IE7500 APPLIED NATURAL LANGUAGE PROCESSING

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)

12/04/2024

Introduction

Objective :

- Discovering relevant research papers among millions of academic publications
- Automatically categorizing papers into appropriate subject areas.

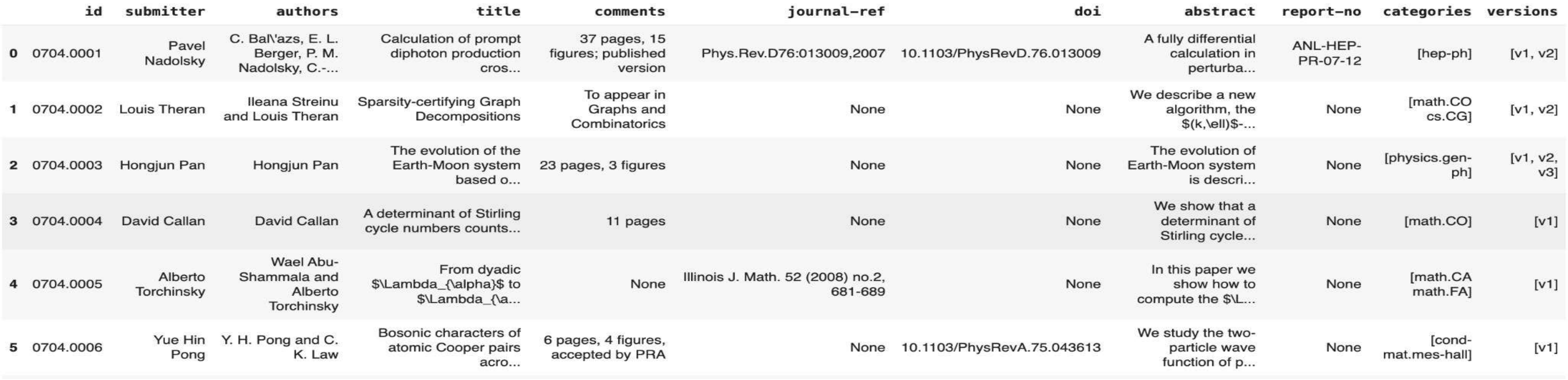
Previous Work:

- **Keyword Matching**: Limited by shallow keyword-based search.
- **Collaborative Filtering**: Relies on user behavior but struggles with cold-start problems and capturing complex semantic relationships.

Proposed Solution:

- **BERT Embeddings**: Leverages deep learning for semantic similarity-based recommendations.
- **Deep Learning Models**: Automated, accurate classification into subject areas

About the Corpus



	id	submitter	authors	title	comments	journal-ref	doi	abstract	report-no	categories	versions
0	0704.0001	Pavel Nadolsky	C. Bal\'azs, E. L. Berger, P. M. Nadolsky, C.-...	Calculation of prompt diphoton production cros...	37 pages, 15 figures; published version	Phys.Rev.D76:013009,2007	10.1103/PhysRevD.76.013009	A fully differential calculation in perturba...	ANL-HEP-PR-07-12	[hep-ph]	[v1, v2]
1	0704.0002	Louis Theran	Ileana Streinu and Louis Theran	Sparsity-certifying Graph Decompositions	To appear in Graphs and Combinatorics	None	None	We describe a new algorithm, the (k,ℓ)-...	None	[math.CO cs.CG]	[v1, v2]
2	0704.0003	Hongjun Pan	Hongjun Pan	The evolution of the Earth-Moon system based o...	23 pages, 3 figures	None	None	The evolution of Earth-Moon system is descri...	None	[physics.gen-ph]	[v1, v2, v3]
3	0704.0004	David Callan	David Callan	A determinant of Stirling cycle numbers counts...	11 pages	None	None	We show that a determinant of Stirling cycle...	None	[math.CO]	[v1]
4	0704.0005	Alberto Torchinsky	Wael Abu-Shammala and Alberto Torchinsky	From dyadic Λ_{α} to $\Lambda_{\a...	None	Illinois J. Math. 52 (2008) no.2, 681-689	None	In this paper we show how to compute the $\L...	None	[math.CA math.FA]	[v1]
5	0704.0006	Yue Hin Pong	Y. H. Pong and C. K. Law	Bosonic characters of atomic Cooper pairs acro...	6 pages, 4 figures, accepted by PRA	None	10.1103/PhysRevA.75.043613	We study the two-particle wave function of p...	None	[cond-mat.mes-hall]	[v1]

Corpus: arXiv Abstracts 2021 (~2M papers, with metadata like title, abstract, authors, categories).

Dominant Fields: Physics, Mathematics, and Computer Science, with hep-ph and hep-th leading.

Insights: Visualizations show the dominance of physics-related categories

Data Preprocessing

To ensure high-quality inputs for the recommendation and classification tasks, a systematic data preprocessing pipeline was implemented. The following steps were applied:

1. **Handling Missing Values:** Removed papers lacking titles or abstracts; retained non-critical metadata.
2. **Text Cleaning:** Standardized text by converting to lowercase and removing noise (e.g., special characters, punctuation).
3. **Stopword Removal:** Filtered out common stopwords using NLTK to retain meaningful words.
4. **Tokenization:** Split text into individual words/phrases for granular analysis.
5. **Lemmatization:** Standardized words to root forms for uniformity (e.g., "running" → "run").
6. **TF-IDF Vectorization:** Converted text into numerical vectors to emphasize key terms unique to each document.

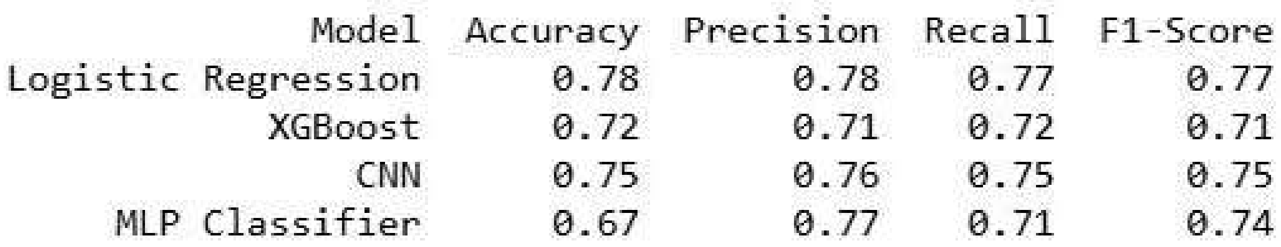
Recommendation Engine

The recommendation engine uses **BERT (Bidirectional Encoder Representations from Transformers)** to derive the semantic meaning of the research paper abstract and find similar papers.

1. **Embedding Creation**:
    - Processes abstracts using the **pre-trained BERT model** (*bert-base-uncased*).
    - Converts text into numerical embeddings that encapsulate meaning.
2. **Similarity Calculation**:
    - Computes **cosine similarity** between the query and paper embeddings.
    - Ranks papers by similarity to the query.
3. **Results**:
    - **Query Example:** "Machine Learning for Physics Simulations"
    - **Result:** A ranked list of semantically relevant research papers, aiding faster literature discovery.

Subject Area Classification

- **Machine Learning Models:**
  - **Logistic Regression:** Best overall performance with **78% accuracy** and balanced metrics.
  - **XGBoost:** Competitive results with **72% accuracy** and strong precision (**71%**).
- **Deep Learning Models:**
  - **CNN (Convolutional Neural Network):** Achieved **75% accuracy**, excelling in pattern detection.
  - **MLP Classifier (Multi-Layer Perceptron):** High precision (**77%**) and competitive F1-score, reducing false positives.



```
              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
```

Model Comparison

Logistic Regression:

- Best performing model
- Highest accuracy and balanced metrics
- Well-suited for text classification tasks
- Handles linear relationships in high-dimensional spaces effectively

XGBoost:

- Slightly lower performance than logistic regression
- May have struggled with high dimensionality of text data

CNN:

- Performed better than XGBoost but slightly worse than logistic regression
- Good at capturing local patterns in text
- Slightly lower accuracy might be due to overfitting

MLP:

- Showed Low accuracy but relatively better precision
- Needs better tuning and architectural enhancement

Conclusion

- Developed Recommendation Engine and Classification Model
- Achieved reliable performance across various model

Future Work

- Perform extensive hyperparameter optimization to boost model performance
- Extend functionality to classify research papers in multiple languages