MIReAD: simple method for learning high-quality representations from scientific documents

Anastasia Razdaibiedina[◇,♠] and **Alexander Brechalov**[◇]
[◇]University of Toronto and [♠]Vector Institute
anastasia.razdaibiedina@mail.utoronto.ca
alexander.brechalov@utoronto.ca

Abstract

Learning semantically meaningful representations from scientific documents can facilitate academic literature search and improve performance of recommendation systems. Pretrained language models have been shown to learn rich textual representations, yet they cannot provide powerful document-level representations for scientific articles. We propose MIReAD, a simple method that learns high-quality representations of scientific papers by fine-tuning transformer model to predict the target journal class based on the abstract. We train MIReAD on more than 500,000 PubMed and arXiv abstracts across over 2,000 journal classes. We show that MIReAD produces representations that can be used for similar papers retrieval, topic categorization and literature search. Our proposed approach outperforms six existing models for representation learning on scientific documents across four evaluation standards. [1]

1 Introduction

A significant increase in the volume of scientific publications over the past decades has made the academic literature search a more challenging task. One of the key steps to improve the recommendation systems (RS) for research articles is to obtain high-quality document-level representations. Recently, transformer-based models have brought substantial progress to the field of natural language processing (NLP), obtaining state-of-the-art results on a variety of benchmarks (Vaswani et al., 2017; Devlin et al., 2018). While transformer models are effective in language modeling and learning sentence representations, deriving document-level representations for scientific articles remains a challenge.

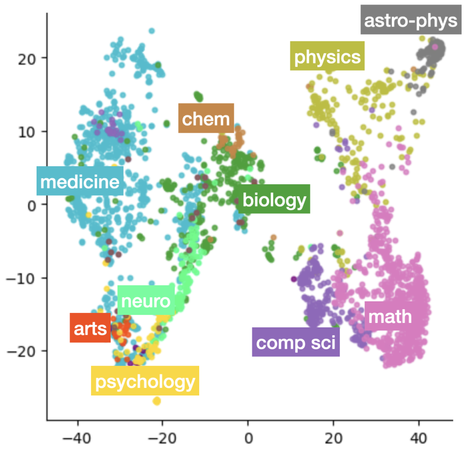

Figure 1: MIReAD representations allow distinguishing abstracts by scientific domain, without using domain or citation information during finetuning. tSNE with abstracts' representations from unseen journals is shown.

Previous transformer-based methods for representation learning on scientific documents are derived from BERT model (Devlin et al., 2018). Classic examples of such approaches are PubMedBERT, BioBERT and SciBERT - scientific domain adaptations of BERT, which were pre-trained with masked language modeling (MLM) objective on PubMed abstracts, as well as full-text articles from PubMedCentral and Semantic Scholar, respectively (Gu et al., 2020; Lee et al., 2020; Beltagy et al., 2019). While MLM objective allows to efficiently capture the context of the sentence, it cannot achieve accurate paper representations that can be used "off-the-shelf" to discover similar articles. To address this problem, recent works explored fine-tuning the pretrained models with supervised objectives based on citation graphs (Wright and Augenstein, 2021; Cohan et al., 2020). Despite their efficiency, citation-based objectives have several disadvantages: (1) citations are not distributed uniformly, with novel

[1] MIReAD model weights are available through HuggingFace at https://huggingface.co/arazd/MIReAD

Abstracts and journal data is available through HuggingFace Hub at https://huggingface.co/datasets/brainchalov/pubmed_arxiv_abstracts_data

Full code is provided at https://github.com/arazd/miread

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 530–539
July 9-14, 2023 ©2023 Association for Computational Linguistics

papers and articles from certain fields being less favoured; (2) citations have a bias related to the increased self-citation and existence of over-cited papers; (3) citation graphs are often large and difficult to preprocess. Hence, there is a gap in representation learning for scientific articles, requiring approaches which would derive high-quality document-level representations, without relying on the citation graphs.

In this paper, we propose **MIREAD**, an approach that requires **M**inimal **I**nformation for **Re**presentation Learning of **A**cademic **D**ocuments. MIREAD combines the SciBERT architecture with novel training objective - a target journal classification. We show that such a simple training objective leads to high-quality representations of academic papers, suitable for RS usage. Figure 1 illustrates how MIREAD representations from unseen abstracts are separated based on scientific domain, even though this information was not accessed during training. We trained MIREAD by predicting one of 2,734 journal classes from the paper's title and abstract for 500,335 articles from PubMed and arXiv. Then we measured the quality of paper representations obtained with MIREAD using three evaluation standards - linear evaluation, information retrieval, and clustering purity scores - on three different datasets. MIREAD substantially outperforms 5 previous approaches (BERT, PubMedBERT, BioBERT, SciBERT, CiteBERT) across all evaluation benchmarks and outperforms SPECTER in most cases.

2 Methods

2.1 MIREAD

MIREAD is based on BERT architecture and we initialize it from SciBERT's weights. We fine-tune MIREAD to predict journal class solely from paper's abstract and title with cross-entropy loss:

$$L(\widehat{y}_i, y_i) = -\sum_{i=1}^{N} y_i \log(\widehat{y}_i)$$

Here $\widehat{y}_i$ and y_i stand for predicted probability and ground truth label of the class i, N is equal to 2734, the total number of unique journal classes.

MIREAD takes as input a concatenation of paper's title and abstract, appended to the `[CLS]` token, and separated by the `[SEP]` token:

$$\text{input} = \texttt{[CLS]}\,\text{title}\,\texttt{[SEP]}\,\text{abstract}$$

Final paper representation v is obtained by passing the input through the transformer model, and taking the representation of the `[CLS]` token:

$$v = \text{forward}(\text{input})_{\texttt{[CLS]}}$$

2.2 Dataset

To achieve a good coverage of different knowledge domains, we constructed a dataset from arXiv and PubMed abstracts and their corresponding metadata (title and journal) (Clement et al., 2019). We limited the number of abstracts per each journal to not exceed 300, and excluded journals with less than 100 abstracts or no publications in year 2021. The final dataset contains 500,335 abstracts (181,967 from arXiv and 318,368 from PubMed), covers 76 scientific fields and 2,734 journals. More details on dataset preparation are in Appendix A.1. We fine-tune MIREAD for one epoch on all paper abstracts using 1e-6 learning rate.

2.3 Baseline models

We compare MIREAD to six baseline approaches based on BERT (Devlin et al., 2018). We use the original BERT model, its three different domain adaptations: BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2020) and SciBERT (Beltagy et al., 2019), as well as two representation extraction models trained with citation objectives: CiteBERT (Wright and Augenstein, 2021) and SPECTER (Cohan et al., 2020). Additionally, we include SentenceBERT (Reimers and Gurevych, 2019) – a modification of the BERT model that includes siamese network structure to find semantically similar sentence pairs.

3 Evaluation of representations

We evaluate the information content of the representations of scientific abstracts produced by different approaches. Ideally, we are interested in representations that contain information about scientific domain, and allow to distinguish specific subdomains within larger fields. We use three common strategies for representation quality assessment: linear evaluation, clustering purity and information retrieval.

3.1 Linear evaluation of representations

We first evaluate representations with commonly used *linear evaluation protocol* (Zhang et al., 2016; Oord et al., 2018; Chen et al., 2020). Under this

Task →	MAG		MeSH		arXiv & PubMed		Unseen journals	
Model ↓	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
BERT	$71.47_{0.82}$	$77.82_{0.49}$	$46.33_{0.41}$	$63.67_{0.33}$	$4.22_{0.21}$	$22.05_{0.92}$	$2.62_{0.19}$	$4.70_{0.25}$
PubMedBERT	$72.65_{0.97}$	$78.25_{0.4}$	$72.45_{0.8}$	$77.80_{0.51}$	$4.07_{0.27}$	$19.00_{0.73}$	$0.71_{0.31}$	$1.41_{0.5}$
BioBERT	$59.43_{0.22}$	$71.63_{0.38}$	$50.60_{1.12}$	$67.87_{0.58}$	$2.53_{0.2}$	$20.00_{0.84}$	$0.65_{0.22}$	$1.95_{0.44}$
SciBERT	$74.84_{0.57}$	$79.47_{0.35}$	$66.67_{0.98}$	$74.19_{0.58}$	$10.75_{0.71}$	$31.30_{0.45}$	$7.90_{1.38}$	$11.46_{1.64}$
CiteBERT	$70.49_{0.58}$	$76.40_{0.21}$	$55.94_{1.23}$	$67.80_{0.30}$	$9.26_{0.49}$	$29.05_{1.07}$	$6.72_{0.73}$	$10.19_{0.79}$
SentBERT*	80.5	–	69.1	–	–	–	–	–
SPECTER	$81.47_{0.18}$	$\mathbf{85.05_{0.14}}$	$86.23_{0.27}$	$87.38_{0.13}$	$30.75_{0.69}$	$44.92_{0.49}$	$18.26_{1.34}$	$23.73_{1.17}$
MIREAD	$\mathbf{81.85_{0.59}}$	$84.85_{0.31}$	$\mathbf{86.71_{0.36}}$	$\mathbf{88.22_{0.19}}$	$\mathbf{34.97_{0.3}}$	$\mathbf{48.95_{0.26}}$	$\mathbf{19.35_{0.49}}$	$\mathbf{25.11_{0.36}}$

Table 1: Linear evaluation of document-level representations obtained from different methods. We report F1-score and accuracy on four standards. Mean and standard deviation across three runs is shown. * denotes results reported by (Cohan et al., 2020).

protocol, a linear classifier is trained on top of extracted representations, and test accuracy is used as a quality metric. Hence, better information content of the representations translates into higher classification accuracy. Details of training the logistic regression are provided in Appendix A.3. In our experiments, we perform linear evaluation of representations derived from the abstracts from four datasets, with varying degree of difficulty:

Academic topics In this task, we predict the research field of the paper using Microsoft Academic Graph (MAG) dataset (Sinha et al., 2015). MAG provides paper labels, which are organized into a hierarchy of 5 levels. We follow SciDocs evaluation framework by Cohan et al. (2020), which provides a classification dataset with labels from level 1 topics (e.g. business, sociology, medicine etc.), and has a train-test split. Overall, MAG dataset consists of 19 classes and covers 25K papers.

Medical subject headings We use Medical Subject Headings (MeSH) dataset by Lipscomb (2000) to classifiy academic paper representations into one of 11 disease classes (e.g. diabetes, cardiovascular disease etc.). Similarly to MAG, we use data with train and test splits provided by SciDocs. This dataset contains a total of 23K medical papers.

PubMed and arXiv categories We constructed a dataset of academic papers and their corresponding PubMed and arXiv categories. For fair comparison, we collected papers solely from journals that were not seen by MIREAD during training. For PubMed data, we used scientific topic identifiers that come with the journal metadata. For arXiv data, we omitted subcategories and used major categories (e.g. CS.ML and CS.CG were labeled as CS). To ensure that each paper is mapped to a single label, we used arXiv papers with all annotations coming from the same major category. This dataset contains 12K papers across 54 scientific field categories (e.g. physics, computer science, bioinformatics, molecular biology etc.).

Unseen journal classification This task evaluates whether the learned representations contain very detailed information that allows to distinguish which journal the paper comes from. Since this task resembles MIREAD training objective, we only used journal classes that were not seen during training. This dataset contains the same 12K papers from PubMed and arXiv as the previous task, and 200 journal labels.

We report test set performance of the linear classifier selected by maximal validation set accuracy, and use 4-fold cross validation.

3.2 Clustering purity

In our subsequent experiments, we evaluate feature performance when they are used "off-the-shelf", without any finetuning. Such scenario is important for measuring quality of the representations, since it more closely resembles paper search with RS. Following pre-trained representations assessment strategy from Aharoni and Goldberg (2020), we first evaluate clustering using *purity* metric, a widely adopted metric of clustering quality based on intra-cluster similarity (Manning et al., 2010). Higher clustering purity indicates model's ability to provide representations that can be more easily grouped into meaningful clusters, such as academic topics. We show results on MAG and MeSH datasets, and perform clustering with k-means algorithm with an increasing number of clusters (10, 20, 50, 100). We compute purity score between ground truth annotations and k-means clustering labels.

	Number of clusters			
Method ↓	10	20	50	100
BERT	29.51	31.51	34.50	37.08
PubMedBERT	32.45	32.70	37.30	40.30
BioBERT	33.45	35.45	41.36	45.30
SciBERT	29.02	31.45	35.22	38.13
CiteBERT	29.22	30.53	33.90	36.73
SPECTER	57.28	**65.07**	70.87	74.21
MIREAD	**57.38**	64.78	**72.15**	**76.26**

Table 2: Clustering purity on MeSH dataset with k-means clustering of frozen representations. Results with 10, 20, 50 and 100 clusters across seven methods are reported.

3.3 Information retrieval

In this final part of our evaluation framework, we measure the quality of representations according to the *information retrieval* perspective. Information retrieval is the process of searching and returning relevant items (in our case scientific documents) based on the input query (Manning et al., 2010). For RS, relevant research papers are found based on similarity score between frozen representations. Hence, evaluating how relevant the recommended documents are based on the query document can indicate the quality of the pretrained representations.

For this experiment, we use arXiv subcategories as more stringent labels to measure relevance of representation retrieval (Clement et al., 2019). We collect arXiv papers with their subcategories metadata from six different fields: Computer Science (CS), Mathematics (Math), Physics (Phys), Electrical Engineering and Systems Science (EESS), Economics (Econ) and Statistics (Stat). We perform independent evaluation of subcategories within each field.

We use a commonly adopted evaluation scheme, when pairs of representations are ranked from highest to lowest based on their Pearson's correlation score. Each pair receives a ground truth label of 0 if no subcategories overlap, or 1 otherwise. We report average precision (AP) and area under curve (AUC) scores as final information retrieval metrics.

4 Results

We compared MIREAD with the original BERT model and 5 other approaches that use BERT architecture: PubMedBERT, BioBERT, SciBERT, CiteBERT and SPECTER.

Table 1 shows results of the linear evaluation of representations obtained from seven different models on four tasks/datasets (See Methods). Overall, **MIREAD shows a substantial increase in accuracy and F1 score on all four tasks**. On MAG and MeSH tasks MIREAD achieved 84.85% and 88.22% accuracy respectively, (81.85 and 86.71 in F1 score). Similarly, MIREAD showed substantial improvement compared to other BERT-based models on 54 PubMed/ArXiv Categories classification and 200 Unseen Journals classification tasks. MIREAD performance is the closest to SPECTER, although MIREAD outperforms SPECTER in F1 scores across all 4 presented datasets, with statistically significant improvement in 3 cases out of 4. To measure significance of improvement, we performed unpaired t-test between scores of both approaches. The p-values of t-test between F1 scores across 5 runs of SPECTER and MIREAD are 0.2, 0.04, 0.0001 and 0.05, for MAG, MeSH, arxiv & PubMed, and unseen journals datasets, demonstrating the significant differences for MeSH, arxiv & PubMed, and unseen journals.

We evaluated the quality of representations with the purity metric of k-means clusters. To compute clustering purity, each cluster is assigned to its "true" class (most frequent class in the cluster), then accuracy is measured by counting the number of correctly assigned documents and dividing by the number of samples. Clustering purity on MeSH (shown in Table 2) and MAG (shown in Appendix A.4, Table 4) datasets has shown that MIREAD achieves the performance better (on MeSH) or equal (on MAG) to the performance of SPECTER. Both MIREAD and SPECTER significantly outperform all other tested models.

Similar results were obtained on information retrieval experiments with arXiv subcategories (Average Precision is shown in Table 3). Although, SPECTER showed better precision for Math and Physics categories, MIREAD outperformed in Economics, Computer Sciences (CS) and Electrical Engineering and Systems Science (EESS) categories of arxiv dataset with the improvement of Average Precision of +12.1% , +11.6% and +4.7%, correspondingly.

Overall, three types of evaluations on various datasets reveal that **MIREAD produces powerful representations of academic papers** whose information content outperforms or matches the performance of the current state-of-the-art feature extraction models.

Method	CS	Math	Phys	EESS	Econ	Stat
BERT	20.86	13.28	21.70	65.41	61.49	61.10
PMBERT	21.00	12.54	22.81	65.79	72.05	63.36
BioBERT	22.98	13.07	23.26	66.28	67.40	**64.70**
SciBERT	23.26	14.97	21.84	67.48	64.71	62.91
CiteBERT	18.75	12.59	17.50	65.70	55.74	60.47
SPECTER	31.97	**27.78**	**37.17**	72.53	69.66	63.91
MIREAD	**35.69**	19.15	34.69	**75.91**	**78.12**	63.99

Table 3: Average precision of the representation pairs ranked by correlation scores across arXiv categories.

5 Conclusions

We present MIREAD, a transformer-based method for representation learning of research articles using minimal information. We fine-tuned MIREAD by predicting the target journal using the paper's title and abstract, and assembled a training dataset spanning over half a million data points from over two thousands journals. We show that this simple training objective results in high-quality document-level representations, which can be used for various applications and are suitable for recommendation systems.

Earlier we have seen this effect on biological data – where the prediction of subcellular localization (dozens of classes) (Razdaibiedina and Brechalov, 2022) or protein (thousands of classes) (Razdaibiedina et al., 2023) from the fluorescent microscopy images allows to obtain high-quality features. These resulting features had higher information content and could be applied for solving various downstream analysis tasks. Similarly to our findings, more classification labels improved feature quality, which was reflected in downstream task performance. We found that journal title is a high-quality label for scientific manuscripts, which can be explained by several reasons. Firstly, scientific journals are often highly specialized and focused on a single topic. Therefore, the journal name can serve as a precise topic label. Additionally, journals with different Impact Factors may accept slightly different types of research works, making journal name a valuable human-annotated label. In our study, the number of journals was determined by available datasets. In a preliminary experiment, we found that increasing the number of labels resulted in better specificity of the representations (data not shown). For example, an increase from 100 to 1000 labels helps the model to learn better separations between sub-fields (e.g.medical sub-domains). We found that lower-level labels encourage the model to learn more fine-grained features to distinguish between journal classes, while high-level labels encourage model to focus on few important features, which may lead to over-simplification of representations content.

Our experimental results show that MIREAD substantially outperforms 6 previous approaches (BERT, PubMedBERT, BioBERT, SciBERT, CiteBERT, SentenceBERT) across three evaluation benchmarks, and outperforms SPECTER, the current SOTA approach for representation learning on scientific articles, in most cases. The major advantage of MIREAD compared to SPECTER is that MIREAD uses solely paper's abstract and metadata, but does not require the additional information, such as the reference graph. Hence, MIREAD can be trained on novel papers that have not obtained citations or papers that have no open access.

6 Limitations

The underlying assumption of our method is that abstract reflects the entire article, creating an unbiased summary of the paper. However, abstract does not guarantee an objective representation of the paper, can often emphasize the main findings while discarding details that the authors deem insignificant. This can lead to potential inaccuracies in paper representations, affecting the results of paper retrieval and recommendation.

Also, in this work we did not exhaust all possible training settings and evaluation strategies due to limited resources. We perform evaluation using three different standards. While we selected the most relevant evaluation tasks, it would be interesting to assess the quality of representations in other ways, such as citation graph reconstruction, predicting reader activity and other clustering-based evaluations. Additionally, with the emergence of large-scale language models, another interesting direction for future research is to investigate the relationship between model size and final performance.

Acknowledgement

We would like to thank Vector Institute for providing computational resources to run the experiments. Anastasia Razdaibiedina is supported by Vector Institute Postgraduate Affiliate Fellowship.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. *arXiv preprint arXiv:2004.02105*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Anastasia Razdaibiedina and Alexander Brechalov. 2022. Learning multi-scale functional representations of proteins from single-cell microscopy data. *arXiv preprint arXiv:2205.11676*.

Anastasia Razdaibiedina, Alexander V Brechalov, Helena Friesen, Mojca Mattiazzi Usaj, Myra Paz David Masinas, Harsha Garadi Suresh, Kyle Wang, Charlie Boone, Jimmy Ba, and Brenda J Andrews. 2023. Pifia: Self-supervised approach for protein functional annotation from single-cell imaging data. *bioRxiv*, pages 2023–02.

Nils Reimers and Iryna Gurevych. 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.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Dustin Wright and Isabelle Augenstein. 2021. Citeworth: Cite-worthiness detection for improved scientific document understanding. *arXiv preprint arXiv:2105.10912*.

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer.

A Appendix

A.1 Dataset preparation

PubMed. Since available PubMed datasets did not contain all the necessary metadata, we created a custom dataset by parsing PubMed artciles. We searched PubMed e-utils interface with the custom Python script. The query contained the journal's ISSN and a year of publication. We run through the list of journals from https://www.scimagojr.com website and performed searches for years from 2016 to 2021. The list of retrieved PMID then was split into batches of no more than 200 items each and used to download the articles in xml format. The xml page then was parsed for PMID, title, abstract, name of the journal and date of the publication. We only saved articles whose abstracts were written in English to a file. Next, the final list of journals was filtered, such that remaining journals had at least 300 publication in the period of 2016-2021 and at least 1 publication in 2021. For the final dataset, we limited number of articles per journal to 300.

arXiv. We used a dataset of arXiv articles https://huggingface.co/datasets/arxiv_dataset available at HuggingFace (Wolf et al., 2019). We limited the number of abstracts per each journal to not exceed 300, and excluded journals with less than 100 abstracts or no publications in 2021. Overall, the arXiv dataset contained >171K abstracts after preprocessing.

A.2 Computing resources

We used resources provided by Vector Institute cluster with 528 GPUs, 6 GPU nodes of 8 x Titan X, and 60 GPU nodes each with 8 x T4, for development and deployment of large-scale transformer-based NLP models.

A.3 Linear probing experiments

For our linear probing experiments, we used multinomial logistic regression with a learning rate of 5e-4 and batch size of 100, which we trained for 5 epochs. We did not add a regularization penalty as we found that the regression model did not overfit due to its simplicity. We used 4-fold cross-validation with early stopping based on the maximal validation set performance, and our final performance is averaged across all cross-validation runs.

A.4 Clustering purity on MAG dataset

We include results for clustering purity experiments on MAG datset in Table 4.

Method	10	20	50	100
BERT	38.92	50.94	57.49	60.43
PubMedBERT	31.64	47.49	58.41	60.48
BioBERT	43.44	56.38	61.63	65.27
SciBERT	46.98	48.83	57.35	60.11
CiteBERT	33.9	45.05	51.73	55.55
SPECTER	61.95	75.03	78.07	78.67
MIReAD	61.03	71.9	75.31	78.63

Table 4: Clustering purity on MAG dataset with k-means clustering of frozen representations. Results with 10, 20, 50 and 100 clusters across seven methods are reported.

A.5 arXiv subcategories

Table 5 includes a description of arXiv subcategories that we used to form a category for article topic classification.

Categories	##	Subcategories
Computer Science (CS)	40	Artificial Intelligence, Hardware Architecture, Computational Complexity, Computational Engineering, Finance, and Science, Computational Geometry, Computation and Language, Cryptography and Security, Computer Vision and Pattern Recognition, Computers and Society, Databases, Distributed, Parallel, and Cluster Computing, Digital Libraries, Discrete Mathematics, Data Structures and Algorithms, Emerging Technologies, Formal Languages and Automata Theory, General Literature, Graphics, Computer Science and Game Theory, Human-Computer Interaction, Information Retrieval, Information Theory, Machine Learning, Logic in Computer Science, Multiagent Systems, Multimedia, Mathematical Software, Numerical Analysis, Neural and Evolutionary Computing, Networking and Internet Architecture, Other Computer Science, Operating Systems, Performance, Programming Languages, Robotics, Symbolic Computation, Sound, Software Engineering, Social and Information Networks, Systems and Control
Mathematics (Math)	32	Commutative Algebra, Algebraic Geometry, Analysis of PDEs, Algebraic Topology, Classical Analysis and ODEs, Combinatorics, Category Theory, Complex Variables, Differential Geometry, Dynamical Systems, Functional Analysis, General Mathematics, General Topology, Group Theory, Geometric Topology, History and Overview, Information Theory, K-Theory and Homology, Logic, Metric Geometry, Mathematical Physics, Numerical Analysis, Number Theory, Operator Algebras, Optimization and Control, Probability, Quantum Algebra, Rings and Algebras, Representation Theory, Symplectic Geometry, Spectral Theory, Statistics Theory
Physics (Phys)	51	Cosmology and Nongalactic Astrophysics, Earth and Planetary Astrophysics, Astrophysics of Galaxies, High Energy Astrophysical Phenomena, Instrumentation and Methods for Astrophysics, Solar and Stellar Astrophysics, Disordered Systems and Neural Networks, Mesoscale and Nanoscale Physics, Materials Science, Other Condensed Matter, Quantum Gases, Soft Condensed Matter, Statistical Mechanics, Strongly Correlated Electrons, Superconductivity, General Relativity and Quantum Cosmology, High Energy Physics - Experiment, High Energy Physics - Lattice, High Energy Physics - Phenomenology, High Energy Physics - Theory, Mathematical Physics, Adaptation and Self-Organizing Systems, Chaotic Dynamics, Cellular Automata and Lattice Gases, Pattern Formation and Solitons, Exactly Solvable and Integrable Systems, Nuclear Experiment, Nuclear Theory, Accelerator Physics, Atmospheric and Oceanic Physics, Applied Physics, Atomic and Molecular Clusters, Atomic Physics, Biological Physics, Chemical Physics, Classical Physics, Computational Physics, Data Analysis, Statistics and Probability, Physics Education, Fluid Dynamics, General Physics, Geophysics, History and Philosophy of Physics, Instrumentation and Detectors, Medical Physics, Optics, Plasma Physics, Popular Physics, Physics and Society, Space Physics, Quantum Physics
Electrical Engineering and Systems Science (EESS)	4	Audio and Speech Processing, Image and Video Processing, Signal Processing, Systems and Control
Economics (Econ)	3	Econometrics, General Economics, Theoretical Economics
Statistics (Stat)	6	Applications, Computation, Methodology, Machine Learning, Other Statistics, Statistics Theory

Table 5: arXiv categories.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.