JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Research Navigator: Automated Research paper Recommendation System

[1]SELVI R, [2]RITHIK HARENDAR M, [3]SETHUMADAV S

[1]Assistant Professor, [2]UG scholars, [3]UG scholars
[1]Department of Information Technology,
[1]Meenakshi Sundararajan Engineering College, Chennai, India

Abstract : In this era of vast scholarly research output, efficient retrieval of relevant journal papers remains a significant challenge. This project introduces a novel approach to journal paper search systems, aiming to not only retrieve papers based on similarity but also enhance originality and minimize plagiarism concerns. The proposed system operates by accepting document abstracts as input and employs a combination of deep learning techniques for classification, Siamese network, and recommendation.The system begins by utilizing an LSTM (Long Short-Term Memory) classifier to accurately categorize the domain of the input document. Following domain classification, a Siamese Network is applied to find similar documents together, thus facilitating a focused search approach. Moreover, the system calculates similarity scores between the input abstract and the base papers, considering both content and publication year.Furthermore, to mitigate the risk of plagiarism and promote originality, the system incorporates a recommendation engine. This engine not only suggests papers with high similarity scores but also provides insightful recommendations for augmenting the user's project. By suggesting additional ideas and perspectives to incorporate into the user's work, the system promotes originality and guards against potential plagiarism pitfalls.In summary, this project presents an innovative journal paper search system that not only retrieves relevant papers efficiently but also fosters originality in academic research by providing tailored recommendations for project enhancement.

IndexTerms - Document Retrieval, Research Papers, LSTM,, Deep Learning, Similarity Score, NLP,

I.INTRODUCTION

In the rapidly expanding landscape of scholarly research, the need for an advanced journal paper search engine has never been more pressing. Researchers are faced with an enormous difficulty in sorting through this large ocean of knowledge as the number of scholarly publications continues to soar. Traditional search engines leave a big hole in the academic discovery process because they can't offer more than basic retrieval.This is where a paradigm shift in journal paper searching comes in, one that has the potential to completely change how academics access and interact with the literature. In order to transform the academic landscape, our ground-breaking work presents a state-of-the-art system that can not only retrieve relevant publications based on similarities but also encourage innovation and prevent plagiarism.At its core, our novel approach reimagines the academic discovery process by utilizing the power of deep learning techniques. Our technology goes beyond traditional search techniques by utilizing cutting-edge machine learning algorithms, providing a dynamic and all-encompassing solution to the problems associated with information overload.Our system's fundamental strength is its capacity to comprehend the complex fields of scholarly writing. We enable our system to precisely classify the domains of incoming documents by using an advanced LSTM classifier, which paves the way for a more intelligent and focused search experience.Our inventiveness doesn't end there, though. Using a Siamese Network, our system builds on this basis and groups texts that are related in a seamless manner, enabling a more targeted and effective search approach. Our method expedites the research process by grouping articles according to their fundamental similarities, allowing researchers to pinpoint pertinent publications with unmatched accuracy. Moreover, our system introduces a novel method for evaluating similarity. We give consumers with nuanced similarity ratings that provide a holistic perspective of document alignment by incorporating sophisticated algorithms that take into account both temporal context and content relevance. By taking a comprehensive approach to similarity assessment, researchers can be guaranteed to find linkages and patterns in a wide variety of scholarly literature.Possibly the most revolutionary feature of our system is its customized recommendation engine. This engine does much more than just retrieve data; it is a spark plug for ideas and originality, providing customized recommendations and different viewpoints to entice researchers and improve project results.

Our recommendation engine helps users avoid plagiarism by pointing them in the direction of fresh ideas and insights. This encourages a more creative and imaginative study environment while also protecting against plagiarism.In summary, our study marks a significant breakthrough in the field of journal paper search engines by going beyond conventional limits to provide a game-changing solution that not only speeds up the finding of pertinent publications but also encourages creativity and authenticity in the academic community. We hope to empower scholars and push the boundaries of knowledge with our innovative methodology, bringing in a new age of academic innovation and discovery

II.LITERATURE REVIEW

G. Mustafa et al. introduced a methodology leveraging genetic algorithms to optimize document classification, aiding various scientific endeavors. While proficient in categorizing research documents into predefined categories, it lacks provision for providing similarity scores between documents.[1]J. Peng and S. Huo presented a few-shot text classification method focusing on feature optimization. Addressing challenges like insufficient feature representation, it optimizes feature representation for text-based document classification. However, it does not offer similarity scores between documents.[2]X. Li, B. Tian, and X. Tian proposed a retrieval model based on Graph Convolutional Network (GCN) and Hesitant Fuzzy Set for scientific documents. Exploring the correlation between document attributes, it lacks provision for providing similarity scores between documents.[3]Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee introduced a similarity measure for text classification and clustering, offering insights into effective methods for assessing similarity between textual documents. However, it primarily focuses on text classification and clustering rather than document retrieval and recommendation.[4]K.-C. Khor and C.-Y. Ting proposed a Bayesian approach for classifying conference papers, aiding in document organization and retrieval. While leveraging Bayesian methods, it may have limitations in handling diverse document types and contexts.[5]R. Zhao and K. Mao introduced a fuzzy bag-of-words model for document representation, potentially enhancing document retrieval and recommendation systems. Despite its novel approach, further validation in real-world scenarios is required.[6]Beel, J., Gipp, B., Langer, S. et al. provided insights into research-paper recommender systems, offering a comprehensive overview of existing approaches and evaluation frameworks. However, it lacks focus on document classification and similarity assessment.[7]Aggarwal, C. C., & Zhai, C. surveyed various text classification algorithms, providing insights into different approaches and techniques. While comprehensive, it primarily focuses on text classification rather than document retrieval and recommendation.[8]Nam, L. N. H., & Ho, B. Q. introduced a comprehensive filter feature selection method for improving document classification, potentially enhancing classification accuracy. However, it may have limitations in handling large-scale datasets and diverse document types.[9]B. Tang, H. He, P. M. Baggenstoss, and S. Kay proposed a Bayesian classification approach utilizing class-specific features for text categorization. While enhancing categorization accuracy, it requires further validation across diverse datasets and document types.[10]

III.MATERIALS AND METHODS

A. OVERVIEW

In the vast expanse of scholarly research, where countless papers are published daily, finding the proverbial needle in the haystack is akin to predicting financial markets' intricate patterns. Our novel approach to journal paper search engines is reminiscent of the complex techniques used to predict financial trends. Our mission to transform academic discovery and promote originality, integrity, and creativity in scholarly research is at the heart of everything we do. Analogous to the painstaking process of collecting data for financial analysis, our technology compiles copious amounts of data covering a wide range of subjects, fields, and viewpoints.This rich dataset provides the framework for our investigation, directing us through the maze of scholarly publications.Raw data, though, is only the beginning. The gathered data is rigorously preprocessed in order to extract valuable insights and enable efficient search. To ensure the data's quality, consistency, and usability, this important phase entails cleaning, normalizing, and arranging it. Our approach prepares scholarly data for analysis and exploration, in the same way that financial analysts carefully clean and preprocess market data to extract relevant insights.Using preprocessed data, our system makes use of sophisticated methods from the fields of natural language processing and machine learning. In order to classify and categorize documents based on their long-term dependencies, Long Short-Term Memory (LSTM) classifiers are used.But categorization is only the beginning of our adventure. We also explore the domain of relevance score and similarity rating. Inspired by financial models that evaluate market trends and asset correlations, our approach computes complex similarity scores that consider temporal context and content relevance. This thorough approach to aligning documents guarantees that consumers will see relevant and intelligent search results.Naturally, creating a system this complex needs a great deal of testing and training. Our models are trained on past academic data and assessed using different datasets, in the same way that financial models are trained on historical market data and verified using out-of-sample testing. Our models gain proficiency in forecasting document effect, relevance, and significance through iterative validation and improvement. After verification, these models drive our search engine, turning it into a lighthouse for exploration and discovery. Scholars can now set out on an intellectual exploration voyage, confidently and easily navigating the huge ocean of scholarly literature. Equipped with cutting-edge tools and processes, individuals may unearth obscure treasures, investigate novel concepts, and participate in academic discussions in a way never seen before.To sum up, our novel approach to journal paper search signifies a revolution in scholarly research. We bridge the gap between data science and academia by utilizing methods from the field of financial analysis in our scholarly study. Using our approach as a roadmap, scholars can successfully negotiate the challenging terrain of scholarly literature, one paper at a time, and discover the undiscovered secrets

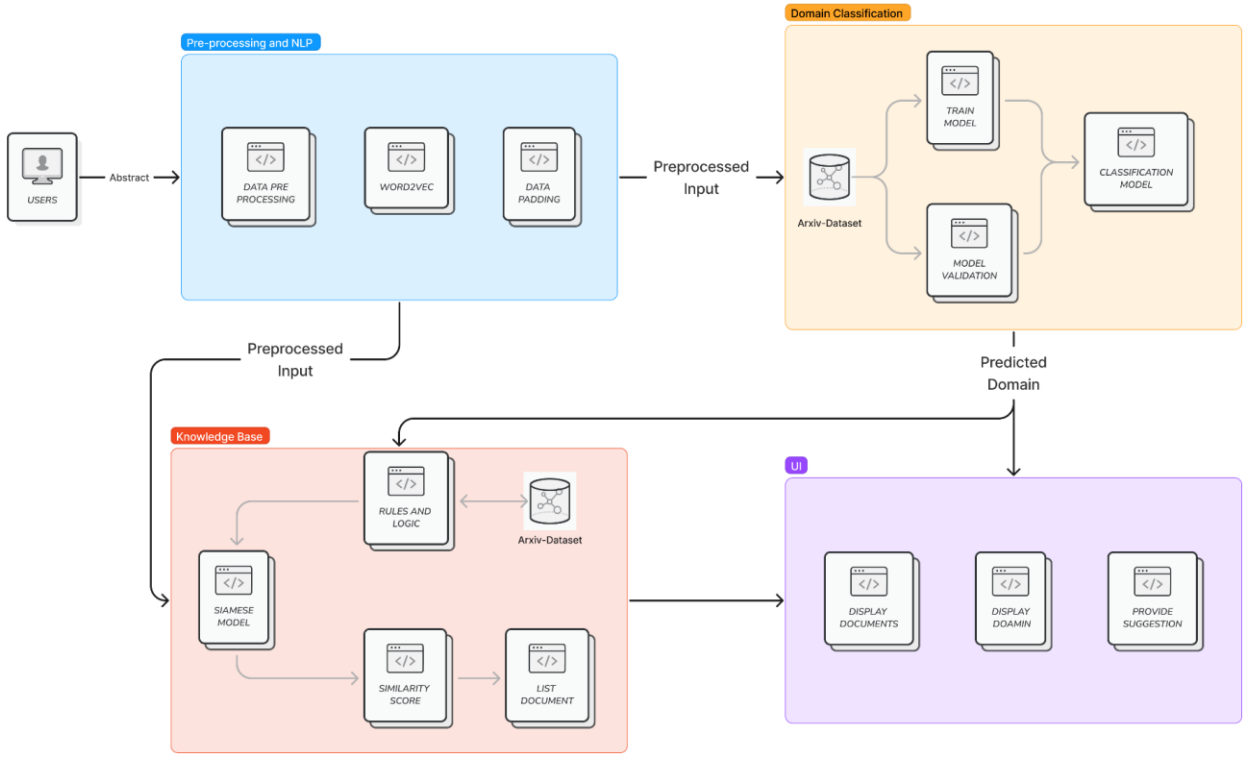

FIG.1 ARCHITECTURE DIAGRAM FOR JOURNAL PAPER SEARCH SYSTEM

B. DATASET

The datasets used for journal paper search systems typically include journal paper published in past year and other data such as abstract,title,authors etc.

For journal paper search systems, the dataset might include abstracts of journal papers , as well as necessary details such as title, authors,publication dates,doi,journal-ref. The data may also include technical indicators such as abstracts,domains,title which can be used as features for deep learning models.

ARXIV DATASET DESCRIPTION

This dataset is a mirror of the original ArXiv data. Because the full dataset is rather large (1.1TB and growing), this dataset provides only a metadata file in the json format. This file contains an entry for each paper, containing ArXiv ID (can be used to access the paper)submitter,authors,title,comments,journal-ref,doi,abstract,categories/ tags in the ArXiv system,versions.

C. LSTM ALGORITHM

LSTM stands for Long Short-Term Memory, and it is a type of recurrent neural network (RNN) used in deep learning. It is designed to overcome the limitations of traditional RNNs, which struggle with long-term dependencies, i.e., the difficulty in retaining information for longer periods of time. It was introduced in 1997 by Hochreiter and Schmidhuber as a solution to the vanishing gradient problem in RNNs, which occurs when the gradient becomes too small during backpropagation, making it difficult for the network to learn and retain information over time.

The LSTM architecture consists of a cell state and three gates - input, forget, and output gates. The cell state acts as a conveyor belt that carries information through time, while the gates regulate the flow of information into and out of the cell state.The input gate determines how much new information should be added to the cell state, based on the input and the previous hidden state. The forget gate decides which information to discard from the cell state, based on the input and the previous hidden state. The output gate determines which part of the cell state should be outputted as the final hidden state, based on the input and the previous hidden state.

LSTM networks are trained using backpropagation through time (BPTT), which involves computing the gradients of the loss function with respect to the network parameters over a sequence of time steps.LSTM networks are well-suited for processing sequential data, such as time-series data or natural language, because they can effectively capture long term dependencies and retain information over long periods of time.

LSTM has been used in various applications, including speech recognition, machine translation, image captioning, and time-series prediction, and has demonstrated state-of the-art performance in many of these tasks.

LSTM uses a memory cell to store and carry forward important information over long periods of time, which is then modified through input and output gates. The input gate decides which information to let into the memory cell, and the output gate decides which information to output. It is widely used in various applications, including natural language processing, speech recognition, and time-series prediction, due to its ability to handle long-term dependencies and effectively retain and process information over long periods of time.

The forward training process of the LSTM can be formulated with the following equations:

$$ft = \sigma\ WF \cdot [ht - 1, xt] + bf \quad (1)$$

$$it = \sigma\ Wf \cdot [ht-1, xt] + bi \quad (2)$$

$$Ct = ft * Ct-1 + it * tanh(Wc \cdot [ht-1, xt] + bc) \quad (3)$$

$$Ot = \sigma(W0 \cdot [ht-1, xt] + b0) \quad (4)$$

$$ht = ot * tabh(Ct) \quad (5)$$

where it , ot , and ft denote the activation of the input gate, output gate, and forget gate, respectively; Ct and ht denote the activation vector for each cell and memory block, respectively; and W and b denote the weight matrix and bias vector, respectively. In addition, σ(◦) denotes the sigmoid function.

IV. RESULTS AND DISCUSSION

PERFORMANCE AND ANALYSIS CHARTS

A training and validation loss chart shown in figure(2) is a graphical representation of the performance of a deep learning model during the training process. The chart displays the loss values (also called error or cost) of both the training and validation datasets over multiple epochs. The loss function measures the difference between the predicted output of the model and the true output. During training, the model tries to minimize the loss function by adjusting its parameters to better fit the data. The validation set is used to evaluate the model's performance on unseen data and prevent overfitting (when the model performs well on the training set but poorly on the validation set)

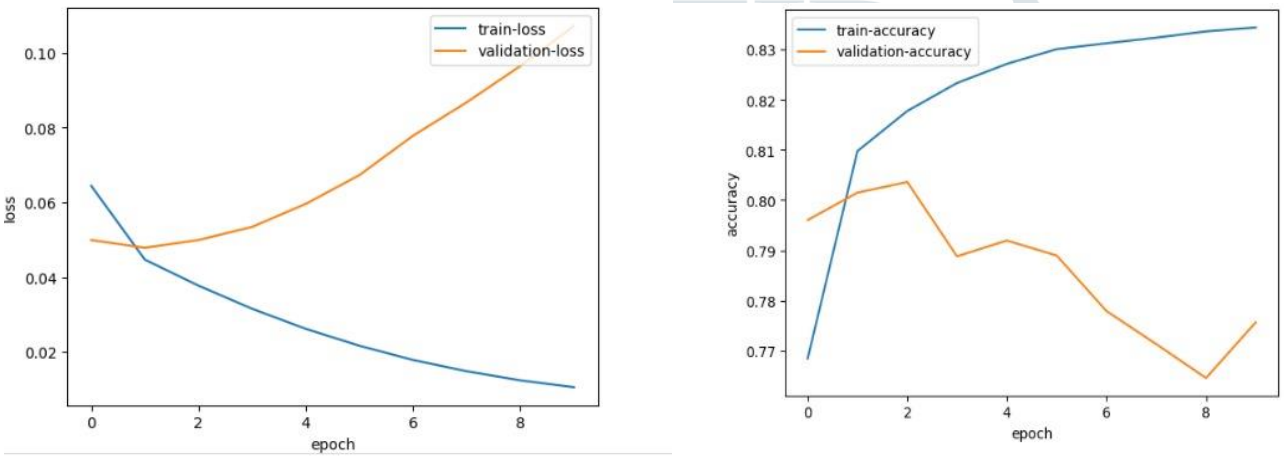

FIG2 TRAINING AND VALIDATION LOSS CHART

V. CONCLUSION

In conclusion, the proposed journal paper search system represents a significant advancement in academic information retrieval, aiming not only to enhance efficiency but also to promote originality and reduce plagiarism concerns.Using deep learning techniques such as LSTM classification and Siamese networks, the system provides a sophisticated approach to document categorization and grouping, allowing users to study linked papers with higher accuracy and relevance.

While the system excels at categorizing documents and making targeted recommendations for project enhancement, it is necessary to recognize some limits. The lack of similarity scores in certain existing methods emphasizes the need for more refining and integration of similarity assessment methodologies. Furthermore, factors such as input data quality, hyperparameter selection, and broader academic landscape dynamics may all have an impact on the system's efficacy.

Despite these challenges, the technology holds promise for future research and development. Continued efforts to improve and optimize deep learning algorithms, together with ongoing evaluation and validation of the system's performance, can lead to more robust and reliable journal paper search engines. This project improves scholarly activity and information dissemination in a number of fields by fostering a culture of originality and integrity in academic research.

VI. REFERENCE

[1] G. Mustafa et al., "Optimizing Document Classification: Unleashing the Power of Genetic Algorithms," in IEEE Access, vol. 11, pp. 83136-83149, 2023, doi: 10.1109/ACCESS.2023.3292248.
[2] J. Peng and S. Huo, "Few-Shot Text Classification Method Based on Feature Optimization," in Journal of Web Engineering, vol. 22, no. 3, pp. 497-514, May 2023, doi: 10.13052/jwe1540-9589.2235.
[3] X. Li, B. Tian and X. Tian, "Scientific Documents Retrieval Based on Graph Convolutional Network and Hesitant Fuzzy Set," in IEEE Access, vol. 11, pp. 27942-27954, 2023, doi: 10.1109/ACCESS.2023.3259234.
[4] Y. -S. Lin, J. -Y. Jiang and S. -J. Lee, "A Similarity Measure for Text Classification and Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 7, pp. 1575-1590, July 2014, doi: 10.1109/TKDE.2013.19.
[5] K.-C. Khor and C.-Y. Ting, "A Bayesian approach to classify conference papers", Proc. Mex. Int. Conf. Artif. Intell., pp. 1027-1036, Nov. 2006.
[6] R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," in IEEE Transactions on Fuzzy Systems, vol. 26, no. 2, pp. 794-804, April 2018, doi: 10.1109/TFUZZ.2017.2690222.
[7] Beel, J., Gipp, B., Langer, S. et al. Research-paper recommender systems: a literature survey. Int J Digit Libr 17, 305–338 (2016).
[8] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer US.
[9] Nam, Le Nguyen Hoai and Bao-Quoc Ho. “A Comprehensive Filter Feature Selection for Improving Document Classification.” Pacific Asia Conference on Language, Information and Computation (2015).
[10]B. Tang, H. He, P. M. Baggenstoss and S. Kay, "A Bayesian classification approach using class-specific features for text categorization", IEEE Trans. Knowl. Data Eng., vol. 28, no. 6, pp. 1602-1606, Jun. 2016.