NRK at FoRC 2024 Subtask I: Exploiting BERT-based Models for Multi-Class Classification of Scholarly Papers

Nguyen Tuan Kiet[1,2] and Dang Van Thin[1,2]*

[1] University of Information Technology - VNUHCM
[2] Vietnam National University, Ho Chi Minh City, Vietnam
21521042@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract. This study presents the system developed by team NRK for SubTask I of the Field of Research Classification (FoRC) - NLSP 2024. The task focuses on single-label, multi-class classification of general scholarly papers. Our approach exploits the capabilities of various pre-trained BERTopology models, combined with a straightforward ensemble voting scheme to enhance classification performance. The proposed system achieved competitive results, ranking within the Top 5 on the final scoreboard.

Keywords: BERT-based models · Paper Classification · FoRC · voting ensemble · FoRC shared task.

1 Introduction

The automated classification of scientific articles into their respective fields of research (FoR) plays a crucial role in various Natural Language Processing (NLP) applications. It facilitates information retrieval, knowledge organization, and facilitates scholarly search engines. While existing repositories often utilize FoR classification systems, these systems face limitations in terms of the employed taxonomy and the underlying classification model.

This paper addresses these limitations by exploring the application of BERT-based [3] models for single-label, multi-class FoR classification (FoRC) of general scholarly papers. We focus on Subtask I of the Field of Research Classification (FoRC) shared task at NLSP 2024, which aims to develop classifiers that accurately assign one of 123 predefined hierarchical classes from the ORKG taxonomy to general scholarly papers based on their available metadata (title, authors, abstract, etc.).

The contribution of this paper is presented below:

* Corresponding author: thindv@uit.edu.vn

- Investigating the effectiveness of pre-trained BERT-based models for single-label, multi-class FoRC: This study examines the potential of BERT-based models in capturing the semantic relationship within scientific articles and their corresponding research fields.
- Contributing to the understanding of automated FOR classification: The findings of this research can provide valuable insights into the effectiveness of pre-trained language models for automated FoRC tasks.

2 Related Work

Automated FoRC has been an active area of research within the Natural Language Processing (NLP) domain. This section explores existing work relevant to Subtask I of the FoRC shared task, focusing on single-label, multi-class classification of general scholarly papers.

Previous works often employed traditional machine learning models like Support Vector Machines (SVMs), Naïve Bayes, K-Nearest Neighbor and Decision Tree introduced in [2] and [6] were employed to classify scientific publications into three categories: Science, Business, and Social Science. These approaches relied on manually extracted features, often utilizing the TF-IDF method to represent the textual content.

To address the challenge of multi-label research paper classification in the face of growing publication volume, [9] propose a joint embedding approach. They utilize separate models for title and abstract processing: a Transformer-based model for title embedding and a combination of GloVe word vectors with a GRU network for abstract embedding. The final joint representation is obtained through a two-tower structure, and their method outperforms baseline models on the CiteULike dataset [8].

In [11], the authors propose a BERT-based graph convolutional neural network (BERT-GCN) model for scientific paper classification. This model leverages the strengths of both BERT and GCNs: BERT, fine-tuned with span masking, learning rate attenuation, and data augmentation, captures the semantic content of paper titles, while the GCN captures the relationships between words within the titles. By combining these elements, BERT-GCN aims to achieve superior classification performance compared to traditional methods.

While prior research has explored various approaches for FoRC, this study specifically focuses on the application of pre-trained BERT models for single-label, multi-class classification of general scholarly papers. We investigate the effectiveness of different BERT models and fine-tuning strategies in the context of Subtask I of the FoRC shared task, contributing to the understanding of their suitability for this specific task and domain. We also compared the performance of our BERT-based model with other techniques, including traditional machine learning approaches and deep learning models, further contributing to the understanding of effective methods for this task.

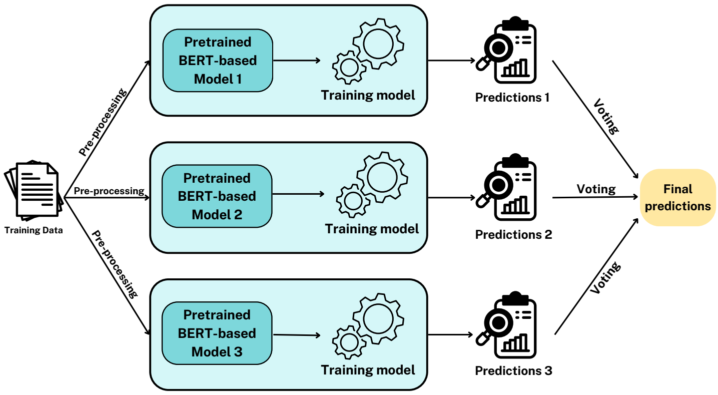

Fig. 1. The overall framework for our system.

3 Approach

We employ multiple pre-trained BERT models from the Hugging Face Transformers library suitable for text classification tasks. These models are fine-tuned on the provided dataset, which allows them to adapt to the specific domain of scientific papers and the classification task at hand. As can be seen in Figure 1, we utilize the power of pre-trained contextual language models, including the SciBERT [1], DeBERTa-V3 [4] and RoBERTa [7].

- **SciBERT**: a BERT [3] model trained on a massive dataset of scientific text, including research papers, scientific articles, and other relevant materials collected from Semantic Scholar. This allows it to understand the nuances of scientific language, such as specific terminology, jargon, and sentence structures.
- **DeBERTa-V3**: a DeBERTa [5] version improved the efficiency of original DeBERTa using ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing [4].
- **RoBERTa**: a pre-trained language model builds upon BERT [3] by addressing its limitations with dynamic masking, removing the potentially harmful NSP objective, and using a full-sentence representation.

Voting Scheme: Our motivation for applying an ensemble approach is to take advantage of the performances of various models. Given predictions $\{\hat{y}_{\theta_1}, \hat{y}_{\theta_2}, .., \hat{y}_{\theta_n}\}$ the n base classifiers. We applied the hard voting technique to merge the predictions of the base models. In hard voting, the class predicted by the majority of models becomes the final predicted class for the given paper.

4 Experimental Setup

4.1 Data

Data: We utilized the official training set for training models. The development set was used to choose which model will be included in the voting scheme.

Pre-processing: We apply different pre-processing steps as below to improve the performance for our system:

- **Cleaning and normalization**: Extensive cleaning and normalization procedures common in more general text datasets (e.g., removing punctuation, converting to lowercase) is likely not required for scientific text data. Maintaining the original text format can be beneficial for preserving domain-specific terminology and nuances within the scientific vocabulary.

4.2 Configuration Settings

The system was implemented using the Trainer API from the Hugging Face library [10] for streamlined training and evaluation of the fine-tuned BERT models. This API simplifies the process by handling data loading, model training, and metric computation. We adopted a sequence classification approach, setting a maximum input length of 512 tokens to accommodate potential variations in paper titles and abstracts. The training process consisted of 5 epochs, utilizing a batch size of 16 papers per batch. To optimize training, we employed the AdamW optimizer, commonly used for deep learning models, and incorporated a linear schedule warm-up technique to gradually increase the learning rate during the initial phase. Notably, we also addressed potential class imbalance within the dataset by employing the Focal loss function with the formula as follows:

$$\textbf{Focal Loss}(p_t) = -\alpha(1 - p_t)^{\gamma}\log(p_t), \tag{1}$$

where:
p_t = Model's predicted probability for the true class.
α = Balancing factor controls the focus of the loss function towards easy or hard examples.
γ = Focusing parameter controls the degree of modulating the loss for well-classified examples.

5 Results and Discussion

5.1 Baseline Performance:

Naive Bayes: The analysis of Naive Bayes reveals poor performance compared to other models on this task. Its performance falls significantly behind all models. This suggests that Naive Bayes might not be well-suited for this specific classification problem due to several limitations.

Support Vector Machine: While still falling short of the BERT-based models, SVM demonstrates a better performance compared to Naive Bayes and BiLSTM and even achieved a competitive result compared to deBERTa-v3-small - a transformer-based model. This suggests that the SVM model is able to capture well features extracted by applying TF-IDF.

We trained both a simple BiLSTM model and a CNN model, leveraging pre-trained word embeddings 'word2vec-google-news-300' provided by [3]Gensim, to achieve moderate performance for both. This performance falls behind the traditional machine learning models (SVM) and the pre-trained BERT models. While pre-trained word embeddings provided some improvement in capturing semantic relationships, the limitations remain:

- **Limited training data:** The available dataset might not have been sufficient for **either the BiLSTM or CNN** to learn the complex relationships needed for accurate classification, even with the additional information from pre-trained embeddings. This limitation likely explains the observed performance gap between these models and the SVM classifier. In tasks with restricted training data, the ability of deep learning models to exploit their full potential compared to traditional machine learning models can be hindered.
- **Lack of pre-trained knowledge:** Compared to BERT, **both BiLSTM and CNN** still lack pre-existing knowledge of language structures and relationships beyond the information captured in the pre-trained word embeddings. Learning these intricacies from scratch remains challenging with limited data.

5.2 Results

The performance of the participant's system is reported by the metrics which are: weighted Precision, Recall, and F1 Score, Accuracy. For individual models, the SciBERT-uncased achieved the highest performance among the individual models, followed by SciBERT-cased, RoBERTa-base, and deBERTa-v3-small. This could be attributed to the specific pre-training data and objectives of each model. SciBERT, being pre-trained on scientific text, might have a better understanding of the domain-specific language used in the papers, leading to its superior performance.

The ensemble model, combining the predictions of SciBERT-uncased, SciBERT-cased, and RoBERTa-base, outperforms all individual models, demonstrating the effectiveness of combining diverse predictions with an Accuracy of 0.7433, Precision of 0.7423, Recall of 0.7433, and F1 score of 0.7391. Among the individual models, SciBERT-uncased performs best, followed by SciBERT-cased, RoBERTa-base, and deBERTa-v3-small. The deBERTa-v3-small model was excluded from the ensemble due to its poor performance, the ensemble was thus

[3] https://radimrehurek.com/gensim/models/word2vec.html

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Naive Bayes	0.2923	0.2325	0.2923	0.1741
SVM	0.6577	0.6537	0.6577	0.6285
CNN	0.6098	0.6184	0.6098	0.6062
BiLSTM	0.5430	0.5344	0.5430	0.5145
SciBERT-uncased	0.7347	0.7301	0.7346	0.7293
SciBERT-cased	0.7279	0.7280	0.7279	0.7251
deBERTa-v3-small	0.6601	0.6592	0.6601	0.6546
RoBERTa-base	0.6984	0.6943	0.6984	0.6941
Ensemble	**0.7433**	**0.7423**	**0.7433**	**0.7391**

Table 1. Results of base models and our ensemble on the test set.

formed using only the remaining models. Consequently, we opted for the ensemble model as the final submission system over the best models based on their performance shown in Table 1 below.

Table 2 showcases the performance of our ensemble model alongside that of the top four teams. Our system ranks fifth in terms of all metrics and shows comparable performance to the fourth-ranked system (”benjwolff”) across all metrics. This suggests a competitive position with potential for improvement. The top three systems (”saliq7,” ”rosni,” and ”flo.ruo”) achieved slightly higher Accuracy and F1 scores compared to ours. This highlights areas where further development can enhance our system's ability to correctly classify examples and maintain a good balance between precision and recall.

Table 2. Results of our best submission compared with four top systems.

User	**Accuracy**	**Precision**	**Recall**	**F1**
saliq7	**0.7572 (1)**	0.7536 (3)	**0.7572 (1)**	0.7500 (3)
rosni	0.7558 (2)	**0.7566 (1)**	0.7558 (2)	**0.7540 (1)**
flo.ruo	0.7542 (3)	0.7545 (2)	0.7542 (3)	0.7524 (2)
benjwolff	0.7476 (4)	0.7438 (4)	0.7476 (4)	0.7426 (4)
kietnt0603 (Ours)	0.7433 (5)	0.7423 (5)	0.7433 (5)	0.7391 (5)

5.3 Error Analysis:

Based on the large number of labels and to focus on insightful cases, the error analysis section will highlight the following classes as shown in 3

- **Biomedical Engineering and Bioengineering**: This class has a precision, recall, and F1-score of 0.00, indicating that the model is unable to correctly classify any instances of this class. With only 1 instance in the

Table 3. Performance metrics and instance count on test set for selected classes.

Class	F1-score	Recall	Precision	# Instances
Biomedical Engineering and Bioengineering	0.00	0.00	0.00	1
Quantitative Finance	0.67	0.91	0.53	11
Computational Engineering	0.07	0.06	0.10	17
Neuroscience and Neurobiology	0.63	0.46	1.00	13

dataset, the model may be underfitting or overfitting this class due to the severe lack of data. Analyzing the misclassified instance(s) and the features used could reveal why the model struggles with this class, potentially due to non-discriminative features or similarities with instances from other classes.

- **Quantitative Finance**: This class has a high recall of 0.91 but a relatively lower precision of 0.53, resulting in an F1-score of 0.67. The high recall suggests that the model is good at identifying instances of this class, but the lower precision indicates that it also tends to incorrectly classify instances from other classes as Quantitative Finance. Investigating the instances misclassified as Quantitative Finance could provide insights into the types of features or patterns that cause the model to make these errors, potentially leading to improvements in the feature set or model architecture.
- **Computational Engineering**: This class has very low precision (0.10) and recall (0.06), leading to a poor F1-score of 0.07. With 17 instances in the dataset, the class may not be well-represented, causing the model to underfit or overfit. Analyzing the misclassified instances could reveal whether there are specific types of instances or features that the model struggles with or whether the issue is more general, guiding strategies for improving the model's performance on this class.
- **Neuroscience and Neurobiology**: This class has a perfect precision of 1.00 but a relatively low recall of 0.46, resulting in an F1-score of 0.63. The high precision indicates that when the model classifies an instance as Neuroscience and Neurobiology, it is highly likely to be correct. However, the low recall suggests that the model misses many instances of this class, classifying them as other classes. Analyzing the instances misclassified as other classes could reveal patterns or features that the model overlooks or fails to capture for this class, potentially guiding improvements in feature engineering or model architecture.

Our error analysis revealed two primary challenges contributing to the model's performance in specific classes:

- **Limited taxonomic structure:** The transformation of the original classes's hierarchical or conceptual taxonomic structure into a single-label, multi-class format (The original structure of classes can be found here) eliminates explicit relationships between classes. This flattening impedes the model's ability to distinguish between categories that share inherent similarities or belong to broader, overlapping groups within the original hierarchy. This is a common challenge in NLP tasks involving multi-class classification with complex and potentially overlapping categories, where inherent relationships between classes might not be readily captured by traditional classification approaches.

- **Imbalanced data distribution:** The limited number of instances for specific classes in the test set hinders the model's capacity to effectively learn the distinctive patterns necessary for accurate classification. This data imbalance is a well-documented issue in various NLP domains, and it can significantly impact the generalizability and robustness of machine learning models, particularly for underrepresented classes.

6 Conclusion and Future Work

This study presents a system for single-label, multi-class classification of scholarly papers. It leverages fine-tuned BERT models, ensemble voting, and the Focal Loss function to address potential class imbalance within the dataset. While achieving a rank of 5 in the shared task, the approach demonstrates promising potential for further advancement. For future work, addressing class imbalance through techniques beyond Focal Loss and extending the approach to classify other types of scientific literature holds significant potential for broader applicability.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund. We also thank the anonymous reviewers for their valuable comments on our manuscript.

References

1. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Hong Kong, China (Nov 2019)
2. Chowdhury, S., Schoen, M.P.: Research paper classification using supervised machine learning techniques. In: 2020 Intermountain Engineering, Technology and Computing (IETC). pp. 1–6 (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (Jun 2019)
4. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations (2022)

5. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentageled attention. In: International Conference on Learning Representations (2021)
6. Kim, S.W., Gil, J.M.: Research paper classification systems based on tf-idf and lda schemes. Human-centric Computing and Information Sciences **9**, 1–21 (2019)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR (2019)
8. Wang, H., Chen, B., Li, W.J.: Collaborative topic regression with social regularization for tag recommendation. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. pp. 2719–2725 (2013)
9. Wei, Y., He, Y., Yang, C.: Jetam: A joint embedding method for research paper mutil-label classification. In: 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). pp. 390–394 (2022)
10. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45 (Oct 2020)
11. Zhang, X., Yu, X., Liu, X., Lyu, X.: Scientific paper classification by fusing bert and gcn. In: 2023 International Conference on Intelligent Education and Intelligent Research (IEIR). pp. 1–6 (2023)