RESEARCH PAPER RECOMMENDER AND SUBJECT AREA PREDICTION

Muppara Vijayaram, Tejesvani (mupparavijayaram.t@northeastern.edu); Moorthy, Hashwanth (moorthy.h@northeastern.edu); Arunkumar, Keshika (arunkumar.k@northeastern.edu)

ABSTRACT

This study presents an advanced system integrating natural language processing (NLP) and deep learning to streamline the discovery and classification of research papers. Leveraging the arXiv Abstracts 2021 dataset, which contains metadata for approximately 2 million papers, the project implements a dual approach: a recommendation engine based on BERT embeddings for semantic similarity and a subject area classification model for automated categorization. This paper outlines the methodologies, including pre-processing, embedding generation, and deep learning-based classification, evaluates the system's performance, and discusses its potential impact on academic research efficiency.

1. INTRODUCTION:

The exponential growth in academic publications has created significant challenges for researchers in efficiently identifying relevant literature. Existing recommendation systems often rely on collaborative or content-based filtering, but they struggle with capturing the semantic nuances of research topics. Similarly, traditional text classification methods lack the sophistication required for accurate categorization. This study addresses these gaps by utilizing cutting-edge NLP models like BERT and advanced deep learning architectures for automated recommendations and subject area prediction.

2. BACKGROUND:

Traditional methods for research paper recommendation have primarily relied on collaborative filtering and content-based filtering. Collaborative filtering makes recommendations based on user interactions, such as ratings or downloads, identifying patterns among similar users. However, this approach often suffers from the cold start problem, where recommendations cannot be made for new users or items with little historical interaction data. Additionally, it is not well-suited for domains like academic research, where user interaction data is sparse.

Content-based filtering, on the other hand, analyses metadata such as keywords, titles, and abstracts to find papers similar to a user's input. While this approach does not rely on user interaction data, it struggles with capturing semantic nuances in text. For instance, it may focus heavily on exact word matches and fail to understand relationships between synonyms or related concepts.

Recent advancements in natural language processing (NLP), particularly the development of Sentence Transformers and BERT (Bidirectional Encoder Representations from Transformers), have addressed these limitations. These models generate dense, contextual embeddings that represent the semantic meaning of text rather than relying solely on surface-level keywords. For example:

- **Sentence Transformers:** Extend BERT to produce fixed-size sentence embeddings, enabling efficient semantic similarity computation.
- **BERT:** Processes text bidirectionally, allowing for deeper contextual understanding and improved performance

in text-based tasks such as classification and recommendation.

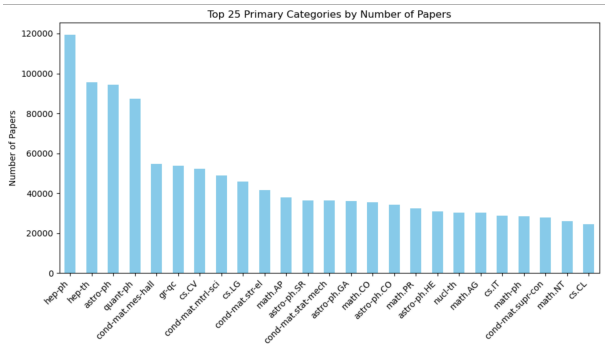
By leveraging these technologies, the proposed system overcomes the limitations of traditional methods. It integrates a BERT-based recommendation engine for identifying semantically similar papers and a deep learning model for automated subject area classification. This combination enhances both the relevance of recommendations and the accuracy of classification, addressing key challenges in academic research discovery.

3. DATA PREPERATION AND PROCESSING

3.1 Data Acquisition

The dataset utilized for this project is the arXiv Abstracts 2021 dataset, consisting of metadata for approximately 2 million research papers. This metadata includes fields such as the title of the research paper, a brief abstract summarizing its content, the list of contributing authors, subject area classifications (e.g., Computer Science, Physics, Mathematics), and additional information such as journal references, submission dates, and version history. The dataset was sourced using the Hugging Face datasets library, chosen for its capability to seamlessly handle and manipulate large-scale datasets. Once loaded, the dataset was converted into a Pandas DataFrame, facilitating easy pre-processing, exploration, and further analysis. The diversity and scale of the arXiv dataset made it an ideal choice for developing robust NLP models for research paper recommendation and classification tasks.

To understand the distribution of papers across different research categories, the top 25 primary categories were analysed based on the number of papers in each category.



3.2 Data Cleaning

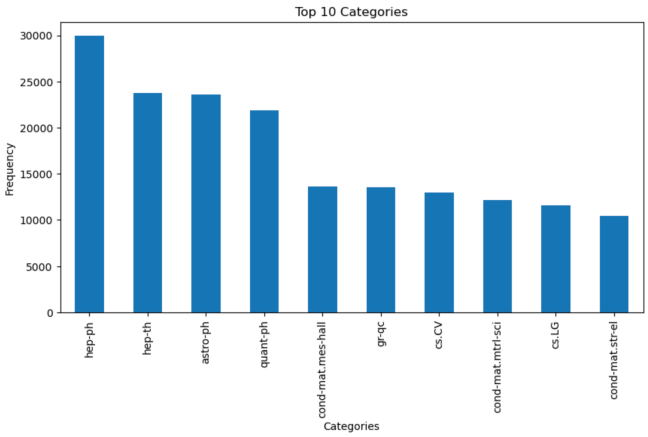
To ensure the dataset's quality, several pre-processing steps were applied to the raw text data, particularly the abstracts and titles. Papers with missing abstracts or titles were removed to ensure that the training data was complete and meaningful. However, rows with missing metadata fields that were not critical to NLP tasks were retained. Textual data was standardized by converting all text to lowercase, thus eliminating inconsistencies due to case sensitivity. Non-alphanumeric characters such as punctuation, symbols, and special characters were removed using regular expressions to minimize noise. Common stopwords, such as "the," "is," and "and," were filtered out using the NLTK library's predefined stopword list, allowing the model to focus on more informative tokens. Tokenization was performed using NLTK's word_tokenize() function, splitting the text into individual words or phrases. Finally, lemmatization was applied to reduce words to their base or root forms, such as transforming "running" to "run" or "better" to "good," ensuring consistency and reducing redundancy in the dataset.

3.3 Feature Engineering

Feature engineering was conducted to enhance the dataset's quality and improve the downstream performance of machine learning models. Abstracts and titles were transformed into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) technique, which assigns higher importance to words that are frequent in specific documents but rare across the entire corpus. Additional features, such as the word count of each abstract and title, were calculated to provide insights into the text length, which might correlate with complexity or relevance. Sentence length, representing the average number of words per sentence in abstracts, was also computed as an auxiliary feature. Furthermore, important keywords were extracted from each abstract using TF-IDF scores and stored for exploratory analysis.

3.4 Exploratory Data Analysis

As part of the data analysis process, an exploratory examination of the dataset was conducted to understand the distribution of research papers across different categories. This analysis provides insights into the dominant research areas and helps ensure balanced representation for classification tasks.



The above figure illustrates the top 10 most frequent categories in the dataset, highlighting the prevalence of physics-related fields such as "hep-ph" (High Energy Physics - Phenomenology), "hep-th" (High Energy Physics - Theory), and "quant-ph" (Quantum Physics). These categories are the most represented, aligning with arXiv's focus on physics and related disciplines. Such insights are valuable for tailoring the machine learning models and understanding the dataset's structure.

3.5 Data Augmentation

To improve model robustness and reduce the risk of overfitting, data augmentation techniques were applied. Synonym replacement introduced variability by randomly replacing words with their synonyms, preserving the semantic meaning while altering the text's surface representation. Random deletion of less significant words simulated diverse writing styles, encouraging the model to focus on critical parts of the text. Noise injection, such as introducing minor misspellings or grammatical errors, was employed to make the model resilient to imperfect real-world data. Additionally, sentence shuffling was performed in the abstracts to reduce reliance on the sentence order, enhancing the model's ability to capture semantic relationships independent of text structure.

3.6 Data Splitting

To ensure unbiased evaluation and effective training, the dataset was split into three subsets: training, validation, and testing. The training set, comprising 80% of the data, was used to train the model and learn patterns and relationships within the data. The validation set, representing 10% of the data, was employed for hyperparameter tuning and overfitting prevention by evaluating metrics such as accuracy, precision, and loss during training. The testing set, also constituting 10% of the data, was reserved for assessing the model's generalization capabilities on unseen data. Stratified splitting was applied to maintain a representative distribution of subject area categories across all subsets, ensuring balanced class representation and consistent performance evaluation.

4. MODEL ARCHITECTURE AND TRAINING:

4.1 Recommendation Engine

The recommendation engine was designed using **BERT (Bidirectional Encoder Representations from Transformers)** to generate semantic embeddings of research paper abstracts. The embedding generation process begins with a pre-trained BERT tokenizer, which tokenizes the text into token IDs while handling truncation and padding to ensure uniform input size. These tokenized inputs are then fed into the pre-trained bert-base-uncased BERT model to produce contextualized embeddings. The last hidden states of the BERT model are averaged to generate a fixed-size representation for each abstract, capturing the semantic meaning of the text.

To calculate similarity between the user query and dataset entries, pairwise cosine similarity was employed. This metric quantifies the closeness of embeddings in high-dimensional semantic space, allowing the system to identify papers with similar themes and content effectively. The recommendation workflow takes a user-provided query, comprising a title and abstract, processes it

through the BERT pipeline, and computes similarity scores with all papers in the dataset. The results are returned as a ranked list of research papers, sorted by similarity scores. For example, when tested with the query "machine learning for physics simulations", the system provided a list of papers that were contextually relevant to the topic.

4.2 Subject Area Classification

The subject area classification task involved training and evaluating multiple machine learning and deep learning models to predict the research category of papers based on their titles and abstracts. The input features for all models were represented using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This method assigns higher importance to words that are frequent in a specific document but rare across the dataset, ensuring the classification models focus on the most informative features.

Four models were explored in this task: Logistic Regression, XGBoost, CNN, and an MLP Classifier. Logistic Regression served as a baseline model, applying one-vs-rest classification to predict categories. XGBoost, known for its robust ensemble learning, used gradient-boosted decision trees to perform multi-class classification. The CNN model aimed to identify spatial patterns in text embeddings through convolutional layers, but its performance was limited by the dataset's characteristics. Finally, the MLP Classifier utilized a deep neural network with multiple dense layers, ReLU activation functions, and dropout layers for regularization.

The models were trained using the Adam optimizer with cross-entropy loss, a batch size of 32, and up to 5 epochs. Early stopping was applied to halt training when validation performance stopped improving. Each model was evaluated on the test set using metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive view of their performance.

4.3 Results and Model Comparison

The performance of the models was evaluated on the test set, and the metrics—accuracy, precision, recall, and F1-score—are summarized below:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.78	0.78	0.77	0.77
XGBoost	0.72	0.71	0.72	0.71
CNN	0.75	0.76	0.75	0.75
MLP Classifier	0.67	0.77	0.71	0.74

The Logistic Regression and XGBoost models demonstrated the highest accuracy, indicating the suitability of simpler, non-sequential models for this task.

CNN showed moderate performance, as its ability to identify local patterns in text embeddings was insufficient for comprehensive classification.

On the other hand, the MLP Classifier displayed high precision, reflecting its ability to minimize false positives and focus on reliable predictions.

4.4 Discussion

The model evaluation highlights the strengths and weaknesses of each approach. Simpler models such as Logistic Regression and XGBoost excelled in accuracy and overall performance, capitalizing on the static, feature-based representation of the dataset. Deep learning models like CNN struggled due to the mismatch between their architectures and the dataset's nature. The MLP Classifier offered a balance, achieving notable precision and a competitive F1-score. These results suggest that while deep learning models have potential, their application to TF-IDF-based features requires additional refinement or alternative feature representations to unlock their full capabilities.

5. CONCLUSION AND PERFORMANCE COMPARISON:

5.1 Conclusion

This project aimed to develop an intelligent system for recommending and classifying research papers using advanced natural language processing (NLP) and deep learning techniques. The recommendation engine, built on BERT embeddings, demonstrated its ability to capture semantic similarities effectively, offering personalized and contextually relevant suggestions. The subject area classification task explored multiple machine learning and deep learning models, each evaluated based on their accuracy, precision, recall, and F1-scores.

The results highlight the effectiveness of simpler models like Logistic Regression and XGBoost for tasks involving TF-IDF features, where they outperformed deep learning architectures like CNN. Despite the challenges faced by these models, the MLP Classifier stood out for its high precision and balanced performance, indicating its suitability for tasks where minimizing false positives is critical.

The system addresses the growing challenge of academic information overload by enabling researchers to efficiently discover relevant literature and categorize papers into appropriate subject areas. While the project achieved significant milestones, there remains room for improvement, particularly in leveraging transformer-based models for classification tasks and exploring more comprehensive feature representations.

5.2 Performance Comparison

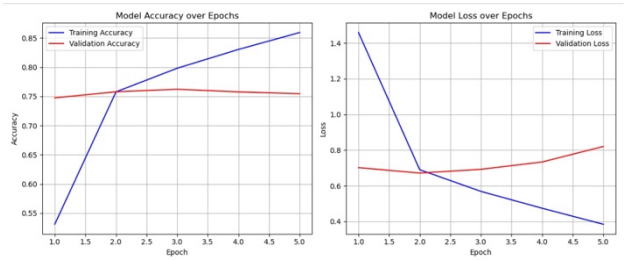
The comparative analysis of the models employed for subject area classification provides critical insights into their performance characteristics:

Logistic Regression and XGBoost

Logistic Regression and XGBoost emerged as the best-performing models, achieving high accuracy and leveraging the effectiveness of TF-IDF vectorization for feature extraction. Logistic Regression slightly outperformed XGBoost in precision, making it the most effective model for classifying research papers into predefined subject areas. These results highlight the strengths of traditional machine learning models in handling feature-based representations like TF-IDF.

CNN

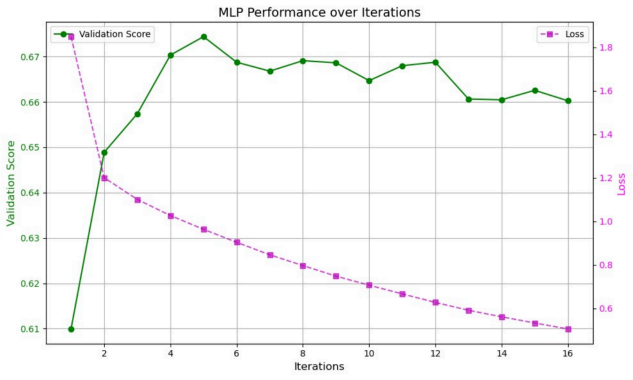
The CNN model demonstrated moderate performance in the classification task. Its ability to detect local patterns within tokenized sequences was limited when applied to static, feature-based representations such as TF-IDF. Despite strong training performance, evidenced by steadily increasing accuracy and decreasing loss, the CNN struggled to generalize well, as reflected in a performance gap between training and validation metrics. Enhancements through hyperparameter optimization or additional techniques like fine-tuning could improve its validation accuracy and loss stability in future iterations.



MLP Classifier

The MLP Classifier achieved competitive precision and recall metrics, offering a balanced approach between traditional machine learning and deep learning. Its dense architecture, coupled with dropout regularization, effectively captured relationships within TF-IDF features, resulting in a strong F1-score.

Visualization of the training process revealed consistent improvement in loss across 16 iterations, with validation scores peaking around iteration 5. Beyond this point, the validation scores exhibited minor fluctuations, indicating the model had reached a performance plateau. This finding suggests the potential benefit of hyperparameter tuning or employing early stopping to optimize training.



6. CONCLUSION

This performance comparison emphasizes the importance of aligning model architectures with the dataset and task requirements. Traditional machine learning models, particularly Logistic Regression and XGBoost, excelled in this study, given the nature of the dataset and the use of TF-IDF for feature extraction. Although deep learning

models like CNN and MLP showed promise, their application in this context revealed areas for improvement. Future work could explore fine-tuning transformer-based models to enhance classification performance, especially for tasks requiring a deeper understanding of semantic relationships.

7. REFERENCE

[1] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, Alexander A. Alemi: "On the Use of arXiv as a Dataset," *arXiv preprint arXiv:1905.00075*, 2019. Available: https://arxiv.org/abs/1905.00075

[2] Nils Reimers, Iryna Gurevych: "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019. Available: https://arxiv.org/abs/1908.10084

[3] Hugging Face: "arXiv Abstracts 2021 Dataset," *Hugging Face Datasets*. Available: https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021