Why Should I Trust You?“. Explaining the Predictions of Any Classifier.“ In: Conference on Knowledge Discovery and Data Mining – KDD2016, ACM. <https://doi.org/10.48550/arxiv.1602.04938>.



# Deliverable 4.1.: Klassen toxischer Inhalte

16.06.2021

Für die Klassifikation der toxischen Inhalte werden in DeTox Annotationsrichtlinien erstellt, in denen die Klassen toxischer Inhalte beschrieben sind. In Klammern sind Beispiele aufgeführt.

## Annotationsschema

Im Folgenden heißt „□“, dass eine Mehrfachauswahl möglich ist, „“ heißt, nur eine Auswahl ist möglich.

### Sentiment (Empfindung, Gefühl)

- Positiv  
("Ich finde Gärtner toll!"  
"Gärtner sind toll!"  
"Ich mag es, dass Gärtner viel arbeiten müssen!")
- Neutral  
("Wir haben einen Gärtner"  
"Der Gärtner ist groß")
- Negativ  
("Gärtner sind die schlimmsten Menschen, die ich kenne!!!"  
"Wieso ist unser Gärtner nie da, wenn man ihn braucht?")

### Hate Speech

- Ja  
("Gärtner sind doch alle gleich: Dumm wie ein Baum und menschenfeindlich!")
- Nein  
("Er sagte: 'Gärtner sind doch alle gleich: Dumm wie ein Baum und menschenfeindlich!'"  
"Der Mörder ist immer der Gärtner!" (Hierbei soll deutlich gemacht werden, dass auch ironische Phrasen erlaubt sind. Die Überlegung dabei war, ob man da vielleicht einen Smiley mit hinzufügt, um deutlich zu machen, dass es ironisch gemeint ist. Falls ja, allerdings nur Smileys (z.B. ";)") und keine Emojis, da diese Word überfordern))

### Strafrechtliche Relevanz

- Ja
- Nein

**Strafrechtliche Einschätzung** (wenn ‚ja‘ bei strafrechtl. Relevanz)



Detektion von Toxizität und

Aggressionen in Postings und

Kommentaren im Netz (**Hass ist Gift**)

---

- § 86 StGB Verbreiten von Propagandamitteln verfassungswidriger Organisationen
- § 86a StGB Verwenden von Kennzeichen verfassungswidriger Organisationen
- § 111 StGB Öffentliche Aufforderung zu Straftaten
- § 126 StGB Störung des öffentlichen Friedens durch Androhung von Straftaten
- § 130 StGB Volksverhetzung
- § 131 StGB Gewaltdarstellung
- § 140 StGB Belohnung und Billigung von Straftaten
- § 166 StGB Beschimpfung von Bekenntnissen, Religionsgesellschaften und Weltanschauungsvereinigungen
- § 185 StGB Beleidigung
- § 186 StGB Üble Nachrede
- § 187 StGB Verleumdung
- § 189 StGB Verunglimpfung des Andenkens Verstorbener
- § 240 StGB Nötigung
- § 241 StGB Bedrohung

#### **Ausdruck**

- Explizit (direkt)  
("Der Gärtner ist groß")
- Implizit (indirekt)  
("Der Gärtner ist ein richtiges Hochhaus. Neulich ist er oben gegen den Türrahmen gelaufen")

#### **Toxizität/Aggressivität**

Skala 1-5

#### **Extremistisch**

- Ja
- Nein

#### **Ziel (vordergründiges Ziel, welches Ziel wird am ehesten angesprochen?)**

- eine konkrete Person/User (interpersonal)  
("Der Gärtner ist groß"  
"Du Gärtner bist groß")
- eine Personengruppe  
("Ihr Gärtner seid groß"  
"Gärtner sind groß"  
"@UserXY Gärtner sind groß" (um zu zeigen, dass es nicht unbedingt auf die User-Mentions ankommt, wenn es um das Ziel geht. Hier wurde auf einen User geantwortet, aber die Aussage an sich richtet sich gegen eine Gruppe))
- Kein spezifisches Ziel (public)  
("Ich habe gelesen, Gärtner seien groß")

#### **Thematischer Kontext** (inkl. Markierung von Stichwörtern bzw. Phrasen)



Detektion von Toxizität und

Aggressionen in Postings und

Kommentaren im Netz (**Hass ist Gift**)

---

- Beruf und/oder Ehrenamt
- politische Einstellung
- Persönliches Engagement und Interesse (z.B. Aktivisten)
- Sexuelle Identität
- Physische, psychische oder mentale Merkmale (z.B. Hautfarbe, Jugendliche, Blinde)
- Nationalität
- Religionszugehörigkeit
- Sozialer Status
- Weltanschauung (z.B. Vegetarismus, Erziehungsstile, Familienform)
- ethnische Zugehörigkeit (z.B. Asylbewerber, Migrationshintergrund)

### **Gefahrenlage**

Skala 1-5

Falls es die Kommentare bisher nirgends in Textform gibt, müssen wir die Texte mit OCR auslesen und können bei der Annotation abfragen, ob der Text korrekt extrahiert wurde.

# Deliverable 4.2: Statistische Daten zur Verteilung der toxischen Inhalte auf die Klassen

25.05.2020

Das Deliverable soll Aufschluss über die Zusammenhänge der annotierten Kategorien und der Toxizität aufzeigen und untersuchen, welche Merkmale in bzw. nicht in toxischen Kommentaren auftreten. Die annotierten und betrachteten Kategorien sind die folgenden:

- Hatespeech
- Sentiment
- Strafrechtliche Relevanz (binär)
- Relevante Strafrechtsparagrafen (14 StGB-Paragrafen)
- Gefahr
- Extremismus
- Ausdruck
- Ziel
- Diskriminierungskategorien

In Abbildung 1 sind in den Diagrammen die Zusammenhänge zwischen jeweils zwischen Toxizität und Hatespeech, Sentiment (-1 negativ, +1 positiv), Strafrechtlicher Relevanz, Gefahr und Extremismus dargestellt. Dabei ist als Linie jeweils der Mittelwert der betrachteten Kategorie für einen Toxizitätswert dargestellt. Im Hatespeech-Diagramm ist beispielsweise ablesbar, dass Kommentare mit einer Toxizität von 1 im Mittel einen Hatespeech-Score von nahezu 0 haben. Kommentare mit einer Toxizität von 3 haben einen Hatespeech-Score von ca. 0.65. Der hellblaue Bereich kennzeichnet die Standardabweichung, d.h. in welchem Bereich sich die Werte im Mittel bewegen, Ausreißer können aber trotzdem vorhanden sein.

Wie zu erwarten, steigt mit zunehmender Toxizität der Hatespeech-Score, d.h. toxische Kommentare beinhalten auch fast immer Hatespeech. Beim Sentiment ist feststellbar, dass der Wert zwar höher liegt, je niedriger die Toxizität ist. Selbst bei einer Toxizität von 1 (nicht toxisch) ist das Sentiment jedoch im Mittel mit ca. -0.3 noch leicht negativ. Das liegt in erster Linie an der Beschaffenheit des Datensatzes (da es um Offensive Sprache geht). Ab einer Toxizität von 3 hat das Sentiment sein Minimum fast erreicht und ändert sich bei Toxizität 4



und 5 nicht mehr. Die strafrechtliche Relevanz steigt relativ kontinuierlich mit der Toxizität. Bei Toxizität 5 sind schließlich nahezu alle Kommentare auch strafrechtlich relevant. Anders sieht es bei der Kategorie Gefahr aus, die bis Toxizität 3 bei 0 liegt, danach steigt der Wert für Gefahr bis auf 0.2 an und hat dabei eine sehr breite Standardabweichung. Das bedeutet, Toxizität und Gefahr haben nur einen sehr schwachen Zusammenhang. Das Ergebnis kann aber auch durch die geringe Anzahl von gefährlichen Kommentaren verfälscht/weniger repräsentativ sein. Auch die Annotation Extremismus steht in dem klaren Zusammenhang mit der Toxizität: Je höher die Toxizität, desto höher ist auch die Extremismuskwahrscheinlichkeit.

Für die Darstellung des Zusammenhangs von Toxizität und dem Ausdruck (explizit/implizit) bzw. dem Ziel (Person/Gruppe/Public) wurden jeweils die Klassenanteile für die Toxizitätswerte dargestellt (Abbildung 2), d.h. die Werte der Graphen für einen X-Wert ergeben in Summe 1.

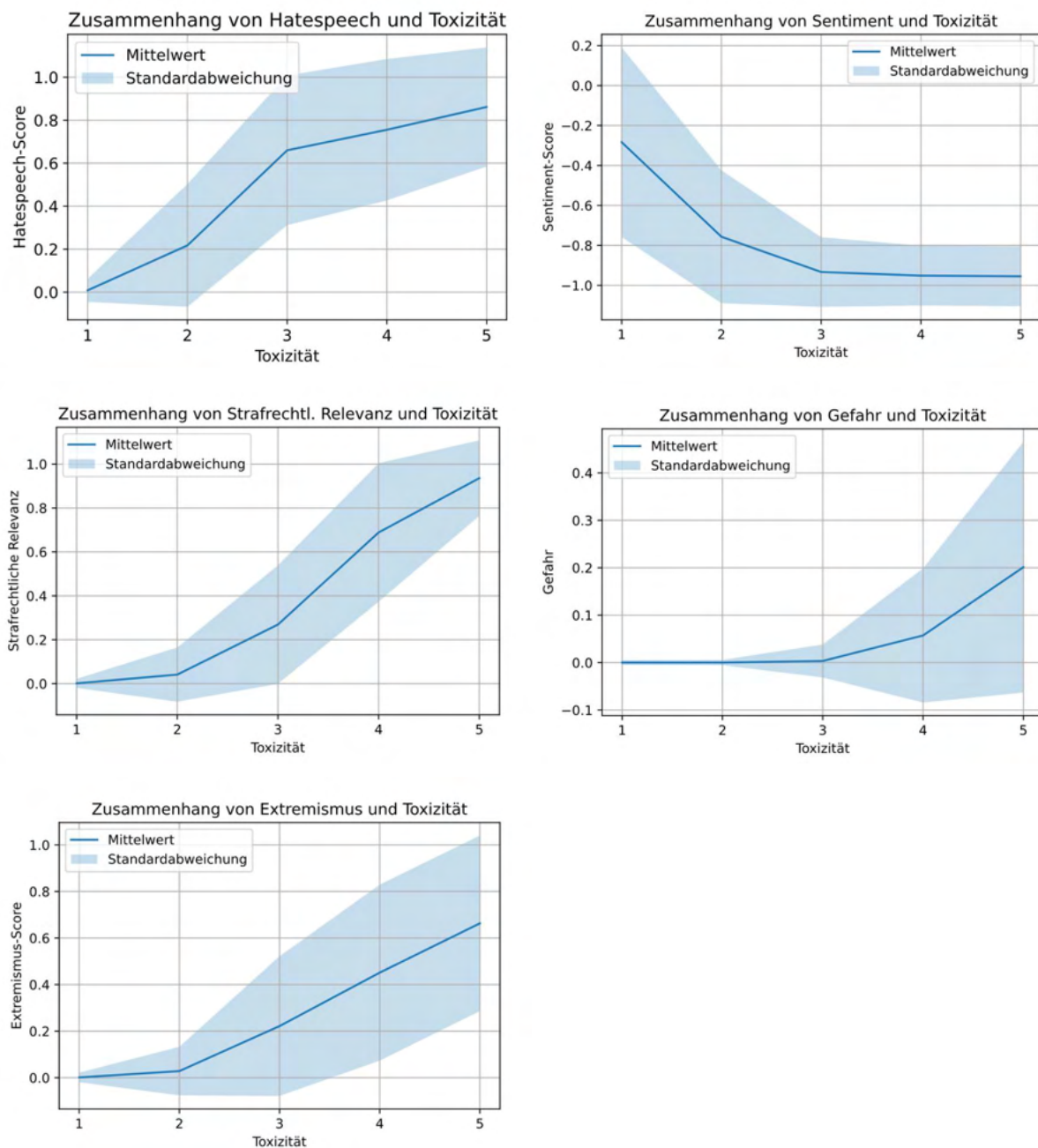


Abbildung 1: Darstellung der Zusammenhänge von Hatespeech, Sentiment, Strafrechtlicher Relevanz, Gefahr und Extremismus mit der Toxizität. Angegeben ist jeweils der Mittelwert und der Bereich der Standardabweichung. Bei Sentiment ist -1 negativ, 0 neutral und +1 positiv. In allen anderen Kategorien bedeutet 0 jeweils „nicht zutreffend“ und 1 bedeutet „voll zutreffend“.

Für den Ausdruck ist erkennbar, dass implizite Ausdrücke deutlich seltener sind (<11 %). Trotzdem ist ein Trend erkennbar: mit steigender Toxizität sinkt der Anteil impliziter Kommentare bis auf fast null bei einer Toxizität von 5. Das bedeutet, Kommentare die z.B.

Ironie oder Sarkasmus enthalten haben tendenziell eine niedrigere Toxizität und haben damit eher humorvollen statt toxischen Inhalt.

Beim Ziel ist die Tendenz zu erkennen, dass mit steigender Toxizität gruppenbezogene Kommentare (Hass gegen Gruppen) zunehmen und persönliche Beleidigungen leicht abnehmen; Kommentare ohne Ziel (Public) mit einer Toxizität von  $> 3$  gibt es nur sehr selten. Das zeigt, dass Kommentare unspezifischen Ziels entweder weniger Hass enthalten oder aber weniger toxisch wahrgenommen werden, weil keine spezifische Person oder Gruppe angegriffen wird.

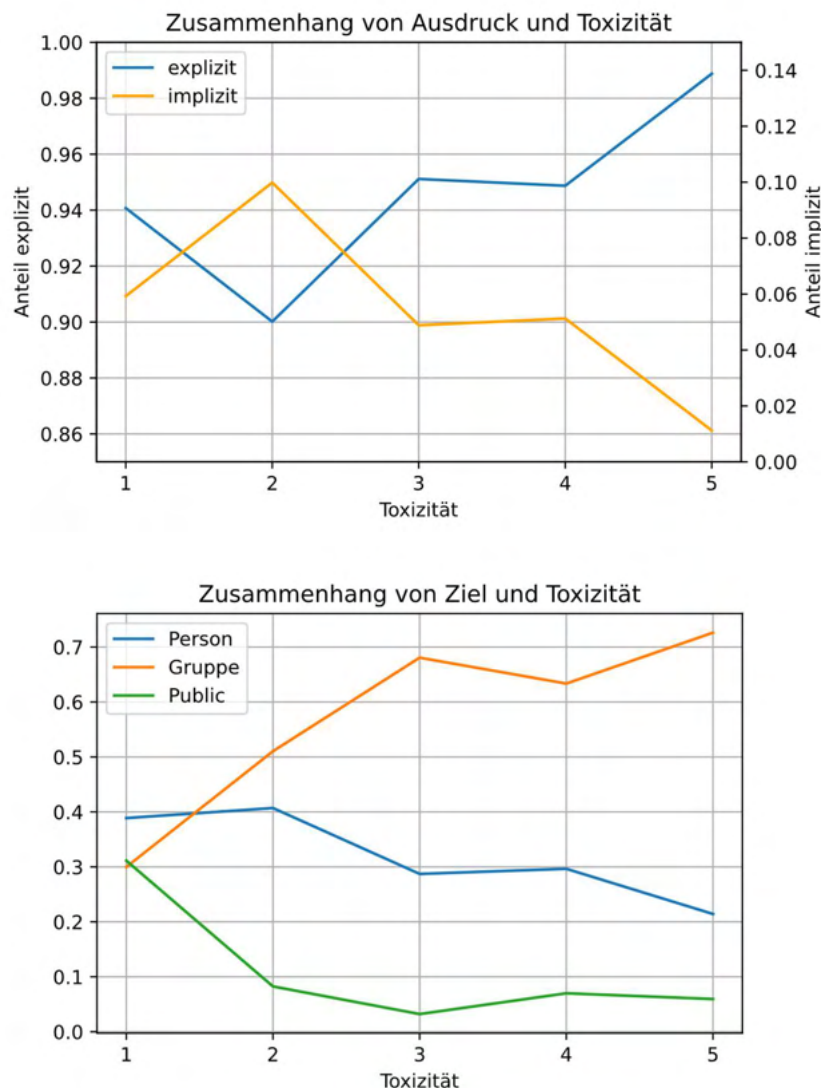


Abbildung 2: Darstellung des Zusammenhangs der Ausdrucksform bzw. des Kommentarziels und der Toxizität.

Weiterhin wurde die Verteilung der Toxizität auf die zehn Diskriminierungskategorien (Abbildung 3) und die 14 Strafrechtsparagrafen (Abbildung 4) untersucht.

Hierbei wurde für jede Klasse (Diskriminierungskategorie bzw. Paragraph) die Häufigkeitsverteilung der Toxizität geplottet. Der Farbton der Balken gibt an, wie oft die jeweilige Klasse im Datensatz vertreten war.

Bei den Diskriminierungskategorien ist erkennbar, dass in allen zehn Kategorien die Toxizität von 3 am häufigsten vorkommt. Das lässt sich damit erklären, dass diese Kategorie nur annotiert wurde, wenn ein Kommentar als Hatespeech eingestuft wurde. Demzufolge ist in den meisten Fällen eine gewisse Toxizität vorhanden. Trotzdem gibt es Unterschiede zwischen den einzelnen Klassen: Kommentare die als diskriminierend hinsichtlich *Religion* oder *Nationalität* gekennzeichnet wurden, sind tendenziell toxischer als beispielsweise Kommentare der Klassen *Politische Einstellung* oder *Persönliches Engagement*.

Die Analyse des Diagramms der Paragrafen zeigt, dass die meisten Paragrafen einen Großteil der Kommentare in Toxizitätsklasse 4 haben. Ausnahmen sind §185 (Beleidigung) und §186 (üble Nachrede), sowie §187 (Verleumdung). Diese Paragrafen können (zumindest für den Laien, wie es unsere Autoren hinsichtlich des juristischen Wissens waren) sehr weit ausgelegt werden und demzufolge auch auf weniger toxische Kommentare zutreffen. Die Paragrafen mit der höchsten Toxizität sind §126, §111 und §241. Diese Paragrafen stehen alle im Zusammenhang mit Bedrohung oder Androhung von Straftaten. Die hohe Toxizität ist demzufolge damit erklärbar.

### Zusammenfassung der Haupteckdaten:

- Mit steigender Toxizität steigt der Anteil von Hatespeech, Strafrechtlicher Relevanz, Extremismus und Gefahr in Kommentaren, das Sentiment sinkt und hat bei Toxizität 3 sein Minimum erreicht (Abbildung 1).
- Auch wenn der Anteil impliziter Kommentare allgemein gering ist (< 11%), ist deutlich erkennbar, dass bei hoher Toxizität der Anteil von 5 - 10% auf fast 0% absinkt (Abbildung 2).
- Nicht toxische Kommentare haben ein ausgeglichenes Verhältnis der betrachteten Zielgruppen. Stark toxische Kommentare sind aber am häufigsten gruppenbezogen (> 70%, Abbildung 2).
- Die Diskriminierungskategorien mit den toxischsten Kommentaren sind *Religion* und *Nationalität*. Am wenigsten toxisch sind die Kategorien *Politische Einstellung* und *Persönliches Engagement* (Abbildung 3).
- Die Paragraphen mit den toxischsten Inhalten (§126, §111 und §241) stehen alle im Zusammenhang mit Bedrohung oder Androhung von Straftaten. §185 (Beleidigung), §186 (üble Nachrede) §187 (Verleumdung) enthalten eher weniger toxische Kommentare (Abbildung 4).

Verteilung der Toxizität von Kommentaren abhängig von der Diskriminierungskategorie

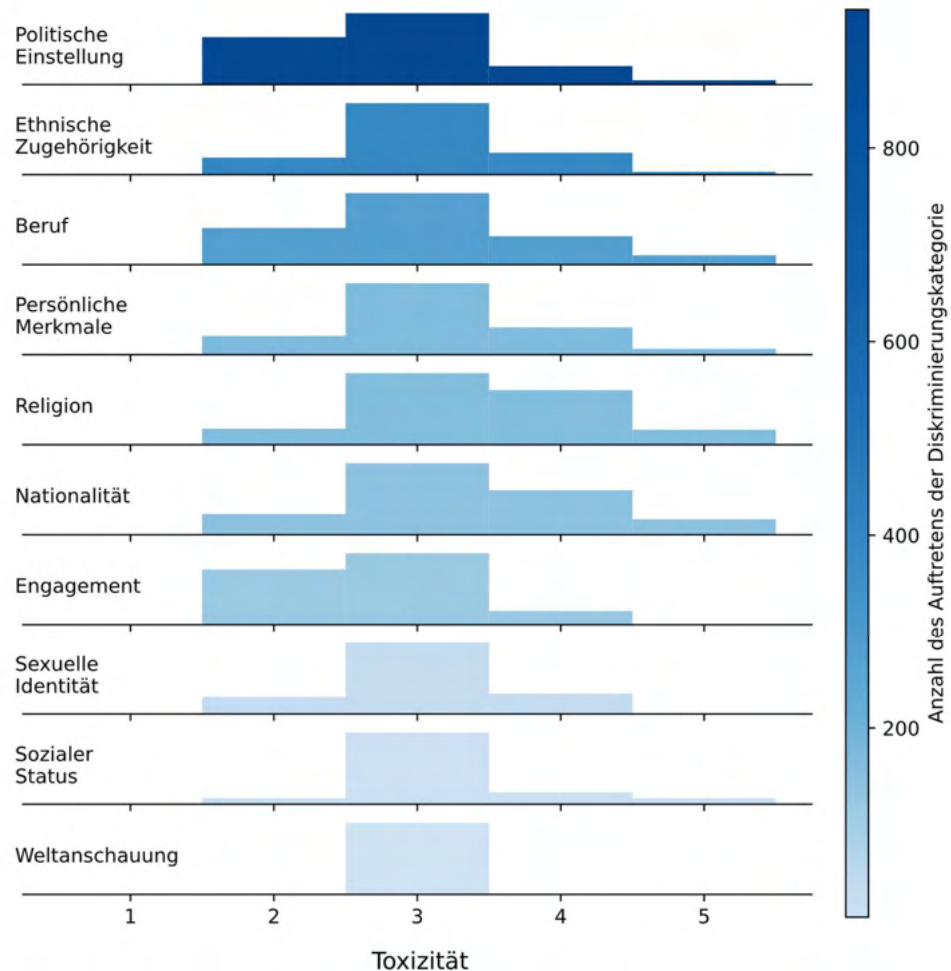


Abbildung 3: Die Abbildung stellt für jede Diskriminierungskategorie die Verteilung der Toxizität dar. Es ist erkennbar, wie toxisch Kommentare sind, für die die jeweilige Diskriminierungskategorie zutrifft. Die absoluten Häufigkeiten, wie oft jede Diskriminierungskategorie zutrifft, ist anhand der Farbskala ablesbar.

Toxizität strafrechtl. relevanter Kommentare unterteilt nach Paragraphen

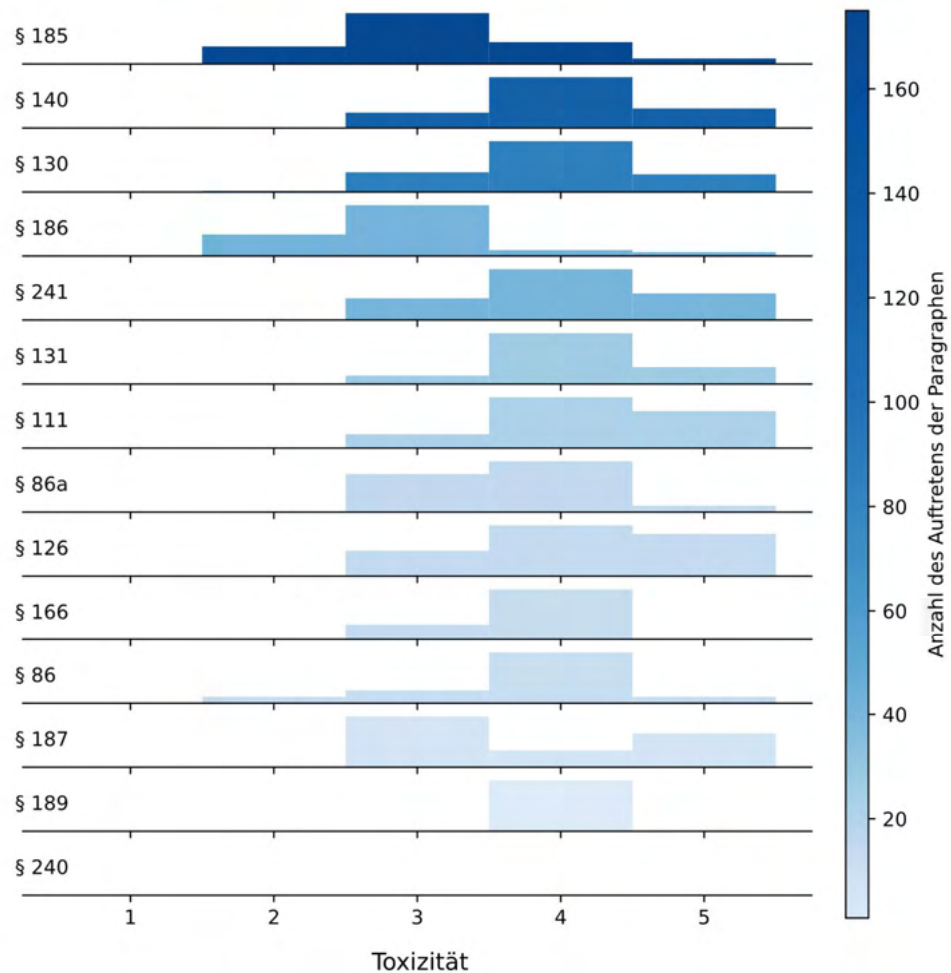


Abbildung 4: Die Abbildung stellt für jeden Strafrechtsparagraphen (StGB) die Verteilung der Toxizität dar. Es ist erkennbar, wie toxisch Kommentare sind, für die der jeweilige Paragraf zutrifft. Die absoluten Häufigkeiten, wie oft jede Diskriminierungskategorie zutrifft, ist anhand der Farbskala ablesbar.

# Deliverable 5.1: Netzwerkanalyse

*5. Januar 2022*

Die Ergebnisse der Netzwerkanalyse sind im Beitrag zum „Handbuch Cyberkriminologie“ der Gruppe beschrieben.



## **Hass im Netz – Aggressivität und Toxizität von Hasskommentaren und Postings, Detektion und Analyse**

Christoph Demus, Mina Schütz, Nadine Probol, Jonas Pitz, Melanie Siegel, Dirk Labudde

Keywords: Hatespeech, Opinion Leader, Netzwerkanalyse, Aggressivität, Toxizität

### **Zusammenfassung:**

Hass und aggressives Verhalten im Netz werden immer größere Probleme. Der bisher etablierte Versuch zur Lösung des Problems ist das Löschen von Kommentaren, doch um dem grundlegenden Problem entgegenzuwirken, müssen Ursachen für die Entstehung von Hass im Netz bekämpft werden. In diesem Kapitel wird daher neben Grundlagen der Hatespeechanalyse insbesondere auf Gruppierungen, Informationsfluss und die Ausbreitung von Hass in sozialen Netzwerken eingegangen. Daraus werden dann Maßnahmen zur Bekämpfung der Ursachen von Hass abgeleitet und diskutiert.

## 1. Einleitung und Motivation

Hass im Netz ist seit der Entstehung der großen Social-Media-Plattformen ein immer größer werdendes Problem. Neben allen Vorteilen dieser Plattformen kommt es immer häufiger zu respektlosem Verhalten, Beleidigungen, Anfeindungen bis hin zu Drohungen gegen Personen. Kommentarspalten auf Nachrichtenseiten müssen häufig gesperrt werden, weil die Moderation aufgrund der Menge der Hasskommentare oft schlicht nicht möglich ist. Ein noch größeres Problem haben die sozialen Netzwerke. Trotz der Verwendung neuester KI-Methoden bleibt die zuverlässige Detektion von Hatespeech ein Problem, denn in vielen Fällen ist die Entscheidung, ob es sich um Hatespeech bzw. strafrechtlich relevanten Inhalt handelt, sehr schwer. Es stellt sich daher auch die Frage, ob diese wichtige Entscheidung einer Maschine überlassen werden sollte, denn die Meinungsfreiheit ist ein sehr hohes Gut. Allgemein stellt sich die Frage, ob die bisher anerkannte Methode des Löschens ausgewählter Kommentare zielführend in Hinsicht auf die Bekämpfung von Hass im Netz ist oder ob es Möglichkeiten gibt, die Ursachen des auftretenden Hasses zu erkennen und diesen entgegenzuwirken. Die Hasskommentare selbst sind letzten Endes nur das Resultat des Hasses und der Aggression, möglicherweise nur die Spitze des Eisberges<sup>1</sup>.

Einen Einblick in dieses Problemfeld soll mit diesem Kapitel gegeben werden. Zu Beginn werden einige relevante Begrifflichkeiten aus dem Bereich der Hatespeech-Analyse definiert und erläutert. Anschließend wird die Herangehensweise an Detektion und Meldung von Hasskommentaren beschrieben und es wird auf allgemeine und textuelle Merkmale betroffener Kommentare eingegangen. Aufgrund der Vielzahl an Publikationen in diesem Bereich wird hier nur ein Überblick gegeben.

Im zweiten Teil dieses Kapitels steht die Netzwerkanalyse und damit einhergehend die Meinungsbildung, Ausbreitung und Veränderung von Hass in sozialen Medien im Vordergrund. Ausgegangen wird von einzelnen Kommentaren. Darauf aufbauend wird das Zusammenspiel der Kommentare in ganzen Konversationen betrachtet und es wird das Verhalten von Nutzer\*innen auf Konversationsebene beschrieben. Eng in Zusammenhang steht damit die Übertragung von Emotionen und Nutzerbeziehungen. Beziehungen zwischen Nutzer\*innen sind ein essenzieller Bestandteil sozialer Netzwerke und wirken sich auf die Gruppenbildung, den Informationsfluss und die Meinungsbildung aus. Darauf aufbauend wird konkret auf die Entstehung und Ausbreitung von Hass im Netz eingegangen. Zum Abschluss werden auf

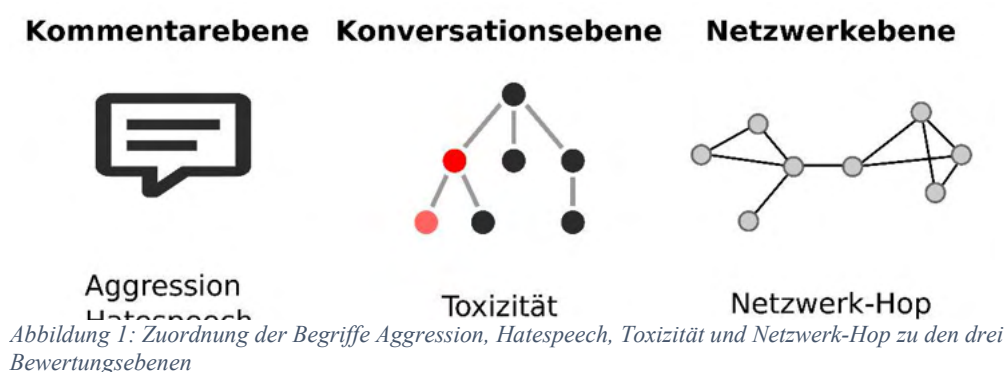
---

<sup>1</sup> <https://www.das-nettz.de/zeig-mir-den-schlüssel-zur-bewältigung-von-hass-eine-psychologische-perspektive>  
(Zugegriffen am 28.12.2021)

Basis der Netzwerkstruktur, Verbreitung von Informationen und der Ursachen von Hatespeech Ideen und Vorschläge für Maßnahmen entwickelt, um Hass in sozialen Netzwerken möglicherweise effizienter als bisher bekämpfen zu können.

## 2. Begrifflichkeiten und Bewertungsebenen von Kommentaren

Bevor Details der Hatespeech-Detektion, -Bewertung und -Ausbreitung im Netz erläutert werden können, müssen zuerst grundlegende Begriffe eingeführt werden. Gerade im Bereich der Hasskommentar-Analyse werden die Begriffe Hatespeech, Aggression und Toxizität häufig unterschiedlich definiert und teilweise als Synonyme verwendet. An dieser Stelle soll jedoch zwischen den drei Begriffen differenziert werden, um verschiedene Analyseebenen trennen zu können (Abbildung 1): Die Kommentarebene, die Konversationsebene und die Netzwerkebene. Auf Kommentarebene sind die Begriffe Aggression und Hatespeech einzuordnen. Dort werden einzelne Kommentare analysiert. Auf Konversationsebene werden ganze Konversationen berücksichtigt. Dadurch ist es möglich, Rückschlüsse auf die Beeinflussung der Kommentare (bzw. Nutzer, die die Kommentare geschrieben haben) untereinander zu ziehen. Noch über diesen beiden Ebenen steht die Netzwerkebene, welche sich damit beschäftigt, wie sich Informationen, Meinungen und auch Hass im Netzwerk ausbreitet. Verantwortlich für die Verbreitung sind die Personen, die Kommentare schreiben, teilen, retweeten etc. Die folgenden Teilabschnitte gehen nun noch einmal detailliert auf die genannten Begriffe ein.



### 2.1. Aggression

Die Aggression, auch Online-Aggression genannt, bezieht sich auf einen einzelnen, konkreten Post oder Kommentar. Der Duden definiert *aggressives* Verhalten als „angriffslustig, streitsüchtig“, „in schädigender Weise auf etwas einwirkend; zerstörend“ und „gezielt-kräftig auf etwas, jemanden richtend“ (Dudenredaktion (o. J.)). Konkret bezieht sich das aggressive

Verhalten in Kommentaren darauf, wie angriffslustig und streitsüchtig ein Kommentar ist und inwieweit er schädigend und beleidigend auf andere Kommunikationsteilnehmer wirkt. Dabei wird unterschieden, ob mit dem Kommentar eine oder mehrere Einzelpersonen, Personengruppen (z.B. religiöse Gruppen) oder aber die Allgemeinheit angegriffen werden. Da aggressives Verhalten als „gezielt-kräftig“ (Dudenredaktion (o. J.)) beschrieben wird, impliziert das die Absicht der Kommentare Schreibenden zur Demütigung, Beleidigung oder sonstigem aggressiven Verhalten im Netz.

Kurzgefasst gibt die Aggression die Stärke des aggressiven Verhaltens an und kann als Grad der Diskreditierung einer Person oder Sache durch den betrachteten Kommentar verstanden werden.

Die Aggression kann anhand objektiver Kriterien gemessen und klassifiziert werden. Für die Strafverfolgung wird eine Bewertung durch Festlegung von Straftatbeständen vorgenommen. Relevant sind dabei meist durch den Kommentar diskriminierte Aspekte wie beispielsweise die Religion, Nationalität, Hautfarbe, Geschlecht, sexuelle Orientierung und sozialer Status.

## 2.2. Hatespeech

In engem Zusammenhang zur Aggression steht der Begriff Hatespeech (Hassrede). Hatespeech kann als Unterklasse der Aggression angesehen werden<sup>2</sup>. Die Bewertung hinsichtlich Hatespeech erfolgt im Gegensatz zur Aggression nur binär, d.h. es wird festgestellt, ob ein Kommentar Hatespeech enthält, jedoch gibt es keine Einschätzung der „Stärke“ von Hatespeech.

Der Begriff Hatespeech ist weder in der Wissenschaft einheitlich definiert noch gibt es eine legale Definition im deutschen Recht. Verschiedene Institutionen (z.B. die Vereinten Nationen<sup>3</sup>, die Amadeu-Antonio-Stiftung<sup>4</sup> oder die zentrale Meldestelle „Hasskommentare“ des hessischen CyberCompetenceCenters (Hessen3C)<sup>2</sup>) definieren den Begriff im Detail unterschiedlich, der Grundgedanke ist aber immer ähnlich. An dieser Stelle wird der Begriff im Detail folgendermaßen definiert: Als Hatespeech gelten Ausdrucksformen, die Personen oder Personengruppen aufgrund von den ihnen zugeschriebenen Gruppenmerkmalen angreifen

<sup>2</sup> <https://hessengegenhetze.de/fragen-antworten/hate-speech> (Zugegriffen am 10.12.201)

<sup>3</sup> <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf> (Zugegriffen am 10.12.201)

<sup>4</sup> <https://www.amadeu-antonio-stiftung.de/digitale-zivilgesellschaft/was-ist-hate-speech/> (Zugegriffen am 10.12.201)

oder herabwürdigen. Aggressive und diskriminierende Äußerungen sind dabei z.B. auf die politische Einstellung, die Religionszugehörigkeit oder die sexuelle Identität bezogen. Auch wenn die Opfer von Hatespeech häufiger Minderheiten angehören, kann Hatespeech ebenso gegen Mehrheiten erfolgen.

Hervorzuheben sind in der Definition die zugeschriebenen Gruppenmerkmale, aufgrund derer eine Diskriminierung stattfindet. Findet keine Diskriminierung eines solchen Gruppenmerkmals statt, handelt es sich nicht um Hatespeech. Nichtsdestotrotz kann der Kommentar aggressiv oder toxisch sein. Zur Veranschaulichung sei der Kommentar „Ich töte Dich!“ genannt. Da in diesem niemand aufgrund von zugeschriebenen Gruppenmerkmalen diskriminiert wird, fällt er nicht unter Hatespeech nach der genannten Definition. Aggressiv, toxisch (siehe nächster Abschnitt) und eine Bedrohung ist der Kommentar aber sehr wohl. Anders sähe es aus, wenn es heißen würde „Ich töte Dich, du N\*!“<sup>5</sup>, weil das Wort „N\*“<sup>5</sup> abfällig eine Gruppe von Menschen beschreibt und somit die Diskriminierung aufgrund von Gruppenmerkmalen gegeben ist.

### 2.3. Toxizität

Der Begriff Toxizität kommt aus dem Gebiet der Medizin und Pharmakologie und bedeutet so viel wie die Giftigkeit eines Stoffes bzw. den „Ausdruck der schädigenden Wirkung eines Stoffes auf den menschlichen, tierischen oder pflanzlichen Organismus“ (Lohs et al. 2008). Die Toxizität eines Stoffes ist von vielen Faktoren abhängig u.a. von der Art der Verabreichung, der Menge des verabreichten Stoffes und der Dauer der Einwirkung. Die Wirkung tritt meist nicht sofort ein, sondern erst nach einer gewissen Latenzzeit (Freissmuth 2016).

In Anlehnung an diese Definition wird der Begriff in den Bereich der Hasskommentaranalyse übertragen: Dort beschreibt die Toxizität die Schädlichkeit/Giftigkeit eines Kommentars für die Konversation und damit auch für andere Konversationsteilnehmer. Es wird betrachtet, ob der Kommentar beispielsweise angreifend und beleidigend ist, aufgrund dessen Teilnehmer vielleicht sogar die Konversation verlassen oder aber ob ein Kommentar harmlos oder freundlich ist. Eine hohe Toxizität geht häufig mit einem negativen Sentiment (Stimmung, Gefühlslage) des Kommentarautors einher. In Abbildung 1 (Mitte) stellt der obere rot markierte Kommentar einen toxischen Kommentar dar, dessen toxische Wirkung sich in der darauffolgenden Antwort widerspiegelt.

Analog zur Toxizität im pharmakologischen Sinne ist auch in der Sprachanalyse die Toxizität eines Kommentars von verschiedenen Faktoren abhängig. Dazu zählen insbesondere der

---

<sup>5</sup> Im Artikel vermeiden wir, diskriminierende Wörter zu verwenden und ersetzen diese durch “\*”

Kontext und das Umfeld (z.B. in einer Gruppe mit bestimmten Interessen und Ansichten), in dem der Kommentar veröffentlicht wird, aber auch der Zeitpunkt kann von Relevanz sein. Aufgrund dieser Einflussfaktoren ist die Toxizität eine sich dynamisch veränderbare Eigenschaft eines Kommentars. In der Hatespeech-Analyse wird die Toxizität deshalb insbesondere in Bezug zur Verbreitung eines Kommentars in einem abgegrenzten Netzwerk beobachtet (Almerekhi et al. 2019; Almerekhi et al. 2020). So kann festgestellt werden, wie sich die Toxizität eines Kommentars während seiner Verbreitung (z.B. durch Retweeten, Teilen, Liken) verändert. Die Veränderung steht wiederum mit der Stimmung im betrachteten Netzwerk in Verbindung.

#### 2.4. Netzwerk Hop

Im Netz gibt zahlreiche Plattformen, wo Nutzer kommentieren können und die untereinander vernetzt sind. Darunter zählen neben sozialen Netzwerken auch Foren und Kommentarspalten von Online-Newspapern. Aufgrund dieser starken Vernetzung sollte der Fokus der Forschung zu Hasskommentaren in Zukunft nicht nur auf der Betrachtung einzelner Kommentare liegen, sondern auch auf der Ausbreitung von Hass in und zwischen wohldefinierten Netzwerken. An dieser Stelle kommt der Begriff des Netzwerk Hops ins Spiel. In Rechnernetzen nennt man einen *Hop* einen Zwischenschritt, eine Etappe, zwischen zwei Netzwerkknoten. Der Begriff kann direkt auf die Ausbreitung von Kommentaren im Netz übertragen werden, denn auch Kommentare verbreiten sich dort über mehrere Zwischenschritte (Hops) im Netzwerk und über Netzwerkgrenzen hinweg. Die Netzwerkknoten sind in diesem Fall schreibende Personen im Netzwerk, statt der Router oder Gateways in der Netzwerktechnik. Durch Personen, die in mehreren Netzwerken oder Netzwerksegmenten agieren, werden Kommentare so von einem Segment in ein anderes weiterverbreitet. Bei diesen Personen muss es sich nicht um den eigentlichen Verfasser des ursprünglichen Kommentares handeln.

### 3. Detektion und Meldung

Die Grundlage, um gegen Hatespeech, aggressive und toxische Kommentare vorgehen zu können, ist, dass jene Kommentare überhaupt erst einmal gefunden werden. Betreiber sozialer Netzwerke und Moderatoren von Kommentarspalten in Online-Newspapern sind damit regelmäßig überfordert, da die Detektion viele Schwierigkeiten mit sich bringt. In diesem

Abschnitt sollen Möglichkeiten und Herausforderungen bei der Detektion und auch der Meldung von besonders drastischen Hasskommentaren an dafür vorgesehene Meldestellen aufgezeigt werden.

### 3.1. Ausgangspunkt und Gegebenheiten für die Detektion von Hatespeech

Die Grundlage für die Analyse und Eindämmungsversuche von Hatespeech ist das Aufspüren von Hasskommentaren, denn erst dadurch werden Ansatzpunkte für mögliche Maßnahmen erhalten. Das kann geschehen, indem Personen konkrete Kommentare, die sie als Hatespeech ansehen, melden, beispielsweise bei der Meldeplattform „Hessen gegen Hetze“<sup>6</sup> des hessischen „CyberCompetenceCenters“ (Hessen3C). Dort werden gemeldete Kommentare gesichtet, bewertet und im Bedarfsfall an Betreiber von sozialen Netzwerken oder Strafverfolgungsbehörden gemeldet. Auf der anderen Seite können Hasskommentare durch intelligente Systeme insbesondere auf Seiten der Netzbetreiber dazu beitragen, Hasskommentare frühzeitig zu entdecken. Kritisch ist es jedoch, wenn Computersysteme allein über die Löschung von Kommentaren entscheiden, da die Meinungsfreiheit ein hohes Gut ist, welches dadurch gefährdet wird.

Für deutsche Behörden kommt die automatische Überwachung von sozialen Netzwerken aus Datenschutzgründen nicht in Frage. Jedoch können intelligente Systeme bei der Klassifizierung von gemeldeten Kommentaren erheblich unterstützen und somit den Personalaufwand verringern. Solche Systeme treffen keine alleinigen Entscheidungen vorbei an Spezialist\*innen, sondern sie priorisieren und nehmen Vorklassifizierungen vor, die es dem Personal erleichtern, Kommentare zu sichten und einzuschätzen.

Im Folgenden sollen deshalb aktuelle Ansätze beschrieben werden, wie Hatespeech automatisch erkannt werden kann. Der Schwerpunkt wird weniger auf die technischen Methoden gelegt, viel mehr sollen Merkmale von Hatespeech aufgegriffen werden, die für die Erkennung relevant sind. Es soll erläutert werden, was Maschinen gut erkennen können und wo die gegenwärtigen Herausforderungen bei der automatischen Erkennung liegen. Dadurch wird deutlich, warum es auch nach mittlerweile vielen Jahren Forschung noch immer Schwierigkeiten bei der Hatespeech-Detektion gibt (Struß et al. 2019).

---

<sup>6</sup> <https://hessengegenhetze.de/hate-speech-melden> (Zugegriffen am 10.12.2021)



In den meisten Fällen stehen einzelne Kommentare zur Verfügung, die zuerst einmal hinsichtlich Hatespeech und Toxizität bewertet werden müssen. Handelt es sich um Hatespeech, kommen meist weitere relevante Kategorisierungen hinzu, um den Kommentar entsprechend bewerten zu können. Zusätzlich kann durch Berücksichtigung des Gefahrenpotentials eine Priorisierung der Kommentare erfolgen.

### 3.2. Allgemeine Merkmale und Metadaten von Hasskommentaren

Die Metadaten, die für einen Kommentar zur Verfügung stehen, sind stark abhängig von der Quelle, aus der ein Kommentar stammt. In sozialen Netzwerken, z.B. Twitter oder Facebook, sind verhältnismäßig viele Metadaten angegeben. Standardmäßig gehören dazu zunächst Datum und Uhrzeit der Veröffentlichung und der Nutzernamen. Je nach Netzwerk sind noch Anzahl der Likes und Dislikes, die Anzahl des Teilens und Zitierens und die Anzahl der Antworten auf diesen Kommentar angegeben. Schaut man sich die Kommentarspalten von Online-Newsseiten an, sind wesentlich weniger Metadaten vorhanden: In der Regel eine Datums- und/oder Zeitangabe, ein Nutzernamen und vielleicht noch die Anzahl der Antworten. Fest steht, je mehr Merkmale zur Verfügung stehen, desto besser ist es für die Auswertung, und desto mehr können die Metadaten zu einer Entscheidung beitragen.

Datum und Uhrzeit sind für die Erkennung von Hatespeech weniger relevant. Die Nutzernamen können jedoch schon Hinweise auf die Einstellung der Schreibenden geben, wenn es sich nicht um Personennamen handelt, sondern beispielsweise Schimpfwörter oder bekannte Codewörter bestimmter Gruppen sind. Letzteres ist häufig zu finden (Jaki und De Smedt 2019).

In sozialen Netzwerken sind insbesondere die Metadaten von Interesse, die anzeigen, wie andere Netzwerkteilnehmer\*innen auf den Kommentar reagieren und wie stark er sich verbreitet. Dazu zählen Likes und Dislikes sowie die Anzahl, wie oft ein Kommentar geteilt und zitiert wurde. Diese Metadaten sind insbesondere hilfreich bei der Abschätzung, wie der Kommentar von der Community angenommen wird und wie kontrovers er diskutiert wird. Erhält ein Kommentar größtenteils Likes bzw. Dislikes, dann wird der Kommentar von der Community mehr oder weniger einheitlich angenommen oder abgelehnt. Im Gegensatz dazu treten kritische Kommentare häufig in Kontexten auf, in denen unterschiedliche Meinungen aufeinandertreffen. Darauf weist ein ausgeglichenes Verhältnis von Likes und Dislikes hin. Selbstverständlich ist nicht jeder Kommentar mit ähnlich vielen Likes und Dislikes ein Hasskommentar, es kommt immer auf das konkrete Umfeld und die Personengruppen an, die



den Kommentar sehen, und auf deren Diskussionskultur. Nichtsdestotrotz ist Hatespeech in Kommentaren mit größtenteils Likes bzw. Dislikes deutlich seltener (Siersdorfer et al. 2014; Wojatzki et al. 2018).

### 3.3. Textuelle Merkmale

Der Haupt-Fokus der automatischen Hatespeech-Klassifikation liegt auf der Auswertung des Kommentartexts. Andere Merkmale unterstützen die darauf basierende Klassifikation lediglich. Die wichtigsten textuellen Aspekte betreffen die Wörter im Text. Das Vorkommen von Wörtern, die direkt Hass ausdrücken, ist dabei ein sicheres Merkmal. Diese Wörter können in Wortlisten gespeichert werden, sodass der Text damit abgeglichen werden kann. Es können Wortlisten verwendet werden, die der Forschung frei zur Verfügung stehen, wie die bei Ross et al. (2016) beschriebene Liste. Eine Liste von Hasswörtern kann auch aus Trainingsdaten automatisch extrahiert werden, indem die Wörter in den verschiedenen Klassen miteinander verglichen werden (Alrehili 2019). Um herauszufinden, welche Wörter in einem Text besondere Bedeutung haben, kann die Term Frequency-Inverse Document Frequency (TF-IDF) verwendet werden. Der Abgleich zwischen den Wörtern im Text und den Wörtern in den Wortlisten geschieht, indem die Wörter im Text in einen „Bag of Words“ (BOW) gepackt werden, wobei die Reihenfolge der Wörter im Text ignoriert wird. Für weitere Details sei an dieser Stelle auf Heyer et al. (2015) verwiesen.

In einigen Fällen werden sogenannte „N-Grams“ verwendet. An die Stelle von Wörtern beim BOW treten bei N-Grams Ketten von Wörtern oder von Zeichen. N-Grams auf Wortebene sind Ketten von N (meist zwei oder drei) Wörtern, N-Grams auf Zeichenebene sind Ketten von N Zeichen (Buchstaben, Satzzeichen, Leerzeichen usw.). Das maschinelle Lernen lernt dabei die Häufigkeit des Vorkommens der N-Grams in den einzelnen Klassen. N-Grams können direkt als Features im maschinellen Lernen verwendet werden, wie zum Beispiel in Roy et al. (2020) oder Wiegand et al. (2018).

Auf der Ebene von Wörtern im Text basiert auch die Idee der „Word Embeddings“ (Mikolov et al. 2013). Die Textdaten werden in numerische Vektoren überführt. In diesen Vektoren ist für jedes Wort kodiert, mit welchen anderen Wörtern es im Kontext (mit welcher Wahrscheinlichkeit) auftreten kann. Es wird dabei der Kontext rechts wie auch links betrachtet. Dadurch kann man semantische Zusammenhänge zwischen Wörtern in den Trainingsdaten

erkennen: Semantisch ähnliche Wörter, die in ähnlichen Kontexten auftreten, und semantisch zusammenhängende Wörter, die häufig gemeinsam auftreten.

Mithilfe von „Part-of-Speech Tagging“ kann bestimmt werden, welcher syntaktischen Kategorie ein Wort angehört. Kombiniert man beispielsweise Part-of-Speech Tagging mit N-Grams, so lassen sich Rückschlüsse auf Wortart-Kombinationen schließen.

Um eine bessere Erkennung von flektierten Wortformen zu erreichen, kann eine „Lemmatisierung“ der Wörter erfolgen. In diesem Prozess werden Wörter auf ihre Grundform zurückgeführt, also z.B. das Wort „Häuser“ auf „Haus“. Für das Deutsche kann auch noch eine Kompositaanalyse hilfreich sein, bei der zusammengesetzte Nomen in ihre Bestandteile zerlegt werden.

In vielen Fällen sind nicht nur einzelne Wörter, sondern Phrasen relevant. Ein Beispiel ist der Ausdruck „kriminelle Flüchtlinge“. Diese Phrasen können an Textbeispielen automatisch gelernt werden, wenn die Wörter der Phrase häufig miteinander auftreten.

Relevant sind aber nicht nur die Hasswörter, sondern auch die Ziele des Hasses und das Thema. Politiker\*innen, Frauen, Flüchtlinge, Angehörige ethnischer Gruppen und Menschen mit jüdischer oder islamischer Religionszugehörigkeit sind besonders häufig Ziele. Aktuelle politische Themen werden besonders häufig aggressiv diskutiert. Bei den Themen gibt es noch die besondere Herausforderung, dass sie sich im Laufe der Zeit wandeln.

Für die Analyse des Sentiments und von Emotionen wird die Satzebene betrachtet. Eine stark negative Stimmung und wütende Emotion, die im Text identifiziert werden können, geben wichtige Hinweise auf potenzielle Hasskommentare. In diesen Analysen ist es auch notwendig, Negationen mit einzubeziehen (Siegel und Alexa 2020).

Bick (2020) zeigt einen anderen Aspekt der Hatespeech-Analyse in deutschsprachigen Social-Media-Daten: Gegenderte Tweets, die z.B. Ausdrücke wie „Muslim\*innen“ enthalten, enthalten seltener Hatespeech und sind seltener negativ in Bezug auf die Zielgruppe.

Weitere Aspekte des Texts, die für die Analyse einbezogen werden, sind Großschreibung ganzer Wörter, Hashtags und Zeichensetzung (z.B. mehrere Ausrufezeichen hintereinander).

#### 4. Bewertung von Hasskommentaren

Bei der Detektion und Analyse von Hasskommentaren ist immer auch die Bewertung solcher Kommentare von Bedeutung. Das beinhaltet sowohl eine gesellschaftlich moralische als auch eine strafrechtliche Bewertung. Die Bewertung ist für die Betreiber sozialer Netzwerke eine große Herausforderung, da Kommentare, die Straftatbestände erfüllen, entfernt werden müssen, gleichzeitig die Meinungsfreiheit aber nicht eingeschränkt werden darf. Es ist eine Gratwanderung. Darüber hinaus ist die Menge an automatisch detektierten oder von Nutzer\*innen gemeldeten Kommentaren immens. Doch nicht nur die Netzbetreiber haben Schwierigkeiten mit der Einschätzung, auch Strafverfolgungsbehörden und Gerichte sind sich immer wieder uneins. Ein Grund dafür ist, dass das Strafgesetzbuch nicht speziell auf Hasskommentare ausgelegt ist. Ein weiterer Grund ist, dass Kommentare oft viel Interpretationsspielraum lassen. Relevante Straftatbestände für Kommentare im Netz sind im deutschen Recht insbesondere die folgenden<sup>7</sup>:

- § 86 StGB Verbreiten von Propagandamitteln verfassungswidriger Organisationen
- § 86a StGB Verwenden von Kennzeichen verfassungswidriger Organisationen
- § 90a StGB Verunglimpfung des Staates und seiner Symbole
- § 90b StGB Verfassungsfeindliche Verunglimpfung von Verfassungsorganen
- § 111 StGB Öffentliche Aufforderung zu Straftaten
- § 126 StGB Störung des öffentlichen Friedens durch Androhung von Straftaten
- § 130 StGB Volksverhetzung
- § 131 StGB Gewaltdarstellung
- § 140 StGB Belohnung und Billigung von Straftaten
- § 166 StGB Beschimpfung von Bekenntnissen, Religionsgesellschaften und Weltanschauungsvereinigungen
- § 185 StGB Beleidigung
- § 186 StGB Üble Nachrede
- § 187 StGB Verleumdung
- § 189 StGB Verunglimpfen des Andenkens Verstorbener
- § 201a StGB Verletzung des höchstpersönlichen Lebensbereichs und von Persönlichkeitsrechten durch Bildaufnahmen
- § 240 StGB Nötigung

<sup>7</sup> <https://hessengegenhetze.de/fragen-antworten/hate-speech> (Zugegriffen am 10.12.2021)

- § 241 StGB Bedrohung

Um Netzwerkbetreibern Richtlinien an die Hand zu geben, wurde 2016 von der Europäischen Kommission der „Verhaltenskodex zur Bekämpfung illegaler Hassreden im Internet“<sup>8</sup> ins Leben gerufen. Darin verpflichten sich mehrere große Unternehmen, u.a. Facebook, Twitter, Microsoft, Youtube und Instagram, freiwillig dazu, illegale Inhalte auf deren Plattformen nach Möglichkeit innerhalb von 24 Stunden nach Bekanntwerden zu entfernen. Hervorgehoben wird aber auch, dass die Meinungsfreiheit gewahrt werden soll.

Wie wichtig es jedoch ist, aggressive Kommentare im Netz zu entfernen, um die Meinungsfreiheit zu wahren, hat eine Eurobarometer-Umfrage<sup>9</sup> aus dem Jahr 2016 gezeigt. Dort gaben knapp die Hälfte derer, die Online-Diskussionen passiv verfolgen und in der Vergangenheit schon Hass im Netz wahrgenommen haben, an, dass sie sich aufgrund von Hass im Netz selbst nicht an den Diskussionen beteiligen. Die andere Hälfte gab an, sich aus anderen Gründen nicht an den Online-Debatten zu beteiligen.

## 5. Reaktionen auf Konversationsebene

Neben der Detektion und Bewertung von Hasskommentaren, die in den vorherigen Abschnitten diskutiert wurden, ist es notwendig zu verstehen, wie auf Hassbotschaften reagiert wird, denn erst dadurch kann sich Hass in einem Netzwerk weiterverbreiten. Von Bedeutung ist dabei nicht nur der Hasskommentar selbst, sondern zusätzlich spielen Nutzereigenschaften und deren Position im Netzwerk eine Rolle: Agiert ein Nutzer nur in einem kleinen „Dunstkreis“ oder ist er in vielen verschiedenen Gruppen aktiv? Ist er nur ein passiver Mitleser oder agiert er als Meinungsmacher? Eigenschaften wie diese und Beziehungen zwischen Nutzer\*innen spielen bei der Analyse von Reaktionen auf Hasskommentare und bei deren Verbreitung eine zentrale Rolle. In diesem Abschnitt soll deshalb versucht werden, ein Verständnis zu entwickeln, wie auf Kommentare reagiert wird, wie sie sich ausbreiten und dabei verändern können.

Dabei wird zuerst die Konversationsebene betrachtet, um zu erläutern, wie ein Kommentar seine Antworten und damit die gesamte Konversation beeinflusst. Später wird die Betrachtung

<sup>8</sup> [https://ec.europa.eu/commission/presscorner/detail/de/QANDA\\_20\\_1135](https://ec.europa.eu/commission/presscorner/detail/de/QANDA_20_1135) (Zugegriffen am 10.12.2021)

<sup>9</sup> [https://ec.europa.eu/information\\_society/newsroom/image/document/2016-47/sp452-summary\\_en\\_19666.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2016-47/sp452-summary_en_19666.pdf) (Zugegriffen am 10.12.2021)

ausgeweitet auf die Reaktion und Weiterverbreitung in einem ganzen Netzwerk, wo insbesondere Relationen zwischen Nutzern und deren Verhalten von großer Bedeutung sind (Saveski et al. 2021).

Bei der Betrachtung von Reaktionen auf Konversationsebene bildet das Verhältnis zwischen Kommentar und der Antwort dazu die Grundlage, denn daraus besteht jede Konversation. Dass bestimmte Kommentare zum Schreiben von Hasskommentaren anregen, förmlich provozieren, wurde in zahlreichen Forschungsarbeiten übereinstimmend festgestellt (Almerekhi et al. 2020; Cheng et al. 2017; Kwon und Gruzd 2017a; Saveski et al. 2021). Damit einher geht auch die Erkenntnis, dass es sich nicht um eine geringe Anzahl an Nutzer\*innen handelt, die sehr viel toxischen Inhalt verbreiten, sondern dass es viele Nutzer\*innen gibt, die nur gelegentlich toxische Kommentare verfassen (Saveski et al. 2021). Aufgrund der Vielzahl der Nutzer\*innen kommt eine große Menge von Hasskommentaren zusammen. Dass diese Nutzer\*innen nur gelegentlich Hatespeech verfassen, zeigt, dass sie in vielen Fällen nicht von sich aus toxisch sind, sondern dass sie in bestimmten Gesprächssituationen dazu verleitet werden, aggressiv zu reagieren. Es stellt sich daher die Frage, wann und was genau Nutzer\*innen zum Schreiben toxischer Inhalte bewegt. Eine wesentliche Rolle spielt die allgemeine Stimmungslage einer Person: Ist die Person in schlechter Stimmung, steigt die Wahrscheinlichkeit, dass sie toxisch reagiert. Kommen negative und provozierende Kommentare (allgemein negativer Kontext) mit schlechter Stimmung zusammen, verdoppelt sich die Wahrscheinlichkeit einer toxischen Antwort im Gegensatz zu positivem Kontext und guter Stimmung etwa (Cheng et al. 2017). Natürlich unterscheiden sich die genauen Zahlen dafür von Person zu Person, klar ist aber, dass die Wahrscheinlichkeit bedeutend ansteigt. Hervorzuheben ist dabei, dass die Stimmung nicht ausschließlich durch den Inhalt in Online-Konversationen bestimmt wird, sondern auch aus der realen Welt, der aktuellen Verfassung der Nutzer\*innen, mitgebracht wird (Terizi et al. 2021).

In der Kommunikation zwischen Personen in der realen Welt geben die Gesprächspartner\*innen freiwillig oder unfreiwillig Emotionen preis. Das geschieht nicht nur durch die Sprache selbst, sondern Emotionen werden vor allem durch Mimik, Gestik, Körperhaltung, Stimmlage u.a. gezeigt und größtenteils unbewusst von den Gesprächspartner\*innen aufgenommen. Adaptiert ein\*e Gesprächspartner\*in die Signale, spricht man von emotionaler Ansteckung (*emotional contagion*). In der Psychologie bezeichnet man diese Adaption als Mimikry (Kwon und Gruzd 2017b; Labudde 2019). In Online-Konversationen fallen diese Wahrnehmungen weg, es bleibt lediglich der Text, über den mit häufig unbekannten Gesprächspartner\*innen kommuniziert wird. Manchmal werden zusätzlich

Emojis, Sticker oder Bilder verwendet, die auch eine emotionale Wirkung haben. Trotzdem fallen die unterbewusst gezeigten Merkmale weg, wodurch Emotionen deutlich schwerer kommuniziert werden können. Nichtsdestotrotz haben mehrere Studien gezeigt, dass Emotionen trotzdem übertragen werden können. Analog zur Adaption von Mimik, Gestik, Körperhaltung u.a. im persönlichen Gespräch übernehmen Konversationsteilnehmer\*innen den Schreibstil oder linguistische Merkmale anderer Nutzer\*innen, wenn sie „angesteckt“ werden. (Kwon und Gruzd 2017b)

### 5.1. Textuelle Mimikry

Auf dieser Basis werden Emotionen aus einer Konversation unterbewusst aufgenommen und in neuen Kommentaren und Antworten widergespiegelt. Ein Beispiel dafür ist Fluchen oder die Benutzung von Schimpfwörtern. Diese können als Beleidigung oder Angriff an konkrete Personen gerichtet (*interpersonal*) oder allgemein (*public*) gehalten sein, d.h. ohne konkretes Ziel (z.B. „So eine Sch\*\*\*\* aber auch“). Von Relevanz ist bei der Übertragung von Emotionen außerdem die Kommentarthierarchie: In diesem Zusammenhang wird bei Kommentaren unterschieden in Elternkommentare und Kindkommentare. Erstere sind Posts oder Kommentare, die nicht als Antwort gepostet wurden, sondern eine neue Konversation initiieren. Letztere sind Antworten auf die Elternkommentare.

Es wurde festgestellt, dass sich toxische Eltern- und Kindkommentare auch dahingehend unterscheiden, dass sie unterschiedlich gerichteten Hass (*interpersonal* und/oder *public*) fördern. Die Ursache ist, dass Elternkommentare Posts sind, die sich an die Allgemeinheit richten, nicht an eine konkrete Person. Kindkommentare hingegen richten sich häufig direkt an die Autorin / den Autor des entsprechenden Elternkommentars. Enthält ein Elternkommentar nun Hatespeech, wurde beobachtet, dass in dessen Antworten auch gehäuft Hatespeech auftrat. Diese war etwa gleichermaßen an die Allgemeinheit (*public*) und an die Autorin / den Autor des Elternkommentars gerichtet (*interpersonal*). Toxische Elternkommentare fördern somit die Entstehung von allgemeinem Hass als auch gegen konkrete Personen gerichteten Hass. Anders sieht es aus, wenn man die Kindkommentare und wiederum deren Antworten betrachtet. Hier wurde festgestellt, dass Reaktionen auf Kindkommentare, die allgemeinen Hass enthalten, größtenteils auch nur allgemeinen Hass enthalten. Dazu kommt, dass Reaktionen auf Kindkommentare, die an konkrete Personen gerichteten Hass enthalten, im Wesentlichen nur Hassreaktionen mit ebenfalls an konkrete Personen gerichteten Hass hervorrufen. Anders ausgedrückt, bedeutet das, wenn man einen Kindkommentar mit allgemeinem (*public*) Hass hat, wird man in dessen Antworten auch hauptsächlich allgemeinen Hass vorfinden, aber weniger an konkrete Personen gerichteten

(interpersonal) Hass. Umgekehrt gilt dasselbe: Hat man einen Kindkommentar mit interpersonal gerichtetem Hass, dann wird man darunter ebenfalls hauptsächlich interpersonal gerichteten Hass vorfinden und kaum allgemeinen (public) Hass. (Kwon und Gruzd 2017b)

Die Erklärung für diese Feststellung liegt darin, dass sich Antworten auf Kindkommentare im Gegensatz zu Elternkommentaren häufig auf konkrete Personen beziehen, demzufolge ist die Kommunikation persönlicher. Das wiederum begünstigt persönliche Beleidigungen. Im Gegensatz dazu wird von allgemein gehaltenen toxischen Elternkommentaren lediglich der Schreibstil übernommen, d.h. es werden beispielsweise auch gehäuft Schimpfwörter verwendet. Der Mimikry-Mechanismus ist also ein leicht anderer, aber noch immer vorhanden. (Kwon und Gruzd 2017b)

Da die Mimikry-Theorie auf der sozialen Interaktion beruht, spielt auch das Verhältnis zwischen Nutzer\*innen eine Rolle. Dieses kommt insbesondere in Online-Konversationen zum Tragen, da es viel ausmacht, ob man die an der Kommunikation teilnehmenden Personen kennt, sei es aus vorherigen Online-Konversationen oder aus dem realen Leben. Abhängig davon sind Reaktionen unterschiedlich: Toxische Kommentare werden häufiger an unbekannte Personen gerichtet als an bekannte, denn weitgehende Anonymität führt oft dazu, dass unbedachter agiert wird. Daher ist das Freundschaftsverhältnis von Nutzer\*innen in sozialen Netzwerken für die Toxizitätsbestimmung von Bedeutung. (Saveski et al. 2021)

Für das Review von Online-Kommentaren kann demzufolge festgehalten werden, dass ein besonderer Fokus auf die Elternkommentare gelegt werden sollte, da diese Auslöser für ganze Hass-Threads sein können. (Kwon und Gruzd 2017b)

## 5.2. Conversation Branching

Jede Online-Konversation kann in Form eines Baumes überführt werden (Abbildung 2), wobei der Ausgangskommentar der Root-Knoten ist. *Conversation Branching* setzt sich mit der Form dieses Baums einer Konversation auseinander. Kritisch in Hinsicht auf Hatespeech sind oft große, kontroverse Diskussionen, in denen Meinungen gegensätzlicher Parteien aufeinandertreffen. Grundlegend gibt es zwei Arten großer Konversationen: Zum einen Konversationen, die lang werden, weil eine bestimmte Gruppe an Menschen intensiv diskutiert – fokussierte Konversationen. Zum anderen Konversationen, in denen viele Personen jeweils nur einzelne Nachrichten schreiben – expandierende Konversationen. Beispiele für expandierende Konversationen sind Gratulationen zu Erfolgen, wo viele Personen



Glückwünsche bekunden. Letztere sind meist ungefährlich in Hinsicht auf die Toxizität. (Backstrom et al. 2013)

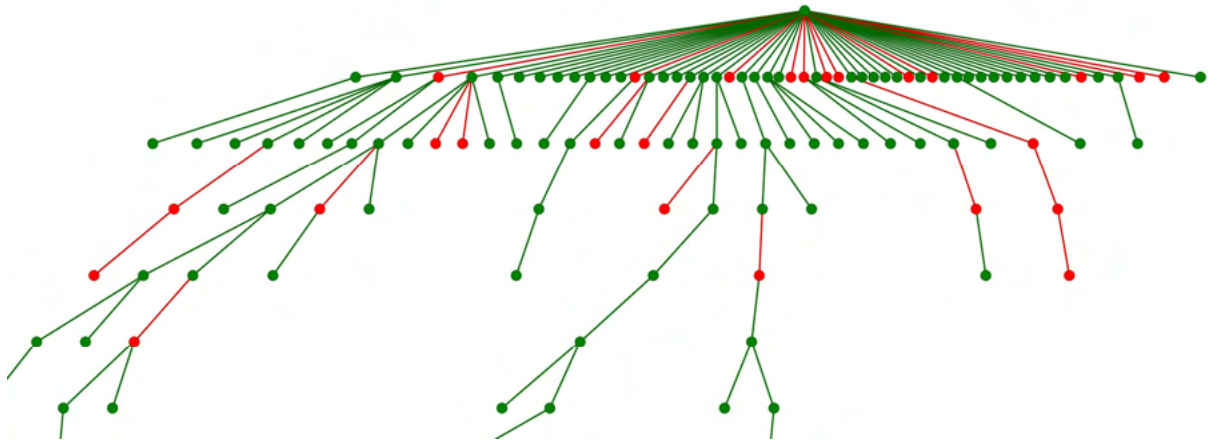


Abbildung 2: Ausschnitt der Baumdarstellung einer Twitter-Konversation mit 187 Kommentaren und 70 beteiligten Nutzern. Der Wurzel-Knoten ganz oben ist der Haupt-Tweet, darunter ist die Hierarchie der Antworten erkennbar. Mit einem Modell zur Hatespeech-Detektion wurden die Kommentare nach Hatespeech/nicht Hatespeech klassifiziert: Rote Knoten enthalten mit hoher Wahrscheinlichkeit Hatespeech, grüne nicht. Als Breite wird die maximale Anzahl von Kommentaren auf einer beliebigen Ebene im Baum bezeichnet. Die Tiefe wird definiert als die Entfernung (Länge des Pfades) zwischen dem Wurzelknoten und dem am weitesten entfernten Blattknoten (die Anzahl der Ebenen).

Es wird daher der Frage nachgegangen, ob die Baumstruktur etwas über die Toxizität der Konversation oder Teile der Konversation aussagt. Dafür liefert schon die Betrachtung grundlegender Eigenschaften dieses Baumes wertvolle Hinweise: Größere Bäume, d.h. Bäume mit vielen Kommentaren sind im Mittel toxischer als kleinere Kommentarbaume. Dasselbe gilt für die Breite (maximale Anzahl Kommentare auf einer beliebigen Ebene im Baum) und die Tiefe (tiefste Ebene/Level) wobei Breite und Tiefe mit der Größe eines Baumes korrelieren. Je größer ein Kommentarbaum, desto breiter und tiefer ist er in der Regel. Diese Beobachtung ist leicht erklärbar: kritische Konversationen, in denen sich Hass entwickelt, sind kontroverse Diskussionen, wo gegensätzliche Meinungen aufeinandertreffen. Dabei entstehen große Diskussionen, die einen großen, breiten und tiefen Baum zur Folge haben. (Saveski et al. 2021) Jedoch können wir noch nicht zwischen den beiden Formen langer Konversationen unterscheiden. Dafür wird die mittlere Tiefe herangezogen, denn in expandierenden Konversationen wird diese niedrig sein, weil Nutzer\*innen größtenteils direkt auf den Elternkommentar antworten. Im Gegensatz dazu haben fokussierte Konversationen eine größere Tiefe, da Nutzer\*innen sich gegenseitig antworten. (Backstrom et al. 2013)

Fokussierte Diskussionen werden nochmals verstärkt durch die zunehmende Anzahl an Konversationsteilnehmer\*innen. Je mehr Personen an einer Konversation beteiligt sind, desto kontroverser wird sie. Die Entwicklung einer Konversation unter Berücksichtigung dieses Verhaltens kann mathematisch modelliert werden. Die Grundlage kann dabei der *Galton-*



*Watson Branching Process* bilden (Kumar et al. 2010). Darauf soll an dieser Stelle jedoch nicht weiter eingegangen werden.

## 6. Nutzerbeziehungen und Gruppierung in Netzwerken

### 6.1. Bedeutung von Nutzerbeziehungen

Wie in den vorherigen Abschnitten bereits angesprochen wurde, spielen Beziehungen von Nutzer\*innen untereinander eine wesentliche Rolle bei der Entstehung und Verbreitung von Hass in sozialen Netzwerken. Bevor die Verbreitung von Hass untersucht werden kann, müssen jedoch erst einmal existente Beziehungen, Gruppenbildung und Wege des Austauschs zwischen Gruppen im Allgemeinen beleuchtet werden.

Grundsätzlich kann jedes soziale Netzwerk als Graph dargestellt werden, in dem die Knoten die Nutzer\*innen sind und (gerichtete) Kanten angeben, welchen Nutzer\*innen der ausgewählte Nutzer folgt bzw. welche Nutzer\*innen befreundet sind. Je nach betrachtetem sozialem Netzwerk gibt es in dieser Hinsicht kleinere Unterschiede, abhängig davon, welche Nutzerbeziehungen in dem Netzwerk möglich sind, also ob Personen sich beispielsweise befreunden können, sich folgen oder ähnliches. Die Betrachtung eines sozialen Netzwerkes als Graph (wie beispielsweise in Spranger et al. (2020) und Spranger et al. (2018)) hat den Vorteil, dass Gruppierungen und Beziehungen der beteiligten Personen abgebildet und damit erkennbar werden. Erkennbar werden beispielsweise Personen mit sehr vielen oder wenigen Followern, Gruppen von Personen, die sich gegenseitig folgen und Personen, die solche Gruppen verbinden, weil sie in beiden präsent sind. Anhand der Vernetzung der Nutzer\*innen kann die Ausbreitung von Informationen analysiert werden, wobei jede Strecke zwischen zwei Nutzern einem Netzwerk-Hop entspricht. Über Netzwerk-Hops verbreiten sich Inhalte im Netzwerk.

### 6.2. Gruppierungstheorien

Die Bildung sozialer Gruppen von Menschen sind seit langem Bestandteil von Forschungsarbeiten. Eine Vielzahl von Untersuchungen wurde in der realen Offline-Welt durchgeführt. Seit soziale Netzwerke immer populärer geworden sind, gibt es auch für die Online-Welt eine große Zahl an Untersuchungen, wie sich Gruppierungen in Netzwerken verhalten. Dabei gibt es Analogien zu Verhaltensmodellen der realen Offline-Welt. Aufgrund anderer Gegebenheiten der Online-Welt können diese aber natürlich nicht 1:1 übertragen werden. Da soziale Gruppen die Grundlage für die Meinungsverbreitung bilden, wird auf diese

im Folgenden eingegangen. (Frenzel und Labudde 2020; Labudde 2019; Spranger und Labudde 2020)

Menschen organisieren sich so gut wie überall in Gruppen: von Freunden und Familie, über den Arbeitsplatz bis hin zu Freizeitaktivitäten in Vereinen zum Beispiel. – All jene sind abgeschlossene, aber über die Zeit veränderbare Gruppen von Menschen, die in diesen Gruppen aufgrund mehr oder weniger vieler Gemeinsamkeiten, gleicher Ziele oder gleicher Interessen zusammenkommen. Es bilden sich sogenannte Mikrokulturen (Stegbauer 2019). Die Basis der Gruppenbildung sind Verbindungen (*Ties*) zwischen Menschen. Diese können verschiedene Ausprägungen haben, beispielsweise eine Stärke, sie können das Verhältnis der Menschen zueinander angeben (Familie, Freunde, Kollegen etc.), gerichtet oder ungerichtet sein. (Rivera et al. 2010)

Auf diesen Verbindungen zwischen Menschen basiert die Gruppendynamik eines sozialen Netzwerks und damit auch der Informationsfluss. Deshalb wird an dieser Stelle auf grundlegende Mechanismen eingegangen, die für die Bildung von Gruppen verantwortlich sind.

Eine im Volksmund als „Ähnliches gesellt sich“ bezeichnete Tendenz ist die *soziale Homophilie*. Sie basiert im Wesentlichen auf der Tatsache, dass Menschen mit Vorliebe den Weg des geringsten Aufwands gehen. Da die Kommunikation mit ähnlichen Personen einfacher ist bauen sich engere Verbindungen zwischen diesen Personen auf. Ähnlichkeiten können gleiche Meinungen und Ansichten sein, aber auch ähnliche Eigenschaften wie Alter, Geschlecht, sozialer Status, Bildungsgrad und Herkunft. In vielen Fällen entstehen Freundschaften oder auch Familienbeziehungen (Rivera et al. 2010). Neben der Ähnlichkeit beruhen die Verbindungen meist auch auf Gegenseitigkeit. Das bedeutet, wenn Person A eine Verbindung zu Person B aufbaut, dann wird B diese mit hoher Wahrscheinlichkeit erwidern. Darüber hinaus gibt es den Effekt der Transitivität. Das bedeutet, wenn Person A eine enge Bindung zu Person B, und Person B bereits eine enge Bindung zu Person C hat, dann ist die Wahrscheinlichkeit hoch, dass auch A eine enge Verbindung zu C aufbaut. Die Anziehung von Personen wird auch als *Attraction* bezeichnet (Abbildung 3). Dadurch, dass sich ähnliche Personen anziehen und Bindungen aufbauen, grenzen sie sich gleichzeitig auch nach außen zu anderen Personen hin ab (*Repulsion*), zu denen keine Verbindungen oder sogar negative Verbindungen bestehen. Einflüsse, die zur Separierung von Gruppen beitragen, werden auch unter dem *Minimal Group Paradigm* zusammengefasst. (Stadtfeld et al. 2020; Törnberg et al. 2021)

Die Homophilie ist der Mechanismus, von dem soziale Netzwerke profitieren und der diese letzten Endes attraktiv macht. Die Empfehlungsmaschinen hinter den sozialen Netzwerken empfehlen Nutzer\*innen Inhalte auf Basis der Ähnlichkeit zu Inhalten, die der/dem Nutzer\*in bereits gefällt (*Collaborative filtering*) (Sahu und Singh 2019). Somit bekommen Nutzer\*innen mit ähnlichen Eigenschaften ähnliche Inhalte vorgeschlagen, über die sie sich dann auch vernetzen können. Es ist festzuhalten, dass die Algorithmen das Zusammenfinden ähnlicher Nutzer\*innen zu Gruppen unterstützen. (Stegbauer 2019)

Dem entgegen steht die soziale Heterophilie oder auch als *Contact Theory* bezeichnet. Diese bezeichnet Verbindungen zwischen ungleichen Personen, also mit unterschiedlichen Eigenschaften, Meinungen, Herkunft usw. Solche Verbindungen entstehen häufig bei der Arbeit oder allgemein in Einrichtungen, die Personen in einem Team zusammenbringen, um ein bestimmtes Ziel zu erreichen. Dabei kommt es nicht auf Ähnlichkeiten zwischen den Personen an, sondern häufig auf deren Wissen oder Können. Die Unterschiedlichkeit der Personen wirkt sich positiv aus, sofern die Gegebenheiten stimmen (Rivera et al. 2010). Darüber hinaus kann Heterophilie auch in dem Aufeinandertreffen zweier Gruppen, z.B. Freundesgruppen, verursacht sein, sofern dabei bereits eine Verbindung zwischen beiden Gruppen existiert (zwei befreundete Personen, von denen jede jeweils eine beiden Gruppen angehört) und die Verbindungsglieder präsent sind, keine Ängste aber Empathie vorhanden sind. Unter diesen Gegebenheiten könne Gruppengrenzen verschmelzen. (Törnberg et al. 2021)

Zusätzlich spielt noch die Popularität von Nutzer\*innen eine Rolle. Nutzer\*innen, die bereits als populär und einflussreich im Netzwerk wahrgenommen werden, ziehen weitere Verbindungen an. Auch hierbei orientieren sich Nutzer\*innen wieder an ihren Gruppen: mögen viele der Gruppenmitglieder ein Mitglied X, werden Mitglieder aus dieser Gruppe Mitglied X wahrscheinlich auch mögen und früher oder später eine positive Verbindung zu ihm bilden. Für negative Bindungen ist das Prinzip dasselbe. (Pál et al. 2015; Stadtfeld et al. 2020)

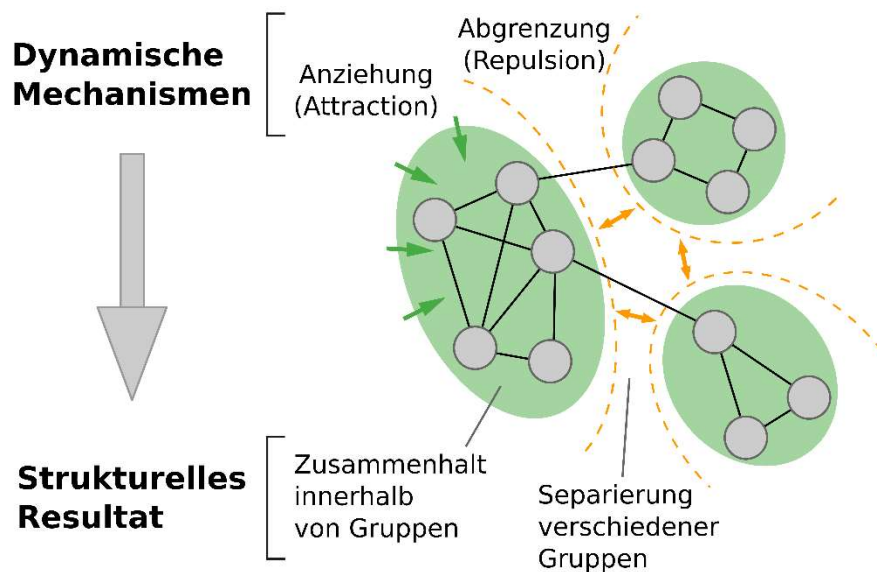


Abbildung 3: Mechanismen der Gruppenbildung in sozialen Netzwerken (in Anlehnung an Stadtfeld et al. (2020))  
Die Gruppenbildung in sozialen Netzwerken ist bedingt durch die Anziehung von ähnlichen Mitgliedern innerhalb von Gruppen und damit einhergehend der Abgrenzung von Gruppen gegeneinander.

Zusammengefasst besteht demzufolge eine hohe Ähnlichkeit zwischen Mitgliedern einer Gruppe, und es gibt Differenzen zwischen Mitgliedern verschiedener Gruppen. Diese Differenzen werden als *Intergroup-Conflict* bezeichnet. Gruppendynamisches Verhalten in sozialen Netzwerken wird in Frenzel und Labudde (2020) weiter vertieft.

Einhergehend mit der Gruppenbildung im Netzwerk verhält sich die Meinungsbildung analog. Da Kommunikation zu einem Großteil innerhalb der Gruppen stattfindet, bilden sich Meinungen heraus, die von den meisten Mitgliedern einer Gruppe vertreten werden. Eine wichtige Rolle für diesen Prozess spielt das soziale Feedback innerhalb einer Gruppe, wodurch sich Meinungen verstärken oder abschwächen. Je homogener die Gruppe ist, desto stärker ist der Prozess und desto extremere Meinungen oder Ansichten können sich herausbilden. Anhaltende extreme Meinungsverschiedenheiten innerhalb von Gruppen sind selten. (Törnberg et al. 2021)

Dieser Prozess der Meinungsbildung findet nicht nur in sozialen Netzwerken statt, sondern auch in der realen Welt. Jedoch bilden sich durch soziale Netzwerke häufig extremere Meinungen aus, da Gruppen dort größer und homogener sind. Im Gegensatz zur realen Welt spielt die geografische Distanz dort keine Rolle, wodurch sich große homogene Gruppen erst herausbilden können. Das ist in der realen Welt schwieriger, da der geografische Aktionsradius eingeschränkt ist. Durch die geringeren Differenzen werden die Meinungen dann extremer und der *Intergroup-Conflict* größer. (Frenzel und Labudde 2020; Törnberg et al. 2021)

## 6.3. Netzwerke – Graphen und Zentralitätsmaße

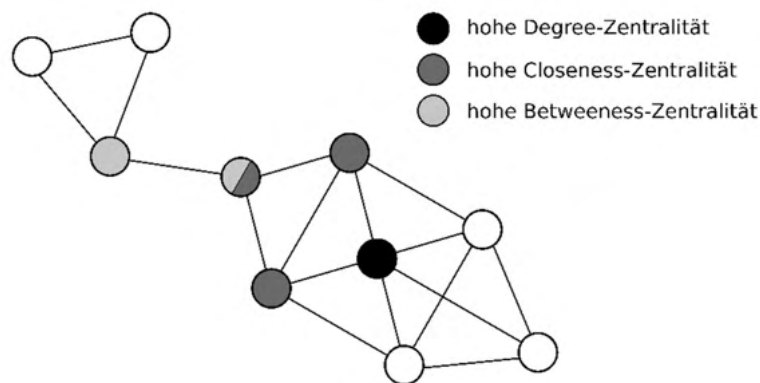


Abbildung 4: Darstellung der verschiedenen Zentralitätsmaße von Mitgliedern (Knoten) in einem Netzwerk (in Anlehnung an Liu et al. (2017))

In einem sozialen Netzwerk, welches als Graph präsentiert ist, gibt es mehrere Maße/Kennzahlen, die etwas über die Position des Knotens, in unserem Fall der Nutzer\*innen, relativ zu anderen Knoten aussagen (Abbildung 4). Zuerst einmal gibt es die Gradzahl (*Degree*) eines Knotens, die angibt, wie viele eingehende und ausgehende Verbindungen zu anderen Knoten bestehen. Je höher der Grad eines Knotens, desto eingebundener ist er im Netz und desto höher ist die Wahrscheinlichkeit, dass ein Knoten ein\*e Meinungsführer\*in (*Opinion Leader*) wird (oder bereits ist). Ein weiteres Maß ist die *Betweenness-Zentralität*. Diese gibt an, wie häufig ein Knoten auf dem kürzesten Weg zwischen zwei anderen Knoten liegt. Knoten mit hoher *Betweenness-Zentralität* liegen daher zwischen verschiedenen Netzwerkclustern und sind wichtig für den Datenaustausch zwischen beiden. Das bedeutet auch, dass diese Knoten in gewissem Maße regulieren können, welche Informationen jeweils in das andere Netzwerkcluster weitergeleitet werden und welche nicht. Zuletzt gibt es noch die *Closeness-Zentralität*, die angibt, wie nah der Knoten zu allen anderen Knoten ist. Liegt ein Knoten zentral in der Mitte des Netzwerkes, hat er zu allen anderen Knoten einen tendenziell kürzeren Weg (weniger Netzwerk-Hops bis zum Ziel) als ein Knoten, der am Rand des Netzwerkes liegt. Demzufolge propagieren Informationen von Knoten mit hoher *Closeness-Zentralität* schneller durch das Netz als Informationen, die von Knoten mit geringer *Closeness-Zentralität* ausgehen. (Liu et al. 2017)

Demzufolge sind *Opinion Leader* meist Knoten, die zum einen gut vernetzt sind (hoher *Degree*) und von denen sich Informationen schnell verbreiten (hohe *Closeness-Zentralität*). Der Informationsfluss geht von den *Opinion Leaders* zu den *Opinion Followers*. Und die *Opinion*

*Leader* werden häufig von den Medien stark beeinflusst. (Dressler und Telle 2009; Liu et al. 2017; Spranger et al. 2018)

## 7. Ausbreitung von Hass im Netzwerk

Wer gelegentlich im Netz unterwegs ist wird es aus sozialen Netzwerken kennen: Gruppen mit gegensätzlichen und extrem gefestigten Ansichten treffen in aggressiven Diskussionen aufeinander. Es entwickeln sich teilweise Wortgefechte, die man mitunter kaum noch als Diskussion bezeichnen kann. Viele Nutzer\*innen beobachten solche Konversationen nur passiv und fragen sich womöglich, was dort eigentlich geschieht. Doch auch wenn man sich nicht selbst an solchen Diskussionen beteiligt, merkt man unter Umständen, wie sich eine gewisse Abneigung oder gar Wut gegenüber der „gegnerischen“ Seite in einem bemerkbar macht. Häufig liegt das an der verzerrten Faktenlage, die in solchen Diskussionen dargeboten wird und natürlich der Diskussionskultur an sich. Und genau das sind sehr gefährliche Situationen, denn es gibt genügend Nutzer\*innen, die sich nicht zurückhalten können und mit eben dieser in sich aufbrodelnden Wut der Konversation beitreten. (Stegbauer 2019)

Wie es erst zu solchen Diskussionen kommen kann, soll in diesem Abschnitt beleuchtet werden.

Wie bereits in den vorherigen Abschnitten erwähnt wurde, gibt es meist nicht einige wenige, sehr toxische und aggressive Nutzer\*innen, sondern viele Nutzer\*innen können durch die Stimmung im Netz und den Kontext einer Diskussion zum Schreiben von toxischen Kommentaren verleitet werden.

### 7.1. General Aggression Model

Ein weitverbreitetes Model zur Modellierung menschlicher Aggression ist das *General Aggression Model* (GAM), welches mehrere Theorien aus der Psychologie und Soziologie berücksichtigt (Allen und Anderson 2017). Dementsprechend werden viele verschiedene Faktoren einbezogen, darunter soziale, kognitive und biologische Faktoren, aber auch Wahrnehmung und Interpretation. Der Grundgedanke ist, dass jede Person eine Art Aggressionsstatus besitzt, der durch Interaktion mit anderen Netzwerkteilnehmern verändert werden kann. Somit unterliegt der Zustand einer Person einer dauerhaften Veränderung, die im GAM als Zyklus modelliert wird (Abbildung 5). (Allen et al. 2018; Terizi et al. 2021)

Den Ausgangspunkt (*Inputs*) stellen Faktoren dar, die den internen Aggressionszustand beeinflussen. Faktoren, die den Aggressionszustand negativ beeinflussen, d.h. die Aggression steigern, sind Risikofaktoren. Analog gibt es Schutzfaktoren, die die Aggression positiv beeinflussen, also senken. Die Einflussfaktoren werden nochmals unterschieden in

persönliche und situationsbedingte Faktoren. Persönliche Faktoren sind psychologische Faktoren, die beeinflussen, wie eine Person auf bestimmte Situationen reagiert, beispielsweise freundlich, kritisch oder abwertend. Persönliche Faktoren sind meist relativ stabil über einen längeren Zeitraum und verändern sich nur langsam. Persönliche Risikofaktoren für hohe Aggression sind z.B. normative Akzeptanz von Gewalt, niedrige Verträglichkeit im psychologischen Sinne u.a. Dazu kommen weiterhin situationsbedingte Faktoren, die zum beobachteten Zeitpunkt eine Person beeinflussen. Das sind wie auch in vorherigen Abschnitten bereits angesprochen z.B. die Stimmungslage, Frustration, Schmerzen, Angst oder auch Alkoholeinfluss. (Allen et al. 2018)

Die zweite Phase (*Routes*) beschäftigt sich damit, wie die Eingangsfaktoren die Entscheidungs- und Beurteilungsprozesse eines Menschen beeinflussen und sich dadurch auf den aktuellen Aggressionszustand auswirken. Je nach aktuellem Befinden eines Menschen verändert sich die Wahrnehmung und das Interpretationsvermögen, wodurch man entsprechend unterschiedlich reagiert. (Allen et al. 2018)

Im dritten Schritt (*Outcomes*) werden die Auswirkungen des veränderten Aggressionszustandes eines Menschen betrachtet und wie diese sich auf die Reaktion des Menschen auswirken, d.h. ob schließlich eine aggressive (Affekt-)Handlung oder eine bedachte Handlung ausgeführt wird. Da das Resultat wiederum die Ausgangsfaktoren beeinflusst, schließt sich an der Stelle der Zyklus und beginnt von vorne. (Allen et al. 2018; Allen und Anderson 2017)

Auf soziale Netzwerke angewendet bedeutet das, dass jedes Mitglied einen aktuellen Aggressionsstatus hat. Dieser wird durch Interaktion mit einem anderen Mitglied (z.B. in einer Konversation) beeinflusst, was sich auch auf dessen Aggressionszustand auswirkt. Im letzten Schritt spiegelt sich der veränderte Aggressionszustand in der Reaktion wider, die den Input für die Kommunikationspartner bildet. Somit schließt sich der Kreis. (Terizi et al. 2021)



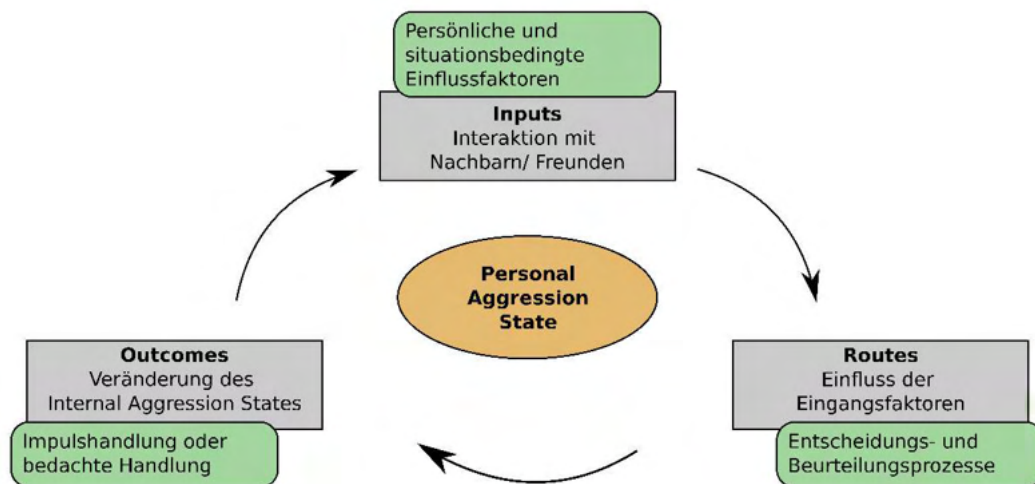


Abbildung 5: Das General Aggression Model beschreibt den Prozess, wie sich der Aggressionszustand von Menschen durch Interaktion mit anderen Menschen verändert. Bestimmte Einflussfaktoren (Inputs) wirken auf einen Menschen ein, beeinflussen dessen Entscheidungs- und Beurteilungsvermögen (Routes) und wirken sich dadurch wiederum auf seine Handlungen (Outcomes) und seinen Aggressionszustand aus (In Anlehnung an Allen et al. (2018) und Terizi et al. (2021)).

Unter der Annahme, dass eine Interaktion zwischen zwei Mitgliedern den Aggressionszustand negativ beeinflusst, d.h. die Aggression steigt, hat das eine Reaktion mit erhöhter Aggression im Vergleich zum vorherigen Zyklus zur Folge. Durch die negative bzw. aggressivere Reaktion ist es wahrscheinlich, dass sich auch bei der anderen an der Konversation beteiligten Person der Aggressionszustand verschlechtert. Beide Kommunikationspartner\*innen geraten so in eine Abwärtsspirale, die den Aggressionsstatus erhöht.

Betrachtet man das Verhalten nun auf einer höheren Ebene, ist es so, dass zwei Netzwerkknoten (Nutzer\*innen), die durch eine Kommunikation verbunden sind, ihren Aggressionszustand gegenseitig verstärken. In vielen Fällen kommuniziert eine Person jedoch mit mehreren Kontakten parallel. Demzufolge ist es wahrscheinlich, dass sich die erhöhte Aggression in weiteren Kommunikationen, die die Person führt, widerspiegelt. Die erhöhte Aggression wird also im Netz gestreut und verbreitet sich so weiter. (Squicciarini et al. 2015)

Dieses Verhalten spiegelt sich u.a. in den Ergebnissen einer Studie zum Aggressionslevel von Nutzer\*innen im sozialen Netzwerk Gab wider. Gab ist ein Netzwerk, das für toxische Inhalte bekannt ist, da dort keinerlei Moderation und Kontrolle stattfinden, wie das beispielsweise auf Twitter und Facebook der Fall ist. Es wurde festgestellt, dass die meisten Personen mit einem niedrigen Aggressionslevel dem Netzwerk beitreten und erst mit der Zeit, die sie in Gab aktiv sind, das Aggressionslevel höher wird, da die „Neulinge“ vom Hass angesteckt werden. (Gallacher und Bright 2021)

## 7.2. Bedeutung einflussreicher Personen und Opinion Leader für die Entstehung von Hass



Ein wichtiger Faktor für die Entwicklung von Strukturen in sozialen Netzwerken ist die Verbreitung von Informationen, denn nicht alle Mitglieder tragen gleichermaßen dazu bei. Der Großteil der Mitglieder liest lediglich die Posts und Kommentare, die im eigenen Newsfeed angezeigt werden und kommentiert oder teilt einige davon möglicherweise noch. Die eigentlichen Informationsquellen im Netzwerk, die neuen Inhalt, Informationen und Meinungen in das Netz einbringen, sind Opinion Leader. Diese sind gut vernetzt und können durch die Medien beeinflusst werden (Liu et al. 2017). Damit ist gemeint, dass aktuelle Trends und politische Themen aus den Medien aufgegriffen werden. Mögliche Medien umfassen dabei nicht nur die deutschen öffentlich-rechtlichen Rundfunksender ARD und ZDF, sondern insbesondere auch soziale Medien und fragwürdige Nachrichtenportale wie beispielsweise „Russia Today“. Im Marketing wird die gute Vernetzung der Opinion Leader auch gezielt ausgenutzt, dort sind diese besser unter dem Begriff Influencer bekannt. Doch würden Opinion Leader lediglich „trockene“ Informationen streuen, wären sie nicht so beliebt, dass sie von tausenden Personen abonniert oder gefolgt werden. Entscheidend für die Follower ist die Wertung durch ihre Opinion Leader, denen sie folgen. Jeder Opinion Leader hat ein bestimmtes Klientel mit ähnlichen Ansichten und Meinungen, das ihm folgt und jede\*r Nutzer\*in wird durch verschiedene Opinion Leader beeinflusst (Spranger et al. 2018). Denn auch an dieser Stelle handeln Nutzer\*innen wieder im Sinne der Homophilie, indem sie eben jenen Opinion Leadern folgen, die ihre Meinung vertreten. Ein Opinion Leader kann somit als eine Art „Oberhaupt“ für die eigene Gruppe, d.h. die Menge der eigenen Follower, angesehen werden. Es bilden sich Meinungsblasen, die durch Opinion Leader gesteuert werden und in denen sich aufgrund fehlender Diversität einheitliche Meinungen herausbilden, die von den Mitgliedern der Meinungsblase stark vertreten werden. Innerhalb der Gruppe ist das Verhältnis daher in der Regel gut. Extremformen solcher nach außen abgeschotteten Gruppen, die lediglich solche Informationen konsumieren, die sie auch sehen wollen, werden als *Echo Chambers* bezeichnet (Törnberg et al. 2021). Als Folge bilden sich extreme Meinungen heraus, die den *Intergroup-Conflict*, d.h. Meinungsverschiedenheiten zu anderen Gruppen, erhöhen.

Für die Detektion von Opinion Leadern in sozialen Netzwerken gibt es Algorithmen, die auf Googles PageRank-Algorithmus basieren (Brin und Page 1998). Dieser berücksichtigt eingehende und ausgehende Verbindungen zu und von Elementen im Netzwerk, in diesem Fall Nutzer\*innen. Vereinfacht gesagt, sind Nutzer\*innen, die viele eingehende Verbindungen haben, d.h. denen von vielen Nutzer\*innen gefolgt wird, einflussreiche Nutzer\*innen (Opinion Leader) und Nutzer\*innen, die viele ausgehende Verbindungen haben, sind eher Opinion Follower. Als Verbesserung für die Bestimmung von Opinion Leadern wurde darauf aufbauend

später der LeaderRank vorgeschlagen (Lü et al. 2011). Beide Algorithmen funktionieren jedoch weniger gut, wenn das Netzwerk eine sternförmige Topologie hat. Das ist der Fall, wenn es nur wenige zentrale und aktive Nutzer\*innen gibt, die Informationen posten. Ein Beispiel dafür ist das Netzwerk, dass durch eine Facebookseite aufgespannt wird. Dort wird hauptsächlich vom Eigentümer der Seite (einer Person oder oft auch Organisation) Inhalt veröffentlicht, jedoch weniger von den einzelnen Nutzer\*innen, die der Seite folgen. Für diesen Fall wurde von Spranger et al. (2018) der CompetanceRank vorgeschlagen. Dieser ist eine Modifikation des LeaderRanks.

Die Opinion Leader selbst sind dabei meist keine Mitglieder solcher *Echo Chambers*, da sie zum einen hauptsächlich unidirektional agieren, d.h. sie geben neue Informationen an ihre Follower ins Netzwerk, aber sie reagieren wenig auf ihre Follower, da sie das auch gar nicht schaffen würden. Zum anderen wollen Opinion Leader selbst nicht in Verruf geraten. (Liu et al. 2017; Stegbauer 2019)

Anfällig für die Verbreitung von Hass sind hingegen Mitglieder der *Echo Chambers*. Diese stehen in der Hierarchie unter den Opinion Leadern, sind aber trotzdem sehr aktiv und dadurch auch gut vernetzt. Durch die Gruppenbildung bekommen Mitglieder dieser Meinungsblasen von der Gruppe starke Unterstützung, was zu einem hohen Selbstbewusstsein führt mit der Annahme, dass sie selbst nicht falsch liegen können. Die Meinung verfestigt sich. Darüber hinaus werden neue Informationen so interpretiert, dass sie zur Gruppenmeinung passen.

Durch eine Weiterverbreitung, z.B. über Retweets, Zitate oder eigene Posts, werden die polarisierten Meinungen im Netz verbreitet. Durch die extremen Ansichten erwecken solche Posts vor allem in Gruppen auf der Gegenseite den Drang, diese Ansichten mit ihren eigenen zu korrigieren. In der Folge entsteht eine Diskussion zwischen Mitgliedern verschiedener Gruppierungen, die zentral im Netzwerk stattfindet und eine große Reichweite besitzt. Da beide (oder auch mehr) Seiten verhärtete Meinungen und kaum einen Willen für eine sachliche Diskussion haben, entstehen keine kontroversen Diskussionen, sondern es entsteht ein hasserfüllter Shitstorm, der von Anfeindungen beider Seiten gegeneinander geprägt ist.

Auch hier sind soziale Netzwerke wieder ein Teufelskreislauf, denn gerade durch die gute Vernetzung und die hohe Zentralität verbreiten sich die Inhalte solcher *Echo Chambers* besonders schnell und weit im Netzwerk. Das wiederum induziert aufgrund der fehlenden Offenheit für Austausch eine Flut von triggernden Kommentaren. Keine der beiden Seiten ist in der Regel willig, einsichtig zu reagieren, ganz im Gegensatz entwickeln sich im Sinne der Regel der sozialen Reziprozität gegenseitige Anfeindungen nach dem Motto „Wie du mir, so ich dir“ (Stegbauer 2019). Über die Ansteckung des aggressiven Verhaltens wie es im GAM

beschrieben wurde, entwickelt sich toxisches und aggressives Verhalten. (Mathew et al. 2019) Es bleibt jedoch dabei, dass sich Gruppen unterschiedlicher Ansichten gegenseitig provozieren und innerhalb von Gruppen die Unterstützung bleibt. Allgemein sinkt die Aggression gegeneinander, je größer die Überlappung sozialer Kontakte und Beziehungen im Netzwerk ist. (Saveski et al. 2021)

Im Allgemeinen wurden auch Zusammenhänge zwischen aggressivem Verhalten und den Zentralitätswerten zweier Kommunikationspartner festgestellt. Es wurde beobachtet, dass Hatespeech häufiger auftritt, wenn beide Kommunikationspartner einen großen Unterschied in ihren Zentralitätswerten haben. Die Person mit der höheren Zentralität ist dann in der Regel diejenige, die Hatespeech schreibt oder aggressives Verhalten gegenüber dem anderen Kommunikationspartner zeigt (Terizi et al. 2021). Das Verhalten ist dadurch erklärbar, dass gut vernetzte Nutzer\*innen mit einer hohen Zentralität meist eine größere Gruppe haben, von der sie Unterstützung bekommen und die die Ansichten untermauert. Zusätzlich können die „Hater“ mit der Verteidigung ihrer Meinung in der Gruppe prahlen und sich somit Ansehen verschaffen. Damit einhergehend fühlt sich eine wenig vernetzte Nutzerin meist stärker angegriffen, wenn sie von einem sehr aktiven Nutzer mit hoher Zentralität beleidigt wird, da die Geschädigte im Prinzip vor der gesamten Gruppe des „Haters“ bloßgestellt wird. (Terizi et al. 2021)

Es bleibt festzuhalten, dass Opinion Leader ein treibender Faktor in sozialen Netzwerken sind. Von ihnen fließen Informationen ins Netz zu den Opinion Followern. Durch mangelnde Diversität bilden sich unter den Opinion Followern sogenannte *Echo Chambers*, Gruppen mit einseitigen und extrem gefestigten Ansichten, heraus. Aufgrund fehlender Offenheit gegenüber anderen Ansichten führt das zu einem hohen *Intergroup-Conflict* und damit häufiger zu aggressivem Verhalten.

## 8. Schlussfolgerungen für Monitoring- und Eindämmungsmaßnahmen

In den vorherigen Kapiteln wurde erläutert, wie sich Hatespeech in sozialen Netzwerken ausbreitet, was für Einflüsse es gibt und welche Folgen aggressives Verhalten nach sich zieht. All dieses sollte zum Verständnis des Informationsflusses und der Meinungsausbreitung beitragen. An dieser Stelle sollen diese Informationen nun genutzt werden, um zu diskutieren, inwieweit uns dieses Verständnis dabei hilft, Ideen und Maßnahmen zu entwickeln, mit denen die Verbreitung von Hatespeech verringert werden kann. Dabei soll aber nicht nur das Unterbinden von Hatespeech und toxischem Verhalten im Vordergrund stehen, sondern es

soll und muss auch die Meinungsfreiheit als ein sehr wichtiges Kriterium dabei diskutiert werden, denn zensierte Medien sind im Interesse keiner der beteiligten Parteien. Wir werden jedoch an vielen Stellen feststellen, dass zwischen der Eindämmung von Hatespeech und der Meinungsfreiheit nur ein schmaler Grat liegt.

Die gängige Praxis ist, dass Hasskommentare mit Warnhinweisen versehen oder gelöscht werden. Voraussetzung dafür ist allerdings, dass sie überhaupt erst einmal als solche erkannt werden. Genau dort liegt das Problem, denn eine manuelle Sichtung der Kommentare ist aufgrund der riesigen Menge nicht möglich. Zwar gibt es mittlerweile viel Forschung zur automatischen Hatespeech-Detektion, jedoch ist es sehr kritisch, allein Maschinen entscheiden zu lassen, welche Kommentare gelöscht werden und welche nicht oder gar Nutzer automatisch zu sperren. Selbst für Menschen ist es nicht einfach einzuschätzen, ob es sich bei einem Kommentar um Hatespeech handelt und ob er so kritisch ist, dass er gelöscht werden sollte. Aus diesem Grund ist es sinnvoll, an der Prävention von Hatespeech zu arbeiten, um gar nicht erst bzw. seltener zu dem Punkt zu kommen, dass entschieden werden muss, ob Kommentare gelöscht werden oder nicht.

Es muss jedoch vorweggenommen werden, dass viele Maßnahmen zur Eindämmung von Hass in sozialen Netzwerken nur durch deren Betreiber selbst umgesetzt werden könnten, da sie auf zugrundeliegenden Algorithmen basieren, die uns neue Inhalte, Freunde und Seiten vorschlagen, die uns gefallen könnten. Seit Oktober 2017 gilt für die Netzbetreiber das Netzwerkdurchsetzungsgesetz<sup>10</sup> (NetzDG), das Netzbetreiber u.a. dazu verpflichtet ein Beschwerdemanagement einzurichten, rechtswidrige Inhalte zu entfernen und regelmäßige Berichterstattung hinsichtlich Hasskommentaren durchzuführen. Es gibt aber auch einige Leitlinien, wie beispielsweise die der Amadeu-Antonio-Stiftung<sup>11</sup> oder der Seite No-Hate-Speech.de<sup>12</sup>, die die Nutzer\*innen selbst umsetzen können, um dem Hass im Netz entgegenzutreten. Eine Schwierigkeit ist außerdem, dass die meisten Ideen zur Prävention in der Literatur schwer getestet werden können, da ein soziales Netzwerk kaum in vollem Umfang simuliert oder nachgebaut werden kann. Daher bleibt es in der Regel bei Vorschlägen, die aus Analysen abgeleitet werden, allerdings nicht in Hinsicht auf ihre tatsächliche Wirkung verifiziert werden können.

Als eine der Hauptursachen für die Entwicklung wurde die Bildung extremer, verhärteter Meinungen in Gruppen identifiziert und der daraus entstehende *Intergroup-Conflict*. Dessen

<sup>10</sup> [https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG\\_node.html](https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html) (Zugegriffen am 28.12.2021)

<sup>11</sup> <https://www.amadeu-antonio-stiftung.de/digitale-zivilgesellschaft/was-ist-hate-speech/> (Zugegriffen am 10.12.2021)

<sup>12</sup> <https://no-hate-speech.de/de/wissen/> (Zugegriffen am 28.12.2021)

Ursache ist wiederum Inhalt mit mangelnder Diversität, den die Nutzer\*innen angezeigt bekommen. Das Problem dabei ist, dass Nutzer\*innen hauptsächlich den Inhalt, der ihnen gefällt, liken, teilen und kommentieren. Auf der Basis dieses Nutzerverhaltens lernen Algorithmen im Hintergrund, was einer Person gefällt und schlagen ihr ähnliche Inhalte vor. Das ist im Prinzip eine geniale Idee, die soziale Netzwerke attraktiv macht: Die Nutzer\*innen sind zufrieden, weil sie Inhalte angezeigt bekommen, die sie interessieren und die Netzbetreiber binden sie dadurch fester an das soziale Netzwerk. Dieses Prinzip hat jedoch auch eine negative Seite: es bewirkt, dass Nutzer\*innen größtenteils lediglich Inhalte angezeigt werden, die ihre bereits vorhandene Meinung widerspiegeln. Das bekräftigt sie in ihrer Meinung oder Ansicht, die Vielfalt und Diversität gehen verloren und der *Intergroup-Conflict* wächst. (Sahu und Singh 2019; Stegbauer 2019)

Dieser Prozess müsste demzufolge unterbrochen werden, indem Nutzer\*innen neben Empfehlungen ähnlicher Inhalte gezielt gegensätzliche Ansichten angezeigt werden. So würde ihnen in den *Echo Chambers* ein realeres Bild der Gesellschaft vermittelt und die Ausbildung und Bestärkung extremer Meinungen würde mit hoher Wahrscheinlichkeit verringert werden. (Sahu und Singh 2019)

Dieser Ansatz, Empfehlungen anzupassen, könnte sogar noch ausgeweitet werden. Einfach ausgedrückt könnten die „Täter“, die gehäuft toxische Kommentare verfassen, abgeschirmt werden von „normalen“ Nutzer\*innen (Terizi et al. 2021). Der Hintergrund ist, dass aggressives Verhalten ansteckend ist, wie im GAM beschrieben wurde. Würde man potenziell toxische Posts nur einem kleinen Personenkreis sichtbar machen, in erster Linie möglicherweise Nutzer\*innen, die dieselbe Ansicht vertreten, könnte die Ausbreitung aggressiven Verhaltens verringert werden. Nach diesem Prinzip könnten auch Gruppen extremer, gegensätzlicher Meinungen, unter denen es häufig Anfeindungen gibt, voneinander abgegrenzt werden. Von der Seite der Meinungsfreiheit betrachtet stellt sich hier die Frage, inwieweit Algorithmen regeln sollten, dass bestimmte Kommentare bestimmten Personen oder Gruppen nicht angezeigt oder vorgeschlagen werden. Ein Kommentar würde so nicht gelöscht werden, aber die Sichtbarkeit und damit die Popularität im Netzwerk verringert. Für einige mag das sinnvoll klingen, da sowieso von Algorithmen gesteuert wird, welche Inhalte man vorgeschlagen bekommt. Andere könnten hingegen eine Art der Zensur in diesem Prinzip sehen.

Eine andere, aber ähnliche Variante ist, dass mit maschinellen Lernverfahren eine Vorhersage dahingehend getroffen wird, ob eine Person auf einen bestimmten Kommentar toxisch reagieren wird oder nicht. Studien haben gezeigt, dass solch eine Vorhersage auf Basis der bereits vorhandenen Konversation und der Nutzereigenschaften durchaus möglich ist (Saveski

et al. 2021). Konversationszweige, auf die diese spezifische Person mit hoher Wahrscheinlichkeit toxisch reagieren würde, könnten dann weit unten in der Konversation angezeigt werden. Je weiter unten ein Kommentar steht, desto wahrscheinlicher ist es, dass der Kommentar nicht gelesen wird, und demzufolge sinkt die Wahrscheinlichkeit, dass darauf toxisch oder aggressiv reagiert wird. (Saveski et al. 2021)

Das Ranking von Kommentaren (positive weiter oben, negative und toxische weit unten) hätte auch noch einen zweiten Nutzen: Lesen die Nutzer\*innen mehr positive Kommentare, weil diese oben im Feed angezeigt werden, verbessert sich die allgemeine Stimmung, im Idealfall im gesamten Netzwerk (Cheng et al. 2017). Da, wie bereits in vorherigen Abschnitten beschrieben, die Wahrscheinlichkeit einer toxischen Reaktion auch stark von der aktuellen Stimmung abhängt, würde eine Verbesserung der Stimmung zu weniger aggressiven Verhalten führen. Ein Nachteil ist, dass aufgrund dieses Rankings negative Kommentare weniger Feedback bekommen. Auf der einen Seite ist das gewollt, um toxisches Verhalten zu verstecken. Auf der anderen Seite kann das Feedback der Community, sofern es sachlich und konstruktiv ist, unter toxischen Kommentaren auch dazu beitragen, dass sich die Einstellung des Nutzers ändert und er in Zukunft bedachter handelt. (Cheng et al. 2017)

Als Zwischenfazit kann bis hier hin festgehalten werden, dass es nicht zielführend genug und nicht ausreichend ist, sich auf das Löschen von Hasskommentaren zu beschränken. Das ist die jetzige Praxis in sozialen Netzwerken, doch obwohl so viele Kommentare gelöscht werden, steigt die Aggressivität in sozialen Netzwerken immer weiter (Cinelli et al. 2021). Durch das Löschen der Hasskommentare wird nicht das aggressive Verhalten an sich bekämpft, sondern es wird vordergründig nur das Resultat, nämlich die Hasskommentare, entfernt. Um das aggressive Verhalten der Nutzer in Griff zu bekommen, müssen Methoden gefunden werden, wie beispielsweise die vorgestellten, um die Stimmung im Netz zu verbessern, einen reflektierten Umgangston zu etablieren und toxische Kommentare, die aus dem Affekt heraus entstehen, zu vermeiden (Cheng et al. 2017).

Hat sich erst einmal eine hasserfüllte Diskussion entwickelt, ist es häufig schwer, diese wieder zu stoppen, denn die treibenden Gesprächspartner sind oft nicht an einer sachlichen Diskussion interessiert und in ihren Standpunkten, so extrem sie seien mögen, gefestigt (Stegbauer 2019). Nichtsdestotrotz ist Gegenrede<sup>11</sup> **Fehler! Textmarke nicht definiert.** (*Counter Speech*) ein Mittel, um zumindest Schadensbegrenzung zu betreiben. Durch sachliche Erklärungen, das Richtigstellen von Fakten und dem Ausdruck der Verachtung von Hassrede und Aggression werden die aktiven Kommunikationspartner selbst in der Regel zwar



nicht beschwichtigt, Betroffene und Mitlesende werden jedoch empfänglich und dankbar für die Gegenrede sein.

## 9. Fazit

In diesem Kapitel wurde auf Möglichkeiten, Probleme und Herausforderungen der Detektion, Bewertung und Ausbreitung von Hatespeech im Netz eingegangen. Bisher beschränken sich viele Forschungsarbeiten auf die Analyse von Postings auf Kommentarebene, d.h. es werden Kommentare im Einzelnen betrachtet. Nicht zuletzt aufgrund der enormen Schwierigkeit einzuschätzen, ob es sich in einem Kommentar um Hatespeech oder sogar strafrechtlich relevanten Inhalt handelt, haben Menschen wie Maschinen damit häufig Probleme. Breitet sich Hass in sozialen Netzwerken jedoch erst einmal aus, wird das aggressive Verhalten der Nutzer immer mehr zum Problem und die Stimmung im Netzwerk verschlechtert sich.

Um die Auswirkungen von Hass im Netz besser verstehen zu können und dessen Ursachen zu untersuchen, wurde die Konversations- und Netzwerkebene betrachtet. Es wurde festgestellt, dass schon Konversationen einen deutlichen Mehrwert im Vergleich zu einzelnen Kommentaren haben. Durch die Darstellung von Konversationen als Baum wird deutlich, ob es sich um fokussierte oder expandierende Konversationen handelt. Darüber hinaus können anhand der Konversationsanalyse Vorhersagen gemacht werden, mit welcher Wahrscheinlichkeit die Antwort eines spezifischen Nutzers Hatespeech enthalten wird und wie sich Emotionen der Nutzer\*innen im Text widerspiegeln und auf andere Nutzer\*innen unbewusst übertragen werden.

Mit von Bedeutung sind darüber hinaus immer auch Beziehungen der Nutzer\*innen untereinander, denn diese geben Aufschluss über Gruppierungen im Netz. Damit einhergehend findet die Meinungsbildung statt: Gruppen bilden sich vordergründig aufgrund ähnlicher Meinungen, Ansichten und Interessen ihrer Mitglieder. Aufgrund der Ähnlichkeiten kommt es zur Verfestigung von Meinungen bis hin zur Herausbildung extremer Meinungen aufgrund fehlender Diversität. Gruppierungen mit solch extremen und gefestigten Meinungen sind gefährlich, denn damit vergrößert sich die Diskrepanz zwischen verschiedenen Gruppen – es verstärkt sich der *Intergroup-Conflict*, fördert die Entstehung und Verbreitung von Hass und verschlechtert dadurch die Stimmung im Netzwerk. Anhand des Informationsflusses im Netzwerk, von den Opinion-Leadern zu den Opinion-Followern, wurden Aussagen zur Ausbreitung von Hasskommentaren gemacht.

Abschließend wurden die gewonnen Informationen genutzt, um mögliche Maßnahmen für eine effektive Bekämpfung von Hass im Netz abzuleiten. Die gängige Praxis, Hasskommentare zu löschen, ist zwar gut, behebt jedoch die eigentlichen Ursachen der Entstehung von Hass nicht. Geeignete Maßnahmen beziehen sich daher auf die Anzeigereihenfolge von Kommentaren in Newsfeeds und auf die Begrenzung der Ausbreitung bestimmter Kommentare und der Reichweite von Schlüsselpersonen.

### Danksagung

Diese Arbeit ist ein Kooperationsprojekt der Hochschule Mittweida, dem Fraunhofer-Institut für Sichere Informationstechnologie Darmstadt, der Hochschule Darmstadt und dem Hessen CyberCompetenceCenter (Hessen3C) gefördert durch das Hessische Ministerium des Innern und für Sport. Im gemeinsamen Projekt „DeTox“ steht die Detektion von Toxizität und Aggressionen in Postings und Kommentaren im Netz im Fokus der Forschung.

### Literaturverzeichnis

- Allen, J. J., Anderson, C. A. & Bushman, B. J. (2018). The General Aggression Model. *Current Opinion in Psychology* (19), 75–80. doi:10.1016/j.copsyc.2017.03.034
- Allen, J. J. & Anderson, C. A. (2017). General Aggression Model. In P. Rössler, C. A. Hoffner & L. van Zoonen (Hrsg.), *The International Encyclopedia of Media Effects* (The Wiley Blackwell-ICA international encyclopedias of communication, S. 1–15). Chichester: John Wiley & Sons, Ltd; Wiley.
- Almerekhi, H., Kwak, H., Jansen, B. J. & Salminen, J. (2019). Detecting Toxicity Triggers in Online Discussions. In C. Atzenbeck (Hrsg.), *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (HT '19, S. 291–292). New York, NY, USA: Association for Computing Machinery.
- Almerekhi, H., Kwak, H., Salminen, J. & Jansen, B. J. (2020). Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. In *Proceedings of The Web Conference 2020*. ACM. <http://www.bernardjjansen.com/uploads/2/4/1/8/24188166/3366423.3380074.pdf>.



- Alrehili, A. (2019). Automatic Hate Speech Detection on Social Media: A Brief Survey. In *IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (S. 1–6).
- Backstrom, L., Kleinberg, J., Lee, L. & Danescu-Niculescu-Mizil, C. (2013). Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. *Proceedings of WSDM 2013*, 13–22.
- Bick, E. (2020). An Annotated Social Media Corpus for German. In N. Calzolari (Hrsg.), *Proceedings of the 12th Language Resources and Evaluation Conference. Twelfth International Conference on Language Resources and Evaluation* (S. 6127–6135). Marseille, France: European Language Resources Association; The European Language Resources Association (ELRA). <https://aclanthology.org/2020.lrec-1.752>.
- Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30 (1), 107–117. <https://www.sciencedirect.com/science/article/pii/S016975529800110X>.
- Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In C. P. Lee (Hrsg.), *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, S. 1217–1230. New York, NY, USA: Association for Computing Machinery; ACM.
- Cinelli, M., Francisci Morales, G. de, Galeazzi, A., Quattrociocchi, W. & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America* 118 (9). doi:10.1073/pnas.2023301118
- Dressler, M. & Telle, G. (2009). *Meinungsführer in der interdisziplinären Forschung*: Gabler, Betriebswirt.-Vlg.
- Dudenredaktion (o. J.). „aggressiv“ auf Duden online. <https://www.duden.de/node/13375/revision/13402>. Zugegriffen: 27. Dezember 2021.
- Freissmuth, M. (2016). Allgemeine Toxikologie. In M. Freissmuth, S. Offermanns & S. Böhm (Hrsg.), *Pharmakologie und Toxikologie: Von den molekularen Grundlagen zur Pharmakotherapie // Pharmakologie und Toxikologie. Von den molekularen Grundlagen*

- zur *Pharmakotherapie* (Springer-Lehrbuch, 2., aktualisierte und erweiterte Auflage, S. 849–866). Berlin, Heidelberg: Springer Berlin Heidelberg; Springer.
- Frenzel, C. & Labudde, D. (2020). Gruppendynamik in Sozialen Netzwerken - Bestimmung und Vorhersage von Gruppendynamiken auf Grundlage von Daten aus Sozialen Netzwerken. In R. Berthel (Hrsg.), *Kriminalistik und Kriminologie in der VUCA-Welt. Kriminalität und digitaler Raum, Gefahren für den Rechtsstaat* (S. 211–240). Rothenburg/Oberlausitz: Eigenverlag der Hochschule der Sächsischen Polizei (FH. <https://www.polizei.sachsen.de/de/dokumente/PolFH/BandX105XRalphXBerthelXXHrsgXXX2020X-XK.pdf>).
- Gallacher, J. D. & Bright, J. (2021). Hate Contagion: Measuring the spread and trajectory of hate on social media. doi:10.31234/osf.io/b9qhd
- Heyer, G., Quasthoff, U. & Wittig, T. (2015). *Text Mining: Wissensrohstoff Text Konzepte, Algorithmen, Ergebnisse* (Informatik): Berlin Dortmund Springer Campus.
- Jaki, S. & De Smedt, T. (2019). Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection. *CoRR abs/1910.07518*.
- Kumar, R., Mahdian, M. & McGlohon, M. (2010). Dynamics of Conversations. In I. Adä & M. Berthold (Hrsg.), *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Modular Data Generators* (KDD '10, S. 553–562). New York, NY, USA: Association for Computing Machinery; Bibliothek der Universität Konstanz.
- Kwon, K. H. & Gruz, A. (2017a). Is Aggression Contagious Online? A Case of Swearing on Donald Trump's Campaign Videos on YouTube. In *Proceedings of the 50th Hawaii International Conference on System Sciences 2017 (HICSS-50). January 4-7, 2017, Waikoloa Village, Hawaii*. Erscheinungsort nicht ermittelbar: Hawaii International Conference on System Sciences; AIS Electronic Library (AISeL).
- Kwon, K. H. & Gruz, A. (2017b). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* 27 (4), 991–1010. doi:10.1108/intr-02-2017-0072
- Labudde, D. (2019). Vorhersage von Gruppendynamiken auf der Grundlage von Daten aus sozialen Netzwerken. In S. Ellebrecht, S. Kaufmann & P. Zoche (Hrsg.), *(Un-)Sicherheiten*

- im Wandel: gesellschaftliche Dimensionen von Sicherheit* (S. 185–204). Münster LIT.  
<https://krimdok.uni-tuebingen.de/Record/1671275993>.
- Liu, W., Sidhu, A., Beacom, A. M. & Valente, T. W. (2017). Social Network Theory. In P. Rössler, C. A. Hoffner & L. van Zoonen (Hrsg.), *The International Encyclopedia of Media Effects* (The Wiley Blackwell-ICA international encyclopedias of communication, S. 1–12). Chichester: John Wiley & Sons, Ltd; Wiley.
- Lohs, K., Elstner, P. & Stephan, U. (Hrsg.). (2008). *Fachlexikon Toxikologie*: Springer Berlin Heidelberg.
- Lü, L., Zhang, Y.-C., Yeung, C. H. & Zhou, T. (2011). Leaders in Social Networks, the Delicious Case. *PLOS ONE* 6 (6), 1–9. doi:10.1371/journal.pone.0021202
- Mathew, B., Dutt, R., Goyal, P. & Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. In P. Boldi (Hrsg.), *Proceedings of the 10th ACM Conference on Web Science - WebSci '19* (ACM Digital Library ). New York, NY, United States: ACM Press; Association for Computing Machinery.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)* (S. 3111–3119).
- Pál, J., Stadtfeld, C., Grow, A. & Takács, K. (2015). Status Perceptions Matter: Understanding Disliking Among Adolescents. *Journal of Research on Adolescence* 26 (4), 805–818. doi:10.1111/jora.12231
- Rivera, M. T., Soderstrom, S. B. & Uzzi, B. (2010). Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology* 36 (1), 91–115. doi:10.1146/annurev.soc.34.040507.134743
- Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N. & Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki & T. Zesch (Hrsg.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (Bochumer Linguistische Arbeitsberichte, Bd. 17, S. 6–9). Bochum.

- Roy, P. K., Tripathy, A. K., Das, T. K. & Gao, X.-Z. (2020). A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. *IEEE Access* 8, 204951–204962. doi:10.1109/ACCESS.2020.3037073
- Sahu, S. & Singh, S. K. (2019). Ethics in AI: Collaborative filtering based approach to alleviate strong user biases and prejudices. In *Twelfth International Conference on Contemporary Computing (IC3)*. IEEE.
- Saveski, M., Roy, B. & Roy, D. (2021). The Structure of Toxic Conversations on Twitter. In J. Leskovec (Hrsg.), *Proceedings of the Web Conference 2021 (WWW '21)*, S. 1086–1097. New York, NY, USA: Association for Computing Machinery.
- Siegel, M. & Alexa, M. (2020). *Sentiment-Analyse deutschsprachiger Meinungsäußerungen: Grundlagen, Methoden und praktische Umsetzung*: Springer Fachmedien Wiesbaden GmbH.
- Siersdorfer, S., Chelaru, S., San Pedro, J., Altingovde, I. S. & Nejdil, W. (2014). Analyzing and Mining Comments and Comment Ratings on the Social Web 8 (3), 1–39. doi:10.1145/2628441
- Spranger, M., Hanke, K.-J., Heinke, F. & Labudde, D. (2020). Measuring Competence: Improvements to Determine the Degree of Opinion Leadership in Social Networks. *International Journal on Advances in Internet Technology* 13, 97–109.
- Spranger, M., Heinke, F., Siewerts, H., Hampl, J. & Labudde, D. (2018). Opinion Leaders in Star-Like Social Networks: A Simple Case? In D. Labudde (Hrsg.), *The Eighth International Conference on Advances in Information Mining and Management (IMMM)* (S. 33–38). Barcelona, Spain: IARIA.
- Spranger, M. & Labudde, D. (2020). Vorhersage von Gruppendynamiken auf der Grundlage von Daten aus Sozialen Netzwerken. In T.-G. Rüdiger & P. S. Bayerl (Hrsg.), *Cyberkriminologie: Kriminologie für das digitale Zeitalter* (S. 653–683). Wiesbaden: Springer Fachmedien Wiesbaden.
- Squicciarini, A., Rajtmajer, S., Liu, Y. & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM.

- Stadtfeld, C., Takács, K. & Vörös, A. (2020). The Emergence and Stability of Groups in Social Networks. *Social Networks* 60, 129–145. doi:10.1016/j.socnet.2019.10.008
- Stegbauer, C. (2019). Massenhafte Wutanfälle im Internet oder kann der Shitstorm jeden treffen? In *Die Digitalisierung der Kommunikation: Gesellschaftliche Trends und der Wandel von Organisationen, Science Policy Paper ; 5*.
- Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M. & Klenner, M. (2019). Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)* (S. 354–365). Erlangen, Germany: German Society for Computational Linguistics & Language Technology.
- Terizi, C., Chatzakou, D., Pitoura, E., Tsaparas, P. & Kourtellis, N. (2021). Modeling aggression propagation on social media. *Online Social Networks and Media* 24, 100137. <https://www.sciencedirect.com/science/article/pii/S2468696421000215>.
- Törnberg, P., Andersson, C., Lindgren, K. & Banisch, S. (2021). Modeling the emergence of affective polarization in the social media society. *PLOS ONE* 16 (10), 1–17. doi:10.1371/journal.pone.0258259
- Wiegand, M., Ruppenhofer, J., Schmidt, A. & Greenberg, C. (2018). Inducing a lexicon of abusive words - a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (S. 1046–1056). New Orleans, Louisiana. [https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/8471/file/Wiegand\\_et\\_al\\_Inducing\\_a\\_Lexicon\\_of\\_Abusive\\_Words\\_2018.pdf](https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/8471/file/Wiegand_et_al_Inducing_a_Lexicon_of_Abusive_Words_2018.pdf).
- Wojatzki, M., Horsmann, T., Gold, D. & Zesch, T. (2018). Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In Adrien Barbaresi, Hanno Biber, Friedrich Neubarth, Rainer Osswald (Hrsg.), *Proceedings of the of the 14th Conference on Natural Language Processing* (S. 110–120). [https://konvens.org/proceedings/2018/PDF/konvens18\\_13.pdf](https://konvens.org/proceedings/2018/PDF/konvens18_13.pdf).



## Deliverable 5.2: Sentimentanalyse

16.06.2022

In diesem Deliverable wird das Sentiment und deren Ausbreitung in Twitterkonversationen betrachtet. Das Sentiment eines Kommentars spiegelt die Stimmung des Autors zum Zeitpunkt, zu dem der Kommentar geschrieben wurde, wider. Es wird auf einer Skala von -1 (maximal negativ) bis +1 (maximal positiv) gemessen, null ist demzufolge neutral. In Abbildung 1 wurden die Kommentare (Knoten) einer Konversation abhängig vom Sentiment eingefärbt. Daran ist die Verbreitung negativen Sentiments gut nachvollziehbar. Alle betrachteten Kommentare wurden von den Hilfskräften manuell annotiert.

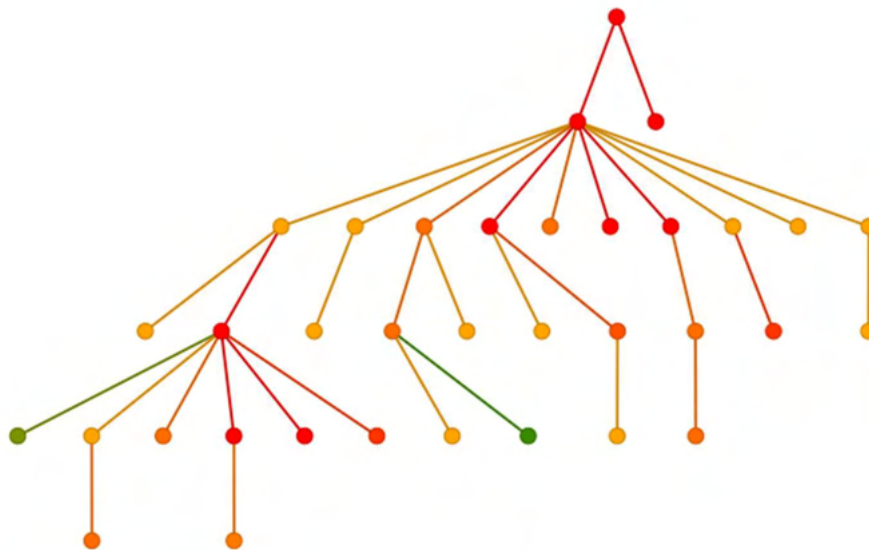


Abbildung 1: Darstellung des Sentiments der Kommentare einer ausgewählten Twitterkonversation (rot = negativ, gelb = neutral, grün = positiv)

Zur Analyse wie sich negatives Sentiment ausbreitet, wurde das Sentiment zum einen in zufällig ausgewählten Kommentaren und zum anderen in Antworten auf toxische Kommentare separat betrachtet (Abbildung 2, Tabelle 1). Tabelle 1 zeigt, dass das mittlere Sentiment in beiden Klassen überraschenderweise nahezu gleich ist, auch der Anteil an Kommentaren mit negativem Sentiment ( $< -0.5$ ) weicht nur geringfügig ab und beträgt in beiden Klassen etwas über 50 %.



	Anzahl negatives Sentiment	Anteil negatives Sentiment	Mittelwert Sentiment
Zufällige Kommentarauswahl	896	53,3 %	-0.48
Antworten auf toxische Kommentare	1919	52,2 %	-0.47
Gesamt	2815	52,7 %	-0,47

Tabelle 1: Verteilung des Sentiments in einer zufälligen Kommentarauswahl aus Konversationen und in Antworten auf toxische Kommentare.

In Abbildung 2 werden dennoch Unterschiede zwischen beiden Klassen sichtbar. Dort ist die genaue Verteilung dargestellt, je breiter die Fläche auf einer bestimmten Höhe ist, desto mehr Kommentare haben diesen Sentimentwert. In der zufälligen Kommentarauswahl ist die Verteilung der Kommentare mit einem Sentimentwert von unter null ausgeglichen, eine leichte Häufung gibt es bei -1. Kommentare mit positivem Sentiment kommen selten vor, sind aber auch vorhanden. Im Gegensatz dazu ist die Verteilung in den Antworten auf toxische Kommentare sehr unausgeglichen: Es gibt deutliche Häufungen bei den Sentimentwerten null und -1. Werte über null kommen nahezu gar nicht vor. Das deutet darauf hin, dass auf toxische Antworten vorzugsweise Kommentare mit stark negativem Sentiment (der Autor fühlte sich angegriffen) oder aber neutrale Kommentare folgen. Neutrale Kommentare könnten von Nutzern kommen, die probieren deeskalierend in eine Diskussion einzugreifen.

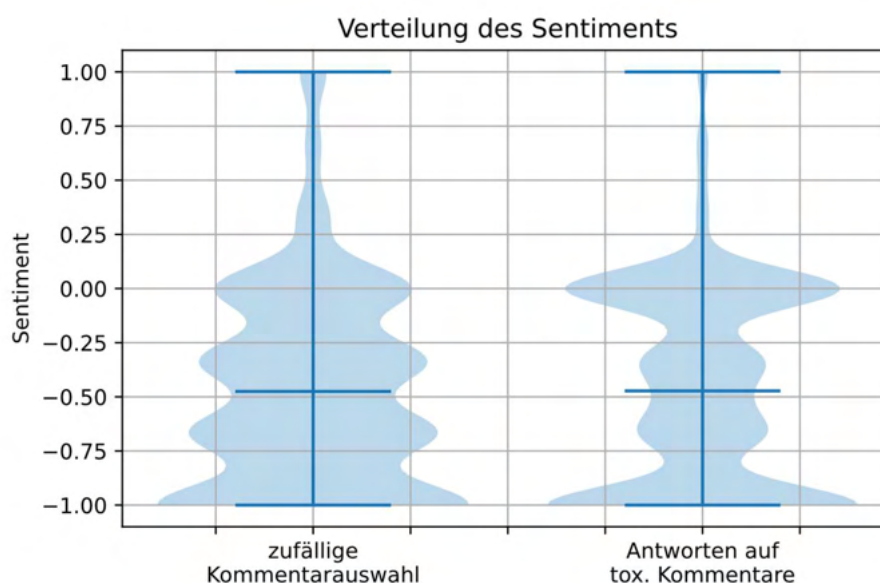


Abbildung 2: Plots der Sentimentverteilung in einer zufälligen Kommentarauswahl aus Konversationen und in Antworten auf toxische Kommentare. Die horizontalen Linien kennzeichnen Maximum, Mittelwert und Minimum.