Beyond original Research Articles Categorization via NLP

Rosanna Turrisi[1,2]

[1]DIBRIS, University of Genova, Genova, 16146, Italy
[2]Machine Learning Genoa (MaLGa) center, University of Genova, Genova, 16146, Italy

Abstract
This work proposes a novel approach to text categorization – for unknown categories – in the context of scientific literature, using Natural Language Processing techniques. The study leverages the power of pre-trained language models, specifically SciBERT, to extract meaningful representations of abstracts from the ArXiv dataset. Text categorization is performed using the K-Means algorithm, and the optimal number of clusters is determined based on the Silhouette score. The results demonstrate that the proposed approach captures subject information more effectively than the traditional arXiv labeling system, leading to improved text categorization. The approach offers potential for better navigation and recommendation systems in the rapidly growing landscape of scientific research literature.

Keywords
Natural Language Processing, Semi-supervised learning, Research Article categorization

1. Introduction

In the past decade there has been a significant increase in the number of published research papers, creating a need for better tools to navigate through the vast literature. ArXiv[1], an open-access archive, has emerged as the most popular platform, housing over two million scientific articles in various fields such as physics, mathematics, computer science, biology, finance, statistics, engineering, and economics. Currently, authors manually assign subject categories to their own articles during submission. However, this process is time-consuming and restricts the labels to sector-based categories. Consequently, inter-disciplinary works focusing on similar topics often receive different labels. For example, two articles studying brain cancer – one using Artificial Intelligence (AI) and the other using statistics – would be assigned to the categories of *Computer Science* and *Statistics*, respectively, even though they investigate the same phenomenon. On the other hand, the AI-based cancer study would share the category label with a work on Operating Systems, resulting in more challenging literature search and less efficient recommendation systems.

Recent advancements in Natural Language Processing (NLP) and the success of pre-trained

Workshop on Human-in-the-Loop Applied Machine Learning (HITLAML), September 04-06, 2023 - Belval, Luxembourg

rosanna.turrisi@edu.unige.it (R. Turrisi)
https://rosanna-turrisi.webnode.it/ (R. Turrisi)
0000-0002-0877-7063 (R. Turrisi)

models [2] have opened up new possibilities for processing text data and performing various tasks such as top modeling [3, 4, 5], text classification [6, 7, 8], and information retrieval [9].

This work leverages NLP to process abstracts of ArXiv papers and classify them into more meaningful and flexible subject categories that go beyond the original labeling. The aim is to create categories providing information about the original subject categories but less restrictive and sector-based. The ultimate goal is to enhance literature search and recommendation systems by providing more accurate and relevant categorization. The proposed approach differs from most studies on text categorization [10, 11] in two main aspects: i) the optimal number of categories (*N*) is unknown, and determining its best value poses a significant challenge; ii) although text categorization is performed in an unsupervised setting, the feature extraction process incorporates knowledge about the original subject labeling resulting in a hybrid approach.

Main contributions This study explores four different abstract embeddings based on the SciBERT pre-trained model [12]. Each embedding is used as input for the K-means algorithm, enabling unsupervised text categorization. The optimal number of categories is determined by evaluating the model performance using the Silhouette score on the validation set. Results demonstrate that this approach effectively captures information from the ArXiv subject categories while providing more meaningful text categorization. For instance, it successfully collapses distinct category labels from ArXiv (e.g., `stat.Th` and `math.ST`) that correspond to the same subject (e.g., *Statistic Theory*) into a single class category. The implemented pipeline was developed using the Python programming language. Its code can be accessed on GitHub.

2. Related work

The automatic classification of research publications is commonly achieved by assigning papers to existing categories within hierarchically-structured vocabularies, such as Medical Subject Headings (MeSH) [13], Physics Subject Headings (PhySH) [14], and the STW Thesaurus for Economics [15]. For instance, [16] introduces three deep learning architectures for article classification, utilizing either the paper title or the full-text as input. Notably, results show that the title-based method performs comparably to the full-text-based approach. Similar investigations are reported in [17, 18], in which the aim is either surpass or achieve results equivalent to the full-text-based approach. This is accomplished by solely utilizing paper titles, which benefit from widespread availability. An intermediate approach is presented in [19], where category classification is accomplished using paper abstracts. This strategy enables the utilization of abundant data while capturing more comprehensive and insightful information about paper content. However, these studies thrive only within a very well-defined category framework. Pushing further, various techniques [20, 21, 22, 23, 24, 25] integrate machine learning methods with background knowledge to identify research topics in documents. Despite enhancing automated paper classification quality, these methods rely on ontologies that are time-consuming and require careful planning and expertise. This manuscript capitalizes on a domain-specific taxonomy, taking advantage of its immediate accessibility and usability. However, it acknowledges that pre-defined categories might not encompass the entire intricacies of concepts and connections present in a formal ontology. To address this gap, an unsupervised machine

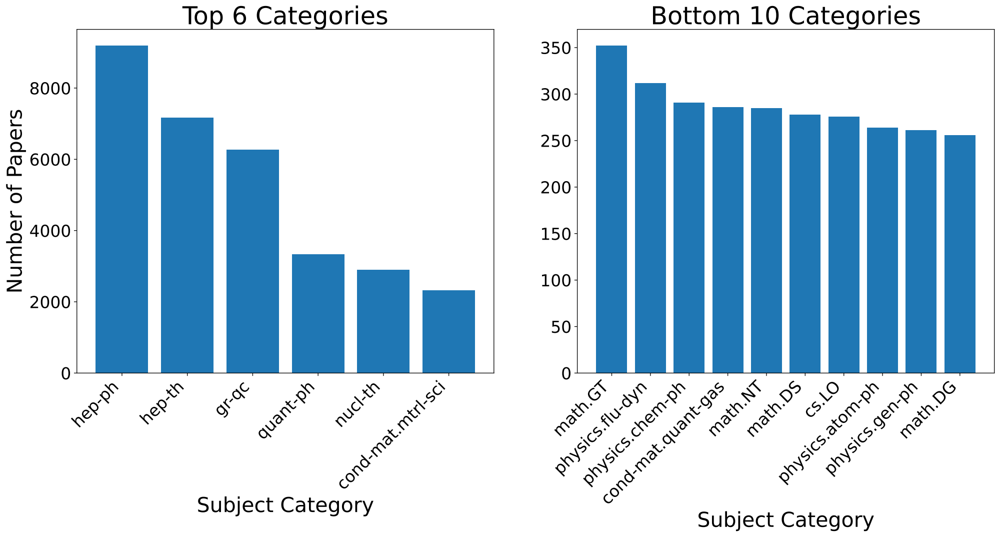


Figure 1: Number of papers in the mostly frequent (left) and the less frequent subject categories (right).

learning algorithm is here proposed to extract complex information from paper abstracts and identify research categories that are related to, but distinct from, the pre-existing categories, offering a more refined representation of the research articles.

3. Dataset

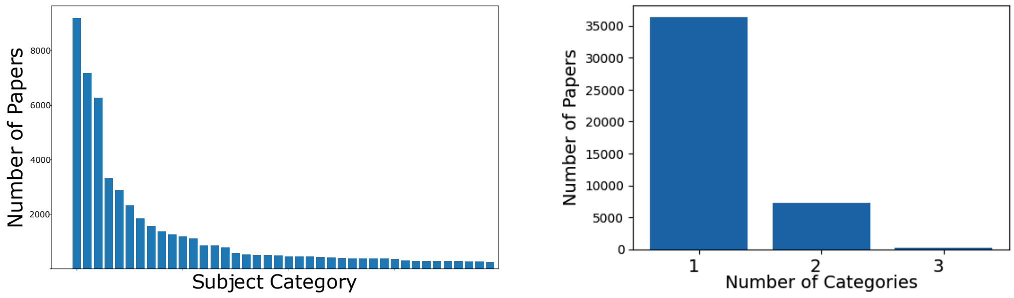
The ArXiv Dataset [26] is a rich corpus of approximately two million articles, including author information, title, journal references, ArXiv categories, and abstract. For computational reasons, the subset of articles published in 2023 was selected. To ensure data quality, duplicated and withdrawn papers were removed from the dataset. Categories with a small number of papers, specifically those containing less than 250 articles, were also excluded. Furthermore, abstracts with fewer than 31 words were not considered in the analysis. This decision was made due to the observation that such short abstracts often contain meaningless or misleading texts, such as sentences indicating revisions or the absence of an abstract for comments. This resulted in a final corpus of 43853 samples and reduced the memory usage from 240.3MB to 2.3MB.

3.1. Data analysis

ArXiv categories. The selected ArXiv subset comprises 40 unique subject categories (e.g., *Algebric Geometry*) from which 15 macro categories (e.g., *Mathematics*) can be retrieved.

Fig. 1 (left) shows the number of papers associated with the 6 most frequent categories. Specifically, it displays the number of papers associated to

1. *High Energy Physics - Phenomenology* (hep-ph);
2. *High Energy Physics - Theory* (hep-th);
3. *General Relativity and Quantum Cosmology* (gr-qc);



(a) **(b)**

Figure 2: ArXiV categories distribution study. (a) Frequency of all categories. (b) Number of papers with one or more categories associated.

4. *Quantum Physics* (quant-ph);
5. *Nuclear Theory* (nucl-th);
6. *Materials Science* (cond-mat.mtrl-sci).

The most frequent category contains more than 9000 papers while the 6th most frequent category includes about 2000 papers. Fig. 1 (right) reports instead the 10 less frequent categories: math.GT (*Geometric Topology*), physics.flu-dyn (*Fluid Dynamics*), physics.chem-ph(*Chemical Physics*), cond-mat.quant-gas (*Quantum Gases*), math.NT (*Number Theory*), math.DS (*Dynamical Systems*), cs.LO (*Logic in Computer Science*), physics.atom-ph (*Atomic Physics*), physics.gen-ph (*General Physics*), math.DG (*Differential Geometry*). As we can see, the number of papers is between 250 and 350 for all of them.

Fig. 2 (a) provides a visual representation of the number of papers within each category. It is evident from the graph that the distribution of category labels is highly unbalanced and that the large majority of categories include less than 1000 articles. Since some papers may be associated with multiple labels, a histogram illustrating the number of categories associated with each paper is presented in Figure 2 (b). This histogram provides insights into the distribution of multiple category assignments for individual papers.

Abstract length. Fig. 3 reports an histogram of abstract length in terms of number of words. As we can see, the abstracts' length typically ranges from 50 to 300 with few exceptions. Table 1 shows average, standard deviation (Std), minimum (Min), maximum (Max) and first, second and third quartiles (25%, 55%, 75%) of the distribution of abstracts' length. This analysis is relevant for the study as the third quartile was used to fix a maximum text length in the tokenization process.

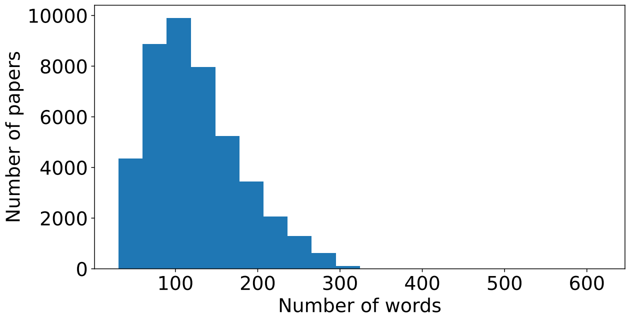


Figure 3: Abstracts length distribution.

Table 1: Abstracts length statistics

N. samples	43853
Mean	124.5
Std	55.3
Min	31
25%	83
50%	115
75%	157
Max	617

3.2. Data processing

Text processing was performed using the `spaCy` [27] package, specifically using the `en_core_sci_lg` component. This component is trained on scientific papers and offers a large vocabulary and 600,000 word vectors. Each abstract in the dataset was first syntactically parsed in order to find linguistic units and their grammar dependencies. Then, the following processing steps were applied: i) the text was converted to lowercase; ii) lemmatization was performed on linguistic units, excluding personal pronouns, to transform them into their base form; iii) punctuation marks and stop words were removed from the text. Finally, processed data was tokenized using the `AutoTokenizer` class from the `Hugging-face/transformers` package [28].

4. Experimental setup

4.1. Embedding estimation

Text embedding was performed by relying on the SciBERT pre-trained model [12], a language model trained on scientific texts and known to effectively capture domain-specific information. Two different text representations were investigated in this study: i) SciBERT-T, obtained by extracting the last hidden layer of the first token in the input sequence; ii) SciBERT-CLS, obtained by extracting the last hidden layer of the classification token in the input sequence. Both of them provide a 768-dimensional dense vector representation of the input text.

PCA-SciBERT. A feature reduction method was applied to the SciBERT embeddings to address the curse of dimensionality and discard irrelevant features. Specifically, Principal Component Analysis (PCA) was applied separately to the SciBERT-T and SciBERT-CLS representations by retaining the 95% of the variance in the data. This reduced the SciBERT-T embedding to a 325-dimensional vector, and the SciBERT-CLS embedding to a 122-dimensional vector.

Fig. 4 and 5 provide a qualitative visual assessment of the resulting embeddings, highlighting the 6 most frequent subject categories. The t-SNE method [29] is employed to project the samples into a 2D space. In both figures, samples belonging to the selected ArXiv category

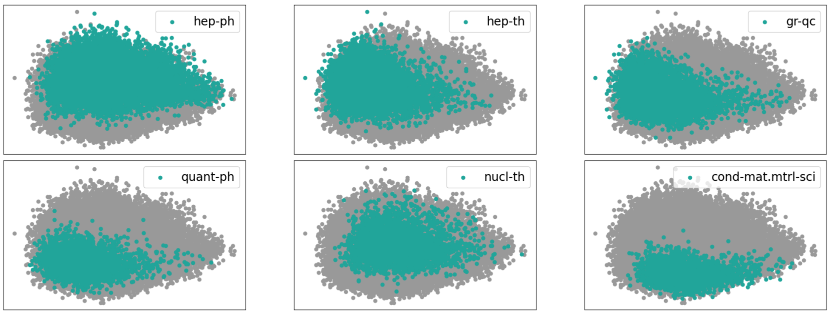


Figure 4: 2D projection of PCA-SciBERT-T embeddings by using t-SNE. Labelled samples belonging to the same arXiv subject category are shown in green, while the others in grey. The graph reports texts from the 6 most frequent categories.

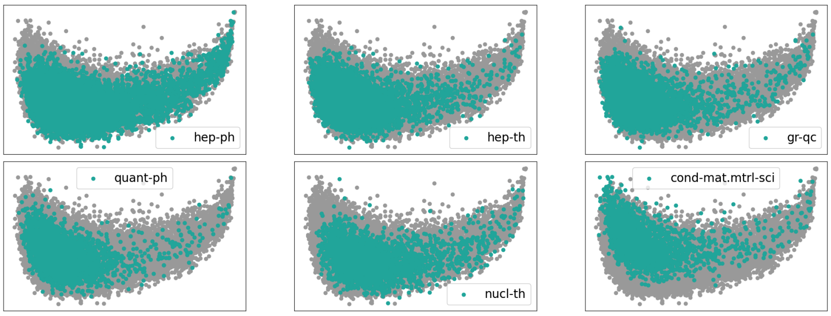


Figure 5: 2D projection of PCA-SciBERT-CLS embeddings by using t-SNE. Labelled samples belonging the same arXiv subject category are shown in green, while the others in grey. The graph reports texts from the 6 most frequent categories.

are represented as green points, while the remaining categories are displayed in grey. The categories are presented in descending order of frequency.

The observed clustering of the green points in the extracted embeddings indicates a tendency for points with similar category associations to group together. However, it is important to note that the 3 most frequent categories exhibit overlapping clusters, particularly in the PCA-SciBERT-CLS representation. This observation should be interpreted considering two factors: i) a paper abstract can have multiple labels, as illustrated in Figure 2 (b); and ii) the representation of distant points in a high-dimensional space may not be accurately captured in a 2D projection.

FT-SciBERT. As reported in [12], NLP tasks performance is generally improved by Fine-Tuning (FT) the SciBERT model for a small number of epochs. Hence, a FT approach was here explored to enhance text representation. The ArXiv subject category associations were used as labels, with the possibility of multiple labels for each sample. To implement the FT, a dense layer with 32 nodes was added on top of the SciBERT model, followed by a classification layer. This additional hidden layer helps incorporate specific prior knowledge about the arXiv dataset into the FT process. While the classification task captures subject information by improving the embedding representation, it should be noted that the embedding will not be a direct representation of the arXiv subject categories due to the presence of the added hidden layer.

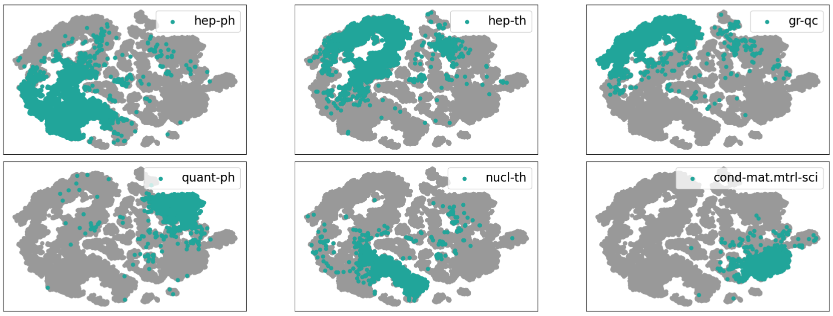


Figure 6: 2D projection of FT-SciBERT-T embeddings of learning set. Labelled samples belonging the same arXiv subject category are shown in green, while the others in grey. The graph reports texts from the 6 most frequent categories.

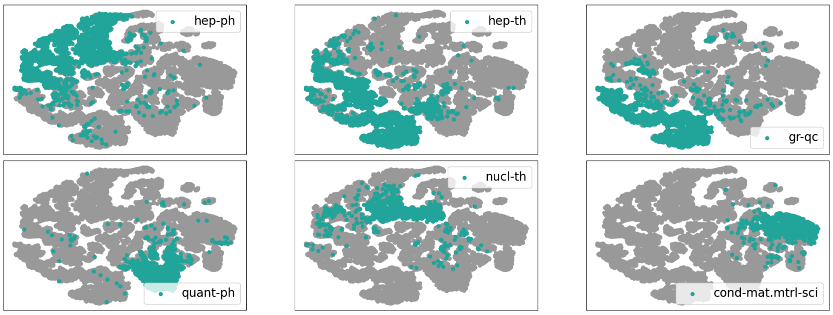


Figure 7: 2D projection of FT-SciBERT-CLS embeddings of learning set. Labelled samples belonging the same arXiv subject category are shown in green, while the others in grey. The graph reports texts from the 6 most frequent categories.

The dataset was split into learning (90%) and testing (10%) set. The learning set was further split into training (80%) and validation (20%) set. Training was performed by using 32 as batch size, 2e-5 as learning rate, and dropout with a dropping probability of 0.1. These parameter values were reported in [12] as the best performing across different datasets and tasks. The maximum number of epochs was set to 4, and early stopping was employed with a patience value of 1, allowing the training process to stop if there was no improvement in performance.

Fig. 6 and 7 illustrate the embeddings obtained from the learning set using FT-SciBERT-T and FT-SciBERT-CLS, respectively. In all figures, green points represent the 2D projections of the embedding points associated with the 6 most frequent subject categories, while grey points represent other categories. The t-SNE algorithm was used to obtain the 2D projection. As anticipated, fine-tuning the model improved the embedding representation, resulting in more distinct clustering of abstracts with the same arXiv category.

4.2. Unsupervised text categorization

Unsupervised text classification was performed using K-Means algorithm for each text representation. The goal was to determine the optimal number of categories, denoted by N, for the

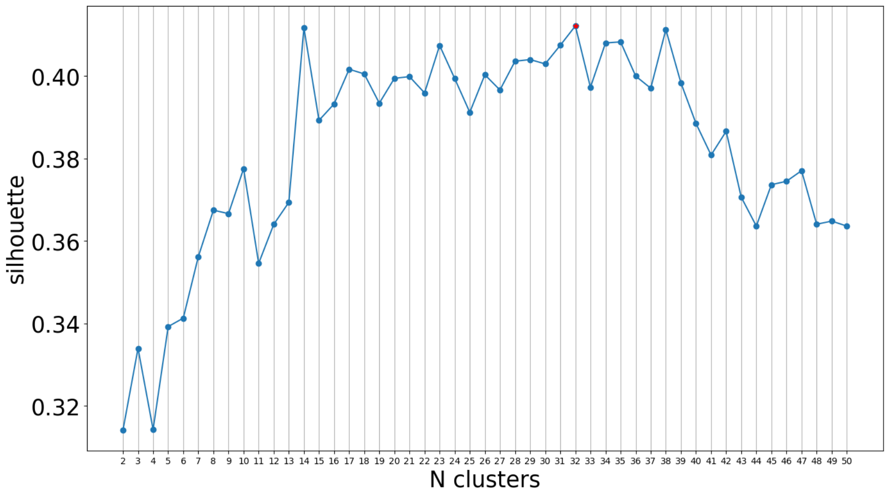


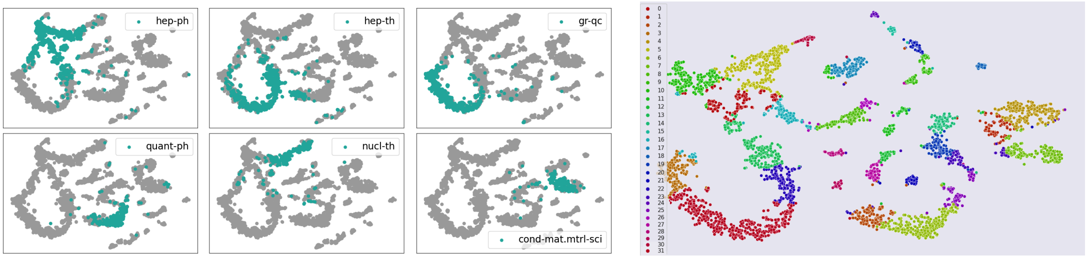
Figure 8: Silhouette score of K-Means algorithm applied to FT-SciBERT-CLS for different number of clusters (N) ranging from 2 to 50. The red point indicates the highest score achieved at $N = 32$.

clustering task. The range of N was set from 2 to 50. To evaluate the clustering performance for different values of N, the algorithm was trained on the training set and then evaluated on the validation set using the Silhouette metric. The Silhouette score ranges from -1 to 1, where a score of -1 indicates that points are wrongly assigned to clusters, a score of 0 suggests overlapping clusters, and a score of 1 indicates that points are perfectly assigned to well-separated clusters. The Silhouette metric was used as a criterion to determine the optimal number of categories. The value of N that yielded the highest Silhouette score on the validation set was taken as the best choice for the number of categories. Alternatively, one may adopt the Within-Cluster Sum of Square (WCSS) curve. The WCSS curve measures the sum of squared distances between each point and its assigned cluster centroid. Typically, the curve exhibits an 'elbow' shape, and the optimal N corresponds to the point where the curve starts to flatten out significantly. However, in this study, the elbow method was not utilized as the WCSS curve was found to be too smooth, making it challenging to identify a clear elbow point.

5. Results

When using PCA-SciBERT representations, the Silhouette scores on the validation set was found to be close to 0 for any value of N. This suggests that these embedding points are difficult to associate with non-overlapping clusters, indicating a lack of clear separation between categories. On the other hand, when using FT embeddings, the K-Means clustering algorithm achieved better results. The Silhouette score for FT-SciBERT-T was 0.36, while FT-SciBERT-CLS achieved a score of 0.41. As the higher Silhouette score for FT-SciBERT-CLS indicated better cluster separation and cohesion, further observations and the evaluation on the testing set were conducted only for the FT-SciBERT-CLS embedding.

Fig. 8 presents the evaluation results on the validation set using the Silhouette metric for



(a) Green points visualizes articles belonging to same arXiv subject category, while the others are plotted in grey.

(b) K-Means clusters for $N = 32$.

Figure 9: 2D t-SNE projection of the testing set with color-coded labels based on (a) arXiv categories (green) and (b) K-means clustering.

different numbers of clusters (N) ranging from 2 to 50. As expected, the Silhouette score initially increases as N increases, reaching a peak value, and then starts to decline rapidly for $N > 38$. The best clustering performance is observed at $N = 32$, which falls between the number of macro subject categories (15) and the number of subject categories (40) of ArXiv. This finding suggests that the arXiv categories do not fully capture the underlying structure of the abstract classes.

Unsupervised text categorization A qualitative evaluation of FT-SciBERT-CLS embedding can be observed in 9 (a). The 2D t-SNE projection of the testing set shows samples colored based on the 6 most frequent arXiv categories (green), while the grey points represent other categories. The clustering of green points indicates that the embedding captures arXiv category information. However, there is some overlap between embeddings of `hep-th` (*High Energy Physics - Theory*) and `gr-qc` (*General Relativity and Quantum Cosmology*) – two related scientific field –, suggesting that the FT-SciBERT-CLS representation goes beyond arXiv categories.

For the final evaluation of the categorization task, $N = 32$ was chosen based on the results obtained on the validation set. The classification model achieved a Silhouette score of 0.4 on the testing set. Figure 9 (b) displays a t-SNE projection of the K-Means results, where the testing samples are color-coded based on their assigned K-Means class. A more detailed analysis of the results on the testing set and the relationship between the identified and the arXiv categories can be seen in Figure 10. The figure presents four bar charts representing the four largest clusters. Each chart displays the distribution of papers across different arXiv categories within the cluster. Interestingly, most clusters exhibit one or a few dominant peaks, indicating a strong association with the arXiv category tags. This pattern is observed consistently across all K-means clusters.

A further examination was conducted to evaluate the top three most frequent subject categories within each cluster. Table 2 presents the list of the clusters sorted by decreasing cardinality, excluding arXiv categories with fewer than 10 samples. Results indicate that 66% of clusters – i.e. the identified categories – correspond to a predominant arXiv category or multiple subject categories within the same macro-category. For example, cluster 8 contains samples associated with *Astrophysics of Galaxies*, *Astrophysics*, and *Cosmology and Non-galactic Astrophysics*, all belonging to the macro-category of *Astrophysics*. According to the results, 21%

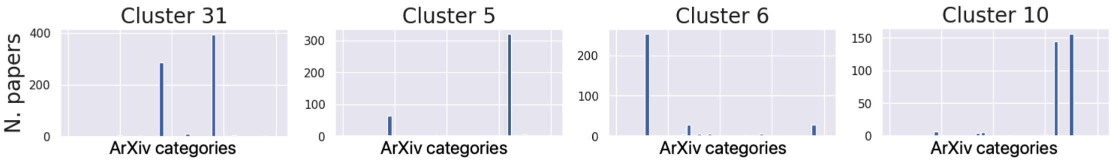


Figure 10: Number of papers associated to arXiv cateogries in the 4 most populated K-Means clusters.

Table 2
Top 3 subject categories per cluster (number of papers per category in parenthesis).

Cluster	1st most frequent	2nd most frequent	3rd most frequent
31	General Relativity and Quantum Cosmology (394)	High Energy Physics - Theory (286)	-
5	High Energy Physics - Phenomenology (320)	High Energy Physics - Experiment (63)	-
6	Quantum Physics (253)	Mathematical Physics (mat.MP) (28)	Mathematical Physics (math-ph) (28)
10	Nuclear Theory (156)	High Energy Physics - Phenomenology (145)	-
13	High Energy Physics - Theory (163)	High Energy Physics - Phenomenology (145)	-
4	Mesoscale and Nanoscale Physics (133)	Materials Science (125)	Strongly Correlated Electrons (10)
3	General Relativity and Quantum Cosmology (167)	Astrophysics (36)	High Energy Physics - Theory (23)
7	Superconductivity (102)	Strongly Correlated Electrons (85)	Materials Science (11)
22	High Energy Physics - Theory (162)	General Relativity and Quantum Cosmology (17)	High Energy Physics - Phenomenology (10)
0	High Energy Physics - Phenomenology (162)	Nuclear Theory (15)	-
16	High Energy Physics - Phenomenology (83)	Astrophysics (74)	General Relativity and Quantum Cosmology (13)
8	Astrophysics of Galaxies (54)	Astrophysics (39)	Cosmology and Non-galactic Astrophysics (34)
2	High Energy Physics - Theory (44)	Mathematical Physics (math-ph) (37)	Mathematical Physics (mat.MP) (37)
17	Nuclear Theory (90)	Nuclear Experiment (36)	-
25	Quantum Physics (40)	Quantum Gases (25)	Atomic Physics (15)
19	Statistical Mechanics (78)	-	-
14	Soft Condensed Matter (77)	-	-
9	Geometric Topology (33)	Algebraic Geometry (28)	Differential Geometry (26)
1	Materials Science (72)	-	-
11	Plasma Physics (33)	Fluid Dynamics (25)	Chemical Physics (10)
23	Strongly Correlated Electrons (24)	Statistical Mechanics (13)	-
27	Analysis of PDEs (42)	Dynamical Systems (15)	-
12	Probability (42)	-	-
20	Number Theory (24)	Algebraic Geometry (11)	-
28	High Energy Physics - Lattice (33)	High Energy Physics - Phenomenology (12)	-
21	Optics (35)	-	-
30	High Energy Physics - Experiment (42)	-	-
26	High Energy Astrophysical Phenomena (25)	Astrophysics (12)	-
29	Statistics Theory (stat.TH) (31)	Statistics Theory (math.ST) (31)	-
18	Computer Vision and Pattern Recognition (35)	-	-
15	Combinatorics (26)	-	-
24	Logic in Computer Science (29)	-	-

of clusters have two main macro-categories, while only 7% have three. Notably, the K-Means algorithm successfully classifies texts that share the same subject area but are labeled with different category tags in arXiv. For instance, the `stat.Th` and `math.ST` papers, which correspond to the same subject of *Statistic Theory* are correctly clustered together in cluster 29. Similarly, cluster 2 and cluster 6 group together abstracts labeled as `math-ph` and `mat.MP` respectively, both representing the subject of *Mathematical Physics*. This demonstrates that the proposed approach effectively identifies and clusters abstracts based on their subject matter, overcoming the limitations of the original category labels and providing a more meaningful classification.

6. Conclusion

This study introduces a novel approach for text categorization in scientific literature using pre-trained language models. By surpassing the limitations of traditional arXiv subject categories, this method enables more meaningful and accurate categorization of abstracts. The results demonstrate the effectiveness of the proposed approach in capturing the underlying subject matter, improving literature search and recommendation systems. This research contributes to the advancement of natural language processing techniques and addresses the need for more efficient tools in navigating scientific literature. Further studies will extend the proposed approach on other datasets from various scientific disciplines.

References

[1] P. Ginsparg, arxiv, 1991. URL: https://arxiv.org/.

[2] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Science China Technological Sciences 63 (2020) 1872–1897.

[3] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, E. Fersini, Cross-lingual contextualized topic models with zero-shot learning, arXiv preprint arXiv:2004.07737 (2020).

[4] D. Ramage, E. Rosen, J. Chuang, C. D. Manning, D. A. McFarland, Topic modeling for the social sciences, in: NIPS 2009 workshop on applications for topic models: text and beyond, volume 5, 2009, pp. 1–4.

[5] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[6] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[7] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18, Springer, 2019, pp. 194–206.

[8] Q. Liu, H.-Y. Huang, Y. Gao, X. Wei, Y. Tian, L. Liu, Task-oriented word embedding for text classification, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 2023–2032.

[9] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2724–2743.

[10] R. Gonzalez-Marquez, L. Schmidt, B. M. Schmidt, P. Berens, D. Kobak, The landscape of biomedical research, bioRxiv (2023) 2023–04.

[11] F. R. Lumbanraja, E. Fitri, A. Junaidi, R. Prabowo, et al., Abstract classification using support vector machine algorithm (case study: abstract in a computer science journal), in: Journal of Physics: Conference Series, volume 1751, IOP Publishing, 2021, p. 012042.

[12] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[13] Medical subject headings (mesh), 1960. URL: https://www.nlm.nih.gov/mesh/meshhome.html.

[14] Physics subject headings (physh), 2016. URL: https://physh.aps.org.

[15] Stw thesaurus for economics, 1998. URL: http://zbw.eu/stw/version/latest/about.

[16] F. Mai, L. Galke, A. Scherp, Using deep learning for title-based semantic subject indexing

to reach competitive performance to full-text, in: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, 2018, pp. 169–178.

[17] L. Galke, F. Mai, A. Schelten, D. Brunsch, A. Scherp, Using titles vs. full-text as source for automated semantic document annotation, in: Proceedings of the Knowledge Capture Conference, 2017, pp. 1–4.

[18] C. Nishioka, A. Scherp, Profiling vs. time vs. content: What does matter for top-k publication recommendation based on twitter profiles?, in: Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries, 2016, pp. 171–180.

[19] B. Kandimalla, S. Rohatgi, J. Wu, C. L. Giles, Large scale subject category classification of scholarly papers with deep attentive neural networks, Frontiers in research metrics and analytics 5 (2021) 600382.

[20] F. Osborne, E. Motta, Mining semantic relations between research areas, in: The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11, Springer, 2012, pp. 410–426.

[21] G. Erétéo, F. Gandon, M. Buffa, Semtagp: semantic community detection in folksonomies, in: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 1, IEEE, 2011, pp. 324–331.

[22] A. A. Salatino, F. Osborne, T. Thanapalasingam, E. Motta, The cso classifier: Ontology-driven detection of research topics in scholarly articles, in: Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23, Springer, 2019, pp. 296–311.

[23] K. Hitha, V. Kiran, Topic recognition and correlation analysis of articles in computer science, in: 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2021, pp. 1115–1118.

[24] S. Al-Shareef, R. Alharbi, R. Alharbi, R. Almfarriji, M. Alsharif, R. Alharthi, L. Althaqafi, Investigating community detection in arabic scholarly network using ontology-based semantic expansion, in: 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2022, pp. 96–103.

[25] A. Sharma, S. Kumar, Machine learning and ontology-based novel semantic document indexing for information retrieval, Computers & Industrial Engineering 176 (2023) 108940.

[26] C. U. Library, 2019. URL: https://www.kaggle.com/datasets/Cornell-University/arxiv.

[27] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[29] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.