Scientific Event Recommendation through Semantic Clustering and Embedding Analysis

Karla S. Silva[1], **Rebeca L. Rezende**[1], **João V. D. Silva**[1], **Gabriel Cardoso**[1], **Thiago M. Ventura**[1], **Allan G. de Oliveira**[1]

[1]Programa de Pós-Graduação em Computação Aplicada (PPGComp) - Universidade Federal de Mato Grosso (UFMT), Cuiabá - Mato Grosso - Brazil

karlasouza1224@gmail.com, beca.rezendel@gmail.com, jvdantasilva@hotmail.com, gabriel.cardoso9811@gmail.com, {thiago, allan}@ic.ufmt.br

Abstract. *This work presents a series of experiments aimed at recommending scientific events within the Brazilian Computer Society (SBC), leveraging semantic embeddings generated by Large Language Models (LLMs) and unsupervised clustering techniques. Using BERTopic for topic modeling and multilingual representations, the experimental approach processes over 12,000 articles from 30 events to align submissions with relevant themes. The best configuration reached an accuracy of 0.91 and demonstrates the potential of LLM-based embeddings to support decision-making in scientific dissemination.*

1. Introduction

Participation in academic events constitutes an essential stage in scientific training, as it contributes to the development of writing skills, the strengthening of scientific production, and integration into collaborative networks, as evidenced by Araújo, Costa, and Lima (2021). According to Jesus et al. (2020), these events significantly promote peer communication and the dissemination of new knowledge, being fundamental sources of learning, especially in professional training contexts.

However, with the expressive growth of scientific production Bornmann et al. (2021) and the increase in the number of events promoted annually by SBC (2023), identifying the most appropriate conference for submitting a paper has become a recurring challenge among researchers, especially in areas with significant thematic overlap. As highlighted by Iyer et al. (2020), many authors face the dilemma of choosing the right conference to disseminate their research, which reinforces the need for systems that support this decision-making process.

Juliani and Donha (2023) point out that recommendation systems based on user profiles can increase engagement with scientific production by facilitating access to relevant content. According to Flicke et al. (2024), personalized recommendations trained from each user's evaluations enhance the relevance of suggested content, contributing to the efficiency of scientific search.

Complementing this perspective, Wang et al. (2022) highlight the role of intelligent recommendation systems in the sustainability of scientific production. Similarly, Asabere et al. (2020) demonstrate that by integrating user behavioral aspects with machine learning techniques, these solutions generate more accurate and personalized recommendations, reducing the time required to find suitable events and facilitating the inclusion of papers in more appropriate scientific forums.

In this context, this work presents a series of experiments focused on recommending events promoted by the Brazilian Computer Society (SBC), based on the semantic analysis of titles and abstracts of submitted papers. The SBC Open Lib (SOL) platform, which gathers a

vast collection of scientific publications, served as the data source for building the experimental models, which employed Natural Language Processing (NLP) and machine learning techniques to extract topics and measure textual similarity.

Studies such as Božić and Grljević (2024) demonstrate the effectiveness of topic modeling and sentiment analysis in personalized recommendation systems, employing NLP and machine learning techniques to enhance the accuracy of suggestions based on textual similarity. Complementarily, Rumadi et al. (2023) highlight the potential of using BERTopic in organizing large volumes of scientific publications by thematic areas, contributing to strategic decision-making in research institutions.

Thus, the experiments conducted in this work aim to support researchers in identifying suitable venues for paper submission, while also offering insights for event organizers in curating content aligned with the themes of their conferences, thereby promoting more strategic and effective participation in SBC activities.

The structure of this article is organized as follows: Section 2 presents the related works, with emphasis on scientific recommendation approaches based on user profiles and machine learning techniques. Section 3 details the adopted methodology, including data collection and preprocessing, topic extraction with BERTopic, and textual similarity calculation for event recommendation. Section 4 presents the results obtained from the experiments conducted. Finally, Section 5 provides the conclusions and directions for future work.

2. Related Work

As related work, Dhinakaran et al. (2022) proposes a recommendation system based on a graph content classification (GCR) model, which exploits relationships between authors and articles to personalize suggestions for scientific studies and academic events. The approach builds collaborative networks (user-article, author-article and article-article) to identify affinity patterns, such as frequently cited authors or recurring co-authors, and uses ranking algorithms to prioritize recommendations.

Also noteworthy is the work of Abinaya et al. (2023), who presented the Time Cluster Personalized Ranking Recommender System (TCPRRS), a recommender system that uses temporal information on user consumption and clustering techniques to generate personalized rankings in multi-cloud environments. The proposal incorporates particle swarm optimization (PSO) to improve the quality of recommendations, overcoming challenges related to scalability and accuracy typical of systems based on collaborative filtering. Although mostly applied to the context of product recommendation, the method offers relevant contributions that can be adapted to the promotion and recommendation of scientific participations, especially due to its ability to deal with large volumes of data and temporal dynamics.

Thus, it can be seen that although the aforementioned works outline user profiles and scientific communities with different approaches, they all share the goal of improving recommendation systems. The different approaches, which range from textual and demographic analysis, through the structuring of collaboration networks, to personalization based on temporal behaviour, demonstrate the diversity and potential of machine learning and artificial intelligence techniques in promoting scientific participation and advancing collaborative knowledge.

Among the research that is in line with this proposal is the work by Juliani and Donha (2023), which presents a system for recommending scientific articles integrated into the Moodle virtual learning environment. The solution aims to personalize the student experience by suggesting scientific readings aligned with the didactic content accessed, using the content-based filtering (CBF) technique and the TF-IDF algorithm. The system analyzes the material consulted by the user and identifies articles with high textual similarity in an open access journal database. Although the focus is on the educational context and reducing dropout in distance learning courses, the approach highlights the potential of user profile-oriented recommendation systems to foster engagement with scientific production. This contribution reinforces the relevance of personalization based on behavior and profile as a strategy to increase participation and interest in scientific activities.

Wang et al. (2022), which presents an intelligent recommendation system for literature and scientific research data, with the aim of supporting a sustainable scientific production environment. Using machine learning techniques and a hybrid recommendation algorithm (CBR-CF), the system integrates aspects of user psychology, such as curiosity, to improve the accuracy and relevance of recommendations. The proposed architecture includes modules for textual representation, similarity calculation, clustering and dynamic updating of the user profile, resulting in improvements of up to 91% in recall and 93% in accuracy for personalized recommendations. The approach shows potential for reducing search time for relevant content and optimizing scientific participation, which aligns with this work's experimental exploration of how user profiles and machine learning techniques can be leveraged to promote engagement in research activities.

3. Methodology

To conduct this study, a series of experiments was carried out to group scientific articles based on thematic and event-related similarities, using Natural Language Processing (NLP) and Machine Learning (ML) techniques. To identify semantic relationships between the articles and the corresponding scientific events, the following steps were performed: data collection and selection from the platform, noise removal through preprocessing, topic extraction, and event recommendation based on the provided title and abstract.

3.1. Data Collection and Selection

The data used for building the model were obtained through web scraping from the SOL platform (SBC - Open Lib), specifically from the sections "Symposia, Conference Proceedings" and "Conferences, Congresses, and Symposia". The titles and abstracts of articles submitted to various editions of scientific events hosted on the platform were extracted.

Since the documents are available in two languages (Portuguese and English), a language distribution analysis was conducted in May 2025 to determine which dataset would be used. A total of 13,073 articles, 6,229 articles in Portuguese, 6,782 in English were identified and 62 unidentified. Given the predominance of Portuguese and the intention to avoid inaccuracies associated with the use of multilingual models, it was decided to use only the texts written in Portuguese.

Although some models, such as RoBERTa, offer support for multiple languages, their performance in contexts with mixed-language data can compromise the quality of semantic groupings [Wu and Dredze, 2023]. Therefore, linguistic homogeneity was maintained to ensure greater precision in the results.

3.2. Topic Extraction with BERTopic

After data collection, the textual data were preprocessed through the following steps: removal of Portuguese stopwords, lemmatization to reduce word variation, and normalization of special characters and punctuation. This preprocessing aimed to reduce noise in the text and standardize it for the semantic vectorization stage [Sivakumar and Gunasundari, 2017]. For the identification and thematic grouping of articles, the BERTopic technique was applied. This method combines BERT-based language models with clustering algorithms to extract topics from short texts.

The BERTopic architecture begins with the generation of embeddings, where titles and abstracts are converted into numerical vectors using pre-trained BERT models. Then, dimensionality reduction is performed to facilitate the clustering process. Finally, topic modeling is carried out through the extraction of representative keywords for each cluster, defining its central themes.

3.3. Event Recommendation by Similarity and Evaluation

After grouping the articles by topic, a series of recommendation experiments was conducted based on the semantic similarity between text vectors. Cosine similarity was used as a metric to compare thematic proximity between articles, enabling the recommendation of events related to a given title and abstract. This approach aims to organize large volumes of publications by theme, while also supporting the discovery of relevant events for both authors and organizers, thereby facilitating the publication process.

The Top-N metric is widely used in recommendation tasks to assess the relevance of the suggestions generated by a model. In this approach, only the top N recommendations produced by the experimental setup are considered, rather than all possible ones, simulating a realistic scenario where users would only view a limited number of options, as shown in Equation (1).

$$Accuray = \frac{n_p}{n_r} \qquad (1)$$

Where:

np: Number of correct recommendations (e.g., events that match the user's paper).

nr: Total number of displayed recommendations.

4. Results and Discussions

The academic event recommendation experiments conducted in this work were structured into three main versions, each incorporating successive improvements in their architecture and achieving progressively better performance, with a particular emphasis on recommendation accuracy. According to Table 1, the first version used the BERTimbau language model, specialized in Brazilian Portuguese, to generate semantic embeddings from the concatenation of the title and abstract of each article. The articles were grouped using the K-means clustering algorithm, configured to form 38 clusters - corresponding to the 38 distinct events present in the Portuguese subset of the dataset used in this version.

This initial dataset consisted of 6,229 articles in Portuguese, which were split into training and test sets in an 80/20 ratio. Embeddings were computed as the mean of the token vectors extracted by the BERTimbau model, resulting in a dense vector representation for each article. After clustering, a mapping was established between clusters and events, based on the most frequent events in each group, which served as the foundation for the recommendation step. Based on the similarity to existing clusters, the model generated the top three most likely events for each test article. Despite its simplicity, this initial version achieved an accuracy of 0.69 in predicting the actual event among the top-3 recommendations, demonstrating the feasibility of the approach even in the presence of limitations such as the absence of dimensionality reduction, lack of language filtering, and a basic text preprocessing stage.

In the second version of the experiments, the approach was restructured through the adoption of the BERTopic framework, which allowed for more interpretable and robust topic modeling, resulting in significant improvements in both semantic organization and overall performance. The BERTimbau model was replaced with the multilingual "paraphrase-multilingual-MiniLM-L12-v2" model, which delivers superior performance in semantic similarity tasks and is computationally more efficient. Additionally, the UMAP dimensionality reduction technique was introduced prior to the application of K-means (now with 30 clusters), which led to more cohesive and semantically distinct groupings.

The preprocessing pipeline was also improved by incorporating text normalization and automatic language detection, reducing noise and increasing dataset homogeneity. Relevant changes were also made to the dataset: from the original 38 events, 8 were excluded due to having fewer than 100 articles each, resulting in a refined dataset with 30 events and approximately 12,629 articles - about 96.6% of the original corpus. These dataset refinements, combined with architectural enhancements, led to a considerable performance gain, with accuracy rising to 0.90 - a 26% increase compared to the first version. Moreover, qualitative evaluation indicated that the recommendations were more thematically aligned with the original article content, confirming the effectiveness of the topic-based approach.

The third and final version introduced an additional preprocessing step aimed at filtering out articles with an unknown language. In this step, texts not automatically detected as either Portuguese or English were removed from the dataset to eliminate noisy examples that could impair the quality of the topic modeling. This filtering proved effective in further refining the generated clusters, reducing semantic dispersion among articles and enhancing topic consistency. As a result, the accuracy observed in the experiments increased slightly to 0.91. Although the numerical gain was modest, this version represents an important qualitative improvement, particularly in data organization and the robustness of the generated recommendations, with notable benefits in topics with a high volume of articles.[1]

Table 1. Comparison of Article Recommendation Experiments

Experiment	Model	Techniques	Dataset	Top-3 Accuracy
1	BERTimbau	K-means	6.229 articles in Portuguese across 38 events	0.69
2	MiniLM	BERTopic + UMAP + K-means	13.073 articles in Portuguese and English across 38 events	0.9
3	MiniLM	BERTopic + UMAP + K-means	12.629 articles in Portguese and English across 30 events	0.91

[1] https://sol.sbc.org.br/index.php/indice

The comparison across the three versions reveals a clear trajectory of technical and methodological evolution, directly reflected in the outcomes achieved. The initial version, based on BERTimbau and K-means, played a fundamental role in validating the approach, despite its limitations in language handling and dimensionality control. The second version marked a significant structural transformation by integrating BERTopic, applying dimensionality reduction, improving preprocessing, and refining the dataset - resulting in the greatest performance gain. The third version consolidated previous advances by explicitly addressing the quality of input data, ensuring greater robustness of the experimental approach. The accuracy comparison - 0.69 in the first version, 0.90 in the second, and 0.91 in the third - clearly demonstrates the impact of model choices and data preparation on final results, highlighting the importance of an incremental and well-founded approach in the development of machine learning-based recommendation methods.

To determine the optimal number of clusters (k) for topic modeling, two evaluation methods were applied: the Silhouette Score and the Elbow Method, as shown in Figure 1. The Silhouette Score plot (left) measures cluster cohesion and separation, with higher values indicating more well-defined groupings. Although the score fluctuates across the tested values, a relatively high peak is observed at k = 18, suggesting strong intra-cluster similarity and inter-cluster separability at this point.

Simultaneously, the Elbow Method plot (right), which analyzes the Within-Cluster Sum of Squares (WCSS), reveals a noticeable change in curvature - or "elbow" - also around k = 18, indicating diminishing returns in cluster compactness beyond this value. The convergence of both criteria supports the selection of 18 clusters as a balanced trade-off between clustering quality and model simplicity. However, for the final experiments, k = 30 was adopted to maintain alignment with the number of events in the refined dataset.

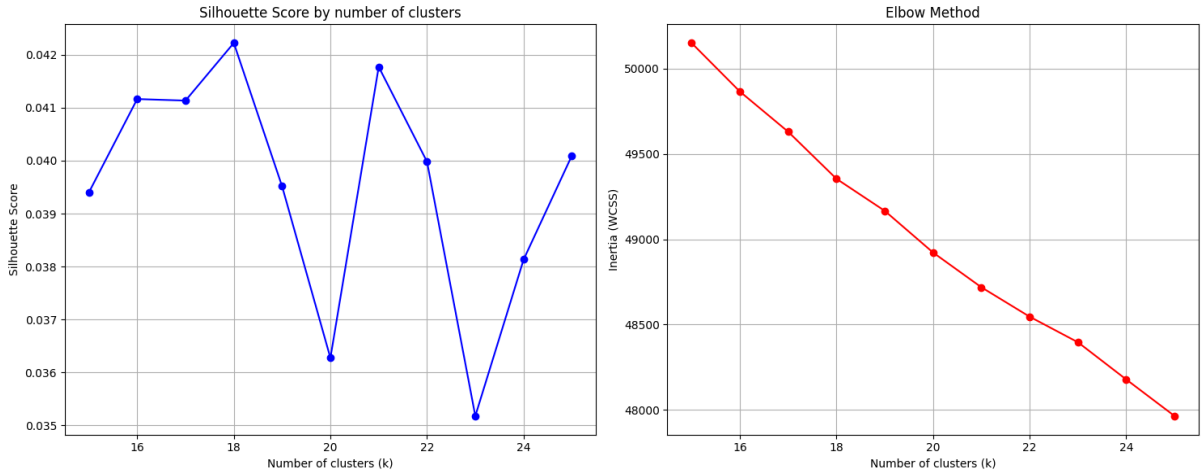

Figure 1. Silhouette Score and Elbow Method for Determining the Optimal Number of Clusters

The two-dimensional visualization (Figure 1), obtained using the UMAP dimensionality reduction technique, provides a graphic interpretation of the semantic structure identified by the model. Although the overall Silhouette Score is relatively low (0.024), indicating overlap between the boundaries of some clusters, the visualization reveals coherent spatial formations and partial separations between thematic areas. This suggests that, although the semantic proximity between certain SBC events naturally generates intersections, the model was able to form identifiable and thematically meaningful groupings.

Denser regions indicate greater internal similarity, while more dispersed clusters may be related to interdisciplinary areas or those with a broader thematic scope.

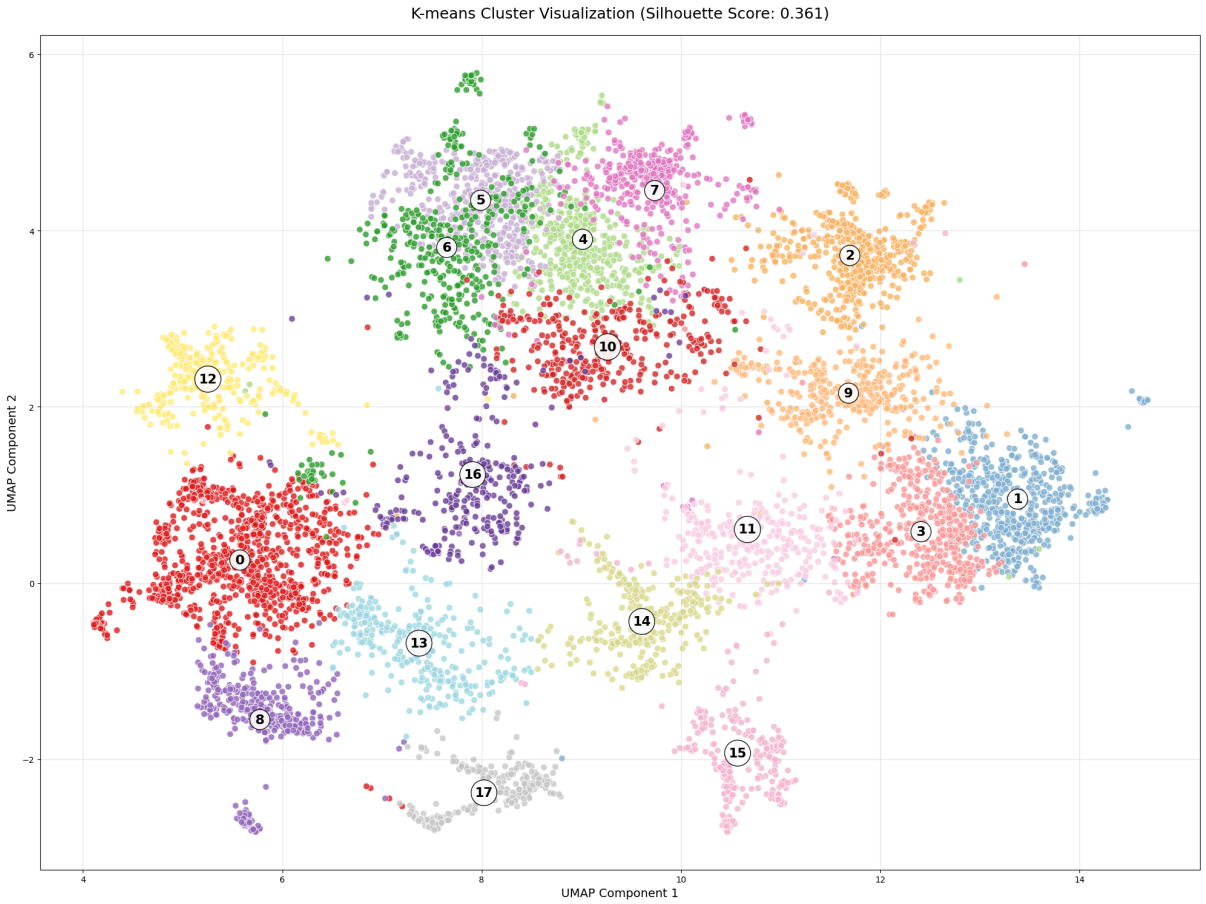

Figure 2. Cluster visualization generated by the K-means algorithm with dimensionality reduction using UMAP

The cluster size distribution graph (figure 2) shows a thematic imbalance in the data set. Larger clusters, between 500 and 800 articles, possibly represent more consolidated events or widely covered topics, while smaller groupings (with less than 300 articles) tend to indicate specialized niches or events with less representation in the database. This distribution is to be expected in heterogeneous academic databases and is in line with the aim of identifying both broad lines of research and specific sub-themes. However, it is important to monitor the occurrence of super-groupings in very large clusters, as this can compromise thematic granularity and reduce the accuracy of recommendations. This imbalance is natural in heterogeneous academic collections, but it also influences recommendation behavior: larger clusters increase recall potential, while smaller ones risk underrepresentation in suggestions.

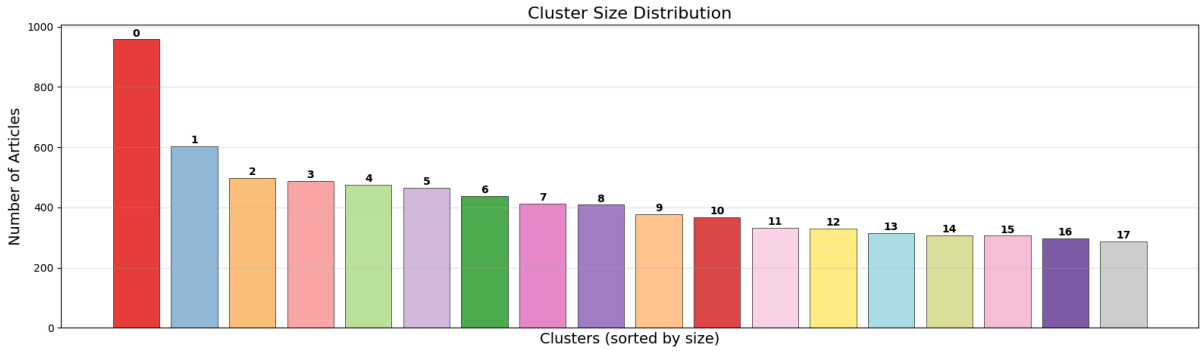

Figure 3. Distribution of cluster sizes (sorted by size)

Finally, the cluster purity analysis (Figure 3) quantifies the degree of thematic cohesion by measuring the proportion of articles belonging to the same event within each cluster. High-purity clusters (above 60%) demonstrate a strong correspondence between the unsupervised grouping and the actual event labels, indicating that the BERTopic model was able to effectively capture domain-specific language patterns and thematic boundaries. On the other hand, clusters with lower purity often emerge from research areas where thematic overlap between events is intrinsic - for example, artificial intelligence and data science, which share terminology, methodologies, and application contexts. Another factor contributing to reduced purity is the limitation of working only with titles and abstracts, which may not contain enough discriminative information to fully separate closely related topics.

Lower-purity clusters do not necessarily represent a model failure but rather highlight the complexity of the SBC event ecosystem. Many conferences within SBC have interdisciplinary scopes, attracting submissions from overlapping research domains, which naturally increases semantic intersection in vector space. In these cases, even a well-performing model may group together papers from different events that are conceptually aligned. This suggests opportunities for refinement, such as incorporating richer metadata (author keywords, conference tracks, or citation networks) or adjusting the number of clusters to better balance thematic granularity and separation.

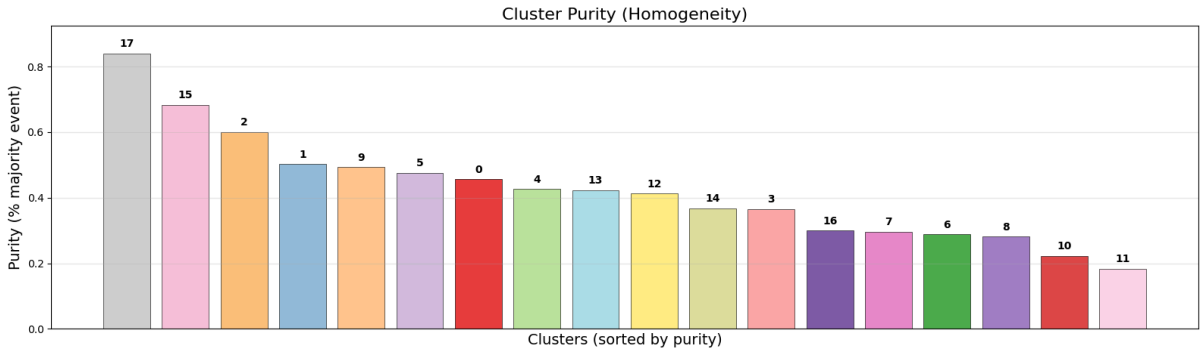

Figure 4. Cluster purity analysis (sorted by purity)

Overall, these visual analyses complement the quantitative accuracy results and reinforce both the interpretability and diagnostic potential of the topic modeling approach. They therefore support the viability of the proposed methodology as a promising experimental basis for decision-support in the context of academic event recommendation.

The experimental development of academic event recommendation strategies explored in this work shows interesting parallels with the research mapping study conducted by Rumadi et al. (2023) in their paper "Research Mapping in Research Organizations Based on Abstracts of Scientific Publications using BERTopic Modeling." Both studies employ advanced natural language processing (NLP) techniques based on BERT models for extracting semantic representations from academic texts, yet they pursue different objectives that lead to methodological differences and distinct outcomes.

The aforementioned article, developed by researchers from Institut Teknologi Sepuluh Nopember (Surabaya, Indonesia), utilizes BERTopic to map research lines at the Aeronautics and Space Research Organization (ORPA) of Indonesia, analyzing 1,107 scientific publications from the Scopus database. Their approach combines BERT embeddings with c-TF-IDF to identify key terms and cluster articles into 29 coherent topics, categorized into areas such as satellite technology, aeronautics, and remote sensing. Topic validation is performed through coherence metrics and relevance analysis for the organization, providing a macro-level overview of research trends. This mapping serves a strategic purpose, assisting the institution in forming research groups aligned with its needs.

In contrast, our work focuses on a more direct experimental application: exploring strategies for recommending suitable academic events for paper submissions. Similar to Rumadi et al. (2023), our approach also evolved to incorporate BERTopic, but with specific adaptations for the recommendation context. While the original study used a standard BERT model, our experiments initially employed BERTimbau (specialized in Portuguese) before migrating to a more efficient multilingual model. Additionally, we implemented preprocessing enhancements, such as language filtering and the selection of events with a minimum article volume, to improve clustering quality. These refinements enabled the experimental setup to reach 91% accuracy in identifying the most suitable events, based on the top-3 suggestions.

A key difference between the studies lies in the granularity of clustering. While Rumadi et al. (2023) aim for broad topics (e.g., "satellite technology"), our experimental approach operates at a more specific level, associating articles with individual academic events. This distinction is also reflected in the evaluation metrics: the compared study prioritizes thematic coherence and topic interpretability, whereas our experiments are assessed based on practical recommendation accuracy. Another relevant contrast is the data scale - the dataset used in our experiments (12,629 articles) is significantly larger than that used in ORPA's mapping, requiring additional strategies to ensure computational efficiency and cluster quality.

Despite these differences, both studies demonstrate the versatility of techniques like BERTopic and transformer-based embeddings in semantically organizing academic publications. Rumadi et al. (2023) illustrate the potential of these tools for strategic analysis in research institutions, while our experimental results suggest their applicability in automated recommendation tasks. The iterative refinement of our approach, with performance gains observed in each version, further emphasizes the importance of adapting the NLP pipeline to the specific problem context, balancing model choices, data processing, and evaluation criteria.

This comparison highlights how similar approaches can be directed toward complementary purposes in academic text analysis. To ensure the reproducibility of our

experiments, all code and data used in this work will be made publicly available after the review process[2].

5. Conclusion

This paper presents a series of experiments for recommending academic events of the Brazilian Computer Society (SBC), based on Natural Language Processing (NLP) and Machine Learning (ML) techniques. Three experimental configurations were tested, each incorporating technical improvements and dataset refinements, which resulted in performance gains (as shown in Table 1), particularly in recommendation accuracy measured by Top-3 Accuracy.

The first version used the BERTimbau model with the K-means algorithm, validating the viability of the approach even with textual pre-processing and the dataset in the Portuguese language, achieving 69% accuracy. The second version, in turn, restructured the procedures, adopting BERTopic, using the MiniLM model, reducing dimensionality via UMAP, improving the pre-processing pipeline and refining the dataset. In addition, the use of a multilingual model helped raise the Top-3 accuracy to 90%

Finally, the third version applied more linguistic filtering, removing texts with undefined languages, which resulted in a slight improvement (91%) compared to the previous model, and consistency in the semantic groupings, as shown in images 1, 2, 3 and 4. Therefore, the complementary analyses - such as UMAP visualizations, distribution of cluster sizes and purity assessment - reinforced the thematic coherence of the topics generated and identified both broad lines of research and specific niches, also pointing out opportunities for refinement.

The comparison with the study by Rumadi et al. (2023) showed that approaches based on embeddings and BERTopic are versatile and effective in both strategic mapping and automated recommendation contexts. While the study prioritizes thematic coherence, the proposed clustering focuses on recommendation accuracy, presenting complementary applications. The limitations found in this article are the reliance on short texts (title and abstract) and the absence of additional metadata, as well as the lack of tests with different models or techniques, which could provide additional results for comparison. Furthermore, more robust experiments, with more practical tools, and the use of other validation metrics with changes in hyperparameters were observed as limitations of this study.

Future work may explore the inclusion of additional keywords, the processing of data in languages beyond Portuguese and English, or the use of datasets from different publication sources. It is also suggested that new experiments be conducted, with validation metrics and hyperparameters different from those presented. This study aims to contribute to the advancement of scientific research in the areas of Data Science, Natural Language Processing, and Intelligent Systems, serving as a foundation for further investigations and teaching support materials. As a potential application, the experimental approach presented here may support future development of tools to assist researchers and event organizers in their activities.

[2] https://github.com/juaodantas/semantic-event-matcher

Acknowledgments

This study was partly funded by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financial Code 001. The authors also acknowledge the support of the Mato Grosso State Research Support Foundation (FAPEMAT), through a postgraduate scholarship, and the Federal University of Mato Grosso for its financial support.

References

Abinaya, S., Indira, K., Karthiga, S., & Rajasenbagam, T. Time cluster personalized ranking recommender system in multi-cloud. Mathematics, 11(6), 1300, 2023.

Asabere, N. Y., et al. Improving smart conference participation through socially-aware recommendations. Applied Sciences, v. 10, n. 24, p. 8823, 2020. DOI: 10.3390/app10248823.

Araújo, J. M. O.; Costa, M. A.; Lima, R. S. A importância do artigo científico na vida acadêmica. Criar Educação, Criciúma, v. 10, n. 1, p. 64–76, 2021.

Božić, M.; Grljević, O. Topic Modeling and Sentiment Analysis for Enhanced Personalization in Recommendation Systems. International Journal of Applied Artificial Intelligence, v. 38, n. 2, p. 245–263, 2024.

Bornmann, L.; Haunschild, R.; Mutz, R. Growth rates of modern science: A latent piecewise growth curve approach. Humanities and Social Sciences Communications, v. 8, n. 224, p. 1–11, 2021. DOI: 10.1057/s41599-021-00903-w.

Ding, J., Qiao, Y. & Zhang, L. Plant disease prescription recommendation based on electronic medical records and sentence embedding retrieval. Plant Methods 19, 91 (2023). https://doi.org/10.1186/s13007-023-01070-6.

Dhinakaran, D., et al. Recommendation System for Research Studies Based on GCR. In: International Mobile and Embedded Technology Conference (MECON). IEEE, 2022. p. 61–65. DOI: https://doi.org/10.1109/MECON53876.2022.9751920.

Flicke, M., et al. Scholar Inbox: Personalized Paper Recommendations for Scientists. University of Tübingen, Tübingen AI Center, 2024.

Iyer, A.; Rao, V. N.; Mangaraj, M. A Correspondence Analysis Framework for Author Conference Recommendations. International Journal of Information Technology, v. 12, p. 1141–1147, 2020.

Jesus, J. M.; Oliveira, R. S.; Santos, F. D. A relevância dos eventos acadêmicos para a formação discente. *Revista Exitus*, v. 10, n. 1, p. 158–179, 2020.

Juliani, J. P. and Donha, R. D. G. Sistemas de Recomendação de Artigos Científicos: Integrando o Moodle com uma Base de Dados de Acesso Aberto. EaD em Foco, v. 13, n. 1, e2027, 2023. DOI: https://doi.org/10.18264/eadf.v13i1.2027.

Ouni, S.; Fkih, F.; Omri, M. N. A survey of machine learning-based author profiling from texts analysis in social networks. Multimedia Tools and Applications, v. 82, p. 36653–36686, 2023. DOI: https://doi.org/10.1007/s11042-023-14711-8.

Rizvi, S. T. R.; Ahmed, S.; Dengel, A. ACE 2.0: A Comprehensive tool for automatic extraction, analysis, and digital profiling of the researchers in Scientific

Communities. Social Network Analysis and Mining, v. 13, n. 81, 2023. DOI: https://doi.org/10.1007/s13278-023-01085-w.

Rumadi; Yuniarno, E. M.; Rachmadi, R. F. Research Mapping in Research Organizations Based on Abstracts of Scientific Publications using BERTopic Modeling. In: *Proceedings of the 7th International Conference on New Media Studies (CONMEDIA 2023)*, Bali, Indonesia, 6-8 Dec. 2023. IEEE, 2023. DOI: 10.1109/CONMEDIA60526.2023.10428435.

Sociedade Brasileira de Computação (SBC). Relatório de Atividades da Diretoria 2022–2023. Porto Alegre: SBC, 2023. Disponível em: https://www.sbc.org.br/images/arquivos/arquivosDestaques/Relatorio_SBC_2022-2023.pdf. Acesso em: 16 jun. 2025.

Sivakumar, A.; Gunasundari, R. A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining. Department of Computer Science, Karpagam University, Coimbatore, 2017. Disponível em: https://acadpubl.eu/jsi/2017-117-20-22/articles/20/68.pdf.

Wang, R., et al. Exploration on Scientific Research Data-Targeted Intelligent Recommendation System Using Machine Learning Under the Background of Sustainable Development. Frontiers in Psychology, v. 13, 788183, 2022. DOI: https://doi.org/10.3389/fpsyg.2022.788183.

Wu, S. and Dredze, M. Are All Languages Created Equal in Multilingual BERT? In: Proceedings of the 5th Workshop on Representation Learning for NLP. Association for Computational Linguistics, 2020. p. 120–130.