

Predicting Tags for Research Articles: An Extensive Approach Using Hybrid Topic Modeling Techniques

Nagashree Jayaram¹, Ravikumar M L²

^{1,2} PG Scholar, M. Tech in Artificial Intelligence,

^{1,2} REVA Academy for Corporate Excellence (RACE)

^{1,2} Reva University, Bangalore, KA, India

Abstract

The exponential growth of scientific literature has made it increasingly difficult for researchers to efficiently locate relevant articles within vast online archives. Topic modeling and tagging offer effective solutions to this problem by providing a clear categorization of research articles, which enhances both the recommendation and search processes. This paper builds upon previous efforts in topic prediction by focusing on the prediction of specific tags for research articles based on their abstracts. The study utilizes advanced machine learning techniques, including Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF) with Count Vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs). The research is conducted in the context of a live hackathon, where the challenge is to predict relevant tags from a predefined set for research articles in four key disciplines: Computer Science, Mathematics, Physics, and Statistics. This approach aims to improve the accuracy and granularity of tag prediction, thereby facilitating more effective information retrieval in academic databases.

Key Words: Tag Prediction, Research Articles, Topic Modelling, Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF), Text Classification, Natural Language Processing (NLP), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), Machine Learning, Text Mining, Abstract Analysis, Multi-Label Classification, Scientific Literature

1. Introduction

In the digital age, the availability of extensive online archives of scientific articles is both a boon and a challenge for researchers. While these archives provide unparalleled access to a wealth of knowledge, the sheer volume of content can make it difficult to find articles that are most relevant to a specific research query. Traditional search methods, which often rely on keyword matching, are increasingly inadequate for navigating this vast landscape. As a result, there is a growing need for more sophisticated methods of categorization and retrieval. Topic modeling and tagging have emerged as powerful tools in this context. By assigning topics or tags to research articles, these methods enable more precise and efficient searches, as well as improved recommendations. However, while topic modeling has seen significant advances, there remains a need for more granular tag prediction that can capture the multifaceted nature of research articles.

Previously, the focus was on predicting the topics of research articles based on their abstracts. Building on this foundation, the current study takes a step further by predicting specific tags associated with each article. Tags provide a finer level of categorization, capturing specific aspects of the content that broader topics might miss. The challenge is to accurately predict these tags from the abstracts of research articles in four disciplines: Computer Science, Mathematics, Physics, and Statistics.

To tackle this challenge, we employ a range of advanced machine learning techniques, including OneVsRestClassifier, Fully Connected Feedforward Neural Network, Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF) using Count Vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF). These techniques are applied to a predefined set of tags, which includes specialized terms such as 'Artificial Intelligence,' 'Cosmology and Nongalactic Astrophysics,' 'Machine Learning,' and 'Optimization and Control.'

The goal of this research is to enhance the accuracy of tag prediction, thereby improving the effectiveness of search and recommendation systems in academic databases. By advancing the state of the art in tag prediction, this study contributes to the ongoing effort to make scientific knowledge more accessible and easier to navigate.

2. Problem Statement

Topic Modeling for Research Articles 2.0 - Researchers have access to large online archives of scientific articles. As a consequence, finding relevant articles has become more and more difficult. Tagging or topic modelling provides a way to give clear token of identification to research articles which facilitates recommendation and search process.

In this context, our research focuses on the problem of accurately predicting the tags for research articles based on their abstracts. Tags serve as identifiers that encapsulate the core topics and themes of a paper, thereby facilitating easier retrieval and recommendation. The ability to predict these tags accurately is crucial for improving the accessibility of relevant research in vast digital archives.

Note that a research article can possibly have multiple tags.

The research article abstracts are sourced from the following 4 topics:

1. Computer Science
2. Mathematics
3. Physics
4. Statistics

List of possible tags are as follows:

[Tags, Analysis of PDEs, Applications, Artificial Intelligence, Astrophysics of Galaxies, Computation and Language, Computer Vision and Pattern Recognition, Cosmology and Nongalactic Astrophysics, Data Structures and Algorithms, Differential Geometry, Earth and Planetary Astrophysics, Fluid Dynamics, Information Theory, Instrumentation and Methods for Astrophysics, Machine Learning, Materials Science, Methodology, Number Theory, Optimization and Control, Representation Theory, Robotics, Social and Information Networks, Statistics Theory, Strongly Correlated Electrons, Superconductivity, Systems and Control]

Discipline	Possible Tags
Computer Science	Artificial Intelligence, Computation and Language, Computer Vision and Pattern Recognition, Data Structures and Algorithms, Information Theory, Machine Learning, Robotics, Social and Information Networks, Systems and Control
Mathematics	Analysis of PDEs, Applications, Differential Geometry, Number Theory, Optimization and Control, Representation Theory, Statistics Theory
Physics	Astrophysics of Galaxies, Cosmology and Nongalactic Astrophysics, Earth and Planetary Astrophysics, Fluid Dynamics, Instrumentation and Methods for Astrophysics, Strongly Correlated Electrons, Superconductivity

Statistics	Methodology, Statistics Theory, Machine Learning, Applications, Optimization and Control
------------	--

Table 2.1: Research Article Disciplines and Possible Tags.

This research seeks to employ advanced machine learning models, including Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and various topic modeling techniques (LDA, LSA, NMF), to predict the correct tags for each research article. By improving the accuracy of tag prediction, this study aims to enhance the discoverability and usability of scientific literature, thereby contributing to more effective research workflows and knowledge dissemination.

3. Methodology

This section outlines the methodology employed for predicting tags for research articles using topic modeling on their abstracts. The process involves multiple stages, from data collection and preprocessing to model training and evaluation. Each step is crucial for ensuring the accuracy and relevance of the predicted tags.

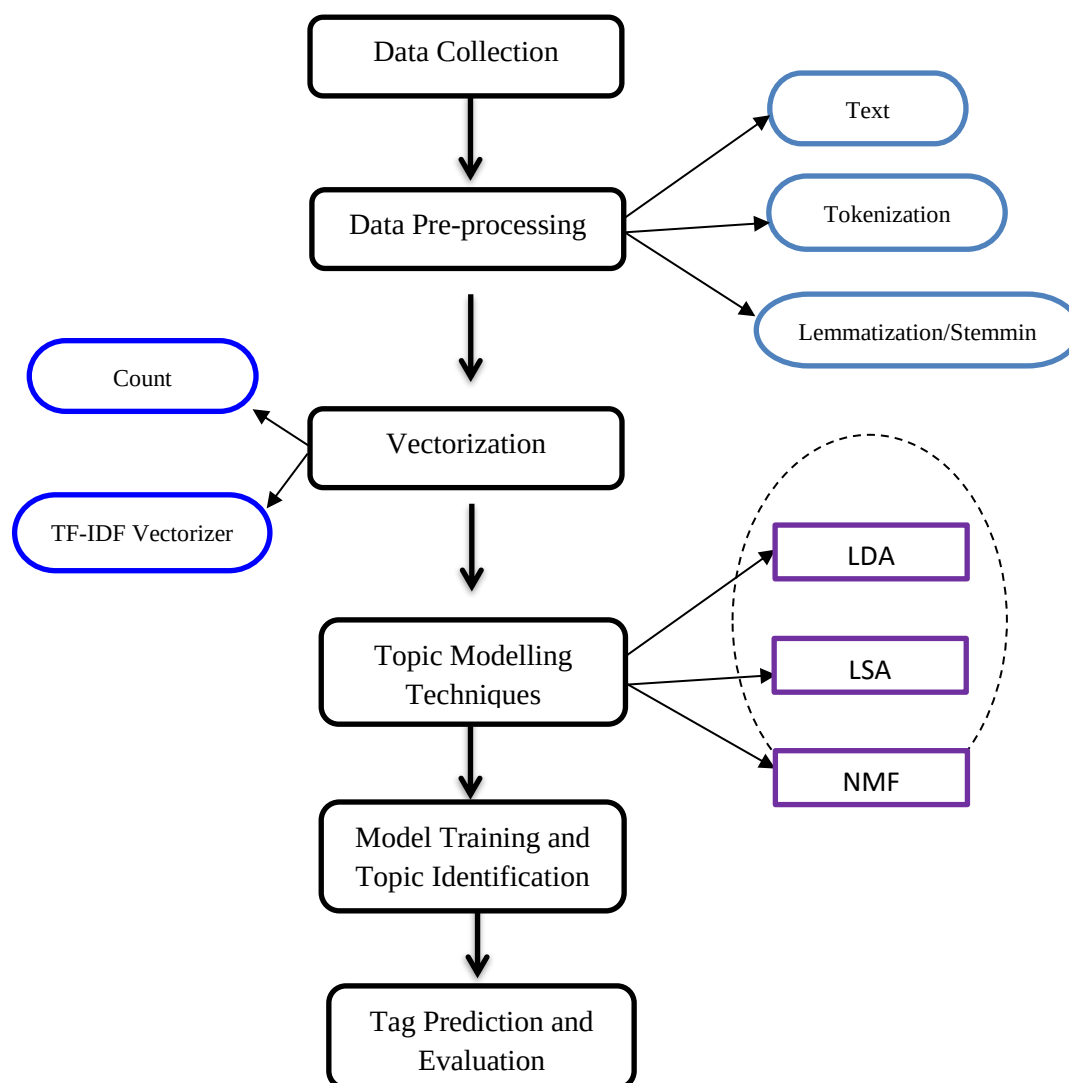


Figure 3.1 Flow diagram of the proposed techniques

3.1.1 Data Collection

The initial step involves gathering research article abstracts from various sources. These abstracts serve as the primary data for the topic modeling process.

3.1.2 Data Preprocessing

Before applying any modeling techniques, the raw text data must be preprocessed to ensure consistency and improve model performance.

- **Text Cleaning:** This step removes unnecessary elements like punctuation, special characters, and stop words.
- **Tokenization:** The text is split into individual tokens (words).
- **Lemmatization/Stemming:** Each token is reduced to its base or root form. For example, lemmatization converts "running" to "run."

Mathematically, given a document d consisting of words $w_1, w_2, \dots, w_n$ tokenization can be represented as:

$$d = \{w_1, w_2, \dots, w_n\}$$

where w_i each is a token.

3.1.3 Vectorization

To apply machine learning models, the textual data must be converted into numerical form through vectorization. Two primary methods are used:

- **Count Vectorizer:** This method creates a matrix χ where each element χ_{ij} represents the frequency of the j -th word in the i -th document.

$$\chi_{ij} = \text{count}(w_j \text{ in } d_i)$$

- **TF-IDF Vectorizer:** This method transforms the text into a matrix where each element χ_{ij} represents the term frequency-inverse document frequency (TF-IDF) score of the j -th word in the i -th document.

The TF-IDF score is calculated as:

$$TF-IDF(w_j, d_i) = TF(w_j, d_i) \times IDF(w_j)$$

where:

$$TF(w_j, d_i) = \frac{\text{count}(w_j \text{ in } d_i)}{\text{total words in } d_i}$$

$$IDF(w_j) = \log \left(\frac{\text{total number of documents}}{\text{number of documents containing } w_j} \right)$$

3.1.4 Topic Modeling Techniques

Three primary topic modeling techniques are employed:

- **Latent Dirichlet Allocation (LDA):** LDA is a generative probabilistic model where documents are represented as mixtures of topics, and topics are mixtures of words. The probability of a word w in a document d is given by:

$$p(w|d) = \sum_{k=1}^K p(w|z_k)p(z_k|d)$$

where z_k represents the k -th topic.

- **Latent Semantic Analysis (LSA):** LSA applies Singular Value Decomposition (SVD) to the term-document matrix X to reduce dimensionality and uncover latent semantic structures.

$$X \approx U \Sigma V^T$$

where U , Σ , and V represent the matrices obtained from SVD.

- **Non-Negative Matrix Factorization (NMF):** NMF factorizes the term-document matrix X into two lower-dimensional matrices W and H such that:

$$X \approx WH$$

where W and H contain non-negative elements, representing the relationships between documents & topics.

3.1.5 Model Training

Each topic modeling technique is trained on the vectorized abstracts to identify the underlying topics. The models are optimized to maximize the coherence score, which measures the semantic similarity of the top words within a topic.

3.1.6 Topic Identification

Once the models are trained, topics are identified as clusters of top words. For each document d , the most relevant topics z_k are assigned based on the highest probability or similarity score.

3.1.7 Evaluation

The coherence score C is used to evaluate the quality of the topics generated by the models. It is calculated as:

$$C = \frac{1}{N} \sum_{m=1}^M \sum_{l=1}^L Sim(w_{ml}, w_{mn})$$

where N is the number of topic words, M is the number of topics, w_{ml} and w_{mn} are the topic words, and Sim represents the semantic similarity between words.

3.1.8 Tag Prediction

Based on the identified topics, relevant tags are predicted for each abstract. The tags correspond to the most prominent topics associated with each document.

4. Result Analysis with metrics

The coherence scores obtained from applying various topic modeling techniques and vectorization methods provide valuable insights into the effectiveness of each approach for tag extraction from research abstracts. The coherence score is a measure of the semantic similarity between the top words in a topic, indicating the quality and interpretability of the topics generated.

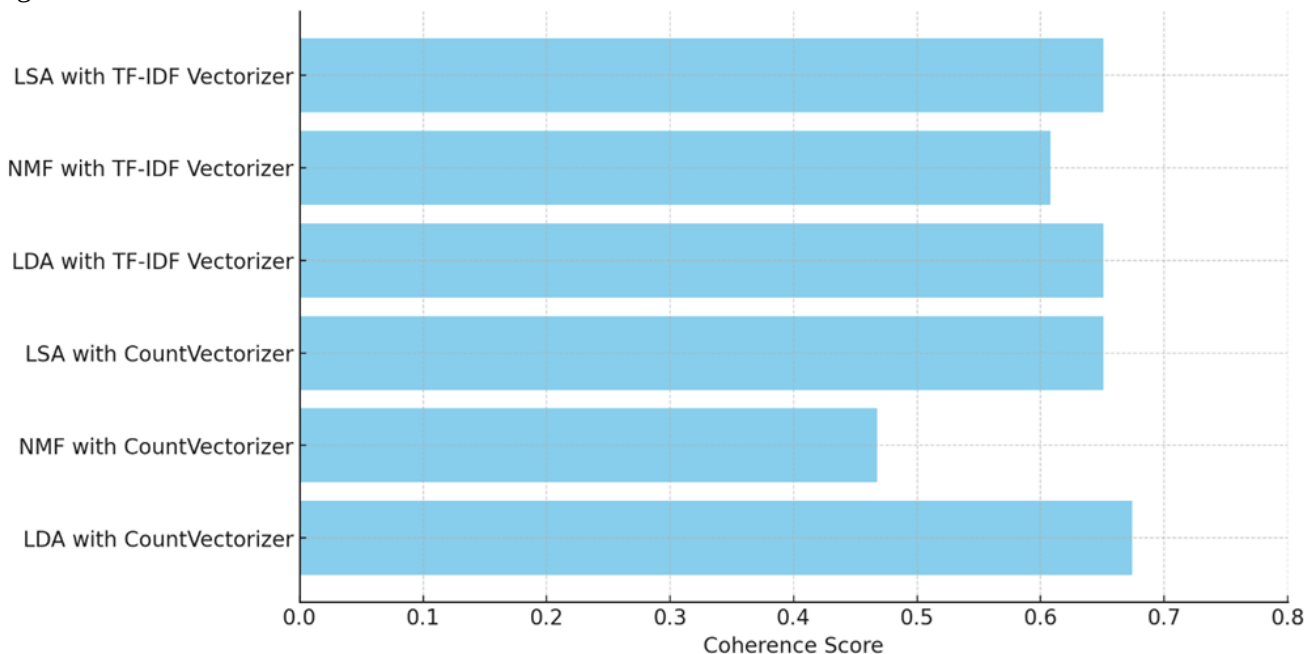


Figure 4.1: Model Performance Metrics

4.1 Model Performance Metrics

Table 4.1.1: Latent Dirichlet Allocation (LDA)

Vectorizer	Coherence Score
Count Vectorizer	0.6743
TF-IDF Vectorizer	0.651

Table 4.1.2: Latent Semantic Analysis (LSA)

Vectorizer	Coherence Score
Count Vectorizer	0.651
TF-IDF Vectorizer	0.651

Table 4.1.3: Non-Negative Matrix Factorization (NMF)

Vectorizer	Coherence Score
Count Vectorizer	0.4678
TF-IDF Vectorizer	0.608



Page 7

This work investigates the application of BERT (Bidirectional Encoder Representations from Transformers) for tagging scientific articles, highlighting its effectiveness in understanding the context of abstracts. <https://arxiv.org/abs/1810.04805>

[6] Using LDA and Word2Vec for Tag Recommendation in Scholarly Documents. Yang, J., Wang, C., & Wang, D. (2018).

This paper presents an approach that combines Latent Dirichlet Allocation (LDA) and Word2Vec embeddings for tag recommendation, offering insights into integrating topic models and word embeddings for better tag prediction.

https://link.springer.com/chapter/10.1007/978-3-319-94965-9_39

[7] Deep Learning Methods for Automatic Tagging in Academic Papers. Liu, Y., & Lee, K. (2020).

The study explores the use of deep learning models, such as LSTMs and attention mechanisms, for predicting tags in academic papers based on abstract text. <https://arxiv.org/abs/2004.03465>

[8] Weihang Ning, Jianhua Liu, Hui Xiong, Knowledge discovery using an enhanced latent Dirichlet allocation-based clustering method for solving on-site assembly problems, Robotics and Computer-Integrated Manufacturing, Volume 73, 2022, 102246, ISSN 0736-5845

[9] Asmussen, C.B., Møller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. *J Big Data* 6, 93 (2019). <https://doi.org/10.1186/s40537-019-0255-7>