Domain-Specific Article Classification using BERT-Based Models

Sharon Kurant **Tal Saphar**

Abstract

In this study, we present an approach for automatically classifying scientific articles to their respective domains based on their titles and abstracts using two BERT-based models; BERT and SciBERT. The motivation behind this study lies in the growing volume of scientific literature, which makes it increasingly challenging for researchers to keep up with the latest findings in their fields. By automating the classification process, researchers can efficiently access articles relevant to their research interests.

The methods used in our study consist of two classification stages using the two BERT-based models. The first method involves the classification of scientific abstracts into broad topic domains i.e. Computer Science, Physics, etc. The second stage further classifies the abstracts into more specific research sub-domains i.e. Deep Learning sub-domain in the Computer Science domain (cs.DL).

The classification of articles into their respective research domains presents several challenges, particularly due to the large number of labels and limited data available in our dataset. In addition, articles that belong to the same general domain can often contain similar terminology and concepts, which can make it difficult to classify them accurately into different research sub-domains. This underscores the importance of model training to improve the accuracy of the classification task.

The results of our experiments demonstrate that our proposed approach achieves high accuracy in classifying articles to their respective domains and sub-domains.

1 Data

With the increasing number of scientific articles available online, it has become more difficult for researchers to find relevant articles. Topic modeling is a technique that can be used to automatically identify topics in a large collection of text documents, such as research articles. In this study, we attempt to predict the domains for a set of research articles based on their abstracts and titles. The research articles are sourced from six different domains, including Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, and Quantitative Finance. Due to the imbalance in the dataset domains which includes a relatively small number of samples for Quantitative Biology and Quantitative Finance compared to Computer Science and Physics. We decided to exclude the underrepresented domains from the dataset in order to prevent a potential impact on the performance and the generalization ability of the models.

1.1 Data Source

The data for this task is sourced from Kaggle, a platform that hosts machine learning and data science competitions. The dataset contains a set of research articles from four different domains, including 20,316 training samples (CS, Math, Physics, and Stat.). The training samples are labeled with one or more domains. The abstract and title for each research article are provided in the dataset.

1.2 Data labeling

The labeling process for the dataset was done automatically by querying Arxiv and extracting the relevant domains. Each research article was labeled with one or more domains. The labeling process was done to ensure that the training dataset is properly labeled and can be used to train a machine learning model.

1.3 Data Statistics

The dataset consists of 20,316 samples. We can observe in Table 1 and Figure 1.1 (figures are in the appendix) the frequency of the articles under each domain. Regarding the number of domains assigned to each article, as shown in Figure 1.2 the

Domain	#Articles
Computer Science	8594
Physics	6013
Mathematics	5618
Statistics	5206

Table 1: Number of Articles per Domain in the Research Article Dataset.

Model	Abstract	Title	Combined
BERT	0.903	0.870	0.904
BERT-FT	0.802	0.763	0.807
SciBERT	0.915	0.891	0.916
SciBERT-FT	0.854	0.831	0.854

Table 2: Accuracy scores of different BERT models with and without fine-tuning on variant data.

majority of articles are labeled with only one domain. However, there exist a considerable number of articles, around 6,000 in total, that are labeled with more than one domain. Such multi-labeled articles may pose an additional challenge in the prediction of domains, as they introduce a degree of complexity in the relationship between the article's content and the domains it belongs to. In Figure 1.3 we can observe that on average, the titles contain around 12 tokens, while the abstracts are more lengthy, containing approximately 200 tokens. Although some abstracts may extend up to 700 tokens, such instances are relatively infrequent.

2 Methods

The proposed method involves two stages of classification. In the first stage, we used the different BERT and SciBERT pre-trained models to classify each scientific abstract into one of the broad domains, such as Mathematics or Computer Science. In the second stage, we further classified each abstract into more granular sub-domains, such as Linear Algebra or Deep Learning.

An additional issue we addressed was of imbalanced sub-domain data within the context of the model's weights. To mediate this issue, a weight adjustment approach was utilized in order to enhance the performance on the minority classes. Specifically, the method involved the incorporation of additional weights to the positive examples.

2.1 Pre-processing

For each article in the dataset, we generated a label vector by extracting the corresponding values from the dataframe columns, where a value of 1 indicates the article's inclusion in a particular domain, and a value of 0 signifies its exclusion from the domain. To enhance the reliability and validity of the experimentation process, we included a concatenation technique by merging the first 20 tokens of the title and the initial 280 tokens of the abstract into a single cohesive text. This unified text is subsequently processed through the BERT tokenizer to ensure optimal performance and accuracy.

2.2 Domain Classification

To perform domain classification on the research articles based on their abstracts and titles, we employed two distinct pre-trained BERT models- Vanilla BERT and SciBERT. Each of these models differs in the domain of texts they are primarily trained on. Incorporating a classification head onto the final layer of pre-trained models was added in order to acquire task-specific compatibility and enhance learning. We used an Adam optimizer to optimize the models' performance with a learning rate scheduler in order to improve the efficiency and effectiveness of the learning process. In addition, we examine the influence of input text length on classification accuracy by experimenting with three input configurations, the title only, the abstract only, and the combination of the article's title and abstract.

We compare the models results with and without fine-tuning to the pre-trained SciBERT and BERT. Fine-tuning the pre-trained models requires a longer learning process but can help optimize their performance on a specific task and improve their ability to accurately represent and classify text in a particular domain.

2.3 Sub-Domain Classification

With the goal of expanding our research scope, we attempted to classify academic articles beyond their general domain and into their specific sub-domain. To achieve this, we used the ArXiv python API to extract relevant sub-domain according to ArXiv Taxonomy pertaining to research domains of Computer Science and Statistics, these two domains were selected as the primary focus for sub-domain classification given the substantial volume of articles within them. Specifically, we utilized the library's querying capabilities to obtain new sub-domain labels, such as cs.DL, cs.AI, stat.ML, etc.,

Model	Abstract	Title	Combined
BERT	0.844	0.787	0.848
BERT-FT	0.585	0.443	0.601
SciBERT	0.866	0.824	0.867
SciBERT-FT	0.735	0.675	0.735

Table 3: F1 micro scores of different BERT models with and without fine-tuning on variant data.

Model	Abstract	Title	Combined
BERT	0.834	0.772	0.838
BERT-FT	0.477	0.342	0.496
SciBERT	0.857	0.811	0.859
SciBERT-FT	0.683	0.607	0.687

Table 4: F1 macro scores of different BERT models with and without fine-tuning on variant data.

which were subsequently assigned to each article. Notably, it was observed that each article had the potential to possess one or more labels in their respective domains, thus highlighting the complexity and diversity inherent within these research domains. To perform the classification, we used the same methods as before. By implementing these methods, we aimed to enhance the F1 macro and accuracy of our article classification process.

3 Experiment

The experimental setup was designed to provide a comprehensive analysis of the different model configurations and training data combinations. The experiments conducted in this study involved a systematic evaluation of different method variants, including the BERT and SciBERT models, with and without fine-tuning.We used the BERT model without fine-tuning as a compared baseline for the results. The data was split into training and test sets, with a ratio of 0.8 for training and 0.2 for testing. To assess the models' performance, we used varying lengths of input data, including only the article's title, only the article's abstract, and a concatenated variant of both fields.

In all the experiments above we conducted hyperparameters optimization to hyperparameters such as the learning rate, batch size, number of epochs, etc. doing this can affect the model's ability to converge and generalize to new unseen data. Therefore, we carefully tuned these hyperparameters to achieve the best performance on the target task.

4 Results

The results in Tables 2-4 indicate that the highest accuracy of 0.916 was achieved using the fine-tuned SciBERT model with both the title and abstract data, as well as F1 scores of 0.867 in F1-Micro and 0.859 in F1-Macro. The superior performance of the fine-tuned models over those without fine-tuning is evident from the results. The fine-tuned BERT model outperformed the SciBERT model without fine-tuning in all performance metrics but in comparison to the fine-tuned SciBERT model, its performance was inferior. Additionally, the results suggest that models trained on the concatenated variant data performed better than those trained on the title or abstract data separately. Overall, these findings provide insight into the importance of fine-tuning and the use of both title and abstract data in an article classification task.

4.1 Fine-tuning Models

In Table 2 we observe that both fine-tuned BERT and SciBERT models outperform non-fine-tuned models in terms of accuracy scores and F1. This finding underscores the significance of fine-tuning our task of classifying articles by their abstract and titles into respective domains. Fine-tuning the models on the task of classifying articles into their respective domains can significantly enhance their ability to discern between articles with similar terminologies and concepts, thereby improving the accuracy of their classifications. Such fine-tuning allows the models to specialize in the task and develop a more nuanced understanding of the characteristics that distinguish articles in different domains.

4.2 Input Data Length

Tables 2-4 demonstrate that the input data length can have a significant impact on the performance of the models where the models trained only on the article titles produced the lowest results, whereas those trained solely on the abstracts showed improved results.

4.3 SciBERT vs BERT

From the results, we observe that fine-tuned SciBERT model which is specifically pre-trained on scientific texts outperformed BERT models with and without fine-tuning. SciBERT has a better understanding of scientific language and is better equipped to extract domain-specific features from

Sub-Domain	Precision	Recall	F1-score	#samples
stat.AP	0.65	0.74	0.69	69
stat.CO	0.26	0.52	0.35	21
stat.ME	0.38	0.82	0.52	55
stat.ML	0.91	0.94	0.92	302
stat.TH	0.63	0.81	0.71	102
Weighted Avg.	0.75	0.87	0.80	572

Table 5: Results for fine-tuned SciBERT model classification report for Statistics sub-domains consisting of precision, recall, and F1 performance metrics. Sub-domains with higher sample counts achieved greater precision, recall, and F1 scores.

research articles. This makes it more effective at classifying research articles than the original BERT model, which lacks this domain-specific knowledge and vocabulary.

5 Discussion

5.1 Importance of Fine-Tuning

Fine-tuning a pre-trained model such as SciBert for a specific task like research article classification is a critical step in achieving high accuracy and reliable results. The reason for this is that pre-trained models like SciBert are trained on a large amount of data and general language modeling objectives, but they may not be optimized for the specific task at hand.

In the case of classifying research articles into their domain and subdomain, fine-tuning SciBert can help to optimize the model for this task by adjusting the weights of the pre-trained model to fit the target task.

5.2 Models' Evaluation Metric

One important consideration when fine-tuning the SciBert model for sub-domain classification is the choice of evaluation metric. Since subdomain classification is imbalanced, accuracy alone may not be the best metric to evaluate the model's performance. Instead, we noticed that metrics such as F1 score Macro can provide a more accurate representation of the model's performance by taking into account the true positive, false positive, and false negative rates of each class.

6 Error Analysis

The SciBert model may have had problems classifying research articles into their domain and subdomain due to several reasons. Firstly, the dataset used for training the model on the sub-domains may have been insufficient in terms of size or diversity, resulting in limited knowledge representation. Secondly, some of the domain-specific terminology used in research articles can appear in multiple domains, or may not have been well-represented in the pre-trained SciBert model vocabulary, leading to insufficient contextual understanding. Thirdly, the hierarchical structure of sub-domains within a domain may have been difficult for the model to capture and distinguish.

The imbalanced nature of the subdomain classification task may have also played a role in the model's performance. Since certain subdomains had fewer examples in the training data, the model may have had difficulty learning to classify articles in these domains accurately. Overall, improving the diversity and quality of the training data, could potentially address these issues and improve the SciBert model's classification accuracy.

7 Future Work

As this study has revealed, a BERT-based model classification of research articles into their domain and sub-domain can be hindered by several factors. To enhance the model's performance, future work could explore several avenues. One potential avenue would be to address the limitations of the dataset used for training, such as its size and diversity. Augmenting the training data with additional examples could provide a more comprehensive knowledge representation for the model. Additionally, incorporating more domain-specific terminology into the model vocabulary could improve the model's contextual understanding. Furthermore, developing a more sophisticated approach to handling the hierarchical structure of sub-domains may be necessary for improved classification performance.

Future work could also explore ways to mitigate

Sub-Domain	Precision	Recall	F1-score	#samples
cs.AP	0.36	0.70	0.47	106
cs.CO	0.69	0.94	0.79	62
cs.ME	0.79	0.93	0.85	40
cs.ML	0.50	0.92	0.65	97
cs.TH	0.50	0.88	0.64	40
cs.CO	0.59	0.88	0.70	57
cs.ME	0.76	0.84	0.80	256
cs.ML	0.83	0.95	0.88	40
cs.TH	0.56	1.00	0.72	73
cs.CO	0.76	0.94	0.84	62
cs.ME	0.54	0.96	0.69	53
Weighted Avg.	0.60	0.88	0.71	886

Table 6: Results for fine-tuned SciBERT model classification report for Computer Science sub-domains consisting of precision, recall, and F1 performance metrics.

the impact of imbalanced sub-domain data by designing new strategies to balance the dataset or to adjust the model's weights to more effectively account for the minority classes. By addressing these issues, more research could further enhance the performance of BERT-based models in classifying research articles into their domain and sub-domain.

8 Related Work

Mustafa et al. proposed a new approach to address the limitations of single-label classification and multi-label classification techniques for research article classification. The suggested approach uses Word2Vec for textual representation to capture semantic and contextual information.

Rivest et al. compared the performance of a deep learning approach with other established techniques for classifying scientific publications and found that the deep learning approach performed as well as the other techniques.

Kandimalla et al. proposed a deep attentive neural network (DANN) to classify scholarly papers by subject domains. The DANN consists of two bi-directional recurrent neural networks with an attention layer, trained on nine million abstracts from Web of Science. Results show that retraining word embedding models, using an attention mechanism, and combining word vectors with TF-IDF improves classification accuracy while also discussing strategies to mitigate imbalanced samples and overlapping domains.

9 References

Arxiv taxonomy.

Kaggle science topic classification competition data.

Jian Wu C. Lee Giles Bharath Kandimalla, Shaurya Rohatgi. 2021. Large scale subject category classification of scholarly papers with deep attentive neural networks. *Frontiers*.

Lisu Yu Muhammad Tanvir afzal Muhammad Sulaiman Ghulam Mustafa, Muhammad Usman and Abdul Shahid. 2021. Multi-label classification of research articles using word2vec and identification of similarity threshold. *Nature*.

Éric Archambault Maxime Rivest, Etienne Vignola-Gagné. 2021. Article-level classification of scientific publications: A comparison of deep learning, direct citation, and bibliographic coupling. *Plos One*.