Course Project Report

Categorizing Knowledge: Approaches For Effective Research Paper Classification

Submitted By

Praveen K (211AI028)
Saliq Gowhar (211AI043)

as part of the requirements of the course

Social Computing (IT480) [Dec 2023 - Apr 2024]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Artificial Intelligence

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

DEC 2023 - APR 2024

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

C E R T I F I C A T E

This is to certify that the Course project Work Report entitled **"Categorizing Knowledge: Approaches For Effective Research Paper Classification"** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Praveen K	211AI005	
Saliq Gowhar	211AI043	

this report is a record of the work carried out by them as part of the course **Social Computing (IT480)** during the semester **Dec 2023 - Apr 2024**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Artificial Intelligence.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S
Associate Professor, Dept. of IT, NITK

DECLARATION

We hereby declare that the project report entitled **"'Categorizing Knowledge: Approaches For Effective Research Paper Classification"** submitted by us for the course **Social Computing (IT480)** during the semester **Dec 2023 - Apr 2024**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Praveen K	211AI028	
2. Saliq Gowhar	211AI043	

Place: NITK, Surathkal
Date:

CERTIFICATE

of Achievement

This is to certify that the team

HALE LAB NITK

led by **Saliq Gowhar**, **Praveen Kempaiah** and **Dr. Sowmya Kamath S** from the **Department of Information Technology, National Institute of Technology Karnataka, Surathkal** has achieved the highest accuracy in **Field of Research Classification (FoRC) - Subtask I** at the **Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024) workshop.**

Raia Abu Ahmad, Ekaterina Borisova, Georg Rehm

FoRC Organizing Committee

Categorizing Knowledge: Approaches For Effective Research Paper Classification

Praveen Kempaiah
Information Technology Department
National Institute of Technology Karnataka
praveenk.211ai028@nitk.edu.in

Saliq Gowhar
Information Technology Department
National Institute of Technology Karnataka
saliqgowhar.211ee250@nitk.edu.in

***Abstract*—Categorizing scientific articles into specific research fields is a challenging problem, exacerbated by the volume and variety of literature published. However, existing classification systems often suffer from limitations in terms of taxonomy or the models used for classification. The Field of Research Classification (FoRC) shared task focuses on assigning research papers to one of 123 predefined classes derived from the Open Research Knowledge Graph (ORKG) taxonomy. In this article, we explore various approaches built on Sentence Transformer embeddings combined with Machine Learning algorithms, Neural Networks, and Transformers to classify articles into these predefined classes. The effectiveness of Large Language Models (LLMs) for generating synthetic data is also experimented with, along with synonym augmentation and employing 1D Convolutional Neural Networks (CNNs) for text classification. The best-performing model, the One vs Rest classifier trained on MP-Net sentence embeddings achieved an accuracy of 75.76datasets.**

***Index Terms*—Document classification, Transformers, Large Language Models, Natural Language Processing**

I. 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]. The exponential growth in scientific articles published every year, 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, employing a variety of machine learning and deep learning models, including transformers and LLMs. In this 2 No Author Given process, text is transformed into numerical representations using diverse techniques tailored to specific applications and then fed into models to yield desired outcomes [2].

Specifically, scientific article classification involves categorizing research papers into specific fields or topics, aiding in efficient information retrieval and knowledge organization. This process employs various techniques, including traditional keyword matching, supervised machine learning, deep learning, ontology-based classification, and ensemble methods. Keyword matching involves identifying predefined categories based on specific terms within the article text but may lack precision. Supervised machine learning utilizes labeled datasets to train models on extracted features like word frequency or embeddings, with algorithms such as Support Vector Machines (SVM) or Naive Bayes. Deep learning techniques, like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), enable learning complex patterns directly from text data, often enhanced by transfer learning with pretrained language models. Ontologybased classification maps article content to hierarchical concepts, allowing for more nuanced categorization. Ensemble methods combine multiple classifiers for improved accuracy and robustness.

These approaches collectively advance scientific literature management by facilitating efficient access to relevant research and insights. 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.

Researchers have attempted to address these challenges through the adoption of diverse approaches G¨undo˘gan 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] examine different term weighting methods within text classification, highlighting their role in enhancing performance. Dien et al. [5] preprocessed and vectorized articles and used SVM to achieve 91accuracy in article classification. Atmadja et al. [6] used the Naive Bayes algorithm to achieve an accuracy of 71This 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.

In this study, the goal is to address the challenge of categorizing scientific papers, with a specific focus on categorizing them into a broad taxonomy of 123 classes given a highly imbalanced dataset.The MiniLM and MP-Net sentence transformers were employed 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 hyperparameter tuning. The remainder of this article is organized as follows. Section 2 presents the detailed methodology employed in the study including all the preprocessing steps, models and embeddings experimented with. Section 3 presents various experiments and the respective observations and results. In Section 4, the conclusion and the possible future work with regards to the study is presented.

II. Related Work

G¨undo˘gan 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 91achieve an accuracy of 71literature 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.

III. Methodology

A. Corpus Description

The dataset used in this study is made available as part of the FoRC Shared task [7] 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 (FoR). 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. Overall, the training and validation sets comprise 41,542 and 8,904 articles respectively.

B. Preprocessing

Several data preprocessing steps were undertaken to prepare the dataset for analysis. First, unimportant columns such as Publishing year, DOI, and Data Index were removed. Next, any missing values within the dataset were eliminated. Additionally, the Title and Abstract columns were concatenated using a separator [SEP], creating a unified input for the model. To ensure consistency in text formatting, HTML tags and non-alphanumeric characters were removed, and all text was converted to lowercase. Finally, the target labels underwent label encoding to facilitate further analysis. These preprocessing steps were crucial in cleaning and structuring the data, ensuring its suitability for subsequent modeling and analysis tasks.

C. Synthesis and Augmentation

To synthesize data for classes with fewer than 100 entries, the Gretel.ai LLM [8] was utilized. About 60 such classes were identified from the dataset, which were divided into six lists, each comprising 10 classes. Using Gretel LLM, about 5,000 synthetic entries were generated for each list, effectively increasing the dataset to 71,540 entries. Subsequently, preprocessing steps were conducted to prepare the data for analysis. To ensure the quality of the augmented dataset, SVM-based testing methods were employed. This process enabled the addressing of the imbalance in class representation in the dataset. For data augmentation, synonym augmentation was employed, 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.

D. Embedding Generation

In our work, separate experiments were conducted on both the authentic dataset and the combined synthetic and original dataset. Sentence Transformers from HuggingFace, specifically MiniLM-L6 [9] and MP-Net [10], were utilized to generate embeddings for both sets. Similar experiments were conducted on these embeddings, emphasizing the 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 was subjected to 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 trained on MiniLM-L6-H384-uncased and fine-tuned on a large dataset using contrastive learning.

E. Classification Models

Machine Learning models. In the machine learning approach, embeddings served as input features, while class labels acted as output. Various models, including SVM, Random Forest, Logistic Regression for multi-class classification, XGBoost, and k-Nearest Neighbors (KNN) were chosen for the experiments due to their robustness, interpretability, scalability, and ability to work with imbalanced datasets given proper hyper-parameter tuning. To enhance the performance of these models, hyper-parameter tuning was conducted via 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.

Deep Neural Networks. In the neural network approach, different architectures were explored using the generated embeddings. This involved experimenting with both increasing and decreasing pyramidal structures, where the number of neurons in layers followed powers of 2. We started with 32 neurons for the increasing pyramidal architecture and went to 512 neurons, and started with 512 neurons and went down to 32 neurons in case of decreasing pyramidal architecture. Sparse Categorical Cross-entropy loss function and the Adam optimizer were used. Additionally, the effectiveness of the Focal loss function, designed for imbalanced datasets, was investigated by assigning higher class weights to underrepresented classes. Furthermore, we also utilized 1D CNN for text classification with the model having 3 convolutional layers with kernel size 4, followed by 2 fully connected layers and finally a softmax layer.

Transformer Models. In the approach utilizing transformer models, BERT (Bidirectional Encoder Representations from Transformers) and its variants were used to perform classification. Among these variants were DistilBERT (Distilled Bidirectional Encoder Representations from Transformers) [11], offering a compressed yet efficient version of BERT while preserving its performance; RoBERTa (Robustly optimized BERT approach) [13], designed to enhance BERT's pretraining process for superior language understanding; SciBERT (Scientific BERT) [14], specialized in comprehending scientific text; and ALBERT (A Lite BERT) [16], incorporating 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.

IV. Experiments and Results

The experiments for validating the proposed work were performed on Kaggle's Jupyter IDE, equipped with a P100 GPU, and Tensorflow, Pytorch, and Scikitlearn frameworks were used. For evaluation, standard metrics like accuracy, precision, recall, and F1-score were considered. Accuracy is calculated as the proportion of correctly classified instances among all instances as per Eq. 1. Precision measures the accuracy of positive predictions, indicating the proportion of correctly predicted positive instances out of all predicted positives (Eg. 2). The ability of the classifier to correctly identify positive instances out of all actual positives is given by recall, which is given by Eq. (3). F1-Score is the harmonic mean of precision and recall, penalizing extreme values of either metric (Eq. 4).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The results of the experimentation on various models and embeddings are given in Table 1. These results are obtained on the validation dataset provided as part of the FoRC shared task. Our team HALE LAB NITK1 achieved the highest accuracy in the task, the official results (Top 3) for which are given in Table 2. The best performance was obtained for machine learning models trained on MP-Net embeddings, especially fine-tuned SVM with specific parameters like polynomial kernel and C value of 1.5. Consequently, the approach was refined to utilize models tailored for handling imbalanced data, such as the One vs Rest Classifier. 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 to predict 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.

In the process of synthesizing data via LLM and synonym augmentation, even though we attained a well-rounded dataset, the outcomes largely mirrored the original dataset. This likeness arose because the generated or augmented samples closely mimicked the originals, resulting in embeddings with minimal divergence. As a result, the models faced repetitive learning scenarios instead of gaining fresh perspectives. Our experimentation revealed that the most optimal outcomes consistently stemmed from utilizing the original dataset, especially with MP-Net embeddings. Promising results were achieved

using a decreasing pyramidal architecture, in the case of neural network experiments. However, it was observed that the focal loss function didn't perform effectively in the scenario. This can be attributed to the highly imbalanced nature of the 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.

TABLE I
OBSERVED PERFORMANCE OF VARIOUS MODEL CLASSES ON THE VALIDATION SET.

Class	Model	Accuracy	Precision	Recall	F1-Score
ML	OVR - MPnet	75%	0.75	0.75	0.75
ML	OVR - MiniLM	73%	0.72	0.73	0.72
ML	SVM	71%	0.71	0.71	0.71
ML	LLM Synthesis	71%	0.71	0.71	0.70
ML	Synonym Replacement	71%	0.71	0.71	0.70
ML	KNN	69%	0.69	0.69	0.69
ML	Logistic Regression	68%	0.67	0.68	0.67
DL	ANN - Decreasing Pyramid	67%	0.67	0.67	0.67
DL	ANN - Increasing Pyramid	67%	0.67	0.67	0.66
TF	SciBERT	67%	0.65	0.67	0.65
DL	ANN - Focal Loss	65%	0.65	0.65	0.64
TF	BERT	61%	0.65	0.61	0.60
TF	DistilBERT	59%	0.64	0.59	0.59
ML	XGBOOST	59%	0.58	0.59	0.58
ML	Random Forest	57%	0.59	0.57	0.54
DL	1D CNN	49%	0.48	0.49	0.48
TF	BERT - Freezed	44%	0.55	0.44	0.44
TF	BERT - Pooler Unfreezed	43%	0.51	0.43	0.45
TF	RoBERTa	23%	0.32	0.23	0.18
TF	ALBERT	20%	0.27	0.20	0.20

TABLE II
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

Using transformers revealed that traditional machine learning algorithms perform better in comparison, which can be attributed to the imbalance in the data, posing challenges for attention mechanism-based transformers in capturing context effectively for each class. Consequently, numerous misclassifications occurred. Moreover, 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. Additionally, precision was observed to be higher compared to accuracy, recall, and F1 in all the transformer models. Machine learning models were able to capture information more effectively, resulting in improved classification outcomes. It was observed 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. The 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.

V. CONCLUSION AND FUTURE WORK

In this study, the challenge of research paper classification was addressed, with the highest performance achieved using a one vs rest classifier employing SVMs. The findings emphasize the effectiveness of traditional machine learning algorithms despite the emergence of sophisticated models like LLMs, Transformers, and neural networks. The robustness of ML algorithms was demonstrated, particularly in handling imbalanced datasets with numerous classes, where attentionbased models might encounter difficulties in accurately extracting context for each class. Additionally, the utilization of LLMs for synthetic data generation was investigated, yielding results similar to the original dataset, albeit with the 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. doi: 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 Con-

ference 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.
7. Rehm, Borisova, Ahmad, "FoRC: Field of Research Classification of Scholarly Publications", NSLP 2024. Available: https://nfdi4ds.github.io/nslp2024/docs/forc shared task.htmlforc-field-ofresearch-classification-of-scholarly-publications