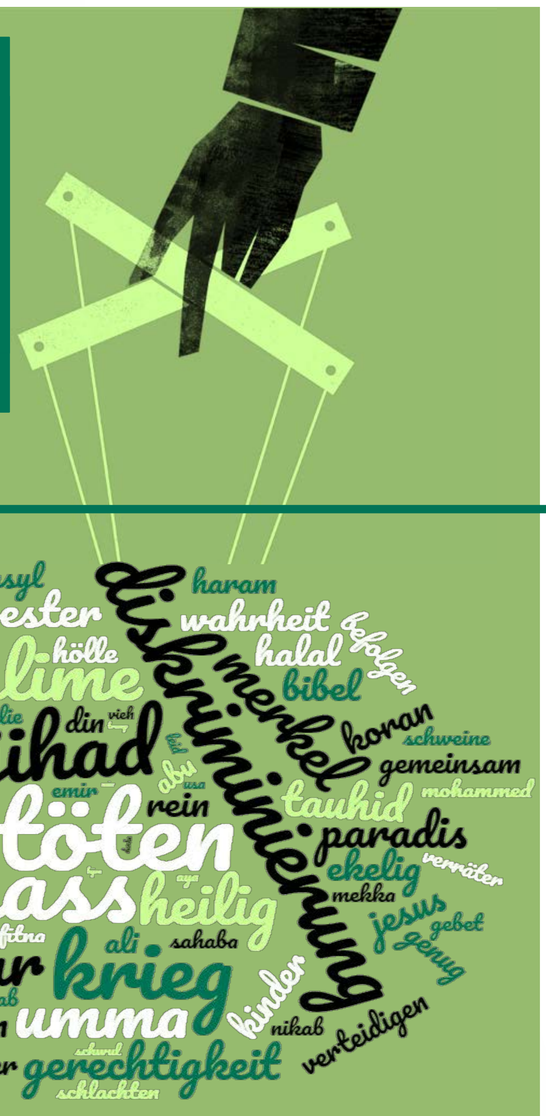
IKG Institut für interdisziplinäre Konflikt- und Gewaltforschung

UNIVERSITÄT BIELEFELD

‚Topic Modelling' von Da'wa & Co.

Semi-automatisierte Ansätze zur Erkennung radikalisierender Sprache im Internet

Annika Hamachers



1 Das Projekt X-SONAR

Das Verbundprojekt „X-SONAR: Extremistische Bestrebungen in Social Media Netzwerken: Identifikation, Analyse und Management von Radikalisierungsprozessen" leistet praxisorientierte, interdisziplinäre **Grundlagenforschung** zum Verständnis **extremistischer Interaktions- und Eskalationsdynamiken** in sozialen **Onlinenetzwerken**. X-SONAR erforscht die Mechanismen der individuellen und kollektiven Gewaltdynamiken sowie die Selbstregulation von Radikalität in sozialen Online-Netzwerken.

Den Forschungsverbund bilden

IKG Universität Bielefeld | LKA Niedersachsen | DFKI German Research Center for Artificial Intelligence | I:P:Bm Institut Psychologie und Bedrohungsmanagement | Deutsche Hochschule der Polizei | Landesinstitut für Präventives Handeln SAARLAND | Fraunhofer SIT

- Projektförderer: das Bundesministerium für Bildung und Forschung (**BMBF**) innerhalb des Programms „Forschung für die zivile Sicherheit 2012 – 2017"
- Förderzeitraum: März **2017** bis Februar **2020**

2 Teilprojektziel: Entwicklung eines Radikalisierungsdiktionärs

Ausgangsüberlegung:

> The **words** we use in daily life **reflect** what we are paying attention to, what we are thinking about, what we are trying to avoid, how we are feeling, and **how we are organizing and analysing our worlds**. (Pennebaker, 2010)

→ Analysen (sog. Word Count Analyses) der Wörter in einem Text lassen Rückschlusse auf Persönlichkeit, Gedanken und Absichten des Verfassers zu
→ i.d.R. erfolgt dabei der Abgleich mit sog. ‚**Diktionären**' (im Vorfeld erarbeiteten Wortlisten für verschiedene – oft psychometrische – Kategorien, z.B. Wut)
→ diktionärsgestützte WC-Verfahren wurden bereits erfolgreich auf radikale Inhalte angewandt (z.B. Chalothorn & Ellmann, 2013; Cohen et al., 2016)

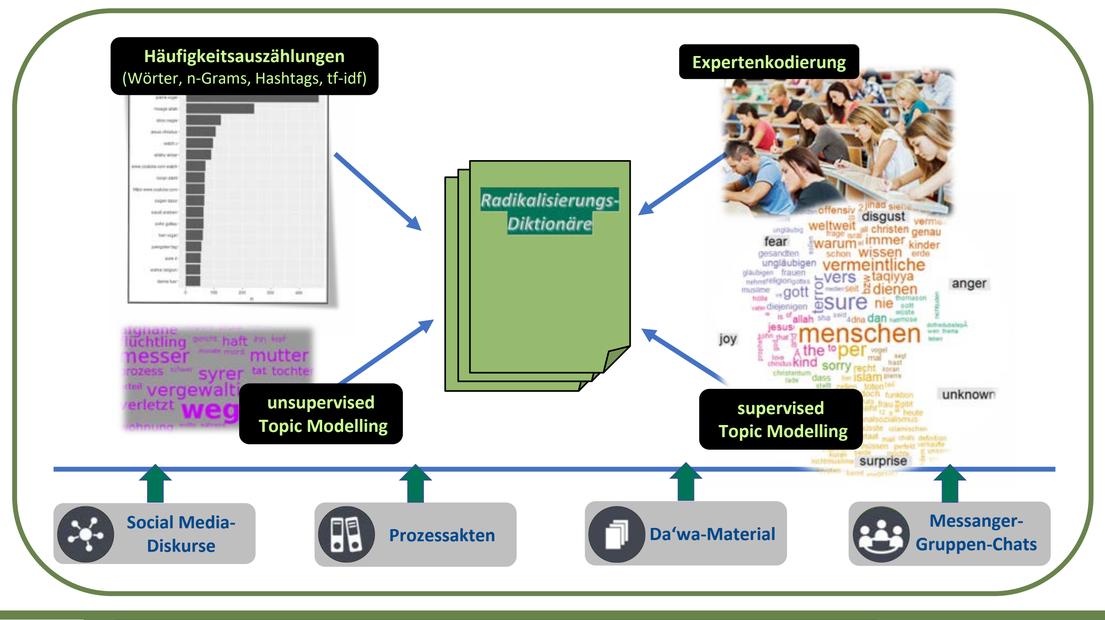
das große ‚Aber'

⛔ bestehende Diktionäre sind häufig nicht valide (geringe Trefferrate)
⛔ liegen für spannende Kategorien oft nur in englischer Sprache vor
⛔ sind nicht ‚domänspezifisch' für Radikalisierung
⛔ sind nicht ‚domänspezifisch' für Sprache im Internet/in Social Media

→Ziel ist es, **theoriegeleitet** bestehende Wortlisten zu ergänzen und eigene Listen zu erarbeiten, um so ein Wörterbuch zu bedrohungsanalytisch relevanten Kategorien und spezifischen Radikalisierungsindikatoren auf den Plattformen **Facebook**, **Twitter** und **YouTube** zu entwickeln

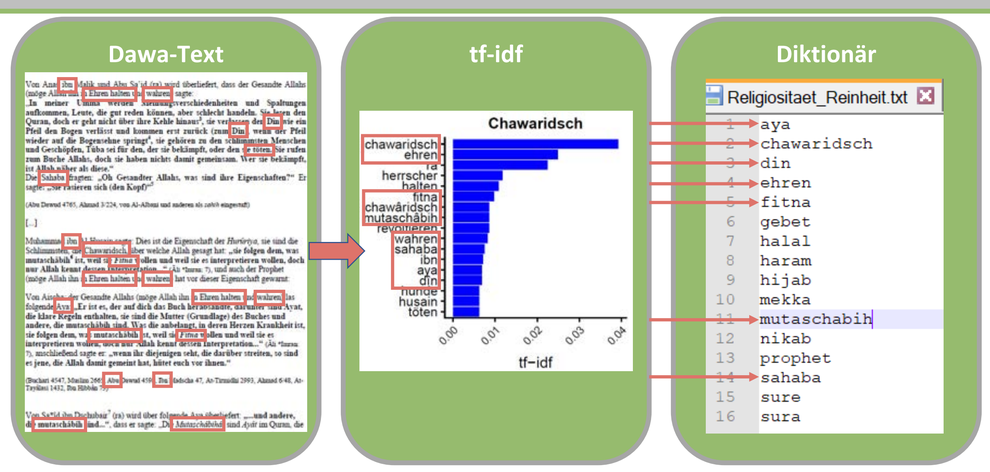
3 Methode & Datenmaterial

- explorativer, **quantitativ-qualitativer Ansatz** zur Modellierung radikal-islamistischer **Themen** (‚topics') und der jeweiligen **Schlüsselbegriffe**
- kombiniert
  - Expertenkodierungen
  - Wort-Häufigkeitsauszählungen
    - n-Grams (Analyse häufiger Wortketten)
    - tf-idf-Analysen (= ‚term frequency – inverse document frequency')
  - sog. *überwachte* Lernalgorithmen (mit gelabelten Trainingsdatensätzen)
    - Naive Bayes-Klassifikator
  - sog. *unüberwachte* Lernalgorithmen (ohne gelabelte Daten)
    - Latent Dirichlet Allocation (LDA) für längere Texte
    - Fuzzy c-means für extrem kurze Texte (z.B. Twitter-Posts)
- wendet diese Methoden nicht nur auf Diskurse aus den Zielplattformen Twitter, YouTube und Facebook an, sondern
- recherchiert und erfasst zusätzliche, potenziell einschlägige Materialen, wie
  - Kommunikationsprotokolle aus **Gerichtsakten** und
  - sog. **Da'wa-Material** (*hier*: islamistische Missionierungs-Schriften)



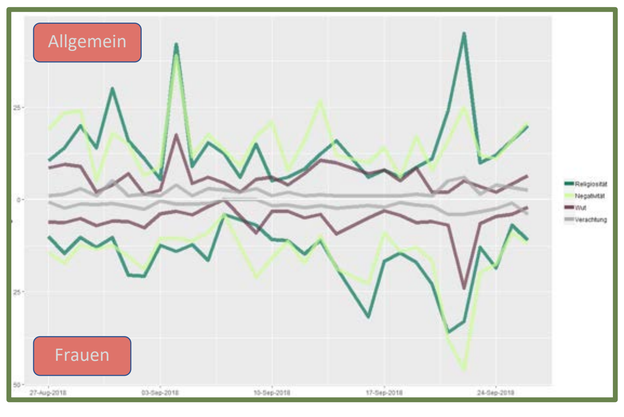
4 Beispiel & weitere Anwendung

Exemplarischer Workaround:
Wortfrequenzanalyse innerhalb eines Da'wa-Textes mittels des tf-idf-Verfahrens (setzt die Häufigkeit der Wörter innerhalb dieses Textes ins Verhältnis zu ihrer allgemeinen Häufigkeit und ermittelt so, welche Begriffe spezifisch für diesen Text sind); anschließend: Entscheidung darüber, welche der gefundenen Begriffe für die jeweiligen Diktionärs-Kategorien relevant sind (hier: Zuordnung zum Diktionär für ‚Religiöse Reinheitsvorstellungen')



Domänspezifische Wortlisten können anschließend nicht nur zur **Früherkennung** radikaler Tendenzen im Netz genutzt werden, sondern – eingebunden in weitere Analysen – auch dazu beitragen, die **Struktur** islamistischer Agitation zu **beschreiben**, besser zu **verstehen** und **Veränderungen nachzuzeichnen**.
So zeigt sich bislang u.a.

- dass vor allem für Frauen erstellte Materialien stark negativ aufgeladen sind,
- dass Themenpräferenzen im Zeitverlauf variieren oder
- dass islamistische Propaganda überraschend wenig Hassrede (Verachtung) beinhaltet.



GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

aktuelle Informationen zum Projekt:

→ www.uni-bielefeld.de