Contribution ID: **7** Type: **Poster**

Auffinden von Open Data in Textpublikationen –Ein systematischer Vergleich von Klassifikationsalgorithmen am Beispiel von Publikationen der TU und des Universitätsklinikums Dresden

Thursday 19 October 2023 14:21 (2 minutes)

Einer der Grundpfeiler von Open Science ist der offene Umgang mit Forschungsergebnissen. Das schließt Daten und Software ein, die innerhalb des Forschungsprozesses entstehen. Die Praxis, diese Forschungsdaten zu veröffentlichen (**Open Data**) ist noch nicht in allen Fachdisziplinen gleichermaßen verbreitet. Umso wichtiger ist es für Institutionen oder Förderer diese Open Data Praktiken zu monitoren und Entwicklungen zu verfolgen, aber auch die Compliance der Forschenden mit Richtlinien zum Umgang mit Forschungsdaten zu überprüfen oder zu incentivieren.

Die vorliegende Arbeit beschäftigt sich mit der Möglichkeit m.H. eines Text Mining Algorithmus Open Data Veröffentlichungen zu identifizieren, die mit wissenschaftlichen Textpublikationen einer Einrichtung verbunden sind. Dazu wurde der an der Charité Berlin entwickelte Algorithmus zur Klassifikation von Open Data und Code (**ODDPub**, Riedel et al., 2020) auf eine Stichprobe von 137 Publikationen der TU bzw. des Universitätsklinikums Dresden angewendet und mit den Klassifikationsergebnissen des **DataSeer** Natural Language Processing Modells (extrahiert aus dem kürzlich veröffentlichten **PLOS Open Science Indicators Datensatz**; PLOS, 2022) und einer **manuellen Kodierung** verglichen.

Der ODDPub Algorithmus identifizierte, ähnlich wie die manuelle Kodierung, Open Data in etwa 57 % der betrachteten Publikationen, das DataSeer Modell fast 20 % mehr. Die beiden automatisierten Klassifikationen stimmen zu einem großen Teil (ca. 80 %) überein und haben vergleichbare F1-Scores (ODDPub: 0,84; DataSeer: 0,83), eine Metrik zur Bewertung von Klassifikationsmodellen. Allerdings hat der ODDPub Algorithmus eine etwas höhere Precision als DataSeer bei gleichzeitig gutem Recall (ODDPub: Precision 0,83, Recall 0,86), während DataSeer durch einen hohen Recall, aber gleichzeitig etwas geringere Precision gekennzeichnet ist (DataSeer: Precision 0,72, Recall 0,99). Das heißt 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 fanden sich für beide Algorithmen u.a. für die Identifikation von Datennachnutzung oder fehlerhafter Verlinkungen. Als Open Code identifizierte ODDPub ca. 13 % der Publikationen, die manuelle Kodierung und der DataSeer Algorithmus zeigten das für fast die doppelte Menge. DataSeer identifizierte mehr von den manuell als Open Code kodierten Publikationen korrekt als ODDPub und erreichte höhere Werte in den anderen Bewertungsmetriken der Klassifikation.

Es konnte exemplarisch gezeigt werden, dass Verfahren wie ODDPub und DataSeer erfolgreich genutzt werden können um Textpublikationen einer Einrichtung automatisiert nach Hinweisen auf Open Data zu durchsuchen. Je nach Nutzungsszenario sollten die Verfahren aber gegeneinander abgewogen werden. DataSeer könnte vorteilhafter sein, wenn möglichst alle Hinweise auf Open Data (unabhängig von der Publikationspraxis) erfasst werden sollen. ODDPub identifiziert präziser die tatsächlich offenen Datenpublikationen, was etwa bei der Nutzung der Ergebnisse für eine Incentivierung innerhalb einer Einrichtung wichtiger sein könnte. Für die Identifikation von Open Code scheint der DataSeer Algorithmus momentan besser geeignet zu sein. Beide Vorgehensweisen sind in der Lage, bei der momentan heterogenen Publikationspraxis für Forschungsdaten, deutlich mehr Forschungsdatenveröffentlichungen aufzufinden als allein über Verknüpfungen von persistenten Identifikatoren und Metadaten-Aggregatoren zu erwarten ist. In Zukunft sollten

aber zunehmend Qualitätskriterien, wie die standardisierte Beschreibung und Ablage von Forschungsdaten in Repositorien mit persistenten Identifikatoren eine größere Rolle spielen, so wie das in der Weiterentwicklung der Verfahren schon angedacht ist.

Public Library of Science. (2022). *PLOS Open Science Indicators* (Version 2) [Data set]. Figshare. https://doi.org/10.6084/m9.figshare.216876
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

Primary author: Dr ZINKE, Katharina (SLUB Dresden)

Presenter: Dr ZINKE, Katharina (SLUB Dresden)

Session Classification: Posterslam und Verleihung des FAIRest Data Award

Track Classification: SaxFDM-Tagung