Categorizing Knowledge: Approaches For Effective Research Paper Classification

Saliq Gowhar*, Praveen Kempaiah*, and Dr. Sowmya Kamath S

Department of Information Technology, National Institute of Technology Karnataka
{saliqgowhar.211ee250,praveenk.211ai028,sowmyakamath}@nitk.edu.in

Abstract. A core application of Natural Language Processing is classifying scientific articles into specific fields of research. Existing classification systems often have limitations in their taxonomy of classification or the classification models. The Field of Research Classification (FoRC) shared task concentrates on classifying research papers into one of the 123 predefined classes from the ORKG taxonomy of research fields. In this study, we explores various approaches utilizing Sentence Transformer embedding coupled with Machine Learning algorithms, Neural Networks, and Transformers, to classify articles into one of 123 predefined classes. Notably our team HALE LAB NITK achieved the top ranking in the FoRC Shared Task Subtask 1. Additionally, we investigate the effectiveness of using Large Language Models (LLMs) for synthetic data generation, Synonym augmentation, and 1D Convolutional Neural Networks (CNNs) for text classification. This study advances research article classification and provides insights into enhancing classification accuracy and efficiency.

Keywords: Transformers, Sentence Transformers, One vs Rest classification, BERT, Large Language Models

1 Introduction

In this rapidly advancing world of scientific research, the sheer volume of published papers continues to skyrocket each year. With the wealth of information available, accessing relevant research on a specific topic has become increasingly time-consuming and cumbersome [1]. According to a study, since 1996, over 64 million academic papers have been published, with this number steadily climbing annually. This exponential growth necessitates the development of automated systems utilizing Natural Language Processing (NLP) and Deep learning techniques to efficiently categorize these papers into their respective research fields. By doing so, we aim to streamline access to vital information for researchers and enthusiasts alike.

Text classification is a well-explored domain in Natural Language Processing (NLP), employing a variety of machine learning and deep learning models,

* These authors contributed equally to this work

including transformers and LLMs. In this process, text is transformed into numerical representations using diverse techniques tailored to specific applications, and then fed into models to yield desired outcomes [2]. However, challenges arise particularly in specialized domains or when there's an imbalance in the number of samples across different classes, with some classes being heavily skewed. This imbalance can lead to the dominance of over represented classes, resulting in erroneous outcomes and reduced accuracy.

In this study, our goal is to address the challenge of categorizing scientific papers. We employed MiniLM and MP-Net sentence transformers for embedding generation coupled with Machine learning algorithms and Neural networks for classification. We utilized models such as SVM, Random Forest, KNN, XGBoost, and Logistic regression. Our neural network experiments involved testing various architectures, including those with increasing and decreasing pyramidal structures with power-of-2 neurons. Furthermore, we explored the use of the Focal loss function, specifically designed for imbalanced datasets, in neural networks. Transformer models such as BERT and its variants (SCIBERT, RoBERTa, DistilBERT, and AlBERT) were also part of our experimentation, where we investigated freezing and unfreezing combinations of layers. Additionally, we explored synthetic data synthesis using Gretel LLM and the use of 1D CNN for text classification. To achieve top rankings in the FoRC subtask, we utilized MPNet embeddings in combination with a One VS Rest Classifier, employing SVM with proper hyper-parameter tuning.

2 Related Work

Gündoğan et al.[1], utilised Word2Vec modeling and community discovery techniques to categorize research papers, where Word2Vec is used to find the similarities between articles followed by network creation based on similarity rates. Chowdhury et al.[3] used ML algorithms including SVM, RF and DTs to classify research papers into three distinct fields: Science, Business, and Social Science. Kadhim et al.[4] perform an examination of different term weighting methods within text classification, highlighting their role in enhancing performance. Dien et al.[5] preprocessed and vectorised articles and used SVM to achieve 91% accuracy in article classification. Atmadja et al.[6] used Naive Bayes algorithm to achieve an accuracy of 71% using a small training corpus of 150 articles. This literature survey conveys the issues related to the Field of research classification such as lack of data, limited classes of articles, and under-performing classification models.

3 Corpus Description

The dataset used in this study is provided in the FoRC Shared task and combines information from various reputable sources, including ORKG, arXiv, Crossref API, and S2AG. It comprises metadata from approximately 59.3K scholarly papers in English, classified into 123 Fields of Research. The metadata includes

crucial details such as titles, authors, DOIs, publication dates, publishers, and FoR labels. Abstracts were collected from different sources and validated against the metadata. Manual mapping of FoR labels was performed to align arXiv's taxonomy with ORKG's. The dataset's strength lies in its meticulously curated FoR labels, chosen by contributors themselves, ensuring high-quality classification. However, the distribution of fields is imbalanced, with some fields having significantly fewer articles than others, posing a challenge in model training. The number of articles per class ranges from 8 to 6.6K with 487 average articles per field. The training set comprises of 41,542 articles whereas the validation set comprises of 8904 articles.

4 Methodology

4.1 Preprocessing

The steps included in preprocessing of the article abstracts are given:

- Removal of unimportant columns from the dataset like Publishing year, DOI, and Data Index.
- Removal of missing values from the dataset.
- Concatenation of Title and Abstract columns using a separator [SEP] which is later used as input for the model.
- Removal of HTML tags, non-alphanumeric characters, and conversion of text to lowercase.
- Label encoding of the target labels.

4.2 Synthesis and Augmentation

We utilized Gretel.ai LLM to synthesize data for classes with fewer than 100 entries, identifying 60 such classes in our dataset. These were divided into six lists, each comprising of 10 classes. Employing Gretel LLM, we generated 5,000 synthetic entries for each list, effectively increasing our dataset to 71,540 entries. Subsequently, we conducted preprocessing steps to prepare the data for analysis. To ensure the quality of the augmented dataset, we employed SVM-based testing methods. This process enabled us to address the imbalance in class representation in the dataset.

For data augmentation, we employed synonym augmentation, replacing synonyms within the text. Augmentation was restricted to entries from classes with fewer than 100 instances, ensuring controlled expansion. This process enlarged the dataset to 80,110 entries. However, the effectiveness of this technique was limited, as the resulting embeddings exhibited similarity, and the model failed to learn substantially new information. Prior to evaluation, the texts underwent preprocessing, and the augmentation method's efficacy was assessed using SVM.

4.3 Embedding Generation

In this study, we conducted separate experiments on both the authentic dataset and the combined synthetic and original dataset. Utilizing Sentence Transformers from HuggingFace, specifically MiniLM-L6 and MP-Net, embeddings were generated for both sets. Similar experiments were conducted on these embeddings, revealing superior performance of MP-Net embeddings.

The MP-Net Sentence transformer maps sentences and short paragraphs into a 768-dimensional vector space. Pretrained on the Microsoft MPnet-base architecture, it underwent fine-tuning on over 1 billion sentence pairs using contrastive learning. The MiniLM-L6-v2 model by Sentence transformers encodes sentences and paragraphs into a 384-dimensional vector space, ideal for tasks like information retrieval. It was pretrained on MiniLM-L6-H384-uncased and fine-tuned on a large dataset using contrastive learning.

4.4 Utilisation of Machine Learning algorithms

In our approach with machine learning algorithms, we employed embeddings as input features and class labels as output for several models, including SVM, Random Forest, Logistic Regression for multi-class classification, XGBoost, and KNN. To optimize the performance of these models, we conducted hyper-parameter tuning using grid search with cross-validation to identify the most suitable parameters. Among these models, SVM emerged as the top-performing model in terms of classification accuracy.

4.5 Utilisation of Neural Networks

In our neural network approach, we explored different architectures using the generated embeddings. This included experimenting with both increasing and decreasing pyramidal structures, where the number of neurons in layers followed powers of 2. We utilized the Sparse Categorical Crossentropy loss function and the Adam optimizer. Additionally, we investigated the effectiveness of the Focal loss function, designed for imbalanced datasets, by assigning higher class weights to underrepresented classes. Furthermore, we conducted experiments with 1D CNN models tailored for text classification.

4.6 Utilisation of Transformers

In our approach utilizing transformer models, we harnessed the power of BERT and its diverse variants to address the challenge of scientific article classification. Among these variants were DistilBERT, which offers a compressed yet efficient version of BERT while preserving its performance; RoBERTa, designed to enhance BERT's pretraining process for superior language understanding; SciBERT, specialized in comprehending scientific text; and ALBERT, which incorporates parameter reduction techniques to facilitate efficient training and deployment.

Our experimentation with these models involved various strategies, including freezing all layers, unfreezing all layers, and unfreezing only the pooler layers. We set a maximum sentence length of 512 to handle the complexities of scientific articles effectively. Additionally, we explored the embeddings generated from different layers of these models and discovered that embeddings from the pooler layer yielded the most favorable results for our classification task. The embeddings extracted from the CLS token were fed into a neural network consisting of two fully connected layers with Rectified Linear Unit (ReLU) activation functions. These layers facilitated the extraction of complex patterns and features from the embeddings. Finally, a softmax layer was employed to obtain the desired output.

4.7 Best Performing Model

We achieved the best outcomes using machine learning models with MP-Net embeddings, especially fine-tuned SVM with specific parameters like polynomial kernel and C value of 1.5. Consequently, we switched our approach to utilize models tailored for handling imbalanced data, such as the one vs rest classifier. By using this classifier with the tuned SVM as its internal model, we secured the top position in the shared task. One vs Rest (OvR) classification is beneficial for imbalanced datasets because it breaks down multi-class classification problems into several simpler binary tasks. This approach helps the classifier deal with underrepresented classes more effectively. Instead of struggling with predicting all classes together, OvR classification treats each class as its own separate task, making it easier to learn the boundaries between different classes and improving accuracy, especially for minority classes.

5 Observations

Below are the observations gathered from the experiments.

- During the synthesis of data through LLM and synonym augmentation, despite achieving a well-balanced dataset, the obtained results remained largely consistent with the original dataset. This similarity stemmed from the fact that the generated or augmented samples closely resembled the originals, leading to embeddings with minimal variation. Consequently, the models encountered repetitive learning patterns rather than acquiring new insights. Through this experimentation, we found that the best results were consistently obtained using the original dataset, particularly with MP-Net embeddings.

- Based on our neural network experiments, we achieved promising results using a decreasing pyramidal architecture, however, we observed that the focal loss function didn't perform effectively in our scenario. This can be attributed to the highly imbalanced nature of our data, where certain classes significantly outnumber the minority classes. In such cases, focal loss struggles

to effectively address the imbalance because it down-weights easy-to-classify samples, but when the majority class overwhelms the minority ones, the imbalance remains inadequately addressed.

- In our experimentation with transformers, we found that they did not perform as well as traditional machine learning algorithms. This can be attributed to the imbalance in our data, which poses challenges for attention mechanism-based transformers in capturing the context effectively for each class. Consequently, there were numerous misclassifications. Moreover, we observed that fully unfreezing the layers produced better results compared to partially unfreezing them or not unfreezing them at all. Specifically, the pooler layer or the last hidden state output yielded the best results among the layers of BERT-based models. Among the transformer variants, SciBERT outperformed the others, likely due to its specialization in understanding scientific articles, followed by BERT, DistilBERT, RoBERTa, and ALBERT. It was also observed that the precision was more as compared to accuracy, recall and F1 in all the transformer models.

- In our experimentation with machine learning algorithms, we discovered that the models were able to capture information more effectively, resulting in improved classification outcomes. We noticed that tree-based algorithms such as Random Forest and Decision Trees did not perform as well, whereas models like KNN, Naive Bayes, and SVM showed superior performance. Our ultimate successful model utilized a one-vs-rest classifier approach with SVM as the internal model, as OVR classifiers are more effective for imbalanced datasets and helped us achieve top rankings.

6 Results

The results of the experimentation on various models and embeddings are given in Table 1. These results are obtained on the validation dataset provided in the shared task. The evaluation metrics used are accuracy, F1-Score, precision and recall. Additionally our team HALE LAB NITK* achieved top rank in the task, the official results (Top 3) for which are given in Table 2.

7 Conclusion

In this study, we addressed the challenge of research paper classification, achieving our highest performance with a one vs rest classifier employing Support Vector Machines (SVM). Our findings underscore the efficacy of traditional machine learning algorithms amidst the rise of sophisticated models like LLMs,

* https://halelabnitk.github.io/

Table 1. Results on validation set

Model	Accuracy	Precision	Recall	F1-Score
OVR - MPnet	**0.75**	**0.75**	**0.75**	**0.75**
OVR - MiniLM	0.73	0.72	0.73	0.72
SVM	0.71	0.71	0.71	0.71
LLM Synthesis	0.71	0.71	0.71	0.70
Synonym Replacement	0.71	0.71	0.71	0.70
KNN	0.69	0.69	0.69	0.69
ANN - Decreasing Pyramid	0.67	0.67	0.67	0.67
Logistic Regression	0.68	0.67	0.68	0.67
ANN - Increasing Pyramid	0.67	0.67	0.67	0.66
SciBERT	0.67	0.65	0.67	0.65
ANN - Focal Loss	0.65	0.65	0.65	0.64
BERT	0.61	0.65	0.61	0.60
DistilBERT	0.59	0.64	0.59	0.59
XGBOOST	0.59	0.58	0.59	0.58
Random Forest	0.57	0.59	0.57	0.54
1D CNN	0.49	0.48	0.49	0.48
BERT - Pooler Unfreezed	0.43	0.51	0.43	0.45
BERT - Freezed	0.44	0.55	0.44	0.44
ALBERT	0.20	0.27	0.20	0.20
RoBERTa	0.23	0.32	0.23	0.18

Table 2. Official results

User/Team	Accuracy	Precision	Recall	F1-Score
HALE LAB NITK	**0.7572**	**0.7536**	**0.7572**	**0.7500**
Rosni	0.7558	0.7566	0.7558	0.7540
Flo.ruo	0.7542	0.7545	0.7542	0.7524

Transformers, and neural networks. We demonstrated the robustness of ML algorithms, particularly in handling imbalanced datasets with numerous classes, where attention-based models might struggle to extract context accurately for each class. Additionally, we investigated the utilization of LLMs for synthetic data generation, yielding results akin to the original dataset, albeit with potential for improvement through enhanced prompts. Looking ahead, future research avenues could explore alternative methods for research classification, leveraging LLMs and NLP concepts such as named entity recognition and proceeding classification to address the challenges posed by vast class numbers and underrepresented data. This highlights the ongoing evolution and diversification of classification methodologies in the realm of research paper analysis.

References

1. E. Gündoğan and M. Kaya, "Research paper classification based on Word2vec and community discovery," 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020, pp. 1032-1036, doi: 10.1109/DASA51403.2020.9317101.
2. V. K. Vijayan, K. R. Bindu and L. Parameswaran, "A comprehensive study of text classification algorithms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 1109-1113, doi: 10.1109/ICACCI.2017.8125990.
3. Chowdhury, Shovan and Schoen, Marco. (2020). Research Paper Classification using Supervised Machine Learning Techniques. 1-6. 10.1109/IETC47856.2020.9249211.
4. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. Artif Intell Rev 52, 273–292 (2019).
5. T. T. Dien, B. H. Loc and N. Thai-Nghe, "Article Classification using Natural Language Processing and Machine Learning," 2019 International Conference on Advanced Computing and Applications (ACOMP), Nha Trang, Vietnam, 2019, pp. 78-84, doi: 10.1109/ACOMP.2019.00019.
6. A. R. Atmadja, M. Irfan, A. Halim and Sarbini, "Classification of Article Knowledge Field using Naive Bayes Classifier," 2020 6th International Conference on Wireless and Telematics (ICWT), Yogyakarta, Indonesia, 2020, pp. 1-4, doi: 10.1109/ICWT50448.2020.9243639.