# Us and Them

## Identifying cyber hate on Twitter across multiple protected characteristics

Herausforderungen des Natural Language Processing:
Hate Speech, diskriminierende Sprache und populistische Rhetorik
SoSe 2024

Alex Froitzheim, 7340859

- Übersicht
- Intersektionalität
- Forschungsstand
- Daten/Annotation
- Features
- Aufbau der Experimente
- Ergebnisse

# Übersicht

- Identifikation von Cyber-Hass gegenüber Gruppen oder Individuen auf Basis von Ethnizität, sexueller Orientierung oder Behinderung
  - Aufbauend auf Modell zur Identifikation von religiös motiviertem Cyber-Hass [1]
- Klassifikation auf Basis lexikalischer und syntaktischer Features

- Welche Auswirkungen hat die Auswahl unterschiedlicher Feature-Sets auf die Klassifikation?
- Lässt sich durch die Kombination von Daten zu mehreren Diskriminierungsformen intersektionale Diskriminierung erkennen?

# Intersektionalität

- Beschreibt, wie durch Überschneidung mehrerer Formen von Diskriminierung weitere, eigenständige Diskriminierungsformen entstehen
- **Konkret:** Die Diskriminierungserfahrungen einer Person, die rassistische und sexistische Diskriminierung erfährt sind mehr als die Summe beider Diskriminierungsformen für sich
- **Beispiel:** Aufgrund der Kombination von Geschlecht (männlich), sozialem Milieu und ethnischer Herkunft wurden die Opfer der NSU-Morde verdächtigt, selbst an kriminellen Aktivitäten beteiligt gewesen zu sein [2].

# Forschungsstand

- Vor der Entwicklung von Transformer-Modellen mit Attention-Mechanismus [3]
  ➜ Beziehungen zwischen entfernten Wörtern schwer greifbar
- Ähnliche Arbeiten:
  - Grundsätzlich relativ kleine Datensätze (< 10.000 Samples) [4]
  - Eher klassische linguistische Features, z.B.  BoW, n-Gramme, POS-Tags... [5]

# Daten

- Pro betrachteter Diskriminierungsform ein Trigger-Event
  - Rassismus: Wiederwahl von Barack Obama
  - Homophobie: Coming-Out des Basketballers Jason Collins
  - Ableismus: Eröffnungszeremonie der Paralympics in London im Jahr 2012
- Daten gesammelt von Twitter im Zeitraum von zwei Wochen nach den Trigger-Events auf Basis von Keyword-Suche
  - Relevante Named-Entities, z.B. „Obama”, „Paralympic”
- Pro Trigger-Event 2000 Ergebnisse zufällig für die Annotation ausgewählt

# Annotation

- Crowdsourcing der Annotator:innen über die Plattform CrowdFlower
  - Keine Garantie, dass alle Beispiele von den gleichen Annotator:innen annotiert werden
  - Mindesten vier Annotator:innen pro Beispiel
- Minimalistische Annotationsrichtlinien
  - „Is this text offensive or antagonistic in terms of race ethnicity/sexual orientation/disability?”
  - Label: *yes, no, undecided*
- Beispiele mit Agreement unter 75 % oder mit *undecided*-Labels (?) nicht berücksichtigt

# Annotation

**Ergebnisse:**

Datensatz *Homophobie*: 1803 Tweets, 183 positiv (10,15 %)

Datensatz *Rassismus*: 1876 Tweets, 70 positiv (3,73 %)

Datensatz *Ableismus*: 1914 Tweets, 51 positiv (2,66 %)

# Features

- **Bag of Words/n-Gramme:**
  - Stemming und Umwandlung in Lowercase der Wörter
  - Extraktion von n-Grammen (1-5 Tokens)
  - 2000 Features ausgewählt (?), Normalisierung an jedem Feature-Vektor durchgeführt
- **Themenspezifische Wortlisten**
  - Listen von Slang-Wörtern aus Wikipedia übernommen
- **Typed Dependencies**
  - Set der Beziehungen zwischen Wörtern in einem Satz mit Art der Beziehung
  - Erstellt mit Stanford Lexical Parser
  - Output in Lowercase umgewandelt und n-Gramme (1-3 Tokens) extrahiert
  - Ebenfalls Auswahl von 2000 Features und Normalisierung an jedem Vektor

# Features

**Typed Dependencies: Beispiel**

**Input:** „Send them all back home.”

**Output:** [root(ROOT-0, Send-1), nsubj(home-5, them-2), det(home-5, all-3), amod(home-5, back-4), xcomp(Send-1, home-5)]

# Aufbau der Experimente

**Modelle:** Support Vector Machine, Random Forest Classifier

**Experimente:**

1. Binäre Klassifikatoren für jeden Datensatz mit unterschiedlichen Kombinationen der Feature-Sets
2. Evaluation binärer Klassifikatoren an Test-Sets der jeweils anderen Datensätze
3. Binärer Klassifikator für gemischten Datensatz (6486 Tweets, 395 Positiv) mit allen Features
4. Klassifikation mit sechs Klassen (zwei pro Datensatz) für gemischten Datensatz mit allen Features

# Ergebnisse

- Precision generell höher als Recall
  - vielleicht wegen ungleicher Verteilung der Klassen?
- Typed Dependencies verbessern Performance bei zwei von drei Arten von Diskriminierung (Rassismus und Homophobie)
- Portabilität von Modellen zwischen Datensätzen ist sehr gering
- Gemischter Datensatz führt zu besserem Recall und F-Measure
  - potentiell wegen Erkennung intersektionaler Diskriminierung

# Literatur

**Burnap, P., & Williams, M.L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* 5, 11. https://doi.org/10.1140/epjds/s13688-016-0072-6**

[1] Burnap, P., & Williams, M.L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 223-242.

[2] Adusei-Poku, Nana (2012). Intersektionalität: „E.T. nach Hause telefonieren“?. *APuZ*, 62. Jg., Nr. 16-17, 47-52. Abgerufen am 16.06.2024. https://www.bpb.de/apuz/130420/intersektionalitaet-e-t-nach-hause-telefonieren.

[3] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Neural Information Processing Systems*.

[4] Schmidt, Anna & Wiegand, Michael. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In Ku, Lun-Wei & Li, Cheng-Te (Eds.), Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1–10). Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/W17-1101. DOI: 10.18653/v1/W17-1101.

[5] Jahan, Md Saroar & Oussalah, Mourad. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232, ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2023.126232.