Reddit als (Text-)Ressource: Erstellung und Nachnutzbarkeit eines deutschsprachigen Reddit-Korpus

Göttel, Sebastian
sebastian.goettel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0002-8590-7730

Körber, Lydia
lydiaekoerber@gmail.com
Universität Heidelberg, Deutschland
ORCID: 0000-0002-8937-3799

Barbaresi, Adrien
barbaresi@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0002-8079-8694

Hintergrund

[1] Reddit [2] gehört mit durchschnittlich über 7 Milliarden Aufrufen pro Monat [3] zu den weltweit am meisten besuchten Webseiten. Die Plattform kombiniert Elemente eines Forums und eines sozialen Mediums und fungiert als Social-News-Aggregator, auf dem Nutzer:innen Texte, Links, Videos und Bilder teilen können. In Subreddits, thematisch spezialisierten Foren, können Nutzer:innen Inhalte teilen und diskutieren. Jedes Subreddit widmet sich einem spezifischen Thema oder Interessenbereich, von allgemeinen Themen wie Nachrichten und Technologie bis hin zu sehr spezifischen Hobbys oder kulturellen Nischen.

Als Korpus weicht Reddit in vielfacher Hinsicht von traditionellen Textkorpora, wie etwa einem Roman- oder Zeitungskorpus, ab. Die Sprache zeichnet sich durch einen hohen Grad konzeptioneller Mündlichkeit aus. Dies äußert sich in einer dialogischen und teilweise umgangssprachlichen Kommunikationsform. Zudem bietet das Korpus eine breite Palette sprachlicher Phänomene und eine erhebliche Variabilität des Wortschatzes, einschließlich Neologismen, fachspezifischer Terminologien und Jargon, die in formellen Texten typischerweise unterrepräsentiert sind. Darüber hinaus lassen sich Belege für regionale und soziolektale Varietäten finden, die von umgangssprachlichen bis hin zu mehr formalisierten Diskursformen reichen, sowie andere pragmatische Aspekte des Sprachgebrauchs. Neben linguistischen Fragestellungen eignet sich das Korpus auch für Forschungen in anderen (Teil-)Disziplinen. Es stellt eine interessante Ressource für die Digital Humanities dar und bietet breite Möglichkeiten für weiterführende Untersuchungen.

Deutschsprachiges Reddit-Korpus

Forschungsstand

In vorherigen Arbeiten wurden aus einem umfassenden Reddit-Datensatz deutschsprachige Kommentare mittels einer zweistufigen Filterung extrahiert (Barbaresi 2015; Blombach et al. 2020). Dabei griffen diese Ansätze wie auch der vorliegende auf die Pushshift-Archive [4] zurück.

Datensatz

Die einzelnen Subreddits und deren Kommentare sind als (ND)JSON-Dateien in zst[5]-Archiven archiviert. Der aktuelle Datensatz reicht von der Eröffnung eines Subreddits (ca. 2006/2007, teilweise auch erst später) bis einschließlich zum 2023-12-31. Vertretene Subreddits sind dabei z. B. r/de (mit über 2 Mio. Abonnent:innen das größte und thematisch allgemeinste Subreddit), Subreddits mit einem speziellen thematischen Bezug, wie r/Finanzen oder r/Studium , aber auch solche, die eine regionale Varietät abdecken, wie r/schwiiz oder r/aeiou .

Die Auswahl umfasst 40 Subreddits, die aus den 100 meistabonnierten Subreddits im DACH-Raum händisch ausgewählt wurden. Ziel war es, eine thematisch möglichst breite Abdeckung zu erreichen, wobei Subreddits ausgeschlossen wurden, die rein pornografische oder rechtsextreme Inhalte aufweisen. Ebenfalls nicht berücksichtigt wurden Subreddits, die trotz ihrer primären Ausrichtung auf den deutschsprachigen Raum einen hohen Anteil englischsprachiger Beiträge enthalten. Die Zahl 40 wurde bewusst gewählt, um eine Datenmenge zu gewährleisten, die sowohl thematisch divers als auch für die Verarbeitung und Analyse effizient handhabbar ist.

Datenaufbereitung und -konvertierung

Die erarbeitete Verarbeitungspipeline transformiert die Ausgangsdaten, komprimierte zst-Dateien, in ein strukturiertes TEI-XML-Format [6] . Dieser Prozess beginnt mit einer Bereinigung der Daten: Hierbei werden nachträglich gelöschte Kommentare, die noch als Fragmente in den Datensätzen vorhanden sind, sowie von Bots generierte Inhalte entfernt. Ebenso werden Einträge, die ausschließlich

aus URLs und/oder Zitaten bestehen, ausgeschlossen [7] . Anschließend werden alle Kommentare, die zu einem zusammenhängenden Thread gehören, in einheitliche JSON-Dateien überführt. Diese Dateien bestehen jeweils aus einem vollständigen Thread, identifiziert durch eindeutig zugeordnete Schlüssel aus den JSON-Objekten. Alternativ haben Nutzer:innen auch die Möglichkeit, die Kommentare nicht in Thread-Gruppen zusammenzuführen, sondern jeden Kommentar in eine separate Datei speichern zu lassen. In der letzten Phase der Datenverarbeitung werden diese JSON-Dateien in TEI-XML konvertiert. Jeder Kommentar wird dabei im XML-Dokument als <item> mit umfangreichen Metadaten, wie Autor, Datum und einer persistenten URL zum Originalkommentar versehen, wodurch die Originalquelle stets im ursprünglichen Reddit-Thread aufrufbar bleibt.

Abbildung 1: Screenshot der TEI-XML-Kodierung eines Threads auf r/Finanzen, URL zum Original-Thread: https://www.reddit.com/r/Finanzen/comments/5wa69r/n26_black_im_ausland/.

Nachnutzbarkeit und Zugang zum Repositorium

Alle für dieses Projekt entwickelten Skripte werden in einem öffentlich zugänglichen Repositorium hinterlegt. Die Verfügbarkeit der ursprünglichen Datensätze aus den Pushshift-Archiven ermöglicht zudem die Reproduktion der Ergebnisse.

Gleichzeitig ist geplant, das Reddit-Korpus in die Plattform des Digitalen Wörterbuchs der deutschen Sprache (DWDS) zu integrieren.

Poster

Das Poster illustriert den Prozess der Datenerfassung, Aufbereitung und Konvertierung des Reddit-Korpus. Es zeigt anschaulich, wie aus den gesammelten Daten strukturierte TEI-XML-Dokumente erstellt werden und wie diese Daten innerhalb der DWDS-Plattform und darüber hinaus analysiert und nachgenutzt werden können.

Fußnoten

1. Contributor Roles: Sebastian Göttel (Conceptualization, Data curation, Software, Writing – original draft), Lydia Körber (Software, Validation), Adrien Barbaresi (Software, Supervision, Writing – review & editing).
2. https://www.reddit.com/
3. vgl.: https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/
4. Pushshift archiviert eine Vielzahl von Reddit-Daten, einschließlich aller Beiträge und Kommentare, die auf der Plattform gepostet werden. Pushshift bietet spezialisierte Daten-Dumps an, die nach einzelnen Subreddits geordnet sind und bis einschließlich Ende 2023 reichen. Für das vorliegende Korpus wurden nur die Pushshift-Dateien verwendet, die ausschließlich Kommentare innerhalb der Threads enthalten (‘_comments’ im Dateinamen) : https://academictorrents.com/details/56aa49f9653ba545f48d-f2e33679f014d2829c10. .
5. zst ist ein komprimiertes Datenformat, das auf dem Kompressionsalgorithmus Zstandard basiert.
6. vgl. die Richtlinien der Text Encoding Initiative (TEI), https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.
7. Zusätzlich zu den beschriebenen Schritten wurden weitere Filtermaßnahmen angewendet, deren detaillierte Auflistung der Dokumentation des Repositoriums entnommen werden kann.

Bibliographie

Barbaresi, Adrien. 2015. "Collection, description, and visualization of the German Reddit corpus." In *2nd Workshop on Natural Language Processing for Computer-Mediated Communication*, 7-11. https://hal.science/hal-01207311v2 (zugegriffen: 24. Juli 2024).

Blombach, Andreas, Natalie Dykes, Philipp Heinrich, Besim Kabashi and Thomas Proisl. 2020. "A Corpus of German Reddit Exchanges (GeRedE)." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6310–6316. https://aclanthology.org/2020.lrec-1.774/ (zugegriffen: 24. Juli 2024).