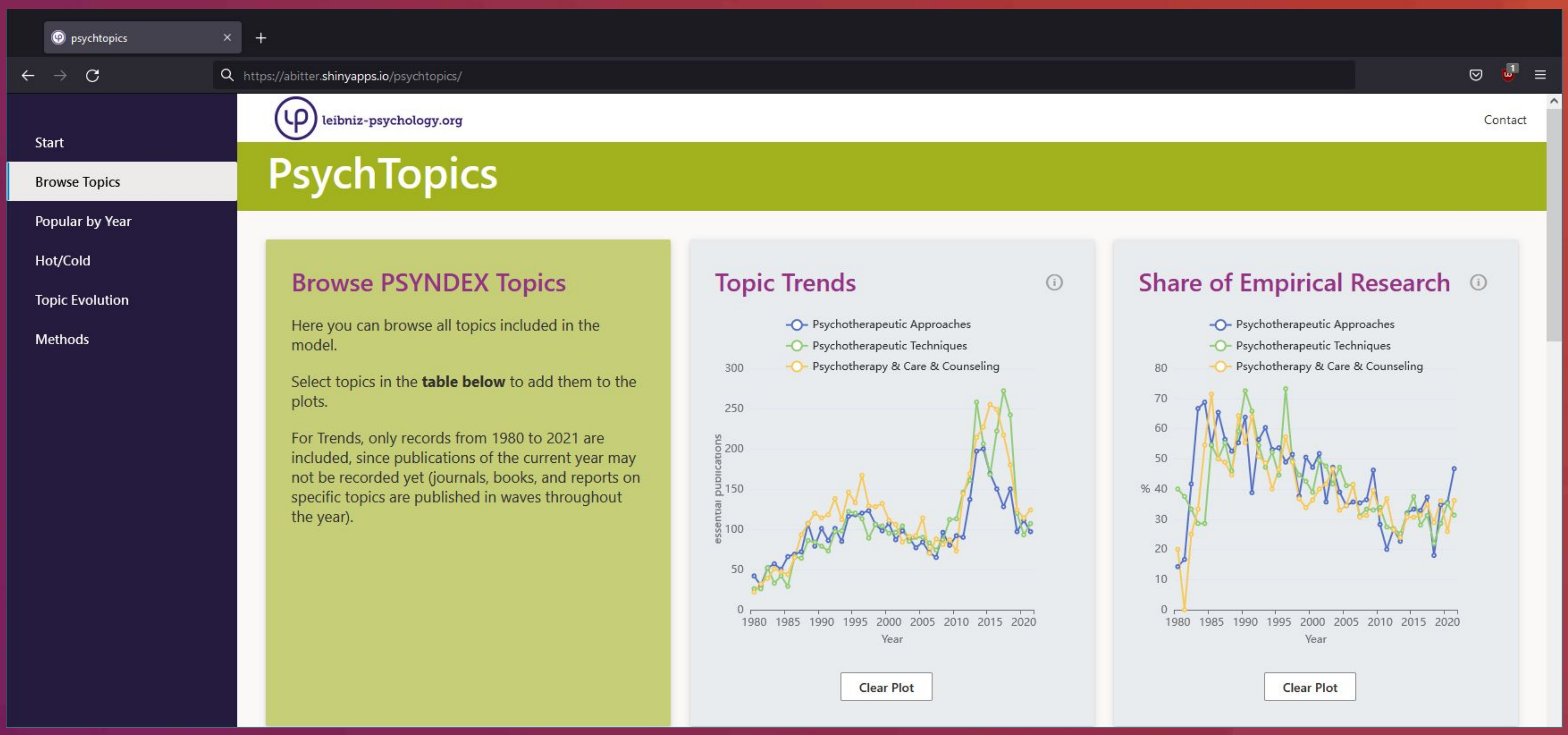
Scientific topics can be monitored with RollingLDA and R Shiny



Browse PSYNDEX Topics

Here you can browse all topics included in the model.

Select topics in the **table below** to add them to the plots.

For Trends, only records from 1980 to 2021 are included, since publications of the current year may not be recorded yet (journals, books, and reports on specific topics are published in waves throughout the year).

How to keep up with the vast amount of scientific information published every day?

We present a framework and app for continuous topic detection in scientific publications.

We employ RollingLDA, a topic modeling variant designed for "living" text corpora.

Our open source Shiny App can be modified to your needs.

Finding Scientific Topics in Continuously Growing Text Corpora

André Bittermann[1] & Jonas Rieger[2] (equal contributions)
[1] Leibniz Institute for Psychology (ZPID), Trier, Germany
[2] Department of Statistics, TU Dortmund University, Germany

Study Design

- We compare the evolution of 42 RollingLDA model variants to a single ldaPrototype reference model (for 2020).
- The best fitting model is determined using cosine similarity to the reference model, topic quality metrics, and external validation:

Start	K	Similarity*	Coherence	Exclusivity	Correlation**	Mean (of z-scores)
2010	**200**	0.623 898	−123.997 870	4.137 017	0.960 064	**0.188 719**
2005	200	0.621 397	−123.516 668	3.949 559	0.962 599	−0.054 622
1995	200	0.621 219	−123.226 158	3.881 941	0.966 658	0.176 869
2010	300	0.621 108	−123.386 484	4.320 748	0.946 135	−0.008 355
2015	200	0.620 810	−123.740 794	4.410 456	0.944 504	−0.302 611

Table 1: Comparison of RollingLDA model variants. The reference model for 2020 (cf. Sect. 3.3.1) comprised 250 topics. The best fitting model variant is printed in bold. Notes: *mean cosine similarity to the topics of the reference model. **correlations between actual classification category frequencies and classification shares in the topics (external validation).

Results

- Using RollingLDA with annual updates of the corpus, 82% of the reference model's topics could be detected.
- Missed topics were either of low prevalence in the reference model (13.6%) or included in other topics (4.4%).

Conclusion

- The model integrates new publications while keeping time series of topic trends consistent.
- This can help various stakeholders like researchers or policy-makers to evaluate how research fields evolve over time.
- The presented framework has a high degree of automation once the initial model is created.

Read the paper:

dx.doi.org/10.23668/psycharchives.8168

Try the app:

abitter.shinyapps.io/psychtopics/

Find the source code:

github.com/leibniz-psychology/psychtopics

3rd Workshop on Scholarly Document Processing
29h International Conference on Computational Linguistics
October 17, 2022, Gyeongju, Republic of Korea

leibniz-psychology.org

technische universität dortmund