

# Illegale Hassrede: Ein Datensatz für die automatische Erkennung KI gegen Online-Hass II (2021)

Johannes Schäfer und Kübra Boguslu

Projekt HASeKI  
Hate Speech und seine Erkennung durch KI

Institut für Informationswissenschaft  
und Sprachtechnologie  
Universität Hildesheim



12. November 2021

# Ziel: Automatische Erkennung von illegaler Hassrede

## Ausgangslage

- Methode **maschinelles Lernen**:  
automatisches System lernt Gewichte für Merkmale und deren Kombinationen anhand von Trainingsdaten
- **Trainingsdatensatz** als Basis zentral
- Bisher verfügbare Datensätze in der Forschung,  
z.B. Germeval Task 2 (2019) für beleidigende Sprache
- Hier Fokus auf **illegale Hassrede**  
→ nachvollziehbare, klar abgrenzbare Kategorisierung

1	Illegale Hassrede - Definitionsversuche	4
2	Annotationsrichtlinien	9
3	Datensammlung	13
4	Ausblick	16


1	Illegale Hassrede - Definitionsversuche	4
2	Annotationsrichtlinien	9
3	Datensammlung	13
4	Ausblick	16


# Vorgehen

## Definition empirisch, aus Beispielen: Ausgangspunkt Gerichtprotokolle

- Justizportale: Nordrhein-Westfalen, Niedersachsen etc.
- Suche nach bestimmten Schlagwörtern
- Dokumentation der gefundenen Protokolle durch:
  - Suchbegriff
  - Tatbestand
  - Paragraph
  - Zitat(e)
  - Urteil und Begründung

# Beispiel: Justizportal NRW

Gericht:  
Alle Gerichtsarten 


Gerichtsort:  
Alle Gerichtsorte 


Entscheidungsdatum:



Aktenzeichen / ECLI:



Suche im Volltext:  
Beleidigung

exakte Wortsuche
  Wortstammsuche
  abgeleitete Suche

Ergebnisse pro Seite:  
10 

Ergebnisse sortiert nach:  
Relevanz 

Suchen  Hilfe 

Neue Suche  Erweiterte Suche 

## Suche

- Stark begrenzte Suchmöglichkeit
- Suche nur nach einem Begriff
- Keine direkte Suche nach Paragrafen

→ Aufbau und Funktion verschiedener Justizportale unterscheiden sich

# Beispiel: Justizportal NRW

In der Rechtsprechungsdatenbank NRWE (NRWEntscheidungen; www.nrwe.de) stehen Ihnen die Entscheidungen der Gerichte in Nordrhein-Westfalen im Volltext zur Verfügung. Bitte beachten Sie unsere Hinweise, insbesondere:

- zur **möglicherweise kostenpflichtigen Nutzung**
- zur **Benutzung der Suchfunktion in NRWE**.

Es wurden **860** Dokumente zu Ihrer Suche gefunden.



## 2 Ss 220/09 - 85 - OLG Hamm - Oberlandesgericht Hamm

Gericht: Oberlandesgericht Hamm

Entscheidungsart: Beschluss

Aktenzeichen: 2 Ss 220/09 - 85 - OLG Hamm

ECLI:DE:OLGHAM:2010:0506.2SS220.09.85OLG.H.00

Entscheidungsdatum: 06.05.2010

## 12 Sa 1651/05 - Landesarbeitsgericht Köln

Gericht: Landesarbeitsgericht Köln

Entscheidungsart: Urteil

Aktenzeichen: 12 Sa 1651/05

ECLI:DE:LAGK:2006:0331.12SA1651.05.00

Entscheidungsdatum: 31.03.2006

Leitsätze:

Kein Leitsatz

## Ergebnisse

- Hohe Anzahl der Treffer
  - Gerichtsprotokolle sehr ausführlich dokumentiert
  - Keine einheitliche Gliederung der Protokolle
- Manuelle Durchsicht erforderlich
- ⇒ Ergebnisse entsprechen meist nicht unseren Anforderungen

# Hassrede: relevante Gesetze

## Situation in Deutschland

- Kein Gesetz speziell für Online-Hassrede
- Relevante Paragraphen des StGB:
  - Öffentliche Aufforderung zu Straftaten § 111 StGB
  - Volksverhetzung § 130 StGB
  - Beleidigung § 185 StGB
  - Üble Nachrede § 186 StGB
  - Verleumdung § 187 StGB
  - Bedrohung § 241 StGB



1	Illegale Hassrede - Definitionsversuche	4
2	<b>Annotationsrichtlinien</b>	<b>9</b>
3	Datensammlung	13
4	Ausblick	16

# Annotationsrichtlinien für illegale Hassrede

Boguslu und Schäfer (2021)

## Aufbau

- Annotationlabels
- Entscheidungshilfe
- Beispiele

## Anwendung

- Als Basis zur manuellen Annotation neuer Daten
- Zur einheitlichen, nachvollziehbaren Markierung mehrerer Annotator:innen

# Annotationslabels mit Beispielen

## Hate Speech

- (A) Öffentlicher Aufforderung zu Straftaten
- (V) Volksverhetzung
- (N) Üble Nachrede und Verleumdung
- (B) Beleidigung

## (OL) Offensive Language

## Other

# Annotationslabels mit Beispielen

## Hate Speech

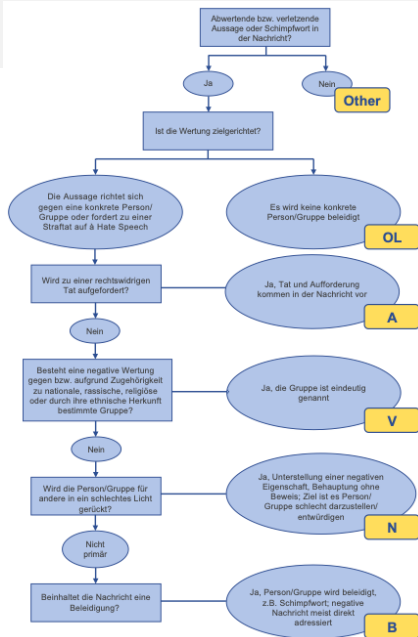
- (A) Öffentlicher Aufforderung zu Straftaten -  
„Meine lieben Migranten, stecht doch bitte die Politiker ab die euch hier rein gelassen haben“
- (V) Volksverhetzung -  
„(...) bist muslim darfst hier gar nicht chatten“
- (N) Üble Nachrede und Verleumdung -  
„(...) die Bundesregierung so wie die Justiz ist so was von Korrupt und arbeiten so fleißig zusammen (...).“
- (B) Beleidigung -  
„Sie ist halt ein kriminelles Dreckstück“

## (OL) Offensive Language

„(...). War eher so ne Scheiss-Schnee-Aktion kurz bevor ich wieder zu Hause war.“

## Other

# Entscheidungshilfe



# Übersicht

1	Illegale Hassrede - Definitionsversuche	4
2	Annotationsrichtlinien	9
<b>3</b>	<b>Datensammlung</b>	<b>13</b>
4	Ausblick	16

# Sammlung von Daten in sozialen Medien

## Sammlung von Twitter-Daten

- Verwendung des offiziellen Twitter-APIs
- Suche mit einzelnen **Suchbegriffen**  
(Ziel: Sammlung von potentiellen Hassbotschaften)
- Problem: **Bias**

# Sammlung von Daten in sozialen Medien

## Sammlung von Twitter-Daten

- Verwendung des offiziellen Twitter-APIs
  - Suche mit einzelnen **Suchbegriffen**  
(Ziel: Sammlung von potentiellen Hassbotschaften)
  - Problem: **Bias**
- Sorgfältige Auswahl von **mehreren** Suchbegriffen (Varianz):
- Suche nach Schlüsselwörtern, die mehrdeutig verwendet werden können  
(für Hass- oder Nicht-Hassbotschaften)
  - z. B. “Schwein” oder “Polizei”



# Annotationsergebnisse

## Annotationsexperiment

- Vorauswahl von 100 Beiträgen (Tweets), gemischte Verteilung vermutet
- Individuelle Annotation von 5 Expert:innen (Forscher:innen zum Thema Hassrede)

# Annotationsergebnisse

## Annotationsexperiment

- Vorauswahl von 100 Beiträgen (Tweets), gemischte Verteilung vermutet
- Individuelle Annotation von 5 Expert:innen (Forscher:innen zum Thema Hassrede)
- **Auswertung Mehrheitslabels**

A	0
V	6
N	16
B	19
OL	31
Other	28
<hr/>	
Total	100
<hr/>	

# Annotationsergebnisse

## Annotationsexperiment

- Vorauswahl von 100 Beiträgen (Tweets), gemischte Verteilung vermutet
- Individuelle Annotation von 5 Expert:innen (Forscher:innen zum Thema Hassrede)
- Auswertung Mehrheitslabels
- **Auswertung Fleiss' Kappa: 0.50 (moderate agreement)**

A	0
V	6
N	16
B	19
OL	31
Other	28
<b>Total</b>	<b>100</b>

# Annotationsergebnisse

## Annotationsexperiment

- Vorauswahl von 100 Beiträgen (Tweets), gemischte Verteilung vermutet
- Individuelle Annotation von 5 Expert:innen (Forscher:innen zum Thema Hassrede)
- Auswertung Mehrheitslabels
- Auswertung Fleiss' Kappa: 0.50 (moderate agreement)
- **Auswertung paarweise: Cohen's Kappa**

	A1	A2	A3	A4	A5
A1	1	0.28	0.33	0.39	0.38
A2		1	0.59	0.35	0.42
A3			1	0.41	0.43
A4				1	0.30
A5					1

# Übersicht

- |   |   |           |
|---|---|-----------|
| 1 | Illegale Hassrede - Definitionsversuche | 4         |
| 2 | Annotationsrichtlinien                  | 9         |
| 3 | Datensammlung                           | 13        |
| 4 | <b>Ausblick</b>                         | <b>16</b> |

# Weiteres Vorgehen

## Anwendung der Annotationsrichtlinien

- Sammlung weiterer Beispiele; Versuch mit Fokus auf Label A
- Annotation eines größeren Datensatzes

## Weiterverarbeitung der annotierten Daten

- Als Trainingsmaterial für automatische Systeme
- Problematik bei Veröffentlichung: Anonymisierung
- Problematik Bias: Eigennamen  
Beispiel: „@USER1 @USER2 bist muslim darfst hier gar nicht chatten“

- Germeval Task 2 (2019) — Shared Task on the Identification of Offensive Language. <https://projects.fzai.h-da.de/iggsa/data-2019/>
- Johannes Schäfer und Kübra Boguslu. 2021. Towards annotating illegal hate speech: A computational linguistic approach. Detect Then Act (DTCT) Technical Report 3. ISSN 2736-6391.  
<https://dtct.eu/wp-content/uploads/2021/10/DTCT-TR3-CL.pdf>
- Kübra Boguslu und Johannes Schäfer. 2021. Annotationsrichtlinien für illegale Hassrede.  
[https://dtct.eu/wp-content/uploads/2021/09/Annotationsrichtlinien\\_iHS.pdf](https://dtct.eu/wp-content/uploads/2021/09/Annotationsrichtlinien_iHS.pdf)
- Projekt HASeKI: “Das Phänomen Hate Speech und seine Erkennung durch KI: interdisziplinär – international – erklärbar? (HASeKI)”  
<https://www.uni-hildesheim.de/fb3/institute/iwist/forschung/forschungsprojekte/aktuelle-forschungsprojekte/haseki/>