

Research stories on Twitter

David G. Lorentzen¹ and Gustaf Nelhans¹

¹ University of Borås, Allégatan 1, Borås, Sweden

Abstract

This paper aims to study what type of research seems to interest the users of a social network platform and then complement the data with data from an open catalogue for research, exemplifying with Twitter and Open Alex. The basic idea is to get an overview of the stories the platform content tells during three months regarding topics, disciplines, and open access status. The findings suggest that the picture look very different between the approaches to map the topics, especially when looking at the articles most mentioned compared to the ones that are most retweeted. The study mainly highlights the methodological opportunities of combining text analysis and link relationships to explore the content and public interest in academic research.

Keywords

Twitter, Open Alex, topics, open access

1. Introduction

The study's relevance is the question of what scientific stories are told on a social media platform. The paper deals with the combination of sources, where one can be categorized as a streaming source, where posts are added continuously, and the other a more static source, where resources can be looked up for complimentary data. The study takes a digital methods perspective with a focus on social science research, which then implies that what we study is the stories that are told by the content [1]. The platform we take off with is formerly known as Twitter, now renamed X. Since data were collected before the rebranding, we use Twitter in this text.


2. Method and data

Data were collected using Focalevents [2]. At the time of data collection, we could use an academic developer account that allowed for searching the archive and streaming in real-time, with a download limit of 10 million tweets a month [3]. Table 1 lists the top base URLs with the number of tweets matching each URL. The data collection period was set to the first three months of 2023, searching for tweets with the base URL `<https://doi.org>`. This also matches URLs such as `<http://www.doi.org/10.51372/bioagro351.1>`

457,775 tweets were collected in this way. We selected all tweets written in English with DOI references (non-retweets) in the next step. Following pre-processing steps in which we unshortened shortened DOIs with Python requests and validated DOIs with python-doi, we ended up with 86,829 unique DOI references. Of these, 623 were invalid, for example `<https://doi.org/10.nuts>`. Using the Open Alex API, we then looked up more data, such as title, publication year, language, text type, open access status, source, keywords, abstracts, connection to sustainable development goals, citation data and retraction status. Data for 84,608 records (97 %) were returned from Open Alex.

We used Word2Vec from the Python gensim library on the abstracts to map topics. Stop words were removed, and words were stemmed using the Porter stemmer. The Word2Vec model was trained on the data for ten epochs. For each of the top 1,000 word stems, we looked up the 100 most similar terms and kept all relationships that were stronger than 0.5. These relationships were used to create term networks

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 david.gunnarsson_lorentzen@hb.se (D. G. Lorentzen); gustaf.nelhans@hb.se (G. Nelhans)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

for further analysis. We did this for the entire dataset, the 1,000 most retweeted DOIs, and the 1,000 most mentioned DOIs.

Apart from the topical maps, we also performed descriptive statistical analyses.

3. Findings

3.1. Descriptive statistics

Of the 84,608 references, the vast majority were articles (82,414). Seven references were retracted. A large share of the references were open access. 25,993 had gold status, 16,291 were hybrid, 10,783 were green, and 5,242 were bronze. This entails that 50% of the references were open access and 69% if we include hybrid. According to the National Library of Sweden, this is quite in line with the share of published research from Swedish academics, which was 70% of the published scholarly articles in 2022. The most cited work was a book titled “Diagnostic and Statistical Manual of Mental Disorders” with 69,177 citations, and another 26 works had citation counts of at least 10,000. 342 were in the range between 1,000 and 10,000, and 3,162 were cited between 100 and 1,000 times. The data covered works from the most recent years, with 67,099 from 2022 and later, but also some historical works, with the oldest being “IV. An account of the tubera terræ, or truffles found at Rushton in Northamptonshire; with some remarks thereon” from 1693. 1,272 works were from 2000 and earlier, of which 26 were from before 1900.

Table 1
Publication types

Publication Type	Count
Article	82,414
Book chapter	981
Book	580
Report	197
Paratext	101
Reference entry	80
Dissertation	74
Dataset	65
Editorial	58
Other	58

As discovered by [4], many sources were from the natural sciences (Table 2). We see a variety of works when looking at the most overall mentioned DOIs, including retweets (Table 3). These are the most visible articles in the dataset across the three months. Most of these are from natural sciences and medicine, but there are also some examples from social sciences and psychology, such as the article about sharing misinformation.

Table 2
Source outlets

Source	Count
bioRxiv (Cold Spring Harbor Laboratory)	1,074
Nature Communications	829
Proceedings of the National Academy of Sciences of the United States of America	693
Scientific Reports	600
eLife	584
PLOS ONE	580
Nature	524
Science	333
Science Advances	309
Cell Reports	292

Table 3
Works grouped by mentions in tweets (including retweets)

Title	DOI	Mentions count
Serious adverse events of special interest following mRNA COVID-19 vaccination in randomized trials in adults	10.1016/j.vaccine.2022.08.036	859
Sharing of misinformation is habitual, not just lazy or biased	10.1073/PNAS.2216614120	629
The management of diabetic ketoacidosis in adults—An updated guideline from the Joint British Diabetes Society for Inpatient Care	10.1111/dme.14788	540
Integrating Molecular Biology and Bioinformatics Education	10.1515/jib-2019-0005	474
The Efficacy and Use of a Pocket Card Algorithm in Status Epilepticus Treatment	10.1212/CPJ.0000000000000922	424
The use of diuretics in heart failure with congestion	10.1002/ejhf.1369	396
2021 World Health Organization guideline on pharmacological treatment of hypertension: Policy implications for the region of the Americas	10.1016/j.lana.2022.100219	388
Metabolic syndrome – a new definition and management guidelines.	10.5114/aoms/152921	369
Management of Hyperglycemia in Type 2 Diabetes, 2022	10.2337/dci22-0034	291
Plant genome sequence assembly in the era of long reads	10.1017/qpb.2021.18	270

3.2. Topics

The first map (**Figure 1**) is based on all works cited in the dataset, where each abstract is treated the same. This map shows the diversity in topics, with several distinct clusters at the bottom showing terms related to molecular medicine, pathogens, climate research, agriculture and ecology. There are methodological and theoretical terms in the centre, while words related to academia and professions, the family, and psychological terms are found in the top right corner.

In the top left corner, different clusters distinguish words with linguistic functions, e.g. the purple cluster contains various types of conjunctions. At the same time, numbers, adverbs or adjectives relating to time and temporal sequencing, comparisons, measurements, and spatial or numerical relationships are found in different clusters. The other two maps zoom in on what the Twitter users find most interesting to redistribute (retweets) (**Figure 2**) and talk about (mentions) (**Figure 3**). Similarities include the focus on medical and clinical terms (bottom left and right, while the natural sciences, especially physics, are pretty well represented in the centre of **Figure 2**). In **Figure 3**, it is harder to distinguish topics, but the bottom cluster seems to relate to clinical medicine. In contrast, the academic and social cluster, including the mention of Chat GPT, is found in the orange cluster to the left.

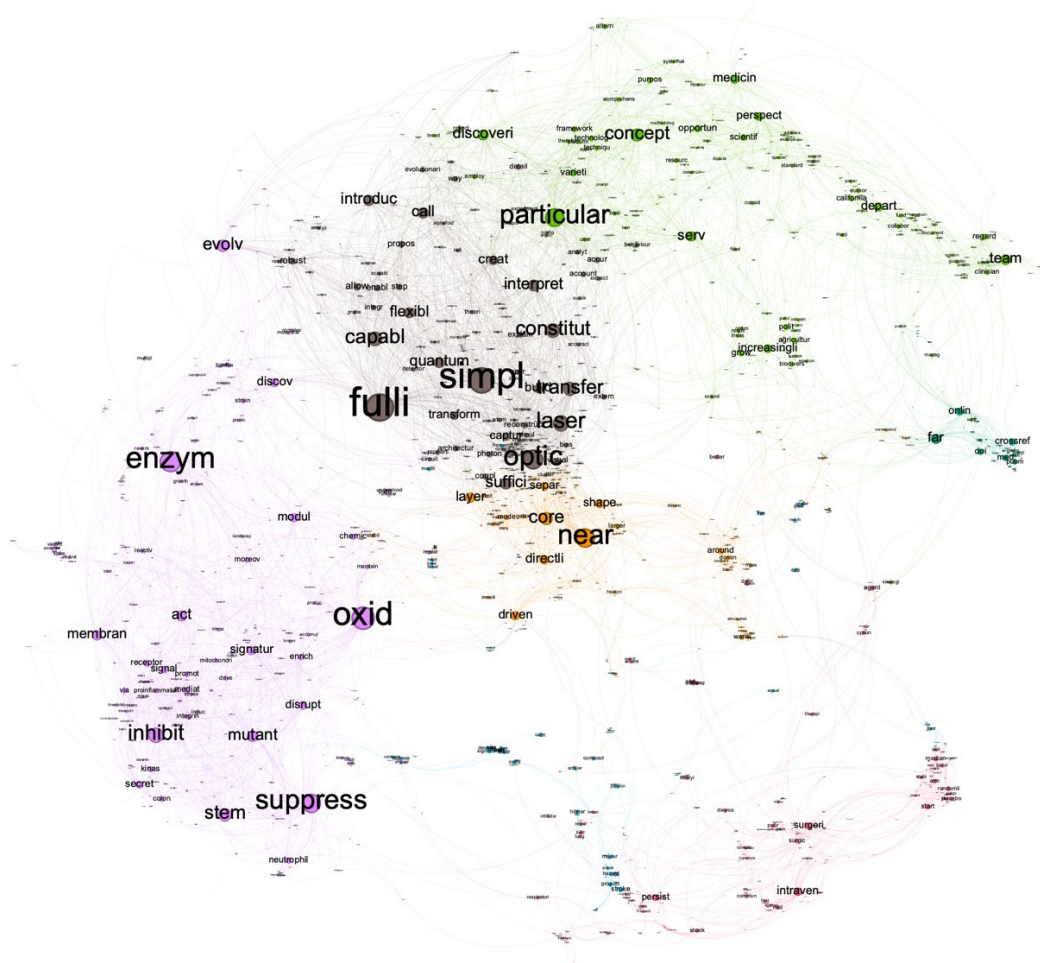


Figure 2: Word2Vec network from abstracts based on retweets.

Regarding data selection, we relied on DOI URLs but by doing so missing out on tweets referring to a direct URL to articles. Perhaps this decision limits the dataset to tweets created by people who are more accustomed to the academia. It is also important to keep in mind the technicalities of the platform at the time of the study. From a researcher point of view it was possible to search in the archive and collect up to ten million tweets a month, and also look up the conversations the tweets are part of. While a study of this type can reveal insights into what research the public is interested in, Twitter is not representative of the general public, and the sharing practices indicate of usage by researchers for self-promotion among other potential purposes [4]. However, when moving beyond the mere mentioning of research, argumentative patterns and practices can be revealed [7]. This paper has shown how one can use digital methods to study sharing practices on a platform in relation to a specific type of artifact, in this case research articles using their DOI URLs, and collecting additional information about what they share using an open data source. The use of digital methods to collect and analyse data makes it possible to uncover patterns that are not apparent when utilising manual analyses. Similar approaches can be used for other contexts in order to enhance the understanding of aspects of human culture.

Acknowledgements

This study was funded by Huminfra.

References

- [1] S. Niederer, *Networked Content Analysis: The Case of Climate Change*. Institute of Network Cultures, 2019
- [2] R. Gallagher, Social Media Focal Events Listener. URL: <https://focalevents.readthedocs.io/en/latest/index.html>, 2023
- [3] Twitter, Academic research access. URL: <https://web.archive.org/web/20230520042703/https://developer.twitter.com/en/products/twitter-api/academic-research>
- [4] G. Nelhans, & D. G. Lorentzen, Twitter conversation patterns related to research papers. *Information Research*, 21(2), paper SM2, 2016. URL: <http://InformationR.net/ir/21-2/SM2.html>
- [5] D. G. Lorentzen, J. Eklund, B. Ekström, & G. Nelhans, On the potential for detecting scientific issues and controversies on Twitter: A method for investigation conversations mentioning research. In *Proceedings of ISSI 2019*, 2189-2198, article-id 375, 2019
- [6] J. Eklund, & G. Nelhans, Probabilistic explorations of citation contexts: Citation roles and subject content of scientific references. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators, STI 2022* (sti22224), 2022.
- [7] A. Foderaro, & D. G. Lorentzen, Argumentative patterns and practices in debating climate change on Twitter. *Aslib Journal of Information Management*, 75(1), 131-148, 2023.