Leveraging machine learning for bibliometric analysis of emerging fields

Claudiu Petrule[1], André Bittermann[2], Viktoria Ritter[3], Anke Haberkamp[4] and Winfried Rief[5]

[1]*cp@leibniz-psychology.org,* [2]*abi@leibniz-psychology.org*
Leibniz Institute for Psychology (ZPID), Universitätsring 15, 54296 Trier (Germany)

[3]*ritter@psych.uni-frankfurt.de*
Goethe University Frankfurt, Varrentrappstr 40-42, 60486 Frankfurt/Main (Germany)

[4]*anke.haberkamp@staff.uni-marburg.de*, [5] *rief@staff.uni-marburg.de*
Philipps Universität Marburg, Gutenbergstr 18, 35032 Marburg (Germany)

Background

Traditional bibliometric methods can be limited in their ability to analyze emerging fields where the object of study resists clear delineations, the boundaries between subfields are porous or the relationships between different subfields are complex. The first obstacles arise at the early stage of literature search. Inconsistent terminology prevents a satisfactory coverage of the construct of interest and leads to biased and distorted representations. Combating such epistemic chambers by broadening the search dramatically increases the noise and the amount of publications needed to be screened.

In contrast to medical translational research, psychological translational research is still in its early phase (Ehring et al., 2022), and as in many emerging fields the terminology of translational research in psychology is inconsistent. One negative effect of inconsistent terminology (cf. Colquhoun et al., 2014), is the challenging database search for eligible studies: an explicit search for "translational psychotherapy" increases the share of eligible studies among the search results at the cost of many missed relevant publications. Vice versa, widening the search query (e.g., "psychotherapy") inflates the share of irrelevant results and renders the already herculean task of screening unfeasible.

Machine Learning (ML) is a promising ally in sifting through data. One popular ML-powered tool is Rayyan (https://www.rayyan.ai), aimed at facilitating the initial screening of abstracts and titles by a semi-automated process (Burgard & Bittermann, 2023).

Aim

The goal of this case study is to map the emerging research landscape of translational psychotherapy and compare the ML-augmented results with those based on a typical search query.

Method

Methodological Rationale

To identify the publications of the emerging field that use unknown terminology, we used the ML feature in Rayyan to automate the screening process and identify eligible records from a large pool of publications. The Rayyan classification model was trained on the screened results of a search query using the field's known terminology (i.e., "translational psychotherapy"). This training data was used to predict the inclusion probability of unseen records, thus reducing the workload for manual screening.

Data

We performed literature searches in the psychology-specific databases PsycInfo and PSYNDEX in September 2022. A total of unique 153,687 records were retrieved. In addition to our search, we also used citation mining to find papers that may not have been detected by the initial queries. The seed for such search was a pool of 22 screened papers which, after the forward and backward citation search, yielded further n = 11,135 publications. We also collected publications as additional sources (n = 6,806) from special issues on the topic of translation, references of authors known to have published on the topic of interest, and other eligible papers we were aware of.

The training data consists of the manually screened publications found with the explicit "translational" search (n = 246) and 40 additional eligible papers.

Assessing the Added Value of ML

We compared publication volume, main actors, journals, regional differences, subfields of psychological science, study methodology and impact indicators such as citations, collaboration networks and social metrics of two datasets: 1) the papers found with known terminology only ("search dataset"), and 2) the papers found by leveraging ML ("ML dataset"). To generate the ML-augmented dataset, we performed two rounds of active learning

and screened results until the inclusion rates of the classifier were above 95%. Next, we included all remaining records predicted by the model. For both datasets, we removed records without DOI (which is needed for retrieving OpenAlex metadata) and secondary research.

Software

For bibliometric analyses we queried the OpenAlex API and used the following R packages: openalexR, bibliometrix, quanteda, semscholar.

Results

Not only the publication volume differs significantly between the search dataset and the ML dataset, but also other metrics (Table 1).

Table 1. Comparison between datasets

Criteria	Search Dataset	ML Dataset
N. of publications	229	683
Annual growth rate	9.33	13.07
M citations	21.56	67.38
Median citations	17	21
M Twitter mentions	4.86	20.58
Median Twitter mentions	1	2
Open Access	34.74%	37.21%

As visible in Figure 1 the most productive authors in the field of translational psychotherapy are either absent (n = 3) or poorly represented in the search dataset (output volume measured as relative frequency).

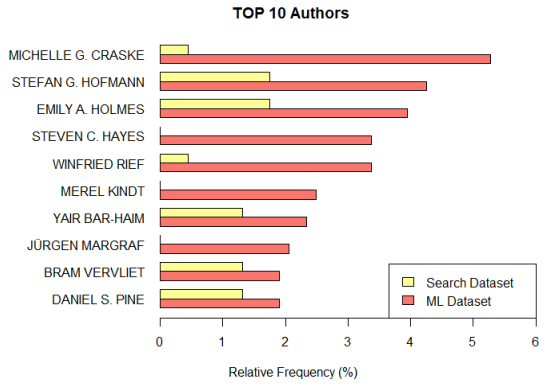

Figure 1. TOP 10 Authors (in ranking order according to ML Dataset).

A similar trend is also observed in the journal distribution, where 5 of the Top 10 journals differ in distribution by 50%. In terms of the regional differences more than half of the Top 25 countries differ in distribution by 50% as well.

As to the differences in subfields of psychological science and study methodology (as recorded by the databases), there are no significant differences.

Discussion

Our study indicates that when researching emerging fields, using a dataset based solely on known vocabulary can lead to different results and biased interpretations. Additionally, due to the lack of specific categorization that matches the field of interest in databases, the screening process, although laborious, is crucial to ensure the validity of the subsequent analyses. However, once the model is trained it can be applied to new data of virtually any size and even used to continuously update the dataset.

Limitations and future research

Generating the pool for the ML classifier was done manually. Ideally, this could be automated by either automatic DOI and metadata retrieval or by database hosts incorporating ML features for record search.

This case study focused on the emerging field of translational psychotherapy, therefore more research is needed to investigate the generalizability of our findings across different emerging fields. Moreover, future studies could examine the characteristics of research fields that would benefit from ML assistance to generate bibliometric datasets (i.e., determine when the vocabulary is too inconsistent for explicit searches).

Another limitation is the use of only one ML classifier. Comparing different classifiers could potentially result in better performance.

Conclusions

While the screening process requires additional time, this case study demonstrates that ML can facilitate bibliometric analyses of emerging fields with inconsistent or partially unknown terminology.

References

Burgard, T. & Bittermann, A. (2023). Reducing Literature Screening Workload with Machine Learning. Systematic Review of Tools and their Performance. *Zeitschrift für Psychologie*, 231(1), 3-15. https://doi.org/10.1027/2151-2604/a000509

Colquhoun, H., Leeman, J., Michie, S. et al. (2014). Towards a common terminology: a simplified framework of interventions to promote and integrate evidence into health practices, systems, and policies. *Implementation Sci* 9, 781. https://doi.org/10.1186/1748-5908-9-51

Ehring, T. et al. (2022). (When and how) does basic research in clinical psychology lead to more effective psychological treatment for mental disorders?. *Clinical Psychology Review*, 95, 102-163. https://doi.org/10.1016/j.cpr.2022.102163

Supplements

We provide datasets and R code on https://github.com/cpetrule/issi23