Je nach Zielsetzung sind zwei verschiedene **Textmining-Algorithmen** unterschiedlich gut geeignet wissenschaftliche Textpublikationen automatisiert **nach veröffentlichten Forschungsdaten** zu durchsuchen.

Auffinden von Open Data in Textpublikationen

Ein systematischer Vergleich von Klassifikationsalgorithmen am Beispiel von Publikationen der TU und des Universitätsklinikums Dresden

Katharina Zinke | Open Science Lab
Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB)

HINTERGRUND

Open Data (d.h. veröffentlichte Forschungsdaten) sind ein Grundpfeiler von Open Science. Diese Veröffentlichungen zu finden und Open Data Praktiken zu monitoren wird (u.a. für Institutionen) immer wichtiger

- um Trends / Entwicklungen zu verfolgen
- um Compliance mit Richtlinien zu prüfen
- als Grundlage zur Incentivierung von Open Data Praktiken (Beispiel Charité Berlin)

FORSCHUNGSFRAGE

Ist ein **Text Mining**-basiertes Vorgehen geeignet um **Open Data zu identifizieren**, die mit **wissenschaftlichen Textpublikationen** einer wissenschaftlichen Einrichtung verbunden sind?

VORGEHEN

Als exemplarische **Stichprobe** dienten 137 Publikationen von Autor:innen der Technischen Universität Dresden bzw. des Universitätsklinikums Dresden in PLOS Journals aus den Jahren 2019 – 2022 (generiert aus einer Scopus-Suche und einem Abgleich mit dem PLOS OSI Datensatz).

Identifikation von Open Data (und Open Code) in der Stichprobe von Publikationen:

- Text Mining Algorithmus **ODDPub** (Riedel et al., 2020) des Charité Dashboards on Responsible Research nachgenutzt und auf eigene Stichprobe übertragen
- Kodierung des DataSeer Natural Language Processing Modells für die Publikationen aus dem **PLOS Open Science Indicators Datensatz** (PLOS, 2022) extrahiert
- **Manuelle Kodierung** der Publikationen orientiert an Kriterien des für das Charité Dashboard durchgeführten Screenings (Bobrov et al., 2023; Iarkeva et al., 2022)

ERGEBNISSE

Vergleich Klassifkation ODDPub und DataSeer

- DataSeer identifiziert ca. 20% mehr Publikationen als Open Data als ODDPub (76% vs. 57%).
- Die Klassifikationen stimmen zu 80% überein.
- Bewertung der Klassifikationsergebnisse mit Metriken ergibt ein differenzierteres Bild:
  - Vergleichbare F1-Scores
  - Precision ODDPub › DataSeer
  - Recall ODDPub ‹ DataSeer

= DataSeer findet zwar fast alle tatsächlichen Forschungsdatenveröffentlichungen, identifiziert dafür aber einige fälschlicherweise als Open Data, während ODDPub weniger fälschlicherweise klassifiziert, dafür aber einige tatsächlich Open Data enthaltende Publikationen verpasst.

Abweichungen von der **manuellen Kodierung** gab es für beide Klassifikationsalgorithmen v.a.

- bei Datennachnutzung
- bei fehlerhaften Verlinkungen

SCHLUSSFOLGERUNGEN

Das erprobte Textmining-Verfahren **ODDPub ist geeignet um wissenschaftliche Textpublikationen automatisiert nach Open Data zu durchsuchen**. Dies ermöglicht bei der momentan heterogenen Publikationspraxis von Forschungsdaten (z.T. ohne korrekte Verknüpfung über persistente Identifikatoren) eine systematische Erfassung.

Die konkrete Nutzung der Algorithmen sollte **je nach Zielsetzung abgewogen** werden.

- Möglichst präzise Erfassung → eher ODDPub
- Möglichst vollständige Erfassung → eher DataSeer

Insbesondere bei der Identifizierung von Open Code gibt es Verbesserungspotential bei ODDPub.

In der Zukunft sollten **gesicherte Qualitätskriterien** (standardisierte Beschreibung der Daten, Nutzung von Repositorien, PIDs) für die Klassifikation als Open Data eine größere Rolle spielen

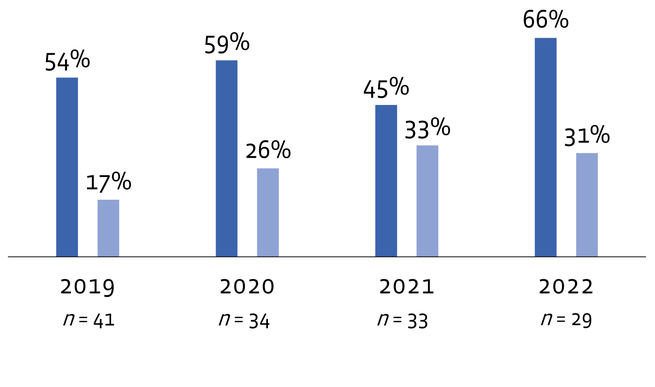
Bewertungsmetriken der ODDPub und DataSeer Klassifikationen für Open Data und Open Code

	Accuracy	F-Score	Precision	Recall	Spezifität
Open Data					
ODDPub	81,8%	0,84	0,83	0,86	0,77
DataSeer	78,1%	0,83	0,72	0,99	0,52
Open Code					
ODDPub	79,6%	0,48	0,72	0,36	0,95
DataSeer	87,6%	0,75	0,79	0,72	0,93

Konfusionsmatrix der manuellen Kodierung und der ODDPub bzw. DataSeer Klassifikationen für Open Data und Open Code (*n* = 137)

		ODDPub			DataSeer (PLOS-OSI)		
Manuelle Kodierung		Ja	Nein	Summe	Ja	Nein	Summe
Open Data	Ja	65	11	76	75	1	76
	Nein	14	47	61	29	32	61
	Summe	79	58		104	33	
Open Code	Ja	13	23	36	26	10	36
	Nein	5	96	101	7	94	101
	Summe	18	119		33	104	

Darstellung der als Open Data und Open Code klassifizierten Publikationen in der Stichprobe über die betrachteten Jahre hinweg



Referenzen:

Bobrov, E., Riedel, N. & Kip, M. (2023, 2. Februar). *Operationalizing Open Data – Formulating verifiable criteria for the openness of datasets mentioned in biomedical research articles.* https://doi.org/10.31222/osf.io/ajh54

Iarkaeva, A., Bobrov, E., Taubitz, J., Carlisle, B. G. & Riedel, N. (2022). Semi-automated extraction of information on open datasets mentioned in articles. *protocols.io.* https://doi.org/10.17504/protocols.io.q26g74p39gwz/v1

Public Library of Science. (2022). *PLOS Open Science Indicators* (Version 2) [Data set]. Figshare. https://doi.org/10.6084/m9.figshare.21687686.v2

Riedel, N., Kip, M., & Bobrov, E. (2020). ODDPub – a Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications. *Data Science Journal, 19*(1), 42. https://doi.org/10.5334/dsj-2020-042

Bilder: undraw.co

SLUB
Wir führen Wissen.

katharina.zinke@slub-dresden.de