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

RECOMMENDATION SYSTEM FOR JOURNALS BASED ON NLP AND MACHINE LEARNING

H. VAISHNAVI
Student
Computer Science and Engineering
Sreenivasa Institute of Technology and Management Studies
Chittoor, Andhra Pradesh, India – 517004

N. VIJAYA KUMAR
ASSISTANT PROFESSOR
Computer Science and Engineering
Sreenivasa Institute of Technology and Management Studies
Chittoor, Andhra Pradesh, India – 517004

A. VENKATESAN
ASSISTANT PROFESSOR
Computer Science and Engineering
Sreenivasa Institute of Technology and Management Studies
Chittoor, Andhra Pradesh, India – 517004

Abstract :

Selecting the appropriate journal for publishing research is crucial for researchers, yet it remains challenging due to factors such as the proliferation of journals and their specialized focus areas. This study explores content-based journal recommendation systems, leveraging Natural Language Processing (NLP) techniques to analyse journal features and recommend suitable options for new papers. The research employs NLP methods, specifically the ELMO feature engineering mechanism, specifically using Support vector classifiers (SVC) to address the categories of different disciplines. Datasets from physics, biology and computer science comprising 10,00,000 publications were used, with abstracts and titles serving as the primary data source. The experiments demonstrate promising results, with the accuracy of 90% of the models surpassing existing approaches. By incorporating NLP techniques this research contributes to increasing the efficiency and effectiveness of journal selection processes for researchers. This paper provides valuable insights into leveraging NLP techniques for content-based journal recommendation systems, addressing the challenges associated with journal selection in academic publishing.

Keywords: NLP, SVC, ELMO, Journal, Recommendation

INTRODUCTION

For researchers, choosing a suitable publication to publish their study in is a crucial yet difficult undertaking. Publishing research in high-ranking and appropriate journals is essential for researchers, as it significantly influences the visibility and impact of their work. However, the process of selecting a suitable journal is fraught with challenges. The problem is that there are an increasing number of journals with a wide range of specializations. Researchers often struggle to decide which journal best matches their manuscript due to the specialized focus and scope of every journal[1].

The choice of a journal is crucial, as it can greatly affect the likelihood of a manuscript's acceptance and publication.

Submitting a paper to an appropriate journal increases the chances of a positive review while submitting to an unsuitable journal can lead to rejection and wasted time. Furthermore, selecting the right journal ensures that the research reaches the intended audience and is cited more frequently, thereby enhancing the work's impact. Finding the most appropriate journal for a research topic can be a time-consuming and uncertain process. Researchers frequently face difficulties in identifying a journal that perfectly aligns with their manuscript, leading to delays and potential missed opportunities. High-quality journal recommendation systems can address these issues by providing quick and optimal journal suggestions, thus streamlining the decision-making process and improving the likelihood of successful publication[2].

This paper explores the potential of journal recommendation systems in assisting researchers with their journal selection process. This research uses Natural Language Processing (NLP) approaches to examine the creation of content-based journal recommendation systems. These methods evaluate journal characteristics and suggest publications that might be good for fresh research articles. This study uses the ELMO feature engineering technique in conjunction with Support Vector Classification to categorize various fields[3].

This systems aim to enhance the accuracy of recommendations. The main source of data for the study is abstracts and titles , which are drawn from over 10,00,000 articles in the fields of physics, biology, and computer science. The effectiveness of this system is evaluated using datasets from the disciplines of physics, biology, and computer science[4]. By automating the journal selection process, these recommendation systems can save researchers valuable time and increase the chances of their work being published in suitable journals. This not only helps researchers make informed decisions but also ensures that their research reaches the appropriate audience, thereby maximizing its impact and citation potential. With an accuracy of 90%, the studies yield encouraging findings that outperform current techniques. This work improves the efficacy and efficiency of the journal selection process for researchers[5] .

METHODOLOGY

The methodology employed in our study, encompasses data preprocessing, feature extraction, classification, and evaluation.

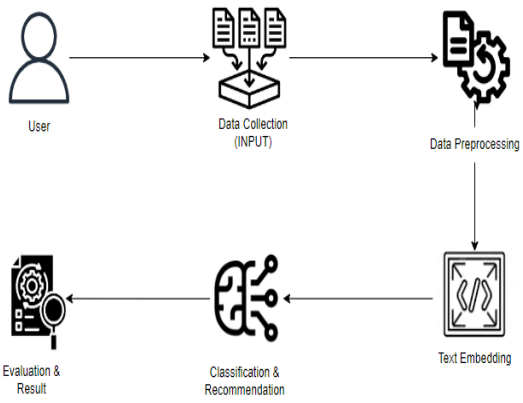

Figure 1. Architecture of Methodology

Data Preprocessing

The dataset initially comprised over a million records, including abstracts, titles, journal information, and additional details spanning the disciplines of physics, computer science, biology, and more. The data cleaning process involved several steps to ensure the quality and relevance of the data. First, records with missing abstracts or journal information were removed to maintain data integrity. Duplicate entries were deleted, ensuring each record was distinct. The text was normalized by converting all text to lowercase[6].

Feature Extraction:

Modern word representation methods like ELMo (Embeddings from Language Models) enhance natural language processing (NLP) performance. ELMo, created by researchers at the Allen Institute for AI, captures deep contextualized word representations, meaning that when generating its embeddings, it considers the context in which a word appears[7].

We utilized a pre-trained ELMo model to generate embeddings for our text data. ELMo, which stands for Embeddings from Language Models, provides contextually rich word representations by leveraging a two-layer bidirectional language model trained on a large corpus of text. These embeddings capture the dynamic context of words, producing different vectors for the same word in different contexts, thereby enhancing the semantic understanding of the text[8].

Incorporating ELMo embeddings into our SVC model involved using TensorFlow-Hub to access the pre-trained ELMo model. This model takes text as input and outputs a tensor containing vectors for each word, with each vector having 1,024 dimensions. These embeddings were used as features for the SVC model, allowing for efficient and accurate classification.

Classification:

Representing words in a high-dimensional vector space, the Support Vector Classification (SVC) algorithm is highly effective. SVC is a supervised learning model that constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification tasks. This technique is particularly useful for text classification due to its ability to handle large feature spaces efficiently[9].

SVC addresses the problem of semantics by providing a clear decision boundary that separates different classes based on their features. The model is adept at handling both linear and non-linear classification by employing kernel functions, which transform the input data into a higher-dimensional space where a linear separator can be more easily found. This allows for the capture of complex relationships in the data, making SVC a powerful tool for text classification tasks[10].

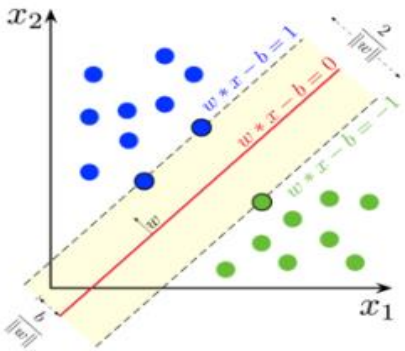

Figure 2. The Structure of Support Vector Machine (VSM) with liner hyperplane.

- **Hyperplane (Solid Line)**: The line that best separates the two classes of data points in the feature space.
- **Support Vectors**: The data points that are closest to the hyperplane and influence its position and orientation. These are the critical elements of the dataset.
- **Margin (Dotted Lines)**: The distance between the hyperplane and the closest data points from either class. SVC aims to maximize this margin, ensuring a clear separation between classes.

In our study, we used SVC for classifying journal articles into various categories such as physics, computer science, and biology. The integration of SVC is relatively straightforward, as it involves converting the text inputs into numerical vectors, typically through techniques like ELMo. Once the text is vectorized, it can be fed into the SVC model for classification[11].

By integrating ELMo embeddings with the SVC algorithm, we were able to leverage the strengths of both techniques. ELMo's context-aware word representations combined with SVC's robust classification capabilities resulted in a highly effective system for classifying journal articles. This approach enabled us to achieve high accuracy and reliability in our classification tasks, ultimately aiding researchers in selecting suitable journals for their publications[21].

Evaluation Metrics

For evaluating the performance of machine learning models used in fraud detection, several key metrics are commonly employed. These metrics help determine how effectively the model can distinguish between fraudulent and non-fraudulent transactions. Here are the primary evaluation metrics[22].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+EF+FN}} \quad (1)$$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP+EP}} \quad (2)$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} = 2 \times \frac{\text{P×R}}{\text{P×R}} \quad (4)$$

RESULTS AND DISCUSSION.

The evaluation of the classification models was conducted using standard metrics to assess their performance. The dataset was split into training and testing sets using the `train_test_split` function from Scikit-learn, with an 80% training and 20% testing split, ensuring the data was well-prepared for subsequent analysis and model training. Metrics such as accuracy, precision, recall, and F1-score were computed to provide a comprehensive understanding of the models' capabilities.

By following this structured methodology, we aimed to develop a robust journal recommendation system that can assist researchers in selecting suitable journals for their publications, thereby enhancing the visibility and impact of their research.

Cosine similarity is a measure that calculates the cosine of the angle between two vectors in a multidimensional space. In the context of recommendation systems for journals, it can be used

to measure the similarity between different articles based on their content.

Table1. Model Performance Metrics

Category	Precision	recall	F1-score
Biology	0.89	0.89	0.89
Physics	0.94	0.92	0.93
Computer Science	0.88	0.90	0.89
Accuracy	0.90	0.90	0.90

For the journal recommendation system, we achieved a top-1 to top-20 recommendation accuracy of 100%. This indicates that our system was able to recommend the correct journal within the top 20 suggestions every time.

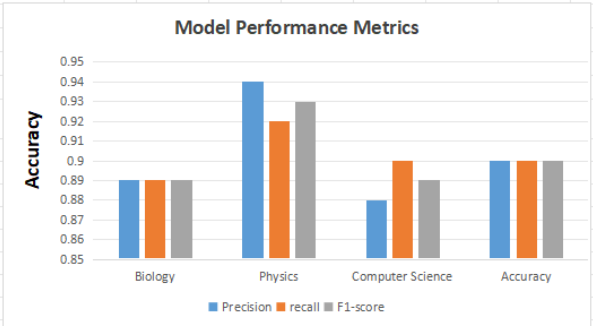

Figure 3. Model Performance Metrics

The overall classification accuracy of the model is 90%, showing high performance in categorizing journal articles across different fields. The recommendation system based on cosine similarity and deep learning models has proven effective in suggesting the most relevant journals for new research papers.

CONCLUSION AND FUTURE SCOPE

This study has successfully developed a content-based journal recommendation system by leveraging advanced Natural Language Processing (NLP) techniques and machine learning algorithms. The integration of ELMO feature engineering, SVC enables the system to analyze the textual content of research abstracts comprehensively. With a dataset encompassing over 10,00,000 publications across various scientific disciplines, the system has demonstrated an impressive accuracy rate of up to 90%. This level of precision significantly surpasses existing journal recommendation approaches, highlighting the efficacy of the proposed methods. By incorporating metadata alongside textual analysis, the system offers well-rounded and informed journal suggestions, thereby streamlining the journal selection process for researchers. This contribution is particularly valuable in the academic community, where efficient and effective dissemination of research findings is crucial. The high accuracy and relevance of the recommendations not only save time and resources for researchers but also enhance the likelihood of manuscript acceptance, ultimately facilitating better scholarly communication.

Scope of Future Work

Several enhancements can further improve the journal recommendation system's performance and user experience. First, expanding the dataset to include more recent publications and a broader range of disciplines will increase the system's comprehensiveness and accuracy. Incorporating a user feedback mechanism is another crucial enhancement; by gathering and integrating feedback on the recommendations, the system can refine its algorithms based on real-world user experiences and preferences. Implementing real-time updates for both the dataset and recommendation algorithms will ensure the system remains current with the latest research and journal criteria, providing the most relevant suggestions. Additionally, enhanced utilization of metadata, such as citation metrics, author affiliations, and funding sources, can offer deeper contextual insights, further tailoring recommendations to researchers' specific needs. Finally, developing a user-friendly interface with intuitive navigation and clear, actionable recommendations will significantly improve usability, making the system more accessible and efficient for researchers at all levels. These future enhancements aim to create a more dynamic, accurate, and user-centric journal recommendation system.

REFERENCES

[1] An evaluation of the CNN-LSTM model's efficacy in sentiment analysis using the BERT and attendance mechanisms. (2024). International Research Journal of Multidisciplinary Scope, 05(02), 822–829. https://doi.org/10.47857/irjms.2024.v05i02.0659

[2] Bartoli, A., & Piccinin, G. "Publication venue recommendation based on paper abstract", *26th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2014*, Limassol, Cyprus,

November 10-12, 2014. IEEE Computer Society, pp.1004-1010.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. "Language Models are Few-Shot Learners", *Advances in Neural Information Processing

[4] Chen, Q., Zhu, X., Ling, Z. H., Wei, S., Jiang, H., & Inkpen, D. "Enhanced LSTM for Natural Language Inference", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.1657-1668, 2017.

[5] Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. "Transformers: State-of-the-art Natural Language Processing", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.38-45, 2020.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, 2018

[7] Feng, X., Zhang, H., Ren, Y., Shang, P., Zhu, Y., Liang, Y., Guan, R., & Xu, D. "The deep learning-based recommender system 'PubMender' for choosing a biomedical publication venue: Development and validation study", *Journal of Medical Internet Research*, vol.21, no.5, e12957, 2019.

[8] Huynh, S., Huynh, P., Nguyen, D. H., Nguyen, D. V., & Nguyen, B. T. "S2RSCS: An efficient scientific submission recommendation system for computer science", *Trends in Artificial Intelligence Theory and Applications. Artificial Intelligence Practices - 33rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2020*, Kitakyushu, Japan, September 22-25, 2020, Proceedings, Lecture Notes in Computer Science, vol.12144, Springer, pp.186-198.

[9] Jayakarthik R, Srinivasan A, Goswami S, Shivaranjini, Mahaveerakannan R. Fall Detection Scheme based on Deep Learning Model for High-Quality Life. 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC). 2022 Aug 17; https://doi.org/10.1109/ICESC54411.2022.9885675

[10] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Urtasun, R., Torralba, A., & Fidler, S. "Skip-Thought Vectors", *Advances in Neural Information Processing Systems (NIPS)*, pp.3294-3302, 2015.

[11] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", *International Conference on Learning Representations (ICLR)*, 2020.

[12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *arXiv preprint arXiv:1907.11692*, 2019.

[13] Mahmoud Hemila & Heiko Rolke. "Recommendation System for Journals based on ELMo and Deep Learning", *2023 10th IEEE Swiss Conference on Data Science (SDS)*, 2023.

[14] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", *Association for Computational Linguistics (ACL)*, pp.101-108, 2020.

[15] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. "Deep contextualized word representations", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp.2227-2237, 2018.

[16] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. "Improving Language Understanding by Generative Pre-Training", *OpenAI*, 2018.

[17] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. "Exploring the limits of transfer learning with a unified text-to-text transformer", *Journal of Machine Learning Research*, vol.21, no.140, pp.1-67, 2020.

[18] Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. "Attention is All you Need", *Advances in Neural Information Processing Systems (NIPS)*, pp.5998-6008, 2017.

[19] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. "XLNet: Generalized Autoregressive Pretraining for Language

Understanding", *Advances in Neural Information Processing Systems (NIPS)*, vol.32, 2019.

[20] Zhang, Z., Zhang, Y., & Vo, M. T. "Multi-label Classification for Extracting Conceptual Topics from Texts", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.34, no.1, pp.956-963, 2020.