Finding Scientific Topics in Continuously Growing Text Corpora

André Bittermann🥑
Leibniz Institute for Psychology (ZPID), Universitätsring 15, 54296 Trier, Germany
abi@leibniz-psychology.org

Jonas Rieger🥑
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
rieger@statistik.tu-dortmund.de

Abstract

The ever growing amount of research publications demands computational assistance for everyone trying to keep track with scientific processes. Topic modeling has become a popular approach for finding scientific topics in static collections of research papers. However, the reality of continuously growing corpora of scholarly documents poses a major challenge for traditional approaches. We introduce RollingLDA for an ongoing monitoring of research topics, which offers the possibility of sequential modeling of dynamically growing corpora with time consistency of time series resulting from the modeled texts. We evaluate its capability to detect research topics and present a Shiny App as an easy-to-use interface. In addition, we illustrate usage scenarios for different user groups such as researchers, students, journalists, or policy-makers.

1 Introduction

In the era of "Big Literature" (Nunez-Mir et al., 2015), the exponentially growing number of research publications (Bornmann et al., 2021) poses a serious challenge to those trying to keep up with the vast amount of scientific information published every day. On the one hand, this affects scientists and students who want to stay up-do-date. Due to the accelerating effects of digitization and globalization (cf. Hilbert and López, 2011), assessing scientific developments in a timely manner has become a challenging endeavor – even for experts in their respective fields. A recent example is the plethora of research papers on COVID-19 that rapidly grew after the outbreak in 2020 (Aviv-Reuven and Rosenfeld, 2021). The exceptionally large number of researchers (Ioannidis et al., 2021) produce scientific output that is arguably too much to be reviewed by individual researchers on a case by case basis. Outside academia, on the other hand, journalists, politicians, and the general public are interested in research processes and findings as well. For instance, policy-makers need to evaluate whether a research field is moving toward the intended direction, e.g., whether funding yields scientific output as expected. Journalists who want to report the latest trends in research often depend on (potentially biased) expert opinions or conferences that take place only once per year or biennially. This hampers trend detection on a timely, large scale, and reproducible basis.

1.1 Related Work

Scientific output that is high in volume and velocity demands statistical methods and tools that assist in processing such amounts of information. One strategy to reduce the overload of information is to condense large volumes of text collections to their main topics. In recent years, bibliometrics enhanced with natural language processing (NLP) has emerged as a promising solution for handling such large text corpora (Atanassova et al., 2019). For finding scientific topics, in particular topic modeling became a standard method in scientometrics (e.g., Colavizza et al., 2021; Griffiths and Steyvers, 2004; Yau et al., 2014). Initially developed for information retrieval purposes (Blei et al., 2003), topic modeling is widely used for gaining insights into the underlying themes of text collections. It reduces high dimensional text data to a few groups of co-occurring terms which are interpreted as topics. Put differently, the goal is to "analyze the words of the original texts to discover the themes that run through them" (Blei, 2012, p. 77). By considering the document metadata, the analyses can get more fine-grained. For instance, by incorporating the date of publication into the model, the topic prevalence over time can reveal patterns of publication trends such as "hot" or "cold topics" (Griffiths and Steyvers, 2004). The main advantage of deriving topics from scholarly texts instead of using

🥑Equal contribution.

Proceedings of the Third Workshop on Scholarly Document Processing, pages 7–18
October 12–17, 2022.

database metadata (such as subject headings or classification codes; Krampen, 2016) is their ability to detect novel topics more flexibly (Suominen and Toivanen, 2016).

In summary, NLP approaches like topic modeling can help in coping with the vast amounts of scholarly documents published every day. From a methodological point of view, however, the integration of new texts into existing models fitted on a previous set of texts poses a major challenge. In particular, it remains an open question how to continuously detect research topics in a "living" corpus of scholarly documents.

1.2 Contribution

The current paper addresses the question of how to keep track of scientific topics and trends. We apply a recent topic modeling method to an annually updated corpus of scholarly documents and present a Shiny App that makes the results accessible to users without prior knowledge of coding or topic modeling. Firstly, we describe how topic modeling works and how traditional approaches deal with the integration of new documents into the model. Secondly, we argue that RollingLDA (Rieger et al., 2021) offers the possibility of sequential modeling of dynamically growing corpora ensuring time consistency of time series resulting from the modeled texts. Thirdly, using publications from the field of psychology as a use case, we investigate whether the RollingLDA approach can detect novel topics by comparing its evolved topics to those from a single topic model fitted on a corpus of publications from the year 2020. Fourthly, we describe a Shiny App that provides a user interface for exploring and analyzing research topics. Finally, we discuss practical implications for different user groups, the assets and drawbacks of our newly presented approach as well as future directions.

2 Methodological Background

Topic modeling is used in many application domains (cf. Blei, 2012), which might be partly due to the intuitive explanation of the model idea: a corpus of documents can be described by distributions of topics over time, where each word in each of these documents is assigned to one of the topics. This in turn yields word distributions for each topic, which are thereby made interpretable.

Probably the best known model among topic models is the latent Dirichlet allocation (LDA, Blei et al., 2003). The underlying probabilistic model (Griffiths and Steyvers, 2004) is given by

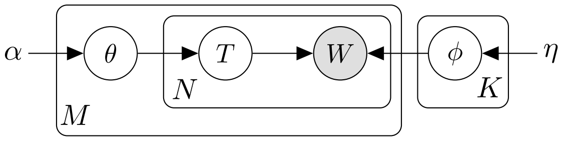


Figure 1: Schematic (plate) representation of LDA.

$$W_n^{(m)} \mid T_n^{(m)}, \phi_k \sim \mathrm{Discr}(\phi_k), \quad \phi_k \sim \mathrm{Dir}(\eta),$$
$$T_n^{(m)} \mid \theta_m \sim \mathrm{Discr}(\theta_m), \quad \theta_m \sim \mathrm{Dir}(\alpha),$$

where α and η are Dirichlet priors and K the number of topics to be modeled chosen by the user and each document $m = 1, \ldots, M$ is considered a bag of words set $\{W_n^{(m)} \mid n = 1, \ldots, N^{(m)}\}$ with observed words $W_n^{(m)} \in W = \{W_1, \ldots, W_V\}$. Then, $T_n^{(m)}$ describes the corresponding topic assignment for each word. Figure 1 gives a schematic representation of LDA. The observable variable W is colored gray, latent variables encircled, while constants are not. The latent word and topic distributions are represented by ϕ and θ, respectively.

For modeling topics in scientific corpora, we use a rolling variant of the classical LDA, estimated with the Gibbs sampler (Griffiths and Steyvers, 2004), named RollingLDA (cf. Sect. 2.2). The main challenge is to update the topic model with new publications while preserving the old time series based on the topic assignments of previous models on the one hand and allowing for the creation and mutation of new topics on the other hand.

2.1 Related Methods

Traditional approaches for this kind of task include the **one model fits all** approach, which consists of assigning new documents to topics of the existing topic model. This type of model is implemented by the online LDA (Zhai and Boyd-Graber, 2013), which is computationally inexpensive but lacks ability to capture new topics.

A second possible approach is to **recalculate the complete model** on the entire corpus for each update. In this way, it is possible that the model also catches more recent themes. However, with this approach, old topics usually change strongly or become unidentifiable. In addition, the consistency of the time series based on previous models is lost. Examples for this type of model are topics over time (Wang and McCallum, 2006) or continuous time dynamic topic model (Wang et al., 2008).

Both methods use information of future documents for modeling past documents.

Instead of calculating the new model on the entire data, it is possible to calculate **separate models** for each time period. In this way, past topics remain consistently interpretable, while the temporal interpretability of topics is lost, so that topics from different time intervals have to be matched in a complex (and tricky) way (cf. Niekler and Jähnichen, 2012) to get a minimum of interpretability.

One way to deal with the aforementioned drawbacks is the **restricted memory** approach. The temporal LDA (Wang et al., 2012), which can be used for monitoring writing styles of individual authors, or the streaming LDA (Amoualian et al., 2016), which is rather suitable for thematically narrower corpora due to a dependence structure between consecutive documents, are specialized models that implement this concept. For the given use case, the RollingLDA (Rieger et al., 2021) implements a more flexible version of the online LDA, whereby knowledge about previous documents is forgotten as time passes, thus allowing for mutations and new topics to be created. For the reasons mentioned above, we use RollingLDA for regular annual updates of the model.

We do not perform a qualitative comparison of the RollingLDA and (for instance) the online LDA, as there is no established evaluation metric for the quality of topic segmentation for the given application. Rather, there is a need for further research that defines task-based evaluation metrics and evaluates their usefulness, cf. Doogan and Buntine (2021); Ethayarajh and Jurafsky (2020) - for example, regarding correlation with human perception of meaningful structured topics, cf. Chang et al. (2009); Hoyle et al. (2021).

2.2 RollingLDA

The rolling version of LDA we use is initially based on one special LDA taken from an user defined initialization period (parameter `init`). Up to this date, a highly reliable run is selected from a set of LDA runs using the LDAPrototype method (Rieger et al., 2022a). Then, RollingLDA models the incoming data in minibatches (parameter `chunks`). For this, only a restricted time directly before each minibatch is considered as `memory`. Based on the topic assignments of the documents within the memory, the topics are reinitialized for each minibatch. By forgetting topic assignments from documents before the memory period, the model allows evolving topics or weakly populated topics to mutate strongly. This allows current topics to be captured by the model as well.

As long as topics are continuously populated, i.e., that there is no extraordinary drop in the topic's frequency, the initialization of the following minibatch ensures that existing topics are preserved. This prevents the problem of matching topics over time (cf. Niekler and Jähnichen, 2012). By the same property, the gradual evolution of topics is made possible by updating the topic initialization with only the most recent documents for every minibatch. In contrast, very weakly populated topics may be replaced by newly emerging topics due to the model architecture.

3 Framework

In order to explore the feasibility of RollingLDA for bibliometric purposes, the goals of the current study are threefold

- to compare the evolved RollingLDA topics to a topic model fitted on a specific year only,
- to show an efficient way of top term lifting in RollingLDA, and
- to illustrate how RollingLDA can be integrated into a Shiny App.

We investigate the eligibility of RollingLDA for topic identification in scholarly documents by setting different temporal lengths for model initialization as well as different numbers of topics and compare their evolved topics of 2020 to an individual LDA model fitted on the 2020 corpus only. We propose a method for time restricted top term weighting that offers additional insights into the evolution of topics. Moreover, we illustrate the integration of RollingLDA in a topic app. Leveraging R Shiny (Chang et al., 2021), we present an easy-to-use interface to the topic model that, among other things, visualizes topic trends and topic evolution, i.e., the change of topic terms over time.

We utilize the approach to the field of psychology as a use case, as psychological research is in most parts empirical, but also comprises theoretical and methodological contributions. This variety in study methodology should favor generalizability of our topic detection approach to other scientific disciplines.

3.1 Data

We extracted publication data from PSYNDEX, the comprehensive reference database for psychology publications from the German-speaking countries. PSYNDEX (`www.psyndex.de/en`) is produced by the Leibniz Institute for Psychology (ZPID) in Germany and has a field structure analogous to the international PsycInfo database, produced by the American Psychological Association. PSYNDEX is accessible for free via PubPsych (`www.pubpsych.eu`). The database was queried in November 2021, including a total of 360,009 publication references (titles, abstracts, and metadata) from the years 1980 to 2021.

3.2 Preprocessing

For finding scientific topics, we build a text corpus that consists of English language titles, abstracts, and standardized keywords. These keywords are the controlled terms of the American Psychological Association (Tuleya, 2007), a thesaurus of central concepts in psychological research similar to the MeSH terms of the National Library of Medicine. In contrast to author keywords, such standardized vocabulary represent the main concepts of the publications while reducing variance due to spelling variants or synonyms. This is especially relevant for methodological terms, as methods like "linear regression" are only indexed with the respective keyword, if the method itself was in focus of the publication, not a mere application for analyzing the data. Abstracts and titles are lemmatized and tokenized, while the keywords are left in their initial form due to their standardization. As suggested by Maier et al. (2018), we transformed all text to lowercase and removed punctuation as well as the stop words of scholarly abstracts provided by Christ et al. (2019) and Bittermann and Klos (2019a).

3.3 Study Design

For selecting a model variant with appropriate parameters, we first build a reliable reference model based only on the data from 2020, aiming for a RollingLDA variant which has a topic structure of the evolved topics in 2020 that is most similar to that of the reference model. In addition, the selected RollingLDA model should satisfy traditional topic quality criteria.

3.3.1 Reference Model for 2020

In order to determine the "actual" topics of 2020, we fit a topic model to documents published in 2020 only. Multiple LDA runs lead to different results, stressing the importance of topic reliability (Maier et al., 2018). We address this issue by applying LDAPrototype (Rieger et al., 2022a), which computes several LDA models and determines the one being the most similar to the other LDA models. For different numbers of topics K, we run 25 replications. Based on Bittermann and Fischer (2018) who found 500 topics in a psychology corpus spanning 37 years, we assume that a single year will have a significantly smaller number of topics. Hence, we inspect $K = 150, 175, \dots, 300$. We set the number of iterations to 500, $\alpha = 0.0001$ and $\eta = 1/K$ (package default), to create a few high probability topics and a lot of close-to-zero probability topics per publication. In order to reduce computation time (Strubell et al., 2019) and most likely without lack of quality (Maier et al., 2020), we exclude terms appearing in less than 15 publications.

To determine the optimal number of topics K, we follow the recommendations of Maier et al. (2018) and focus on topic interpretability. As proposed by Roberts et al. (2014), we jointly use two statistical metrics of topic quality: Semantic coherence as defined by Mimno et al. (2011) and topic exclusivity using LDAvis relevance score with $\lambda = 0$ (Sievert and Shirley, 2014). Subsequently, we manually inspect top words and the most representative documents of the three models with highest quality, leading to a final 2020 reference model with 250 topics.

3.3.2 RollingLDA Candidate Models

For RollingLDA, three model-specific parameters have to be set: `chunks`, `memory`, and a threshold for vocabularies to be considered, `vocab.limit`. The memory parameter determines how much information from prior years is used to model the documents from the new publication year. Setting memory to a larger value has the effect of topics remaining rather stable, while smaller values let topic terms vary more from year to year. For the present corpus, years are the smallest available unit of time. Fixing all other parameters for RollingLDA, we inspect the results of setting `memory` to the last two years, the last year, and a random sample of 30% of last year's documents. While the random sample produce topics that are hard to interpret, using the documents from the last two years yield only minor changes in topic terms over time. Hence, as we were looking for flexibility while preserving the overall topic structure over time, we decide to use all last year's publications as memory for the RollingLDA topic assignments.

The vocabulary threshold controls which new terms are integrated into the overall vocabulary: Words that occur more than `vocab.limit` times in a minibatch are added, otherwise discarded for modeling the topics of the new publication year. We set it to ten, as we find this to be the best compromise of flexibility and computation time (after inspecting thresholds ranging from 5 to 25, cf. Strubell et al., 2019; Maier et al., 2020). The `chunks` parameter cuts the corpus into intervals, which is set to yearly updates in the present case. We inspect $K = 200, 250, \dots, 500$ (cf. Bittermann and Fischer, 2018), taking into account that modeling topic evolution will result in a lower total number of psychology topics in the RollingLDA model. The remaining parameters (α, η, and number of iterations) are set analogously to the LDAPrototype model for 2020 (cf. Sect. 3.3.1).

Another important parameter for the model evaluation is the date until which the documents are used for the initial model, because the RollingLDA updates are based on these initial topic structures. For a continuous tracking of scientific topics, we evaluate whether the topics evolve correctly in the long term. If the initial model is based on too little data, the RollingLDA might not be able to incorporate future changes adequately. Indeed, this is especially true when a scientific discipline has broadened its thematic spectrum over the years – which might be the case for psychology from the German-speaking countries: In PSYNDEX, the number of documents is rather low in the 1980s (cf. Bittermann, 2022, Fig. 14). This suggests that taking only documents from this period of time into consideration for the initial model won't provide enough information to let the RollingLDA evolve to the "actual" topics of 2020. Hence, we test several variants for the initial model, i.e., different starting points for RollingLDA, namely $1990, 1995, \dots, 2015$. All initial models start with the publication year 1980 and include terms that appear in at least 25 publications.

3.3.3 Model Comparisons

In total, we try seven values for K and six different starting years. The resulting $7 \times 6 = 42$ RollingLDAs are evaluated using the following criteria:

- Cosine similarity to the reference model,
- topic quality metrics, and
- external topic validation.

We consider similarity to the 2020 reference model as the most crucial factor, as it helps to assess whether sequential modeling can lead to topic results comparable to static modeling. Specifically, we compute the mean cosine similarity between all possible pairwise combinations of word distributions of the topics from the 2020 reference model and each rolling variant's 2020 topics. We decide to use cosine similarity as Rieger et al. (2021) propose this measure to be superior to other metrics for monitoring topic stability or topic self-similarities. In order to emphasize this first criterion, we select the five most similar RollingLDA model variants for subsequent analysis of topic quality and external validation of topic contents.

Despite being able to reflect the semantic contents of the "actual" 2020 topics, high quality topics are still an important issue. Hence, for topic quality metrics, we calculate semantic coherence and topic exclusivity (cf. Sect. 3.3.1). Maier et al. (2018) stresses the importance of topic validity. While intra-topic semantic validity (Quinn et al., 2010) via inspecting the top terms and most representative documents for each of the model variants is not feasible (especially w.r.t. change of top terms over time), we employ a strategy of external validation. Here, we use the concordance of topics with the database classification system (cf. Griffiths and Steyvers, 2004). For each topic, we determine the share of the APA classification categories (https://www.apa.org/pubs/databases/training/class-codes) in those publications where the topic was the overall most dominant one (i.e., document's topic probability > 0.5). By doing so, we retrieve a distribution of classification category shares for each topic, which we then correlate with the actual frequency distribution of these categories in the corpus metadata: The higher the resulting correlation coefficient, the more similar the category distributions of the RollingLDA variants are to the actual distributions. For determining the overall best fitting model, we standardize all values to z-scores and calculate the mean for each RollingLDA variant.

3.4 Shiny App, Term Lifting, and Topic Labels

Building upon the LDA-based Shiny App developed by Bittermann (2019), we design a novel

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).

user interface that visualizes RollingLDA topics while keeping it reasonably simple. In order to be both easy-to-use by novices and adaptable by the research community, we find R Shiny (Chang et al., 2021) to be a suitable solution: A slim user interface allows even users without programming skills to explore the topics, and the widespread R programming language (Muenchen, 2019) lets data analysts easily modify the app to their needs. Our topic app "PsychTopics" is updated quarterly, licensed as open source software, and made available on GitHub (`https://github.com/leibniz-psychology/psychtopics`).

In topic modeling, topics are characterized by groups of words that tend to co-occur. These so-called global top terms are determined according to the occurrence probabilities of the words over the entire time horizon. In addition, the RollingLDA approach lets topic terms vary over the years. In the PsychTopics app, we call these year-specific words evolution terms. Here, the occurrence probabilities of the words in the topic are determined for a specific year and weighted for disproportional occurrences in this topic compared to other topics (cf. Rieger et al., 2022a, Formula 9), which allows mapping particularly characteristic topic alignments in individual years. By distinguishing between global and year-specific evolution top terms, it is possible both to classify them in the global topic structure and to identify temporary shifts.

Since the absolute frequency and the exclusivity of a word for a specific topic can vary greatly, determining the overall theme of a topic is not trivial. To facilitate topic interpretation, we manually assign labels to the topics by adopting best-practice recommendations by Maier et al. (2018). Specifically, two researchers independently inspected the evolution of top terms, the most representative publications, and the most frequent journals that published articles on this topic. In addition, for each topic we take the most frequently observed classification categories into account. In case of topic shifts, i.e., new or diverging contents in the topic starting in a specific year, we assign arrows to the label. For instance, the topic label "Miscellaneous Disorders → Trauma" indicates that over the years, a rather broad topic on psychological disorders became specialized on trauma.

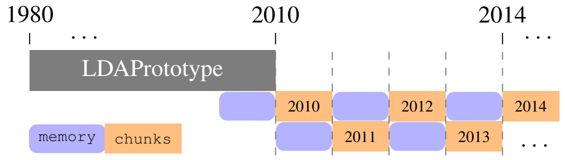


Figure 2: PsychTopics modeling scheme for the best fitting model (start = 2010).

4 Analysis

The five model variants with highest cosine similarity to the reference model (cf. Sect. 3.3.2 and 3.3.3) comprise either 200 or 300 topics, while their RollingLDA starting years ranged from 1995 to 2015. Table 1 shows the metrics used for comparison. The cosine similarities are rather close, but the variants differ in topic quality metrics (especially exclusivity) and correlations with the metadata classification categories. The five models' overall high correlation coefficients (0.95 to 0.97) underline their high external validity. The mean z-scores indicate that the variant with $K = 200$ topics and the starting year of 2010 for RollingLDA is the overall best fitting model (cf. Figure 2), so we choose this for integration in the topic app. All analysis scripts were executed in R (R Core Team, 2022) and can be found in the supplementary material.

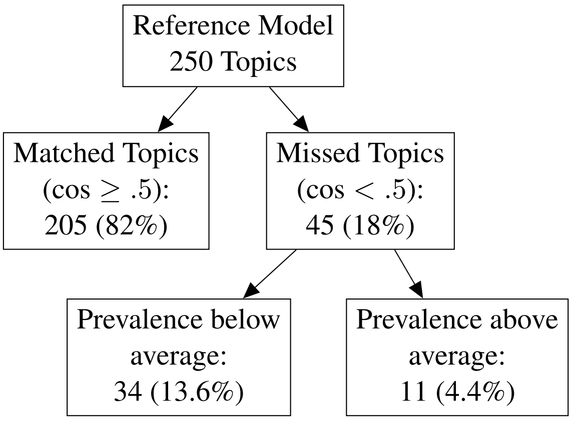


Figure 3: Matched and missed topics of the reference model for the best fitting model ($K = 200$, start $=$ 2010).

4.1 Matched and Missed Topics

The best fitting model ($K = 200$, start $= 2010$) is not perfectly aligned to the reference model (cos $= .62$), which is not surprising, as the number of topics in the models differ (200 vs. 250) and as the variants are initialized with data from 1980 to 2009. The individual topic similarities range from .30 to .91 ($\sigma = .13$, $x_{0.25} = .52$, $x_{0.5} = .62$, $x_{0.75} = .72$). Of the 250 topics in the reference model, 45 (18%) get a similarity value of less than .5, realizing prevalences $\theta_{m,k}$ ranging from .19% to .46%, with 11 topics having a prevalence above the model's average ($1/K = 1/250 = 0.4\%$). That is, 205 (82%) topics can be detected satisfactorily by the RollingLDA, whereas eleven (4.4%) of the more prevalent topics in 2020 are missed as individual topics (cf. Figure 3). Despite being not matched satisfactorily, characteristic terms of these topics (e.g., dreams, climate, tinnitus) can be found in other topics, so these themes are not lost, but just less prevalent. The remaining 34 (13.6%) topics are negligible due to their low prevalence in the reference model.

A moderate correlation between cosine similarity and topic prevalence in the reference model ($r = .34$) indicates that topics without match in the variant model (i.e., low similarity) have the tendency to be less prevalent. Indeed, nine of the ten most common topics in the reference model (e.g., psychotherapy, psychoanalysis, mental disorders, memory, group therapy), can be matched to the most similar variant topics (ranging in cosine similarity from .64 to .88). The only exception is a topic on refugee psychotherapy. The highest value of cosine similarity has a variant topic on psychotherapy. Nevertheless, six refugee-related topics are included in the variant model, however, scoring lower as they focus on refugees in context of trauma, COVID-19, social issues, or health services. In the supplementary material, we provide tables with global top terms of the reference and evolution terms of the variant model, as well as a table including the cosine similarities.

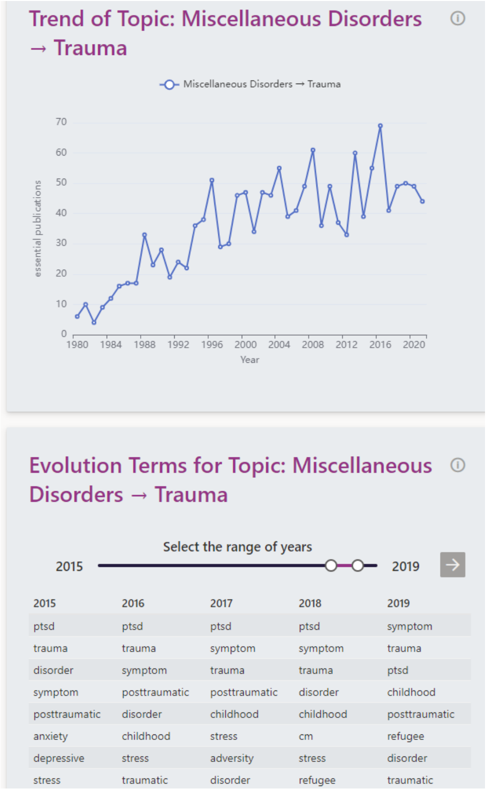
4.2 Topic Interpretability and Topic Shifts

Focusing on the variant's 200 topics, there is one topic to be too diverse for a coherent interpretation (global top terms: "theory, social, process, model, concept, behavior, development, psychology, group, system"). These are rather generic terms in psychological research, which is why we regard this as a "background topic". For 20 (10%) topics, top terms vary within an overarching theme (e.g., "Miscellaneous Disorders") and/or within a specific period in time (e.g., "Miscellaneous Disorders → Trauma"). In total, shifts are found for 34 (17%) topics, while the remaining 83% of all topics evolve within the same semantic scope. In nine cases (4.5%), topic shifts are limited to a relatively close semantic space (e.g., "Child Psychopathology → Trauma") or refined the topic (e.g., "Experimental Psychology → Decision Making"). Eight (4%) topics "disappear", as their top terms over time become too diverse for coherent interpretation (e.g., "Learning Environments → Miscellaneous"). Interestingly, for 17 topics "hard shifts" can be detected, as their their top terms change drastically (e.g., "Psychoanalysis → COVID-19"). Such shifts reflect the RollingLDA model's ability to integrate rising topics (e.g., COVID-19) and to neglect declining topics. This finding does not mean that these topics became irrelevant to the scientific community; rather, they are subsumed under broader topics or they no longer contribute to the main research topics of the field.

4.3 Topic App

Our associated app is called "PsychTopics" (`https://abitter.shinyapps.io/psychtopics/`) and features

- "Start" – a general overview of the overall most prevalent topics as well as the preliminary topics of the current year,
- "Browse Topics" – a detailed list of topic characteristics (such as the number of essential publications or the share of empirical research within these publications),

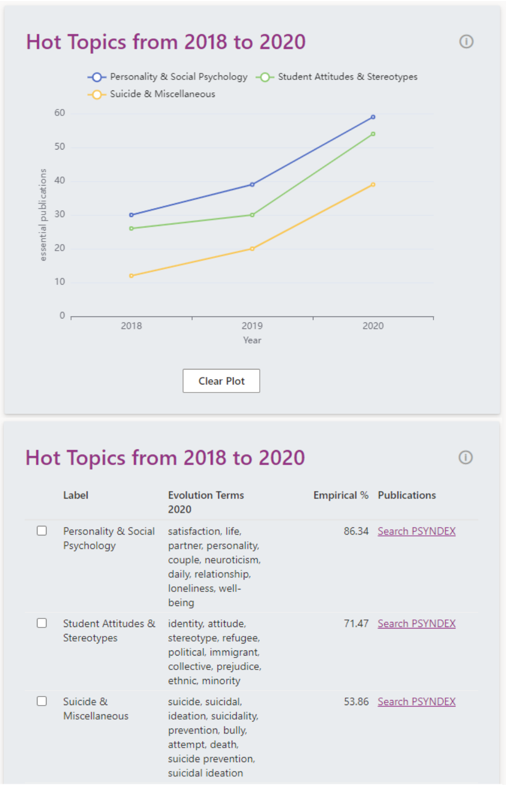

Trend of Topic: Miscellaneous Disorders → Trauma
Miscellaneous Disorders → Trauma
essential publications
70
60
50
40
30
20
10
0
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
Year

Evolution Terms for Topic: Miscellaneous Disorders → Trauma

Select the range of years
2015
2019

2015	2016	2017	2018	2019
ptsd	ptsd	ptsd	ptsd	symptom
trauma	trauma	symptom	symptom	trauma
disorder	symptom	trauma	trauma	ptsd
symptom	posttraumatic	posttraumatic	disorder	childhood
posttraumatic	disorder	childhood	childhood	posttraumatic
anxiety	childhood	stress	cm	refugee
depressive	stress	adversity	stress	disorder
stress	traumatic	disorder	refugee	traumatic

Figure 4: Screenshot of the evolution for topic “Miscellaneous Disorders → Trauma”.


Hot Topics from 2018 to 2020
Personality & Social Psychology
Student Attitudes & Stereotypes
Suicide & Miscellaneous
essential publications
60
50
40
30
20
10
0
2018
2019
2020
Year
Clear Plot

Hot Topics from 2018 to 2020

	Label	Evolution Terms 2020	Empirical %	Publications
☐	Personality & Social Psychology	satisfaction, life, partner, personality, couple, neuroticism, daily, relationship, loneliness, well-being	86.34	Search PSYNDEX
☐	Student Attitudes & Stereotypes	identity, attitude, stereotype, refugee, political, immigrant, collective, prejudice, ethnic, minority	71.47	Search PSYNDEX
☐	Suicide & Miscellaneous	suicide, suicidal, ideation, suicidality, prevention, bully, attempt, death, suicide prevention, suicidal ideation	53.86	Search PSYNDEX

Figure 5: “Hot” topics with the greatest publication gradient between 2018 and 2020.

- “Popular by Year” – the most prevalent topics for a specific year,
- “Hot/Cold” – the topics with the largest increase or decrease in publications,
- “Topic Evolution” – the evolution of lifted top terms across publication years, and
- “Methods” – describing technical details and links to further literature.

Figure 4 shows a screenshot of the “Topic Evolution” view with the example of the topic “Miscellaneous Disorders → Trauma”. The line chart depicts the number of essential publications (i.e., $\theta_{m,k} > .5$) for this topic over time. The table below the chart lists the “evolution terms” for the years 2015 to 2019. The topic is less prevalent in the 1980s and at the same time more characterized by publications addressing neurological conditions, schizophrenia, and depression in a more general way. Over the years, and especially from 2001 onwards, there has been a greater specialization of the topic. This topic shift is accompanied by a more prominent appearance of the terms “posttraumatic”, “PTSD” and “trauma”, from 2012 additionally “childhood” and from 2018 additionally “refugee”. In the German-speaking countries, psychology has increasingly addressed the topic of “flight and migration” as a result of the so-called “refugee crisis” in 2015 (Bittermann and Klos, 2019b). A time lag in the appearance of the topic can be explained by a “publication lag” between the initial study idea and the publication of the paper (cf. Björk and Solomon, 2013).

Besides inspecting the evolution of topics, another way to use PsychTopics is to examine trends in the research literature. The “Hot/Cold” view in Figure 5 shows the topics with the strongest rising and the strongest falling linear trend (cf. Griffiths and Steyvers, 2004). Here it can be seen that between the years 2018 and 2020 “Personality & Social Psychology” is the hottest topic. By clicking on the respective points of the lines in the diagram, details of the topics can be accessed. Moreover, clicking the “Search PSYNDEX” link automatically queries the evolution terms in the PubPsych portal and provides relevant publication references.

5 Discussion

In this paper, we applied RollingLDA to a continuously growing corpus of scholarly documents. Using the field of psychology as an use case, we found that RollingLDA is capable of integrating the annual updates of the database to meaningful topics. The framework can be easily applied to any scientific discipline or even to multiple fields. For this, the text input should at least consist of titles and abstracts. In addition, we recommend controlled keywords (e.g., MeSH terms), as they provide the main contents of the articles in a standardized manner. Regarding metadata, we used the year of publication, the classification category, and the study methodology (e.g., empirical research, theoretical discussion). This allows to analyze temporal trends, to validate topic contents, and to highlight topics that might be suitable for meta-analyses. However, our approach is not limited to these metadata and many other additions are conceivable. For instance, the share of open access articles or study preregistrations over time could be compared between topics and research fields. The model is implemented as a Shiny App that lets users explore and analyze the topics and trends without the need of programming skills, while the open source code facilitates the mentioned modifications to the PsychTopics app.

5.1 Practical Implications

The PsychTopics app encourages exploration and thus provides an overview of the variety of scientific publications to researchers, students, policymakers, and the interested public. For journalists and policy-makers, it might be of interest to determine the extent to which publications address topics of social relevance. A corresponding topic in PsychTopics is "Psychology & Society", which is increasingly dedicated to climate change from 2019. The hyperlink to the free literature search in PSYNDEX helps students in finding reading material for class. Furthermore, PsychTopics lists the three journals that have published the most on the topics. This can guide early career researchers in finding suitable journals for their own research papers. In addition, the proportion of empirical studies indicates topics that be suitable for quantitative research syntheses (meta-analyses). In particular, hot topics with very high publication activity and a large share of primary studies may be of relevance for living research syntheses (e.g., Burgard et al., 2022) to keep the meta-analytic evidence as up-to-date as possible.

5.2 Limitations and Further Research

Like most topic modeling techniques, the presented approach focuses on texts written in the English language, but is easily adaptable to other monolingual corpora. In contrast, multilingualism in topic modeling can lead to different topics despite the same content (e.g., English "Therapy" topic and German "Therapie" topic) or lower the semantic coherence of topics (Mimno et al., 2011). Hence, the handling of multilingual text input in sequential modeling of dynamically growing corpora represents a target for future research (e.g., based on Mimno et al., 2009; Vulić et al., 2015).

Topic shifts, i.e., changes in top terms over the years that imply the ending of the prior and the beginning of a new topic, were detected manually and indicated in the topic labels using an arrow symbol. For instance, "Experimental Psychology → Decision Making" means that the topic became more specialized over the years. Topics with an abrupt shift to completely different contents (e.g., "Psychoanalysis → COVID-19") are split into separate topics in the app. In this way, misleading interpretations of topic names are avoided (such as psychoanalysis became concerned with COVID-19). However, the different types of changes (e.g., abrupt, flowing) remain to be investigated. Moreover, the current manual detection of shifts is labor intensive. This process could be automated by change detection within topics (cf. Rieger et al., 2022b).

It is methodologically interesting to split topics including shifts into two temporal topics, so that the model would have a dynamic number of topics over time. Naturally, it is reasonable to assume that some years of research lead to more different topics, others to less. An approach for a dynamic number of topics might be to delete topics from the initialization of a following minibatch that are characterized by both few document assignments and incoherent top words. This specific topic would end, and the empty topic "slot" could develop a new topic. Unless this newly emerged topic develops a coherent context in the following minibatch, the topic would be neglected. However, as soon as it develops its own meaning, it is taken up as a new topic and also detached from the previous meaning, so that it is considered as an individual topic for the interpretation.

We tested a total of 42 RollingLDA variants, using different settings for the number of topics and starting years of the sequential RollingLDA modeling. We found 200 topics and an initialization model for the publication years 1980 to 2019 yielding the best results in terms of evolving to topics in 2020 comparable to a single 2020 reference model. As we argued, our corpus shows a strong increase in publication volume during the 1980s with a steady increase onwards (cf. Bittermann, 2022, Fig. 14). Other research fields might show a different pattern in publication activity over the years, making different parameters necessary. Thus, the generalizability of the specific model parameters presented might be limited, but our framework and model selection procedure can give guidance to find the best parameters for an application to other corpora of scholarly documents.

The transfer of the framework to other domains requires the major manual effort for the initial preparation of the model. During the routine updates there is some monitoring effort (e.g., whether new subtopics have emerged, whether topics have strongly mutated), which can be kept to a minimum by automated procedures. Optimal model parameters (in particular K, `init`, `memory`) for other domains will depend on the publication volume over time, the desired update intervals and the topical variety of the modeled texts. With our proposed procedure for finding the optimal parameters (cf. Sect. 3.3.3 and Table 1), the resulting manual effort can also be kept to a minimum.

5.3 Conclusion

Taken together, RollingLDA is a suitable method for an ongoing monitoring of scientific topics. It is capable of reducing information overload by summarizing a plethora of publications by means of their main topics. A major benefit of the presented framework is the high degree of automation once the initial model is created. Updates can be produced efficiently and thus timely with regard to runtime and manual effort. Importantly, the model integrates new publications while keeping time series of topic trends consistent. This, in contrast to standard LDA methods, can help various stakeholders like researchers or policy makers to evaluate how fields of research evolve over time. The presented topic app makes these insights easily accessible.

Acknowledgements

The present study is partly supported by the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University.

References

Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. 2016. Streaming-LDA: A copula-based approach to modeling topic dependencies in document streams. In *Proceedings of the 22nd SIGKDD-Conference*, pages 695–704. ACM.

Iana Atanassova, Marc Bertin, and Philipp Mayr. 2019. Editorial: Mining scientific papers: NLP-enhanced bibliometrics. *Frontiers in Research Metrics and Analytics*, 4(2).

Shir Aviv-Reuven and Ariel Rosenfeld. 2021. Publication patterns' changes due to the COVID-19 pandemic: a longitudinal and short-term scientometric analysis. *Scientometrics*, 126:6761–6784.

André Bittermann. 2019. Development of a user-friendly app for exploring and analyzing research topics in psychology. In *Proceedings of the 17th Conference of the International Society for Scientometrics and Informetrics*, pages 2634–2635. Edizioni Efesto.

André Bittermann. 2022. Publikationstrends der Psychologie zu Themen gesellschaftlicher und fachlicher Relevanz: Juni 2022. *ZPID Science Information Online*, 22(2).

André Bittermann and Andreas Fischer. 2018. How to identify hot topics in psychology using topic modeling. *Zeitschrift für Psychologie*, 226(1):3–13.

André Bittermann and Eva Maria Klos. 2019a. Code zu: "Ist die psychologische Forschung durchlässig für aktuelle gesellschaftliche Themen? Eine szientometrische Analyse am Beispiel Flucht und Migration mithilfe von Topic Modeling". *PsychArchives*.

André Bittermann and Eva Maria Klos. 2019b. Ist die psychologische Forschung durchlässig für aktuelle gesellschaftliche Themen? *Psychologische Rundschau*, 70(4):239–249.

Bo-Christer Björk and David Solomon. 2013. The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, 7(4):914–923.

David M. Blei. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(224).

Tanja Burgard, Michael Bosnjak, and Robert Studtrucker. 2022. Psychopen cama: Publication of community-augmented meta-analyses in psychology. *Research Synthesis Methods*, 13(1):134–143.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS: Advances in Neural Information Processing Systems*, volume 22, pages 288–296. Curran Associates Inc.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. *shiny: Web Application Framework for R*. R package version 1.7.1.

Alexander Christ, Marcus Penthin, and Stephan Kröner. 2019. Research general stop words for: Big data and digital aesthetic, arts and cultural education: Hot spots of current quantitative research. *PsychArchives*.

Giovanni Colavizza, Rodrigo Costas, Vincent A. Traag, Nees Jan van Eck, Thed van Leeuwen, and Ludo Waltman. 2021. A scientometric overview of CORD-19. *PLOS ONE*, 16.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 NAACL-Conference*, pages 3824–3848. ACL.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 EMNLP-Conference*, pages 4846–4853. ACL.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Martin Hilbert and Priscila López. 2011. The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Lee Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *NeurIPS: Advances in Neural Information Processing Systems*.

John P. A. Ioannidis, Maia Salholz-Hillel, Kevin W. Boyack, and Jeroen Baas. 2021. The rapid, massive growth of COVID-19 authors in the scientific literature. *Royal Society open science*, 8(9).

Günter Krampen. 2016. Scientometric trend analyses of publications on the history of psychology: Is psychology becoming an unhistorical science? *Scientometrics*, 106:1217–1238.

Daniel Maier, Andreas Niekler, Gregor Wiedemann, and Daniela Stoltenberg. 2020. How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research*, 2(2).

Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. 2018. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 EMNLP-Conference*, pages 880–889. ACL.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 EMNLP-Conference*, pages 262–272. ACL.

Robert A. Muenchen. 2019. The popularity of data science software [blog post]. Accessed 2022-07-04.

Andreas Niekler and Patrick Jähnichen. 2012. Matching results of latent Dirichlet allocation for text. In *Proceedings of ICCM*, pages 317–322.

Gabriela C. Nunez-Mir, Basil V. Iannone III, Keeli Curtis, and Songlin Fei. 2015. Evaluating the evolution of forest restoration research in a changing world: a "big literature" review. *New Forests*, 46:669–682.

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2021. RollingLDA: An update algorithm of latent Dirichlet allocation to construct consistent time series from textual data. In *Findings Proceedings of the 2021 EMNLP-Conference*, pages 2337–2347. ACL.

Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2022a. LDAPrototype: A model selection algorithm to improve reliability of latent Dirichlet allocation. *Preprint available at Research Square*.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. 2022b. Dynamic change detection in topics based on rolling LDAs. In *Proceedings of the Text2Story'22 Workshop*, volume 3117 of *CEUR-WS*, pages 5–13.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70. ACL.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th ACL-Conference*, pages 3645–3650. ACL.

Arho Suominen and Hannes Toivanen. 2016. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10):2464–2476.

Lisa Gallagher Tuleya, editor. 2007. *Thesaurus of psychological index terms*, 11th edition. American Psychological Association.

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147.

Chong Wang, David M. Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *Proceedings of the 24th UAI-Conference*, pages 579–586. AUAI.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th SIGKDD-Conference*, pages 424–433. ACM.

Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. TM-LDA: Efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th SIGKDD-Conference*, pages 123–131. ACM.

Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. 2014. Clustering scientific documents with topic modeling. *Scientometrics*, 100:767–786.

Ke Zhai and Jordan Boyd-Graber. 2013. Online latent Dirichlet allocation with infinite vocabulary. In *Proceedings of the 30th ICML-Conference*, Proceedings of Machine Learning Research, pages 561–569. PMLR.

A Supplementary Material

The analysis code and the mentioned topic and similarity tables are provided on GitHub (`https://github.com/abitter/sdp22_supplements`)