gesis Leibniz Institute for the Social Sciences

Social-Media und Text-Mining mit R

Veronika Batzdorfer

Teil IV

Ausblick Social-Web-Data Sampling und Ethik, 29.09.2022

Member of Leibniz Association

gesis Leibniz Institute for the Social Sciences

Schedule

	Uhrzeit	Inhalt
	09:00 - 10:30	Konzepte & Herausforderungen bei der Analyse von Social-Web-Daten (Twitter-API)
	10:30 - 11:00	*Kaffeepause*
	11:30 - 12:30	Getting Started mit Twitterdaten: (i) Sampling, (ii) Pre-Processing & (iii) Grundlagen der Textanalyse (Häufigkeiten, Co-Occurences, Netzwerke)
	12:30 - 13:30	*Mittagspause*
	13:30 - 15:00	Twitter Demo & Exkurs Crawling Social-Web-Data
	15:00 - 15:30	*Kaffeepause*
*	15:30 - 17:00	Ausblick: Social-Web-Data-Collection & Fortgeschrittene NLP-Techniken; Bias und Ethik im NLP

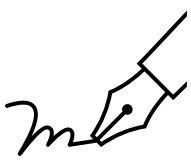
Member of Leibniz Association

Dictionary/ Regelbasierte Ansätze

- Ein strenger Satz von Regeln, wie ein Computer eine Messung durchführen sollte.
- Codebuch für den Computer

Z.B. für **Sentimentanalyse**:

- R-Package: *Syuzhet* basiert auf Romanen, errechnet einen Mittelwert über die Wörter des Textes
  - beinhaltet auch *afinn* dictionary
- R-package: ***tidytext***
  - beinhaltet „bing“, „nrc“ (Plutchik’s wheel emotions)
  - Oder custom dictionary wie *NRC-VAD*
- R-package: *VADER* (unsupervised Sentimentanalyse für Social Media Texte)



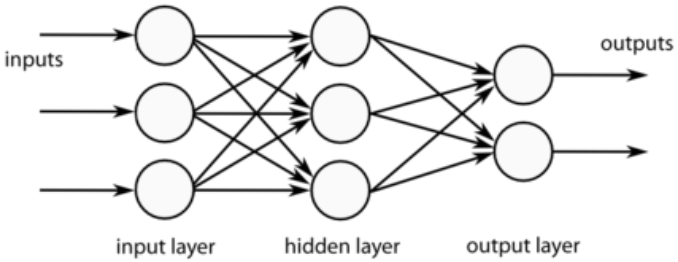
Supervised Machine Learning

Auf der Grundlage von annotierten Texten (z.B. manuell kodierten durch Crowd-Worker) ein Machine-Learning-Modell trainieren, das neue Texte annotieren kann.

Z.B.:

- Naive Bayes
- Decision trees
- Support Vector Machines

Neuronale Netze



+

- Ermöglicht nicht-lineare Klassifizierung
- Kann mit großen Anzahl von Daten trainiert werden
- Können die Wortfolge/Syntax berücksichtigen, ohne n-Gramme zu verwenden.
- Fortgeschrittene Modelle (deep, konvolutional und/oder rekursive neuronale Netze)

⊖

- Training erfordert viel Rechenleistung
- Modellgewichte sind so gut wie uninterpretierbar
- Für einige Aufgaben funktionieren einfachere Modelle (fast) genauso gut

Bildquelle: https://de.wikipedia.org/wiki/Deep_Learning

Un-Supervised Machine Learning

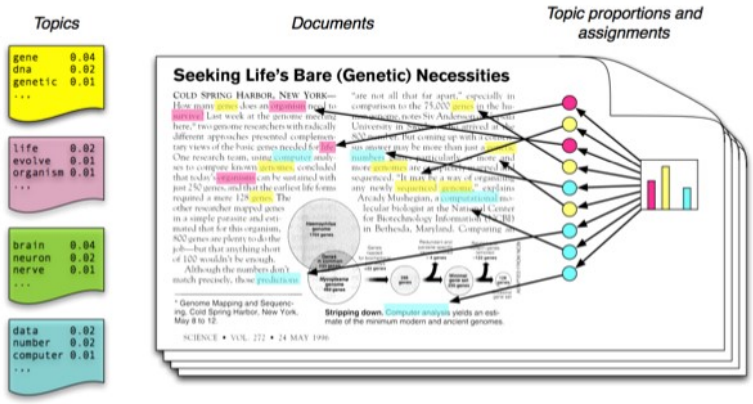


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

- Das beliebteste Modell ist die *Latent Dirichlet Allocation* (**LDA**)
- *Probabilistische Topic Models*
- Neben der reinen „Vanilla“ LDA gibt es viele Varianten:
  - dynamische Topic Models (rolling LDA)
  - Structural Topic Models (R-Package: *stm*) (LDA + metadata)
  - BERTopic

a) jedem Wort wird ein Thema zugewiesen
b) Dokumente können Mischungen von Themen enthalten
c) dasselbe Wort kann in verschiedenen Dokumenten unterschiedliche Themen zugeordnet sein

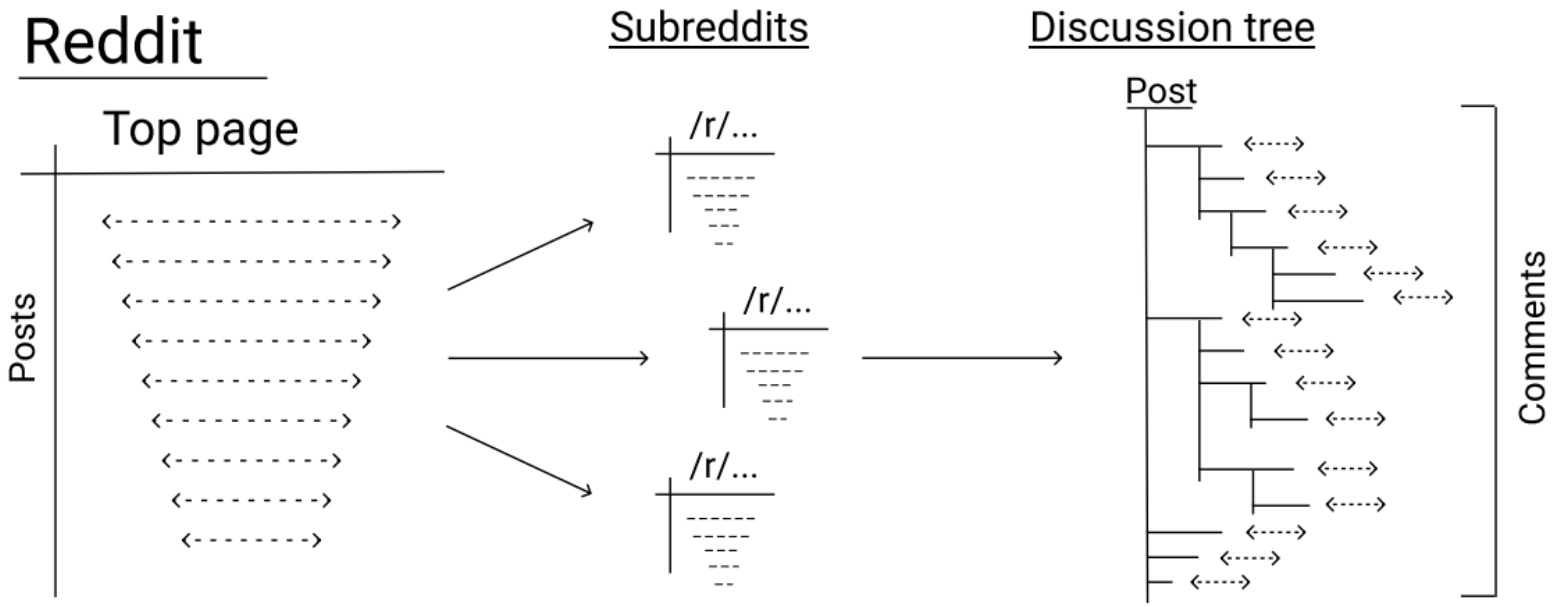
Social-Web-Data Erhebung

Data Donations:

- ***Browser-plug-in*** für Facebook-Daten
  Haim, M., & Nienierza, A. (2019)
  - sammelt öffentliche Beiträge (+ einige Metadaten) aus den Feeds der Nutzer

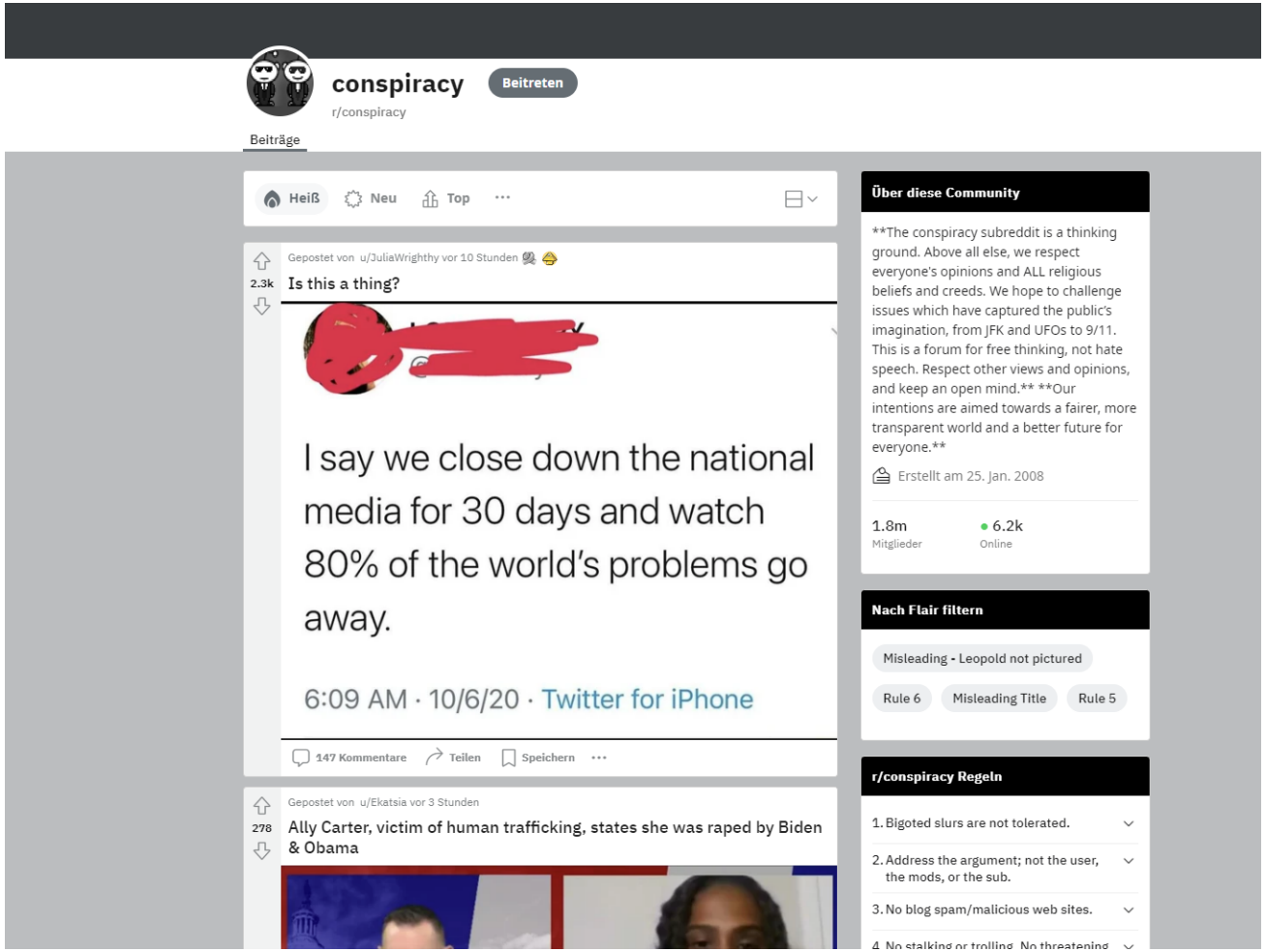
- **+** Informed Consent, Transparenz, Terms of Service Limitationen umgangen
- **-** Anonymisation (z.B. “Freunde”-Tags in Postings), systematischer Bias im Dropout, Aktivitätsqualität, Veränderungen in der Feedstruktur

Social-Web-Data Erhebung



Quelle: Medvedev, Lambiotte, & Delvenne (2020)

Social-Web-Data Erhebung

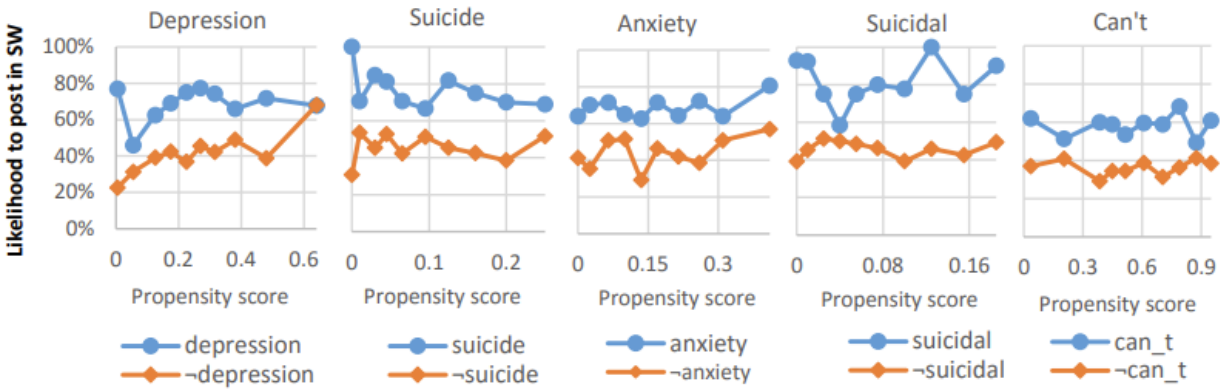


r/conspiracy Screencapture: 09/26/2022

Social-Web-Data Erhebung

{r/SuicideWatch} und Mental Health

- Übergang vom Diskurs über psychische Gesundheit zu Selbstmordgedanken



Quelle: De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016, May). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2098-2110).

Social-Web-Data Erhebung

{r/}

- Modellierung der Posting-Strategien der Nutzer im Hinblick auf das Feedback der Gemeinschaft (Das & Lavoie, 2014)
- Experiment zu Herdeneffekt durch sozialen Einfluss (Muchnik et al., 2013)
- Verbot Hate-Speech-Subreddits (Chandrasekharan et al., 2017)

Data Linking

- Deterministisch vs. Probabilistisch
- **Ex-ante** (Daten zusammen erhoben) vs. **Ex-post** (Linking mit bestehenden Daten)
- Individual-Level vs. Aggregate-Level

- A. Social-Web-Data Erhebung -> Rekrutierung Probanden über Platformen
- B. Survey-Erhebung -> Einverständnis Social-Web-Daten zu erheben

Ethik und Open-Science

- Welche ethischen Fragen haben sich in Ihrer Forschung mit Social-Web-Daten ergeben?
- Falls Sie selbst noch nicht mit Social-Web-Daten geforscht haben: Welche ethischen Aspekte/Fragen/Herausforderungen sind Ihnen beim Lesen von Publikationen aufgefallen, in denen diese genutzt wurden?

Ethik und Open-Science

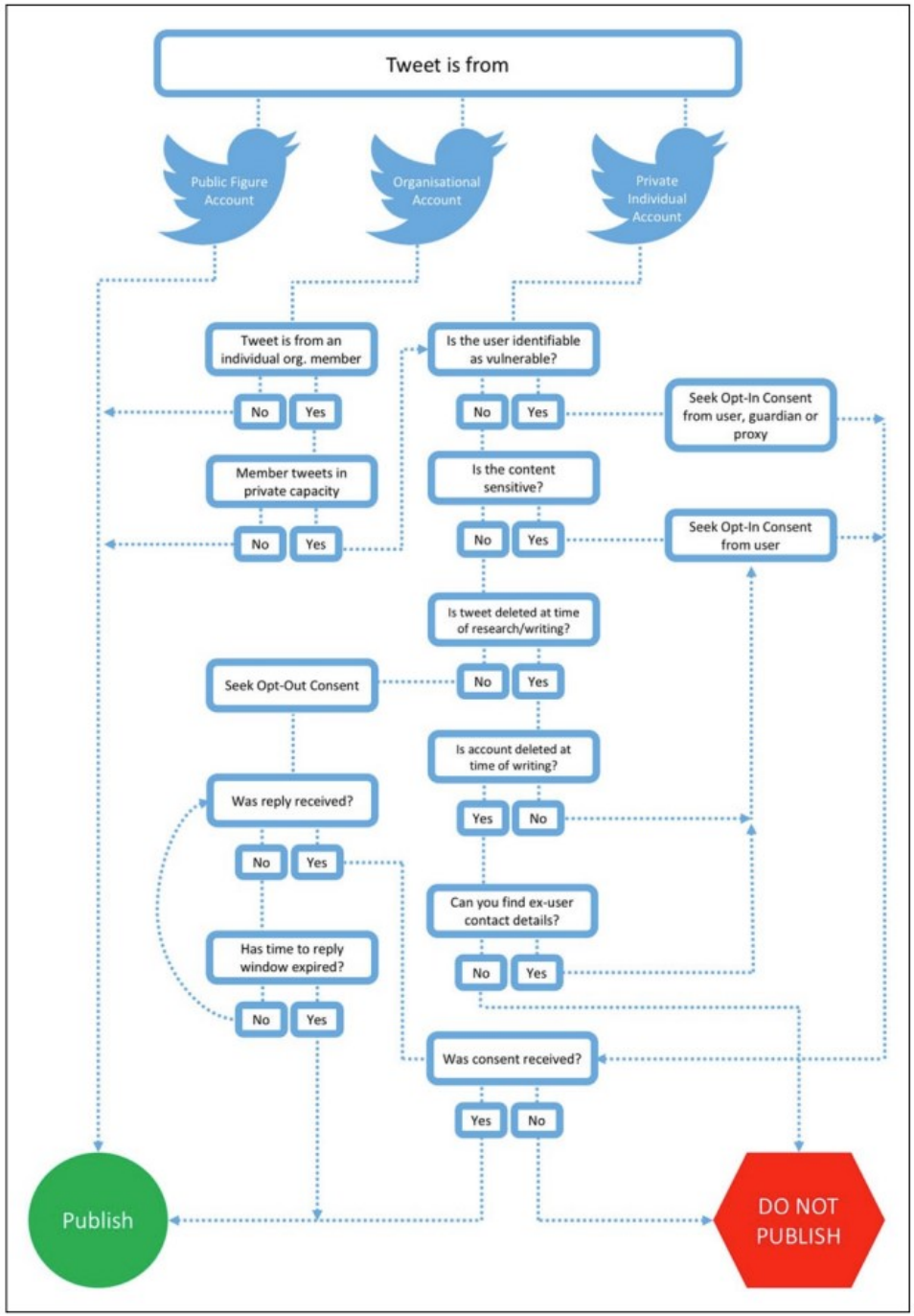
- Minderjährige/ vulnerable Gruppen
- Datenstruktur (Texte, Bilder)
- sensible Themen

Struktur und Inhalt der Daten beeinflussen:

- Anonymisierung/Pseudonymisierung (keine direkten Identifier)
  - Bei Tweets Meta-Daten indirekter Identifier (Lokationsinformationen)
- *Data Minimization*: Möglichst nur erheben, was man tatsächlich benötigt

Welche negativen/unerwünschten Folgen kann die Veröffentlichung von Social-Media-daten haben, wenn Personen in diesen identifizierbar sind?

Ethik: Twitter Decision-Flow



Quelle: Williams et al., 2017