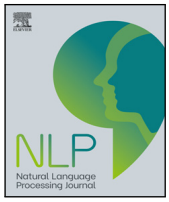
Natural Language Processing Journal 6 (2024) 100044

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp



Benchmarking topic models on scientific articles using BERTeley ☆

Eric Chagnon *, Ronald Pandolfi, Jeffrey Donatelli, Daniela Ushizima

Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, 94720, United States

ARTICLE INFO

Keywords:
NLP
Topic modeling
Scientific articles
Transformers

ABSTRACT

The introduction of BERTopic marked a crucial advancement in topic modeling and presented a topic model that outperformed both traditional and modern topic models in terms of topic modeling metrics on a variety of corpora. However, unique issues arise when topic modeling is performed on scientific articles. This paper introduces BERTeley, an innovative tool built upon BERTopic, designed to alleviate these shortcomings and improve the usability of BERTopic when conducting topic modeling on a corpus consisting of scientific articles. This is accomplished through BERTeley's three main features: scientific article preprocessing, topic modeling using pre-trained scientific language models, and topic model metric calculation. Furthermore, an experiment was conducted comparing topic models using four different language models in three corpora consisting of scientific articles.

1. Introduction

Being able to discover the underlying themes and patterns in a collection of documents is necessary in order to parse the overwhelming amount of information that is available today. Topic modeling is a Natural Language Processing (NLP) technique that was designed specifically for this task and has found applications in various domains, ranging from social media analysis (Ramondt et al., 2022) to content recommendation systems (Bansal et al., 2017) to transportation safety analysis (Oliaee et al., 2023).

Early topic models utilized a statistical approach with popular models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) which is one of the most used topic modeling techniques to this day. These early statistical approaches rely on the assumption that a document is a bag of words or an unordered collection of words. Although this has the benefit of simplifying the topic modeling process, the context of the words is lost, and thus the resulting topics may not be the most accurate representation of the true latent topics within the document.

The introduction of the transformer architecture (Vaswani et al., 2017) marked a fundamental shift in topic modeling approaches. The contextual document embedding enabled by the transformer architecture creates the opportunity for a topic model that can utilize the context of a word in a way that statistical topic models cannot. Most transformer-based approaches to topic modeling follow a similar structure: create document embeddings via a transformer then cluster the document embeddings to create topics, then use a word weighting scheme to extract the topic words from the topics.

BERTopic, a topic modeling technique that uses bidirectional encoder from transformers (BERT), was introduced in early 2022 and presented a highly modular and customizable workflow for transformer-based topic models (Grootendorst, 2022). In addition to the workflow a new word weighting scheme, the class-based Term Frequency - Inverse Document Frequency (cTF-IDF), was introduced. BERTopic was compared to two other transformer-based techniques and two statistical-based techniques, including LDA, on three different corpora. While BERTopic did not have the best topic modeling metrics on every corpus, it remained competitive with the other state-of-the-art transformer-based topic models while consistently outperforming the statistical-based topic models. The strong topic modeling metrics along with the inherent flexibility of BERTopic due to its modularity make it ideal for topic models that want to iterate as new improvements are released.

When conducting topic modeling on a corpus consisting of scientific articles a unique challenge becomes apparent. The inherently stringent formatting requirements of scientific articles cause the final topics to be riddled with structural words that dilute the potential learnings about the corpus such as *abstract*, *introduction*, *appendix*, etc. Similarly, scientific articles are all documents with the same goal: to present the findings of a research project. Thus, words that are used to convey this goal have an abnormally high frequency that once again dilute potential topic words that accurately represent the latent topics in the corpus such as *use*, *study*, *show*, etc. Although conventional topic modeling approaches may offer satisfactory outcomes at times, there exists a substantial opportunity for enhancement by tailoring techniques to accommodate the intricacies inherent in applying NLP to

☆ This document is the results of the research project funded by the National Science Foundation.

* Corresponding author.
E-mail address: echagnon@lbl.gov (E. Chagnon).
URL: https://camera.lbl.gov (E. Chagnon).

https://doi.org/10.1016/j.nlp.2023.100044
Received 19 October 2023; Received in revised form 21 November 2023; Accepted 27 November 2023

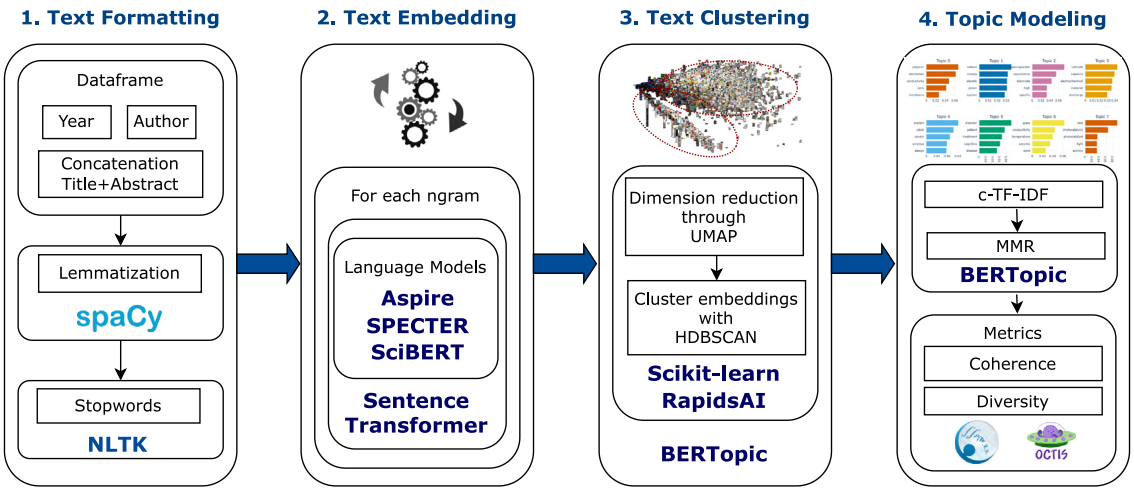


Fig. 1. Diagram of the BERTeley workflow.

scientific articles, as described in Section 2. Addressing this challenge is paramount to advance the effectiveness of topic modeling within the realm of scientific discourse (Hall et al., 2008).

This paper presents BERTeley, a Python package built upon BERTopic to address This unique challenge. Through modifications and enhancements to the BERTopic methodology, BERTeley provides a simple, yet powerful tool for topic modeling on a corpus consisting of scientific articles. This tool would allow users to expedite the literature review process on a project in a new domain that may be outside their area of expertise. When able to understand the types of content that is published in the area, researchers can focus on understanding documents in the relevant areas.

2. Software description

BERTeley has three main features: scientific article preprocessing, topic modeling, and metric calculation. BERTeley's preprocessing suite addresses the aforementioned challenge by removing specific words from the input data. The topic modeling and metric calculation add quality-of-life features such as pre-selected language models trained specifically on scientific articles, and one-line topic modeling metric calculation. The BERTeley workflow can be seen in Fig. 1. BERTeley can be installed by running the command `pip install berteley` at the command line (Eric Chagnon and Ushizima, 2023).

2.1. Preprocessing

To ensure that the resulting topics reflect the latent theme within the corpus, it is essential to filter out high-frequency but low-value words, which are words that do not provide information related to the content of the document. After reviewing the results of different topic models over various corpora, a curated stopword list was created that includes both structural words and common words found in nearly every scientific article multiple times. To maximize the recall of stopword removal, preprocessing steps are carried out, particularly the lemmatization step, which reduces words to their base or dictionary form to normalize variations and facilitate analysis. The key steps in the preprocessing suite are listed below:

- remove empty strings,
- remove html tags,
- expand all contractions,
- remove all punctuation,
- remove excess whitespace created by the previous step(s),
- lemmatization,
- standard and scientifically-irrelevant stopword removal,
- removal of documents with length smaller than 10 tokens.

The operations in each step are offered as individual routines, allowing users to execute a subset of the available preprocessing operations as needed. They are also packaged in a single routine, `preprocess`, if a user wishes to utilize the entire preprocessing suite. If users are working on a machine with sufficient memory, then the `preprocess_parallel` function is recommended instead, as parallelization greatly reduces the time to preprocess large corpora. While parallelization does decrease the runtime of the preprocessing overall, the lemmatization step can be time-consuming if the corpus size and average document length are large as each word in the corpus is tokenized in this step. For users with less domain knowledge about their corpus, there is an option to remove abbreviations from the final results, as acronyms may be confusing without proper context.

2.2. Topic modeling

The topic modeling step uses BERTopic as the topic modeling engine. Here, the BERTopic package was modified to expose a few of its key hyperparameters to the users. Specifically, the selection of the language model can have a significant impact on the final results. Deciding what language model to use (and possibly also deciding on its size) can be a tedious process that is filled with technical jargon that some users may not be familiar with.

BERTeley includes additional capabilities that present users with a choice of three language models pre-trained specifically on scientific articles: Specter (Cohan et al., 2020), Aspire (Mysore et al., 2021), and SciBERT (Beltagy et al., 2019). BERTopic requires little to no setup when the selected language model comes from the SentenceTransformer library (Reimers and Gurevych, 2019). However, the Aspire and SciBERT language models are not available in this package, so several extra steps are required to obtain a model in the proper form to interface with BERTopic. Instead of requiring the user to take these steps, we offer a flexible hyperparameter for the language model. The language model argument can take in either a `SentenceTransformer` object if the user wishes to use a different language model, or it can also take in a string containing the name of one of the pre-selected models. If a proper string is passed, the model is downloaded and built for the user ready to interface with BERTopic.

2.3. Metric calculation

Topic modeling can be subjective in nature and typically requires human input to judge whether or not the results are acceptable. However, this does not imply that topic modeling metrics are useless, as

they can provide insight when directly comparing topic models with a large number of topics. Human evaluation paired with topic modeling metrics is the recommended course of action for topic modeling evaluation (Doogan and Buntine, 2021).

The most common metrics used to evaluate topic models are Topic Coherence and Topic Diversity (Röder et al., 2015). Topic Coherence is a measure of the word similarities of the top words within a given topic (Rosner et al., 2014), and it works as a coefficient to gauge intra-cluster correlation. This measure ranges from $[-1, 1]$ where 1 indicates a perfect correlation between the topic words and -1 indicates that the topic words are not related at all. There are several variations of Topic Coherence which use different formulas for calculating the metric (Lisena et al., 2020). Here we use the C_v measure as it has the highest correlation with human interpretation when evaluating topic models. This metric calculates the normalized pointwise mutual information (npmi) coherence value over all the topic words within a sliding window. Then the cosine similarity between these npmi values is calculated as the C_v measure (Rosner et al., 2014).

Topic Diversity indicates the percentage of unique topic words and measures the repetitiveness of a topic model, with values ranging from $[0, 1]$, with 1 indicating that all topic words are unique and 0 indicating that there are no unique topic words (Dieng et al., 2019). This metric works as a coefficient to gauge inter-cluster correlation. The `Octis` library provides a wealth of topic modeling metrics and means to calculate them. These topic modeling metrics can be calculated in BERTeley for a given topic model using the `calculate_metrics` function.

3. Topic modeling experiment

In this section, we detail an experiment with BERTeley, on topic models using different language models to create the document embeddings. Here, the scientific language models Specter, Aspire, and SciBERT, along with a generic language model, all-MiniLM-L6-v2 (MiniLM), were used to create document embeddings for BERTeley. The experiment consists of three use cases on scientific articles with varying corpus sizes. The document sizes in each of these corpora should be considered to be relatively equal since the documents themselves were made up of a concatenation of the titles and abstracts of scientific articles.

3.1. Data sources

The first use case consists of articles published at the Advanced Light Source (ALS) at Lawrence Berkeley National Lab. The articles are separated by the respective beamline used in the publication (alspubs, 2022). Overall there are data for 47 beamlines each with the number of documents ranging from 17 to 1167. The second use case is larger and consists of articles retrieved using the Springer API (Nature, 2023b). The search parameters consisted of a start year of 2010, an end year of 2022, and a search term of *lithium ion battery*. The corpus consisted of 9,142 documents. The final use case consists of 2,187,967 articles from the Arxiv dataset maintained by researchers at Cornell (arXiv.org submitters, 2023). This use case is far larger than the rest, and it was intended to see how the different language models perform topic modeling at scale.

3.2. Experimental design

Language models have been created to complete a variety of scientific tasks, such as completing citations (Hinde and Spackman, 2015) and identifying named entities in articles such as proteins or chemical compounds (Eltyeb and Salim, 2014). In order to do this, these language models are trained solely on scientific articles. Theoretically, when trained on scientific articles, these tailored language models are able to produce semantically superior embeddings that traditional language models cannot because of their lack of exposure to the nuanced and sometimes complex content contained within scientific articles compared to traditional textual data.

The goal of this experiment is to try and determine how the embeddings created by different language models affect the results of topic modeling on scientific articles. The models being compared are Specter, Aspire, and Scibert, all of which were trained on scientific articles, and MiniLM, which is a more generic language model and is the default embedding model in BERTopic. The performance of the topic models will be measured by two topic modeling metrics: topic coherence and topic diversity.

3.3. Procedure

In the first two use cases, BERTeley's default hyperparameters worked appropriately as the corpora were within thousands of articles. In those cases, the experiment was carried out on a dedicated desktop with an Intel Xeon W-2245 CPU at 3.90 GHz, 128 GB RAM, and an NVIDIA GeForce RTX 3080Ti; furthermore the first two use cases utilized BERTeley's default hyperparameters.

However, the Arxiv dataset consisted of over 2 million documents, and as a result hyperparameters were tuned and different GPU-accelerated versions of the dimension reduction and clustering algorithms to conduct the analysis at this scale. The main bottleneck of this experiment is the UMAP and HDBSCAN steps, as they are used by default using the sklearn implementation that utilizes the CPU of the machine (Pedregosa et al., 2011). Rapids AI created GPU-accelerated versions of these algorithms that greatly reduced the training time of the topic models.

In order to make use of these functions, slight modifications to the experiment had to be made. Substituting these functions is a complicated process that involves the creation of a specific environment, making it difficult to distribute. As a result the functionality for adding these GPU-accelerated methods to BERTeley was not developed. To carry out the experiment using these methods, a BERTopic object was used instead of a BERTeley object. The preprocessing and metric calculation steps contain code that is verbatim to their implementation in BERTeley, but in this case, they were not used as part of the package and were instead called on their own.

Along with these new components replacing their CPU counterparts, some hyperparameters were adjusted for corpus size (Raschka et al., 2020). Most importantly, the HDBSCAN parameter `min_cluster_size` which decides the minimum number of data points in a group that can be considered a cluster. When the corpus is of this size, having topics with only a small amount of documents in them is not helpful for learning about the corpus. This value went through a few iterations, and the value of 100 provided both reasonable topic sizes and did not produce significant memory issues. This solves the issue of large computation times, but with a corpus this large, the default hyperparameters also need to be changed, specifically the minimum number of documents in a cluster for it to be considered a topic. With a relatively small dataset, topics with only a handful of documents are reasonable, but with a dataset of millions of documents, it is not ideal to have topics that only have a few dozen documents in them, as it creates too many unnecessary topics. This is solved by increasing the min_cluster_size in Rapids AI cuml HDBSCAN to 100. This value was determined after testing large values starting at around 500 until the topic model produced both quality results and did not have memory issues. In addition to the changes to the algorithm itself, this project used the resources of the National Energy Research Scientific Computing Center (NERSC). Most of the computations took advantage of NERSC Perlmutter, more specifically, a single 4-GPU Perlmutter node, using the distributed data-parallel functionality in PyTorch (Nature, 2023a).

Table 1
BERTeley results for ALS Beamline 8.3.2.

Language model	Coherence	Diversity	Number of topics
Specter	0.6920	0.96	5
Aspire	0.7163	0.95	8
SciBERT	0.5707	1	4
MiniLM	0.7433	0.9333	9

4. Results

4.1. Use Case 1: ALS

Considering that the Advanced Light Source (ALS) offers multiple beamlines, our choice has been made with a specific one: beamline 8.3.2. This selection is primarily attributed to the fact that this beamline boasts a collection of 325 articles authored by beamline users from distinct backgrounds, ranging from biologists to civil engineers. Despite using default hyperparameters for topic modeling, different language models produced diverse outcomes. The topic coherence and diversity metrics can be seen in Table 1. Ideally, the topic model would create topics that have a clear theme, and contain words that are related, while avoiding significant overlap between topics. Despite using default hyperparameters for topic modeling, different language models produced diverse outcomes. A topic model that achieves both of these will have both coherence and diversity values close to 1. According to Table 1, three of the topic models perform similarly in terms of coherence, while the topic model with the SciBERT language model has a slightly worse performance. All of the models have a high diversity score which is expected when the number of topics is low.

As an example, Fig. 2 shows the bar charts for the nine most relevant topics when using MiniLM as the embedding model to create the document embeddings. Each bar chart lists 5 topic words on the y-axis and their respective c-TF-IDF scores on the x-axis. Observing the top themes makes it possible to understand how the instrument was used in most of those articles. For example, Topic 1 evidences instrument users focused on imaging plants and their resilience to drought, Topic 2 clusters articles about soil experiments in which one measures permeability and pore structure, and Topic 3 relates to users studying fuel cells, where water acts as both a transport medium and a catalyst within the proton exchange membrane, facilitating the efficient movement of ions across the layers. Topic 4 concentrates text from users imaging bones to check for toughness and properties associated with fractures. Topic 5 groups work on battery quality control where studies focus on detecting dendrites that can potentially pierce the electrolyte and connect the electrodes. Topic 6 refers to users studying ceramic matrix composites reinforced with fibers to mitigate cracks. Topic 7 shows the most general topic words associated with the instrument, which uses synchrotron X-ray to acquire 3D image data. Topic 8 lists terms associated with archeological concrete samples that contain volcanic ash. Finally, Topic 9 shows evidence that the instrument was also used to study arthropod animals.

The distinctions in the outcomes of these topic models can be better understood by examining the topics themselves. Bar charts representing the topic models for Specter, SciBERT, and Aspire are provided in Appendix. One way to assess these barcharts is to compare the prevalence of certain topics in each graph. This means that different topic models capture the same topic(s). In this case, there are several topics that have been manually labeled, which overlap between all topic models, as illustrated in Table 2. Here, the MiniLM and Aspire topic models manage to capture all six shared topics, with Specter capturing four, and SciBERT only capturing two.

Based on this new information, it is likely that the MiniLM or Aspire topic models are the best for this use case. However, the slightly higher coherence score of MiniLM does not necessarily imply that it is better than the Aspire topic model. Each topic model has some extra topics as well; these can be unique topics that are only captured by one topic model; an example of this is Topic 3 in Fig. 2 with the water topic, or topics that are generic and do not provide much information (topic 7 in Fig. 2). When these additional topics are not repetitive they positively contribute to the coherence values of the model without providing insights into the corpus. This is where a trade-off between topic models is introduced: if one wishes to capture only certain topics, then a topic model that captures all the desired topics but has an overall lower number of topics may prove beneficial. On the other hand, if one wishes to see a broader view, then a topic model with more topics would likely be superior, but this could potentially assign documents to topics that do not provide relevant information to the user. Combining the topic modeling metrics along with the priorities of the use case will allow for a heuristic approach on which language model will provide the best results when used for topic modeling.

Table 2
ALS common topics.

Language model	Plant	Bone	Battery	Spider	Materials science	Volcanoes
Specter	X	X	X	X	–	–
Aspire	X	X	X	X	X	X
SciBERT	X	X	–	–	–	–
MiniLM	X	X	X	X	X	X

Table 3
Springer lithium ion battery results.

Language model	Coherence	Diversity	Number of topics
Specter	0.6284	0.7565	86
Aspire	0.5824	0.7222	108
SciBERT	0.4886	0.6875	80
MiniLM	0.5964	0.7270	115

4.2. Use Case 2: Springer

Table 3 shows the results from topic modeling on the Springer corpus, which is 20 times the size of the corpus in Use Case 1, while maintaining the same hyperparameters. Here all the topic models have a slightly lower coherence score and a lower diversity score in comparison with Table 1.

Since there are a large number of topics that result from this use case, only the top 20 bar charts are displayed for each language model. Fig. 3 shows the bar charts after running a topic model using Specter, and the resulting bar charts of the other language models are included in Appendix. Since all articles are already related by the common search term *lithium ion battery*, it is reasonable to assume that each topic model will contain one or more topics specifically about batteries.

Table 4 displays the common topics found within the top 20 topics for each of the models. These topics were chosen and labeled by manual inspection. The majority of common themes are captured in all four topic models. The Healthcare, Electrodes, Electrolytes, and Storage topics are prevalent areas of battery research in the past 10 years (Wilberforce et al., 2022), so these topics are expected to be captured by every topic model. In addition to the prevalent topics, there are three others that are not present in every topic model. The Specter and SciBERT topic models captured topics about electric vehicles and thermal research, while the Aspire and MiniLM topic models captured topics related to dielectric conductivity. Interestingly, topic models with fewer topics (SciBERT and Specter) capture similar common topics. Similar conclusions can be drawn for topic models with a larger number of topics (Aspire and MiniLM).

For a better understanding of the differences between topic models, we conduct a direct comparison of the content of specific topics (the topic words and their summary) and analyze how each topic model represents a particular idea. To do this, we create search terms, which can be any word or phrase, and embed them in the vector space

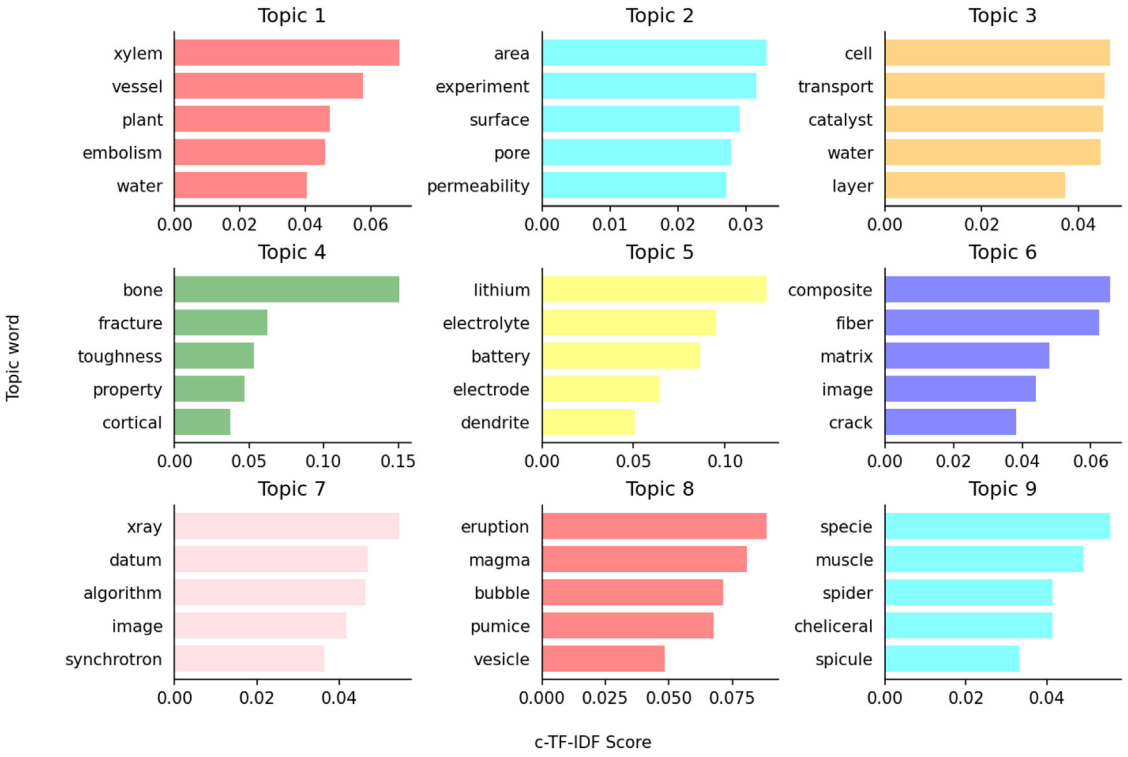


Fig. 2. ALS corpora: Topic words and corresponding c-TF-IDF scores when using MiniLM as the topic model to create the document embeddings.

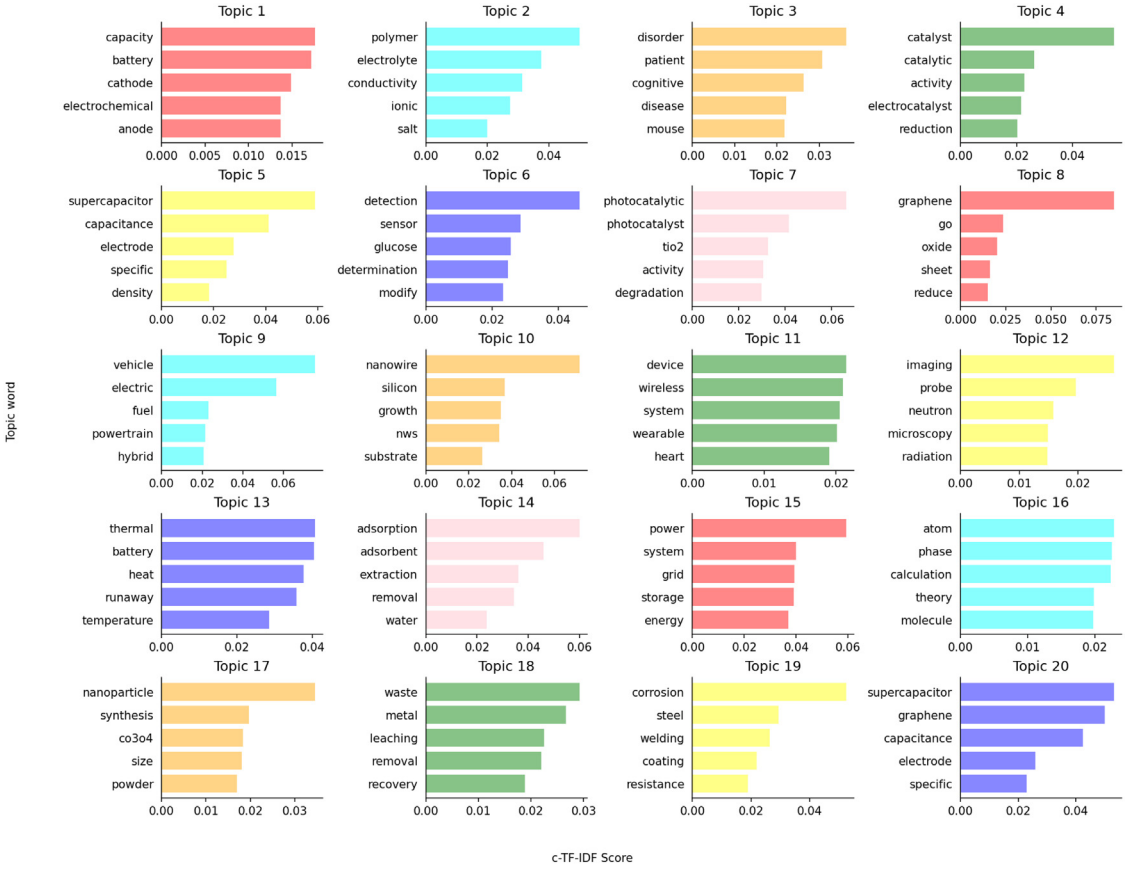


Fig. 3. Springer corpora: Topics for the topic model using Specter to create document embeddings.

Table 4
Springer common topics.

Language model	Healthcare	Electrode	Electrolyte	Storage	Electric vehicles	Thermal	Ceramics
Specter	X	X	X	X	X	X	–
Aspire	X	X	X	X	–	–	X
SciBERT	X	X	X	X	X	X	–
MiniLM	X	X	X	X	–	–	X

Table 5
Topic models and their most similar topics to the search term: *electric vehicle* along with manually labeled summaries of the Topic Summary.

MiniLM		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 46 (30)	Vehicle, electric, evs, cost, car, bev, emission, mobility, adoption, policy	Electric vehicle policy
Topic 32 (45)	Vehicle, electric, powertrain, bus, hybrid, drive, fuel, consumption, driving, optimization	Electric vehicle performance
Topic 84 (20)	Battery, liion, energy, cathode, lithiumion, allsolidstate, storage, chair, lithium, electric	Batteries
Aspire		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 32 (42)	Vehicle, electric, powertrain, hybrid, bus, drive, consumption, optimization, fuel, simulation	Electric vehicle performance
Topic 27 (54)	Vehicle, electric, emission, evs, energy, impact, cost, policy, assessment, electricity	Electric vehicle policy
Topic 15 (83)	Energy, battery, power, flow, storage, cost, density, fuel, technology, development	Batteries
Specter		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 8 (131)	Vehicle, electric, fuel, powertrain, hybrid, bus, drive, power, cost, system	Electric vehicle technology
Topic 30 (43)	Vehicle, battery, energy, electric, flow, 48v, cost, system, cell, technology	Electric vehicle power systems
Topic 29 (44)	Prediction, estimation, noise, rul, soc, failure, degradation, algorithm, filter, circuit	Failure prediction in circuits
SciBERT		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 25 (50)	Nobel, innovation, briefing, discovery, technology, researcher, day, year, daily, news	Scientific innovation
Topic 55 (17)	Battery, electrolyte, advance, energy, lithium, microelectronic, rechargeable, magnesium, voltage, hold	Rechargeable battery technology
Topic 60 (16)	Purpose, assessment, life, car, vehicle, electric, passenger, ecoinvent, inventory, impact	Electric vehicle impact

generated by the BERT model. After the search term is embedded, the cosine similarity of the search vector is calculated with respect to the topic vectors within the model, and the three most similar topics are returned. This process is repeated for each topic model. Tables 5 and 6 show the three most similar topics for each topic model with respect to the given search term along with a short description of each topic's content.

The search results in Table 5 show an overall similar performance for the MiniLM, Aspire, and Specter models. Each of their top three topics has content relative to the search term. The MiniLM and Aspire topic models contain topics with nearly identical content, while the Specter topic model contains topics that are still related to electric vehicles, but have topic words with less overlap than the other two topic models. On the other hand, the SciBERT topic model only has two relevant topics in its top three. This is exacerbated by the fact that the most similar topic in terms of cosine similarity is the irrelevant topic indicating the SciBERT topic model is not creating sufficiently similar embeddings in relation to the term *electric vehicles*.

The search results in Table 6 are similar to those in Table 5. Here the top three most similar topics found in MiniLM, Aspire, and Specter are relevant to the search term *renewable energy*. However, it is a bit subjective as to which of these three topic models best represents the search term in the results. On the other hand, once again the SciBERT topic model has an irrelevant topic in its second most similar topic to the search results while capturing two relevant topics that were also captured by the previous topic models. The poor search results in both cases alongside the much lower coherence score indicate that this use case is not well suited to the SciBERT topic model.

4.3. Use Case 3: Arxiv

Table 7 shows the results of topic modeling on a corpus of more than 2 million Arxiv documents. Although the exact algorithm is slightly

Table 6
Topic models and their 3 most similar topics and corresponding topic words to the search term: *renewable energy* along with manually labeled summaries of the Topic Summary.

MiniLM		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 27 (55)	Power, microgrid, grid, storage, system, energy, electricity, renewable, wind, cost	Energy storage and wind energy
Topic 84 (20)	Battery, liion, energy, cathode, lithiumion, allsolidstate, storage, chair, lithium, electric	Lithium battery performance
Topic 46 (30)	Vehicle, electric, evs, cost, car, bev, emission, mobility, adoption, policy	Electric vehicle policy
Aspire		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 15 (83)	Energy, battery, power, flow, storage, cost, density, fuel, technology, development	Battery performance
Topic 27 (54)	Vehicle, electric, emission, evs, energy, impact, cost, policy, assessment, electrcity	Electric vehicle policy
Topic 23 (65)	Power, grid, storage, system, energy, microgrid, electricity, load, wind, optimization	Energy storage and wind energy
Specter		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 14 (62)	Power, system, grid, storage, energy, microgrid, electricity, renewable, wind, cost	Energy storage and wind energy
Topic 79 (15)	Energy, sustainable, policy, raw, ece, world, supply, renewable, face, decarbonization	Renewable energy policy
Topic 30 (43)	Vehicle, battery, electric, flow, 48v, cost, system, cell, technology	Electric vehicle technology
SciBERT		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 55 (17)	Battery, electrolyte, advance, energy, lithium, microelectronic, rechargable, magnesium, voltage, hold	Battery performance
Topic 25 (50)	Nobel, innovation, briefing, discovery, technology, researcher, day, year, daily, news	Scientific innovation
Topic 60 (16)	Purpose, assessment, life, car, vehicle, electric, passenger, ecoinvent, inventory, impact	Electric vehicles impact

Table 7
Arxiv results.

Language model	Coherence	Diversity	Number of topics
Specter	0.6932	0.7213	47
Aspire	0.7295	0.7149	47
SciBERT	0.6949	0.448	200
MiniLM	0.7289	0.7299	67

different from the previous use cases, since it uses different dimension reduction and clustering algorithms and a `min_cluster_size` of 50, the overall process remains the same. The value of 50 for the `min_cluster_size` leads to fewer overall topics, as topics that do not meet this minimum size requirement will have their documents allocated to other topics that have the required size.

Overall, the Specter, Aspire, and MiniLM topic models perform similarly in terms of topic modeling metrics, with Specter having a slightly lower coherence value. On the other hand, the SciBERT topic model has a significantly larger amount of topics than the other models, and, as a result, a lower diversity score. Although the SciBERT topics may be more repetitive, the individual topics maintain a competitive coherence score with the other topic models.

Fig. 4 shows the resulting topics from the MiniLM topic model. From the bar chart, it is evident that some domains are more prevalent than others in the corpus. For example, there are multiple topics related to Quantum Physics and Astronomy. These topics being repetitive and in the top 20 indicate that the corpus is dominated by documents related to these areas of research. Upon visual inspection, this trend continues in the corresponding attached bar charts for the other topic models.

Table 8 displays the prevalence of common topics found within the top 20 topics for each of the topic models. These topics were chosen and labeled via manual inspection. The topics captured by the most topic models (Celestial Bodies, Quantum Physics, and Cosmology) are within expectations as the articles within the Arxiv corpus are primarily from these domains.

There are some interesting observations to make from the table. The Aspire topic model has only a single topic containing a defined area of mathematical or statistical research. Instead, it has a generic topic with math-related topic words that contain documents that were assigned to more specific topics in other topic models. For example, the Specter topic model contains topics on Graph Theory, Algebraic Geometry, Financial Risk Analysis, and Statistics. In fact, if the expected topics (Cosmology, Astronomy, and Quantum Physics) are not considered, the Specter and Aspire topic models focus on very different areas of mathematics and science while having the same number of topics and similar topic modeling metrics.

The SciBERT topic model's top 20 topics contain either multiple topics spanning the same theme or generic topics that cannot be attributed to a specific research field. For example, there are five topics related to Quantum Physics. The large amount of generic topics in its top 20 along with a very low overall diversity score indicate that, while the amount of topics is large and the coherence score is reasonable, the topics themselves are likely repetitive and do not provide additional information about the corpus after the initial set of topics.

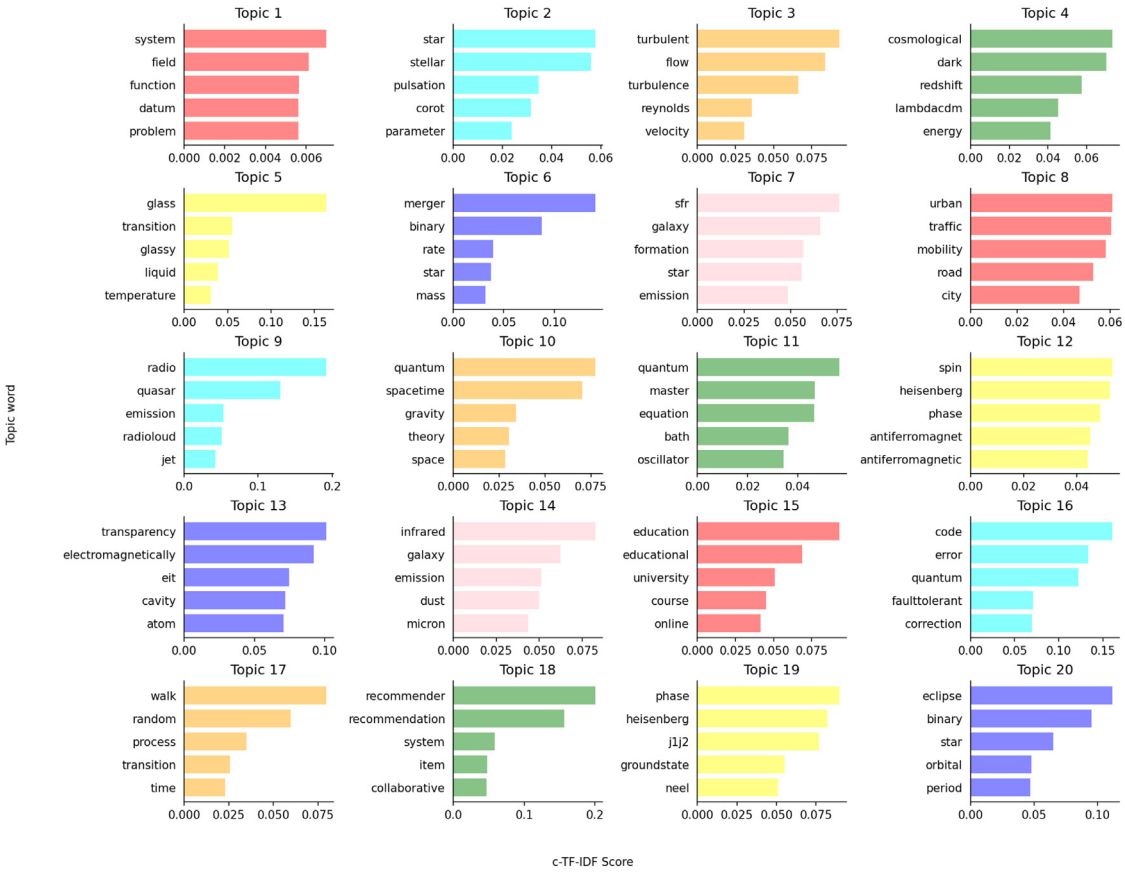


Fig. 4. Arxiv corpora: top 20 topics for the topic model using MiniLM to create the document embeddings.

Table 8
Arxiv common topics.

Language model	Celestial Bodies	Dark Matter	Cosmology	Quantum Physics	Urban Planning	Graph Theory	Algebraic Geometry	Topology	Finance
Specter	X	–	–	X	–	X	X	–	X
Aspire	X	X	X	X	–	–	–	–	–
SciBERT	X	–	X	X	–	–	–	X	–
MiniLM	X	–	X	X	X	–	–	–	–

Language model	String Theory	Information Theory	Fluid Dynamics	Super-conductors	Condensed Matter Physics	Machine Learning	Statistics	Chaos Theory
Specter	–	X	–	–	X	–	X	X
Aspire	X	X	X	X	X	–	–	–
SciBERT	X	–	–	–	–	–	–	–
MiniLM	–	X	X	–	X	X	X	–

Just as in Use Case 2, we can take advantage of the BERT-based approach used for topic modeling and embed a search term in the vector space occupied by the topics for each topic model. Tables 9 and 10 display the three most similar topics to each respective search term.

Table 9 shows the results for the search term *Earthquake detection using machine learning*. This search term was selected due to its lack of overlap with the topic words contained in each topic model. This should remove obvious results that simply match topic words to the search term and instead try to find topics that are semantically similar to the search term. In this case, the MiniLM topic model did not return a relevant topic in the search. However, as we know from Table 8 that the MiniLM model contains topics on Machine Learning and Statistics, it just did not find those topics similar in the search. This is likely a drawback to using a language model that is not trained on scientific text. On the other hand, both the Aspire and Specter topic models yield topics related to both Machine Learning and Statistics. The SciBERT topic model has particularly unique results, as its most similar topics are related to waves, which can be related to seismic waves generated from earthquakes.

Table 10 shows the results for the search term *superposition and states*, which describes an area of research in quantum physics that examines the fundamental principles governing the simultaneous existence of multiple quantum states. Similar to the previous search, this term was selected because it does not overlap with the topic words in each of the models. Unlike the previous search, each topic model contains results that are relevant to the search term. The Aspire and MiniLM topic models contain three relevant topics, while the Specter and SciBERT topic models only contain one. The second and third topics in the results from the Specter model are interesting as the model only found a single relevant topic while it contains several topics

Table 9
Topic models and their most similar topics to the search term: *Earthquake detection using machine learning* along with manually labeled summaries of the Topic Summary.

MiniLM		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 57 (?)	Percolation, network, degree, scalefree, transition, node, random, correlation, distribution, threshold	Network Analysis
Topic 37 (?)	Detonation, explosion, white, deflagration, ignition, dwarf, supernovae, thermonuclear, burn, type	Stars
Topic 42 (?)	Flare, reconnection, loop, magnetic, solar, coronal, ribbon, rope, euv, flux	Stars
Aspire		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 0 (42)	System, field, function, datum, quantum, problem, find, present, theory, time	Gereric topic
Topic 20 (54)	Fractal, function, dimension, attractor, set, selfsimilar, iterate, system, space, selfsimilarity	Geometry
Topic 28 (83)	Program, language, type, programming, verification, static, semantic, code, check, analysis	Programming
Specter		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 13 (131)	Spacial, dataum, regression, process, bayesian, multivariate, estimation, estimator, covariance, likelihood	Statistics
Topic 35 (43)	Feature, selection, mutual, select, classification, information, subset, dataset, predictive, datum	Machine Learning
Topic 18 (44)	Information, entropy, algorithmic, theory, complexity, shannon, kolmogorov, probability, statistical, notion	Information Theory
SciBERT		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 82 (50)	Metamaterial, wave, electromagnetic, medium, dielectric, scatter, mode, casimir, numerical	Electromagnetic Waves
Topic 99 (17)	Algebra, group, homology, geometry, complex, knot, lie, curve, commutative, form	Algebraic Geometry
Topic 192 (16)	Wave, equation, medium, dispersive, shock, propogation, elastic, nonlinear, casimir, solution	Waves

spanning several areas in the domain of Quantum Physics. This is likely due to the search embedding being too specific and there being a lack of articles in the corpus about superposition. Thus the generic Topic 0 and Topic −1 (the outlier topic) are the next most similar topics.

4.4. Discussion

Three of the language models tested in this experiment were pre-trained on scientific articles, while the MiniLM model is an all-around generic language model that has historically good performance on NLP tasks (Wang et al., 2021). The expectation here is that topic models using a scientific language model to compute the embeddings would create superior topics in terms of topic modeling metrics and human evaluation. Overall, the SciBERT topic model performed worse than the other topic models in terms of topic modeling metrics. In Use Case 1, when all diversity scores were high, the SciBERT topic model had a much lower coherence score. In Use Cases 2 and 3 the SciBERT topic model had consistently lower diversity scores than the other topic models. Surprisingly, the generic MiniLM topic model performed similarly to the Specter and Aspire language models, which are pre-trained on scientific data.

In terms of human inspection of the topic models, the searches shown in Use Cases 2 and 3 provide a way to directly compare the performance of topic models by analyzing the validity of their search results. In these searches, the SciBERT topic model contained several topics that were completely irrelevant to the search term. The other topic models have comparable search results and present topics that are very relevant to their respective search terms.

5. Conclusion

When using modern topic modeling techniques on a corpus consisting of scientific articles a unique challenge arises. Due to their rigid structural requirements and a shared underlying goal of presenting research, certain words will overwhelm the topic modeling results and nothing will be learned about the corpus. BERTeley was developed in order to overcome this challenge, while also providing some quality of life improvements to BERTopic; these were accomplished through three features: a preprocessing suite for scientific articles, preloaded scientific language models, and topic model metric calculation. These features in an easy-to-use, yet powerful topic modeling package allow users to easily understand the underlying themes and ideas within their corpus of scientific articles.

The topic modeling experiment was carried out to analyze the performance of scientific language models and compare the results to a generic language model with historically good performance when it comes to performing NLP tasks. Through the three use cases, the Specter, Aspire, and MiniLM models had similar performances in terms of both topic modeling metrics and human evaluation. On the other

Table 10
Topic models and their most similar topics to the search term: *superposition and states* along with manually labeled summaries of the Topic Summary.

MiniLM		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 39 (?)	Interference, photon, quantum, twophoton, entanglement, entangle, multiphoton, singlephoton, experiment, fourphoton	Quantum Optics
Topic 26 (?)	Condensate, boseeinstein, atom, lattice, bose, optical, trap, atomic, gas, interference	Ultra Cold Atomic Physics
Topic 22 (?)	Localization, anderson, disorder, localize, lattice, potential, atom, quantum, mobility, ultracold	Condensed Matter Physics
Aspire		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 9 (42)	Quantum, decoherence, system, field, evolution, matrix, classical, initial, density, time	Quantum Mechanics
Topic 11 (54)	Quantum, thermodynamic, mechanic, system, classical, statistical, equilibrium, ensemble, theory, dynamic	Quantum Mechanics
Topic 18 (83)	Entropy, quantum, information, classical, channel, theory, von, system, neumann, shannon	Quantum Information Theory
Specter		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 5 (131)	Quantum, mechanic, bell, epr, theory, argument, experiment, interpretation, theorem, probability	Quantum Mechanics
Topic 0 (43)	System, field, datum, function, quantum, problem, find, present, theory, time	Generic Science topic
Topic -1 (44)	Quantum, system, field, theory, find, present, datum, function, time, energy	Outlier topic
SciBERT		
Topic number (Number of documents)	Topic words	Topic Summary
Topic 172 (50)	Graphene, spin, conductance, fermi, superconducte, dirac, spinorbit, band, fermion, majorana	Condensed Matter Physics
Topic 82 (17)	Metamaterial, wave, electromagnet, medium, magnetic, dielectric, scatter, mode, casimir, numerical	Electromagnetic Materials
Topic 171 (16)	Black, hole, bps, solution, dirac, case, dimension, generalize, equation, monopole	Black Holes

hand, the SciBERT topic model had consistently lower statistics and created lower quality topics than the others.

At the time of writing, these new scientific languages models have either been released or are in development. When a new model is made available on HuggingFace, for example, Specter 2.0 (AllenAI, 2023), we will conduct a test similar to the one in this paper to understand how well it performs topic modeling on scientific articles. Although there is no downside to having an extensive list of all scientific language models, some due diligence is required before the addition of a new model to our package.

CRediT authorship contribution statement

Eric Chagnon: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ronald Pandolfi:** Software. **Jeffrey Donatelli:** Funding acquisition, Writing – original draft, Writing – review & editing. **Daniela Ushizima:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the US Department of Energy (DOE) Office of Science Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) under Contract No. DE-AC02-05CH11231 to the Center for Advanced Mathematics for Energy Research Applications (CAMERA) program. It also included support from the DOE ASCR-funded project Analysis and Machine Learning Across Domains (AMLXD), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Appendix

See Figs. 5–13.

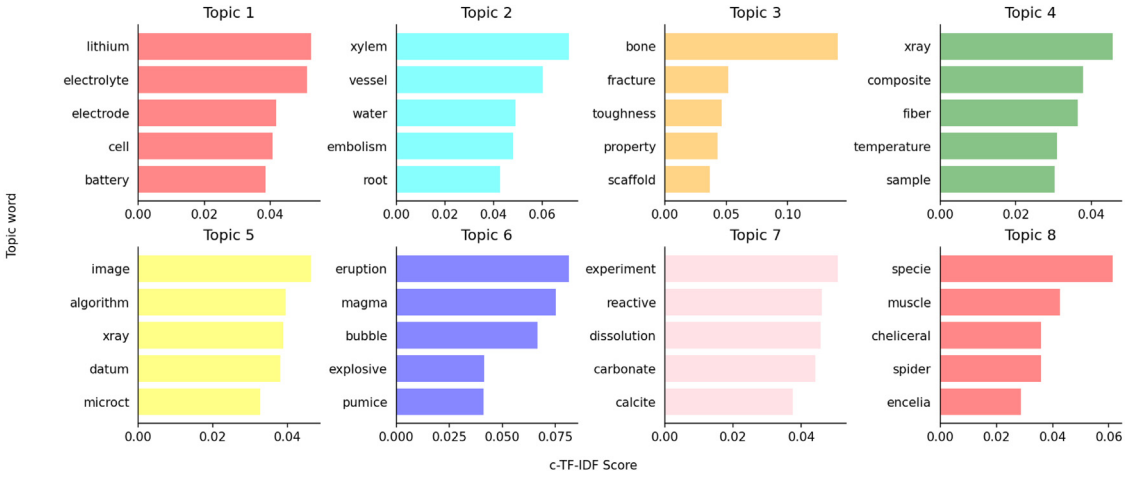


Fig. 5. ALS corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Aspire to create the document embeddings.

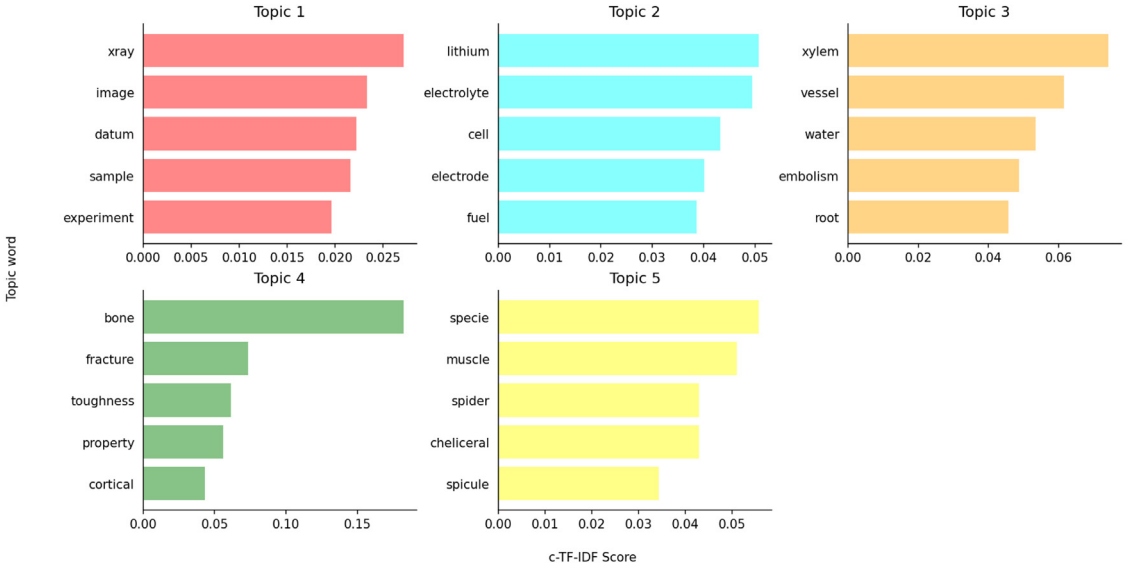


Fig. 6. ALS corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Specter to create the document embeddings.

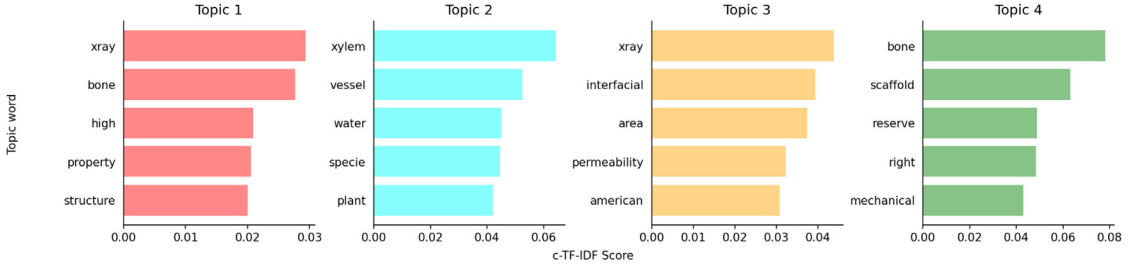


Fig. 7. ALS corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Scibert to create the document embeddings.

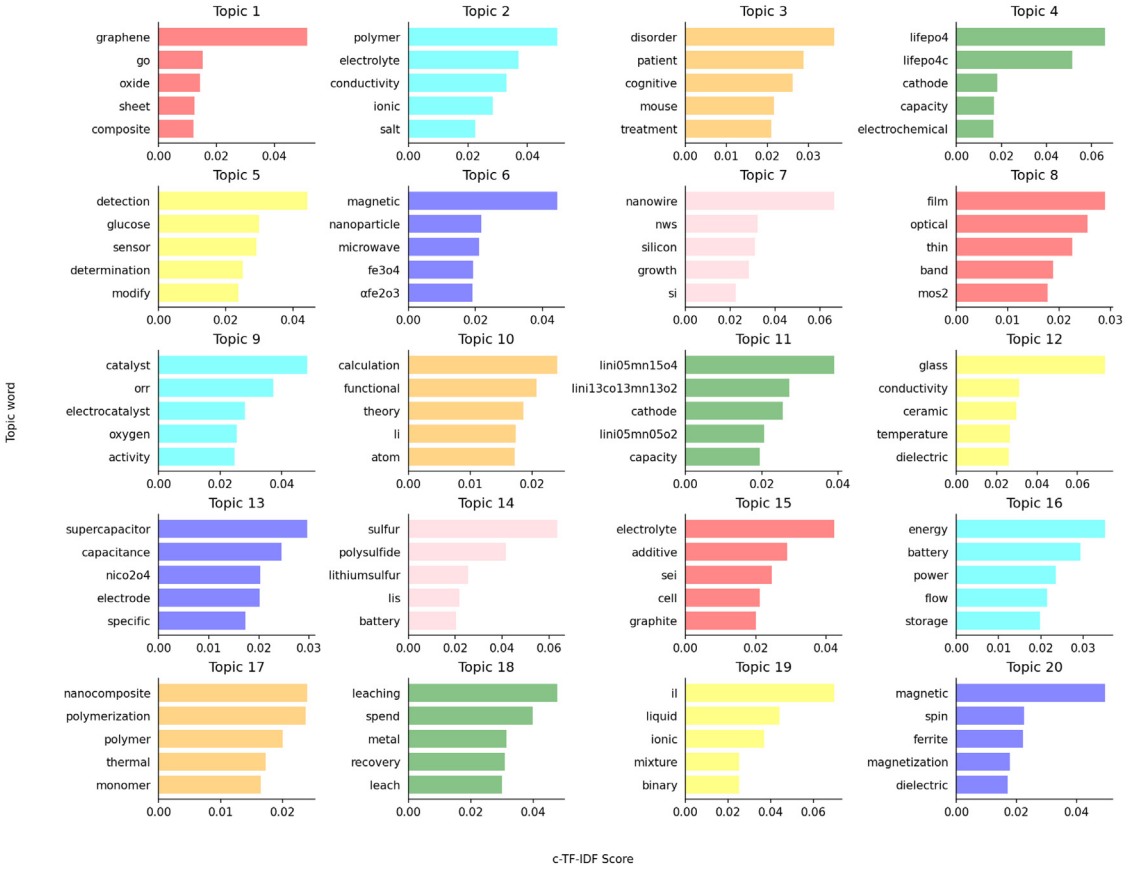


Fig. 8. Springer corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Aspire to create the document embeddings.

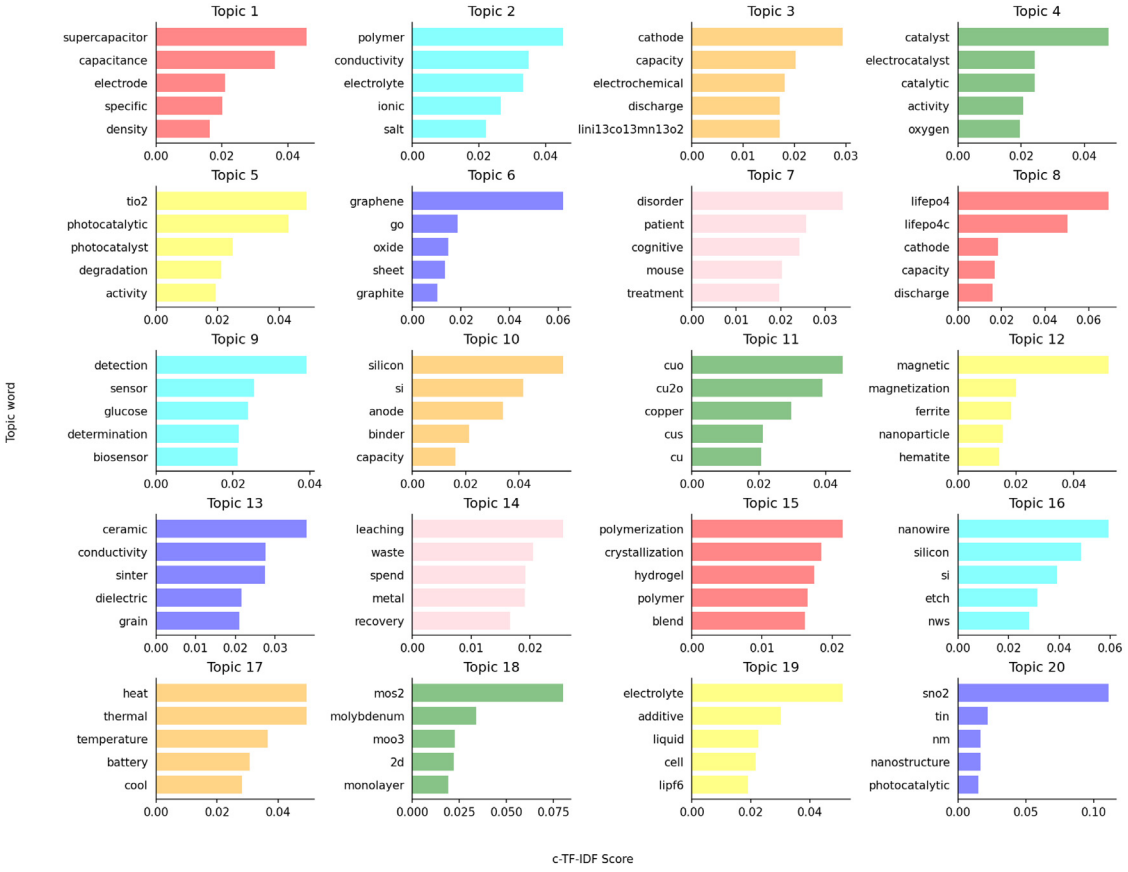


Fig. 9. Springer corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Mini-LM to create the document embeddings.

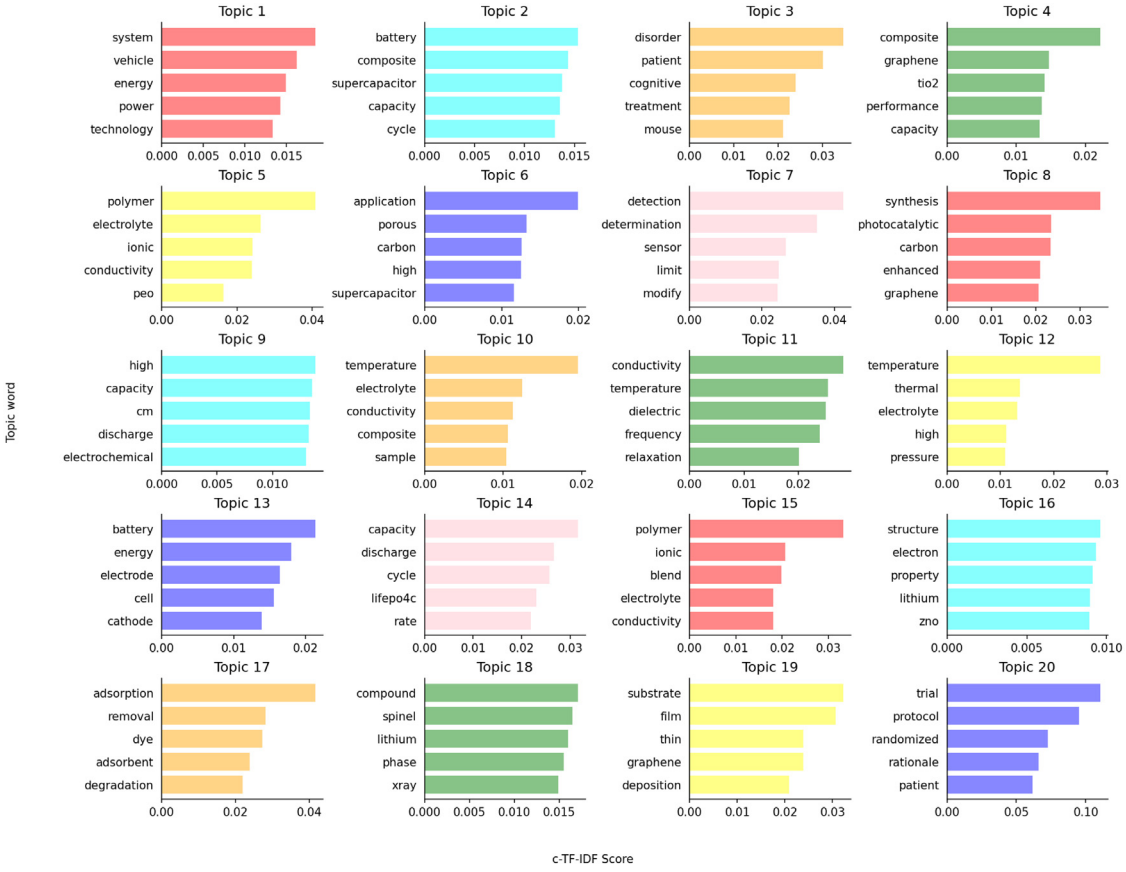


Fig. 10. Springer corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Scibert to create the document embeddings.

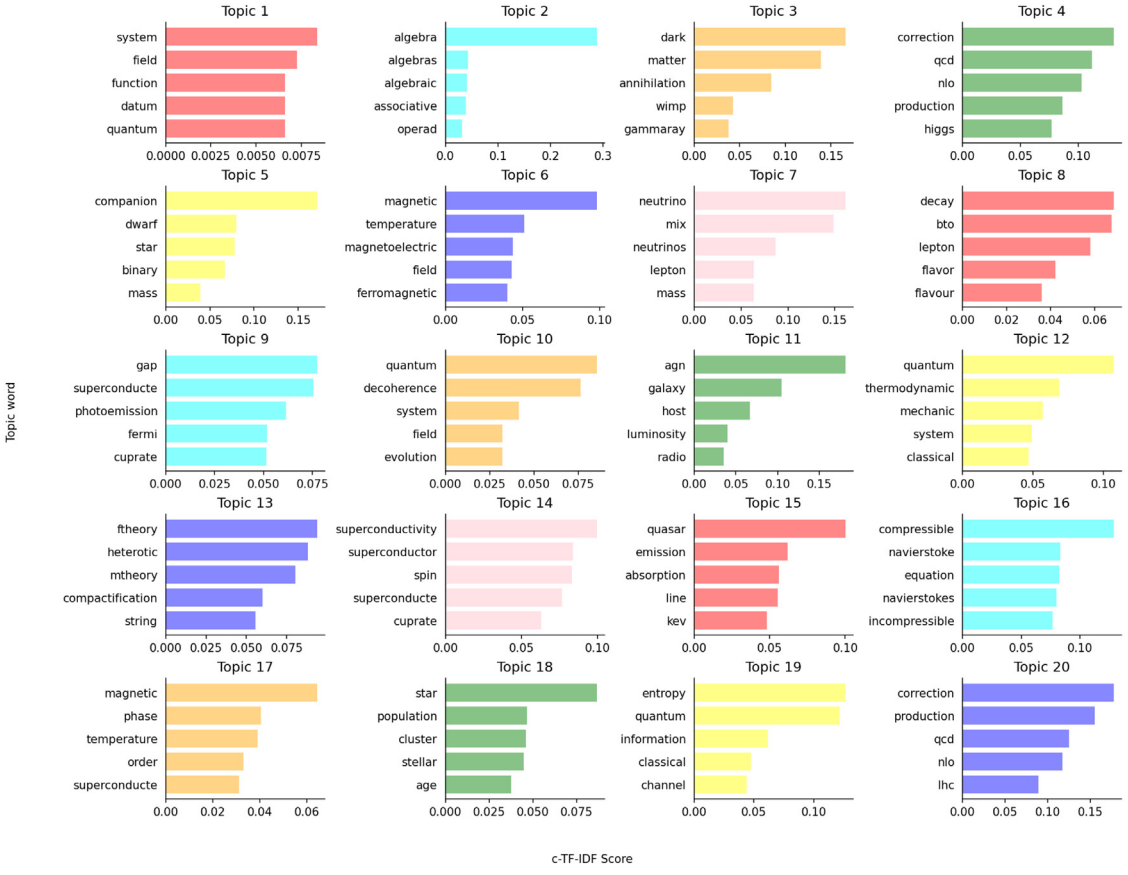


Fig. 11. Arxiv corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Aspire to create the document embeddings.

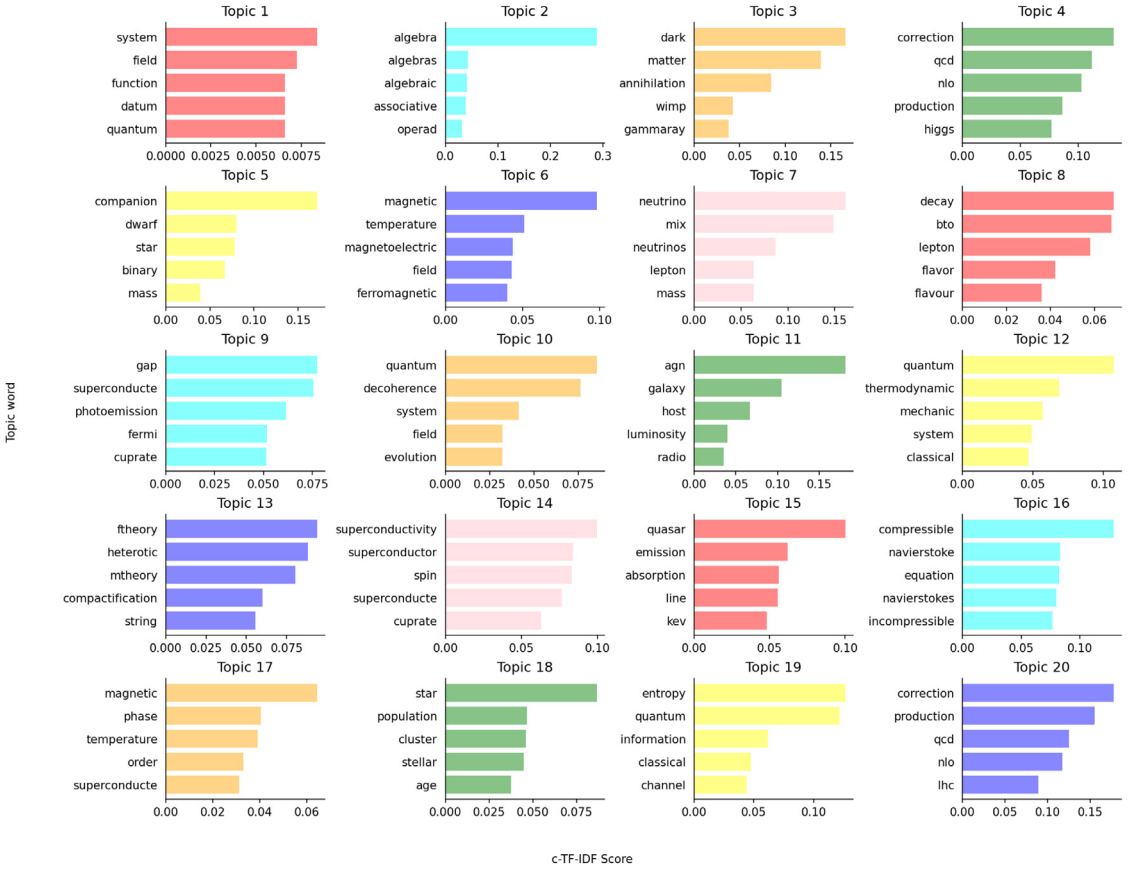


Fig. 12. Arxiv corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Mini-LM to create the document embeddings.

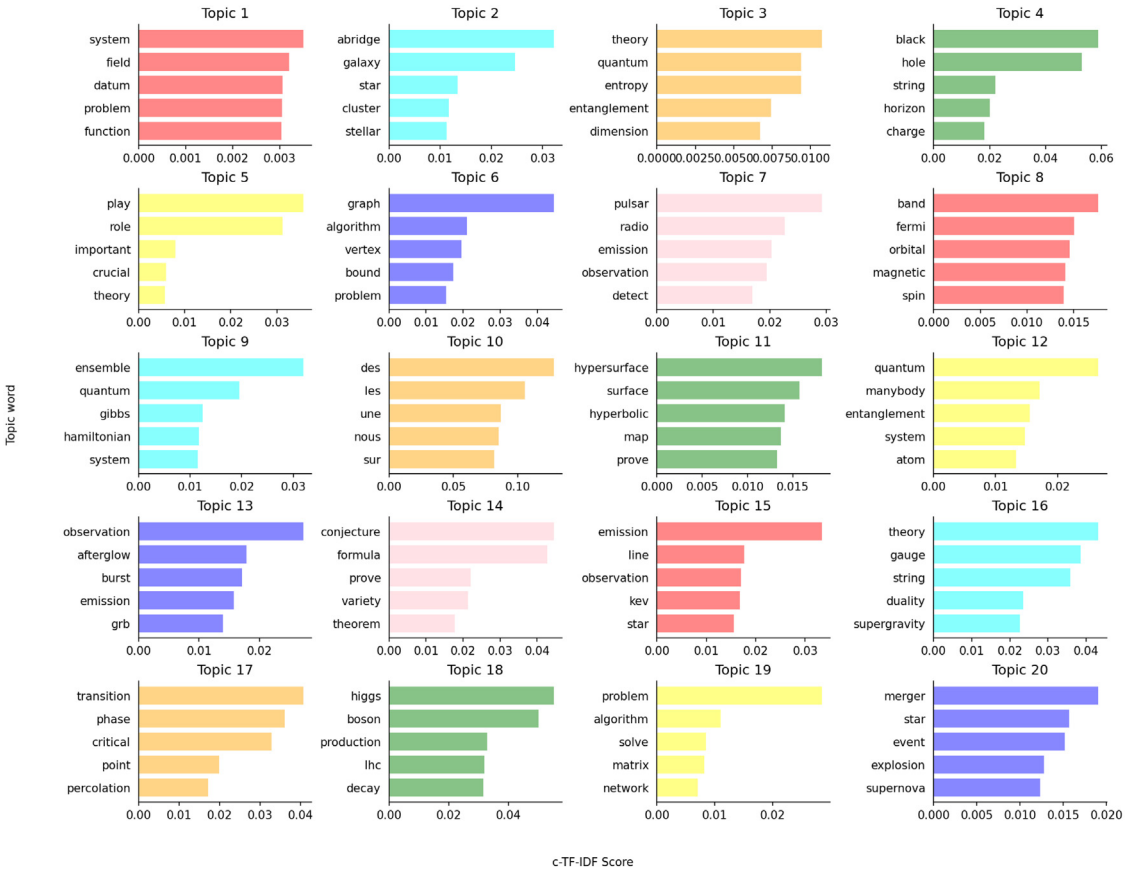


Fig. 13. Arxiv corpora: Topic words and corresponding c-TF-IDF scores for the topic model using Scibert to create the document embeddings.

References

AllenAI, 2023. SPECTER 2.0. URL https://huggingface.co/allenai/specter2.

alspubs, 2022. ALS publication search. URL https://alsusweb.lbl.gov/4DCGI/WEB_GetForm/PublicationSearch.shtml/Initialize.

arXiv.org submitters, 2023. Arxiv dataset. http://dx.doi.org/10.34740/KAGGLE/DSV/5895943, URL https://www.kaggle.com/dsv/5895943.

Bansal, S., Srivastava, A., Arora, A., 2017. Topic modeling driven content based jobs recommendation engine for recruitment industry. Procedia Comput. Sci. 122, 865–872. http://dx.doi.org/10.1016/j.procs.2017.11.448, URL https://www.sciencedirect.com/science/article/pii/S1877050917326960, 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Beltagy, I., Cohan, A., Lo, K., 2019. Scibert: Pretrained contextualized embeddings for scientific text. CoRR, abs/1903.10676, arXiv:1903.10676, URL http://arxiv.org/abs/1903.10676.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S., 2020. SPECTER: document-level representation learning using citation-informed transformers. CoRR, abs/2004.07180, arXiv:2004.07180, URL https://arxiv.org/abs/2004.07180.

Dieng, A.B., Ruiz, F.J.R., Blei, D.M., 2019. Topic modeling in embedding spaces. CoRR, abs/1907.04907, arXiv:1907.04907, URL http://arxiv.org/abs/1907.04907.

Doogan, C., Buntine, W., 2021. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp. 3824–3848. http://dx.doi.org/10.18653/v1/2021.naacl-main.300, URL https://aclanthology.org/2021.naacl-main.300.

Eltyeb, S., Salim, N., 2014. Chemical named entities recognition: a review on approaches and applications. J. Chem. 6, 1–12.

Eric Chagnon, R.P.J.D., Ushizima, D., 2023. Berteley. URL https://github.com/lbl-camera/berteley.

Grootendorst, M., 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Hall, D., Jurafsky, D., Manning, C.D., 2008. Studying the history of ideas using topic models.. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Hinde, S., Spackman, E., 2015. Bidirectional citation searching to completion: an exploration of literature searching methods. Pharmacoeconomics 33, 5–11.

Lisena, P., Harrando, I., Kandakji, O., Troncy, R., 2020. TOMODAPI: A topic modeling API to train, use and compare topic models. In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS). Association for Computational Linguistics, Online, pp. 132–140. http://dx.doi.org/10.18653/v1/2020.nlposs-1.19, URL https://aclanthology.org/2020.nlposs-1.19.

Mysore, S., Cohan, A., Hope, T., 2021. Multi-vector models with textual guidance for fine-grained scientific document similarity. arXiv:2111.08366.

Nature, S., 2023a. Perlmutter architecture. URL https://docs.nersc.gov/systems/perlmutter/architecture/#gpu-nodes.

Nature, S., 2023b. Springer nature API. URL https://dev.springernature.com/.

Oliaee, A.H., Das, S., Liu, J., Rahman, M.A., 2023. Using bidirectional encoder representations from transformers (BERT) to classify traffic crash severity types. Nat. Lang. Process. J. 3, 100007. http://dx.doi.org/10.1016/j.nlp.2023.100007, URL https://www.sciencedirect.com/science/article/pii/S2949719123000043.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Ramondt, S., Kerkhof, P., Merz, E.-M., 2022. Blood donation narratives on social media: A topic modeling study. Transfus Med. Rev. 36 (1), 58–65. http://dx.doi.org/10.1016/j.tmrv.2021.10.001, URL https://www.sciencedirect.com/science/article/pii/S0887796321000572.

Raschka, S., Patterson, J., Nolet, C., 2020. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. arXiv preprint arXiv:2002.04803.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, URL https://arxiv.org/abs/1908.10084.

Röder, M., Both, A., Hinneburg, A., 2015. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web

Search and Data Mining. WSDM '15, Association for Computing Machinery, New York, NY, USA, pp. 399–408. http://dx.doi.org/10.1145/2684822.2685324.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., Both, A., 2014. Evaluating topic coherence measures. CoRR, abs/1403.6397, arXiv:1403.6397, URL http://arxiv.org/abs/1403.6397.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, W., Bao, H., Huang, S., Dong, L., Wei, F., 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. arXiv:2012.15828.

Wilberforce, T., Thompson, J., Olabi, A.-G., 2022. Classification of energy storage materials. In: Olabi, A.-G. (Ed.), Encyclopedia of Smart Materials. Elsevier, Oxford, pp. 8–14. http://dx.doi.org/10.1016/B978-0-12-803581-8.11762-X, URL https://www.sciencedirect.com/science/article/pii/B978012803581811762X.