Processing of computer algorithms for traceability identification in scientific research

Processamento de algoritmos computacionais para identificação da rastreabilidade em pesquisas científicas

Procesamiento de algoritmos computacionales para identificación de rastreabilidad en investigaciones científicas

Hilda Carolina de Jesus Rios Fraga 1
Vagner de Oliveira Machado 2
Julian Reina 3
André Lucas Coelho dos Santos 4
Bruno Santos Oliveira 5
Antonio Carlos dos Santos Souza 6

Abstract

Considering that scientific research is an essential part of the development of new knowledge and a multidisciplinary, time-consuming and error-prone task, it must be conducted under verifiable conditions in order to contribute to safe decision-making. The aim is to extract quality information from scientific articles automatically, presenting reliable, traceable and safe knowledge. To this end, this study investigates the perspective of identifying traceability

1 Postgraduate from Programa de Pós-Graduação em Difusão do Conhecimento (PPGDC). Universidade Federal da Bahia (UFBA). Salvador, Bahia, Brazil. E-mail: hildacarolina21@gmail.com Orcid: https://orcid.org/0000-0002-1191-5789

2 Post-doctorate in Knowledge Dissemination from Universidade Federal da Bahia, Instituto Federal de Educação, Ciência e Tecnologia da Bahia, Universidade do Estado da Bahia (UFBA/IFBA/UNEB). Universidade do Estado de Minas Gerais. Frutal, Minas Gerais, Brazil. E-mail: biolmach@gmail.com Orcid: https://orcid.org/0000-0002-2156-6623

3 Graduated in Oil and Gas Engineering. Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA). Salvador, Bahia, Brazil. E-mail: julianreina@gmail.com Orcid: https://orcid.org/0009-0004-5610-9600

4 Graduating in Systems Analysis and Development. Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA). Salvador, Bahia, Brazil. E-mail: andrebr889@gmail.com Orcid: https://orcid.org/0009-0007-7309-8407

5 Graduated in Information Systems. Universidade do Estado da Bahia (UNEB). Salvador, Bahia, Brazil. E-mail: brunosantos19@gmail.com Orcid: https://orcid.org/0009-0005-3008-2523

6 PhD in Knowledge Diffusion from Universidade Federal da Bahia, Instituto Federal de Educação, Ciência e Tecnologia da Bahia, Universidade do Estado da Bahia (UFBA/IFBA/UNEB). Salvador, Bahia, Brazil. E-mail: acsantossouza@gmail.com Orcid: https://orcid.org/0000-0001-6593-3192
and reproducibility patterns, using algorithmic Natural Language Processing methods, to demonstrate the identification of information contained in scientific articles, regardless of the research area. Therefore, in this work, the languages Naive Bayes (NB), Cosine Similarity, Bag of Words (BOW) and Neural Networks (RN) were used for this purpose. As a result, it was possible to identify nine traceability patterns in the articles analyzed and propose an Artificial Intelligence model using algorithms with a minimum accuracy of 70%, demonstrating the traceability and reproducibility of the scientific articles analyzed.

**Keywords:** Reproducibility. Traceability. Natural Language Processing. Research.

**Resumo**

Considerando ser a investigação científica uma parte essencial do desenvolvimento de novos conhecimentos e uma tarefa multidisciplinar, demorada e sujeita a erros, deve ser conduzida em condições verificáveis de modo a contribuir para a tomada de decisões seguras. Objetiva-se extrair, informações de qualidade de artigos científicos de modo automático, apresentando um conhecimento confiável, rastreável e seguro. Para tanto, este estudo investiga a perspectiva de identificação de padrões de rastreabilidade e reproduzibilidade, utilizando métodos algorítmicos de Processamento de Linguagem Natural, para demonstrar a identificação de informações contidas em artigos científicos, independentemente da área de pesquisa. Desse modo, neste trabalho, foram utilizadas as linguagens Naive Bayes (NB), Cosine Similarity, Bag of Words (BOW) e Redes Neurais (RN) para esta finalidade. Como resultado, foi possível identificar nove padrões de rastreabilidade nos artigos analisados e propor um modelo de Inteligência Artificial utilizando algoritmos com precisão mínima de 70%, demonstrando a rastreabilidade e reproduzibilidade dos artigos científicos analisados.


**Resumen**

Considerando que la investigación científica es una parte esencial del desarrollo de nuevos conocimientos y una tarea multidisciplinaria, lenta y propensa a errores, debe realizarse en condiciones verificables para contribuir a una toma de decisiones segura. El objetivo es extraer información de calidad de artículos científicos de forma automática, presentando conocimientos fiables, trazables y seguros. Para ello, este estudio investiga la perspectiva de identificar patrones de trazabilidad y reproducibilidad, utilizando métodos algorítmicos de
Processing of computer algorithms for traceability identification in scientific research

Procesamiento del Lenguaje Natural, para demostrar la identificación de información contenida en artículos científicos, independientemente del área de investigación. Por ello, en este trabajo se utilizaron para este fin los lenguajes Naive Bayes (NB), Cosine Similarity, Bag of Words (BOW) y Neural Networks (RN). Como resultado, fue posible identificar nueve patrones de trazabilidad en los artículos analizados y proponer un modelo de Inteligencia Artificial utilizando algoritmos con una precisión mínima del 70%, demostrando la trazabilidad y reproducibilidad de los artículos científicos analizados.


Introduction

Infectious diseases continually threaten public health, socioeconomic activity and global biosecurity, as demonstrated by Covid-19, which has become the largest epidemic of the 21st century (LI T et al., 2023). Scientific research, a fundamental part of the chain of development of new knowledge, must generate scientific work conducted under controlled and verifiable conditions, in order to ensure a solid basis in the dissemination of information and decisions. There is also growing recognition that the quality of research depends on the rigor of methodological practices and the resulting analyzes managed by scientists (LANDIS S et al., 2012) and for some researchers, science is based on the possibility of critically evaluating statements and conclusions that are produced, in order to guarantee the reproducibility of the investigation. Data traceability is essential to confirm or correct discoveries in the scientific context and for the transparency of the methods used in research (WANG J et al., 2021). However, for this to be possible, the evidence must be validated by criteria that guarantee the verification, improvement and reproducibility of empirical results. Controlling unstructured data is enabled by natural language processing (NLP) and allows for in-depth analytical tracking of structured data (AYDINOGLU AT & KUSHCHU I, 2022). In this article, our contribution mainly includes presenting an Artificial Intelligence model with algorithms aimed at recognizing traceability and reproducibility patterns in scientific articles. Although our study of analytical recognition of scientific traceability patterns was analyzed in scientific articles in the area of COVID-19, the methodology can be used to monitor research in different areas of knowledge, particularly in health, generating a significant impact.
on the validation of knowledge in the scientific community and society.

Theoretical Framework

2.1 Artificial Intelligence at the Service of Epidemiological Investigation

Following the evolution of Artificial Intelligence (AI) in different areas, its presence in human processes has been noted, with its services proving to be effective in relation to health in different aspects, and could advance in the future, to “medicine of precision” (JEAN, A, 2020). In the field of medicine, the use of AI tools has impacted clinical research, improving exams, hospital practices and ensuring medical conduct that contributes to improving the diagnosis and prognosis of diseases as a whole. However, understanding what AI is, what it is capable of doing and especially what it will never do is the safe way to anticipate possibilities and especially prepare for future challenges. Today, activities such as digitalization and data privacy, the applicability of algorithms, the inclusive development of AI systems and their reproducibility are still considered challenges, topics that still require an in-depth approach so that trust in technology is reinforced (JOHRI et al., 2021).

Natural language, in addition to being constantly changing, is made up of a set of rules and definitions that make broad and general understanding difficult. The function of Natural Language Processing (NLP), a technology considered a field of AI, is to create algorithms that can make sense of language. To make this possible, NLP, which analyzes large volumes of textual data in a consistent and unbiased way and organizes unstructured data in an understandable way at a syntactic and semantic level (JOHRI et al., 2021), uses deep learning techniques and large data sets that can, among other things, enable the extraction of information through tasks, solve issues such as text classification, named entity recognition, meaning disambiguation and language modeling (OLIVEIRA, BS et al., 2022).

With the emergence of the COVID-19 pandemic, the potential of PLN to improve public health was realized, assisting in the analysis of texts from different sources, in the early detection and monitoring of infectious diseases, in addition to allowing the adaptation of various processes of business (RAZA S & SCHWARTZ B, 2023). However, even after the pandemic was brought under control, PLN developments have not stopped and the previously simple approaches are constantly evolving. Extracting information for data recognition is very useful for building databases, classifications, service diagnoses, creating sales, analyzing
conflicts and solutions, although there are many challenges in its application, which involves, for example, evaluating the data quality and accuracy (RAZA S & SCHWARTZ B, 2023; JURAFSKY, D.; MARTIN, J. H., 2023).

2.2 The Use of Natural Language Processing Algorithms to Track Scientific Research

According to Jean (2020), algorithms are “the set of processes incorporated in digital tools used to carry out a series of operations with a specific purpose”. They can be differentiated between explicit algorithms, complex systems that contain only conditional structures to perform operations that lead to a result, or implicit algorithms, which describe the logical connections established for the model, defined based on real scenarios. These include statistical learning algorithms, categorization algorithms, and neural network learning algorithms. While categorization algorithms are implicitly organized by a set of data according to the values they take and statistically representative classes, neural network learning algorithms operate on a neural network built on data representing different scenarios for which the resulting decisions are known, which “learn”, from so-called labeled training data (JEAN, A, 2020; QASSIM, 2019).

Within AI, the objective is to allow computers to learn, through data analysis, to perform tasks, whether programmed or not. In the field of Machine Learning (ML), the concentration occurs through algorithms and techniques that allow the construction of systems and that are capable of learning from data and, therefore, improving their performance in a specific task, making predictions or taking decisions. During the pandemic period, several algorithms were used for geographic analysis and the spread of the COVID-19 disease and are currently capable of making diagnoses, using data and extracting resources from images, signals and audio data, among others, although their applicability still remains today being considered a challenging field (JEAN, A, 2020; OLIVEIRA, BS et al., 2022; 11. PAUL, S.G. et al., 2023).

The algorithm learning process can occur from the construction of data that represents different scenarios, with an optimization calculation of the statistical weights of the network nodes that designs the logic for solving the problem for which the algorithm will provide a solution. At this point, the algorithms “learn” from so-called labeled training data and when this happens, it is called supervised (or controlled) learning. In this supervised learning, which has the keywords classification and regression, the system receives the desired input and
output data, which is labeled for classification to provide a learning basis for future data processing. In unsupervised learning, which is based on clustering and association, the training algorithm takes information that is neither classified nor labeled and allows the algorithm to act on this information without guidance based on the similarity between them (JEAN, A, 2020; QASSIM, 2019).

In this article, the choice to use computational algorithms linked to NLP was mainly due to the breadth of the topic – scientific traceability standards – which requires dynamic analyses, with the probability of change with each use, a criterion fully compatible with the use of the chosen techniques. As the presence of different patterns will be evaluated at the same time, the choice of NLP techniques and neural networks used for recognition and classification of traceability and reproducibility patterns was applied satisfactorily, with those that best met the proposed objective of identifying proposed metrics: Naive Bayes (NB), Cosine Similarity, Bag of Words (BOW), ANN using Keras and Recurrent Neural Networks (GRU). Although the analyzes focus on a specific area (COVID-19), as the methodology is used, it is possible that other researchers interested in working with scientific quality management will make adjustments consistent with other diseases.

**Methodology**

The development of the model used to recognize traceability patterns, through the use of computational algorithms, was carried out in three successive stages: identification of initial criteria and validation of patterns, definition of algorithms for pattern recognition and implementation of the model as described in the steps from 3.1 to 3.3 below:

**3.1 Identification of Initial Criteria and Validation dos Padrões**

The identification and validation of traceability standards were carried out by selecting criteria in high-impact journals that could be used for this purpose and presenting them to a scientific community of researchers so that they could validate them by establishing the percentage of importance of each of them. During the validation of the patterns and to determine the percentages of importance of each of them, the basis of the Bayesian network model was used.
3.2 Definition of Algorithms for Pattern Recognition

Once the traceability and reproducibility standards and their individual percentages were validated, the modeling was defined, using NLP and Neural Networks techniques, according to the following steps:

3.2.1 Definition and preparation of the techniques to be used

The techniques Naive Bayes (Gohari K *et al.*, 2023), Cosine Similarity (Chen, W., Sun, Y. & Shi, K, 2023), Bag of Words (BoW) (Kshirsagar S, Falk TH, 2022) and artificial neural networks (Mashraqi AM & Allehyani B, 2022). The preparation of each of the techniques used followed the flows established below:

**Figure 1**

*Modeling of the Naive Bayes (NB) technique*

In the NB technique, the process involves importing the necessary libraries by downloading them through codes, preparing the training data by making sure its characteristics (or independent variables) are in numerical format; Afterwards, the data is divided into training and testing sets using the Sklearn library, with the train_test_split function, to define the percentage of data used for training and testing and separating them accordingly to evaluate the model's performance; At this point, it is necessary to create and triage the NB model by importing the desired classifier, instantiating and adjusting the training data, with the .fit method; After training the data, predictions are made and the model's
performance is evaluated using the established accuracy. The Cosine Similarity technique followed the flow below:

**Figure 2**

*Modeling of the Cosine Similarity (SC) technique*

The SC technique was built by importing the necessary libraries and downloading them through codes; Afterwards, net input vectors were prepared and the sentences were processed, removing words with no semantic value; After calculating the similarity between the vectors, using a specific formula, the results were visualized. The result is evaluated between 0 and 1, where the closer it is to 1, the more similar they are.

**Figure 3**

*Modeling the Bag of Words (BOW) technique*

For the BOW technique, the technical process was done by importing the necessary libraries and preparing the database by creating the desired words; the BOW vector was
created and the database converted into a numeric vector; At this point the entire article has been converted to lowercase to avoid errors; After processing, each word in the database is compared with the article and after comparison, validation is carried out with the database.

The analyzes carried out using the Neural Networks technique, using the keras library, were validated as follows:

Figure 4
*Modeling of the Artificial Neural Networks technique using Keras (RNA-K)*

The technical process of Neural Networks, using the keras library, was modeled by importing the necessary libraries, defining the model architecture; Since keras allows you to create sequential or functional models, the sequential model using the `add()` method was used in the initial tasks; through the compilation method and defined the optimizer, the loss function and the desired metrics were established and the training and test data were prepared, dividing it into two parts, one for training and the other for performance evaluation; training is done through the `fit` module, minimizing losses, where the evaluation occurred through the metrics defined during the compilation of the model, and the test data, previously separated; other predictions were made using new input data.

3.2.2 Technique training

At this stage, after all the techniques had been established, English phrases related to each of the validated standards were used for the training stage. After performing sentence tokenization strategies, morphological and syntactic analysis on the sentences, the models were trained using processing and machine learning technology with the English sentences to
teach the techniques to accurately recognize the validated traceability patterns. At this time, a computational tool and graphical interface was used for software development on the .NET 6.0 platform, using the C# programming language (C-Sharp) and the Javascript language as source code, to assist in data display and validations of users.

3.3 Model Deployment

A sampling of scientific articles in the area of COVID-19 was selected from the SCOPUS database from January 2020 to December 2023, and registered in a virtual system so that the analysis could be carried out. The choice to use the Scopus database to collect scientific articles is justified because it allows a comprehensive view of global research production in different areas and because it is one of the largest databases of citations and abstracts of peer-reviewed literature (Figueiredo A et al., 2017). At this time, the four chosen PLN techniques were used through a computational tool to identify previously validated traceability patterns and proved capable of identifying whether the analyzed articles have traceability patterns in their writing and providing the percentage of the same for each article consulted.

3.4 Characterization of the Research and Sample Data

This research, characterized as applied, generated a practical solution with the provision of a PLN model, with a qualitative-quantitative approach and exploratory objective, which allowed the interpretation of phenomena and construction of hypotheses, based on the bibliographical survey on the topic and interviews with researchers who have practical experience on the subject. As this is an exploratory study, the objective is not to provide a definitive answer, but rather to present proposals for improvement and discussion. Because it involves human beings and meets ethical aspects, the research project was presented to the IFBA Research Ethics Committee (CEP IFBA), where it was assessed and approved by reasoned opinion nº. 079685/2022.

The sample population was made up of scientific journals in the identification and validation stage of traceability standards, by a scientific community of researchers, since it was necessary to apply a questionnaire to validate the initially proposed criteria, transforming them into traceability standards and establish a percentage of importance for them, to be used
by PLN and scientific articles in the area of COVID-19, which were used to test and identify previously validated traceability standards. These were collected in the CAPES - SCOPUS database from January 2020 to December 2023 and belonged to the COVID-19 area.

**Results and Discussions**

The main objective of this article was to present an Artificial Intelligence model with algorithms aimed at recognizing traceability and reproducibility patterns in scientific productions. Concern with the topic arose from curiosity in verifying the conditions under which scientific articles are produced combined with the considerable increase in the number of scientific articles published and made available to society during the pandemic. Considering the preparation of articles, the first stage of a fundamental process, which is the construction of scientific knowledge, quality factors such as those studied can influence the adoption of important scientific evidence. Furthermore, it is beneficial to recognize the increasing use of AI tools in various fields of research, which has reinforced the desire to align the two topics.

As scientific traceability demonstrates the mapping of the entire path that must be followed during the processing of a research (SILVA AR da, 2017) and reproducibility retraces the path taken by a product, during and after its production phase, so that the history is preserved (RAMALHO TS *et al.*, 2020), the use of these quality standards can be essential for a discovery to occur, as science is a cumulative process that depends on the repeated investigation of the same questions from different angles. In the health area, analyzes such as those presented can help identify and disseminate the knowledge generated and be used to support decision-making, which will be influenced by scientific evidence of new knowledge (NSF, 2021; SUMMER J *et al.*, 2020).

After analyzing 1,770 scientific articles using the proposed model, some results were fundamental to conclude that the four NLP and Neural Network techniques used for this purpose proved to be completely effective in analyzing the identification of patterns and thus being able to build an AI model capable of recognizing them in scientific analysis.
4.1 Standards Used

The validated traceability and reproducibility standards are described in the figure below. The validation of these standards by the scientific community was necessary because, to date, there is no information on clear criteria that can be considered traceability and reproducibility standards. More recent data on the topic COVID-19 (SUMMER J et al., 2020; PLOS, 2021) that establish their own conditioning criteria for publication, partially presented the identified parameters, corroborating the scientific evidence of the chosen standards and contributing to implementing the choice of standards proposed in this work, which can serve as a global research protocol.

**Figure 5**

*Standards were validated by the scientific community*

The standard referring to the description of the methodology was subdivided into 1. materials, 2. method and 3. artifacts or collection instruments, as well as the shared data standard was subdivided into 4. which data were used and 5. where the data can be found. Both the description of the methodology and the referencing of data are considered essential for the rapid dissemination of studies, new knowledge and new research, which can improve health and contribute to accelerating new scientific discoveries, in addition to allowing data from previous research to be used and validated for future research (ARAGÃO, 2017; NIH, 2023). Standard 6. bibliographic references, contributes to the validation of the research carried out by positioning readers in the existing context on the topic and assists in understanding the topics covered, allowing the reader to increase their knowledge of the
subject covered (NATURE, 2023); standard 7. technical terms or technical vocabulary, identifies the main concepts and ideas presented, facilitating the understanding of the work and saving time when reading and helping to determine a common language that communicates clearly and precisely the concepts linked to the theme (FINATTO, 2017); standard 8. software or tools, refers to computer programs used in writing the article for various purposes, such as document creation, image editing, project management, communication, that is, that perform specific tasks on a computer or electronic device; results, standard 9, provides validated quality data associated with associated results, materials, data, codes, and protocols (NATURE, 2023).

These results made it possible to find a set of criteria, previously non-existent as combined standards, to be used as a potential reference for writing scientific articles based on the traceability and reproducibility of the data generated.

4.2 Modeling Used

After validating all the patterns in Fig.5, we decided to combine NLP modeling and Neural Networks since they allow, through their different techniques, to identify different texts, select patterns of interest, without necessarily knowing the exact terms and the information contained and used. in the search. Table 1 presents the modeling used for each pattern analyzed, considering viable processes to perform data processing tasks, identify texts, perform sentiment analysis, translate and automatically extract information.
Table 1

Defined analysis standards

<table>
<thead>
<tr>
<th>Technique</th>
<th>Pattern(s) analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLN - NATURAL LANGUAGE PROCESSING</td>
<td></td>
</tr>
<tr>
<td>Naive Bayes - NB</td>
<td>Materials (equipment/reagents);</td>
</tr>
<tr>
<td>Cosine Similarity</td>
<td>Presence and access to bibliographic references;</td>
</tr>
<tr>
<td>Bag of Words (BOW)</td>
<td>Artificial neural networks</td>
</tr>
<tr>
<td></td>
<td>Software;</td>
</tr>
<tr>
<td></td>
<td>Artificial neural networks</td>
</tr>
</tbody>
</table>

Based on the relationships available in table 1 and observed analyzes of technical data, the ideal computational model was selected, aiming to identify and select the pattern of interest; a minimum accuracy of 70% for each of the analyzed patterns was established as an acceptance criterion for using the model. The performance results of these parameters were demonstrated in Figure 6.

Figure 6

Accuracy of the models

The two patterns that presented the highest and lowest accuracy respectively were “Materials” and “Artifacts”, which were analyzed using the same Naive Bayes technique - NB. The high performance in the classification of the “Materials” standard can be explained
by the clarity of the information in the scientific texts. Sentences that mention this pattern generally contain specific terms and clear contexts, facilitating pattern identification using the Naive Bayes technique, as consistent patterns in the training data contribute to better identification accuracy. On the other hand, the “Artifacts” pattern presents challenges due to its ambiguous nature. The sentences related to this pattern contained vague terms, varied contexts and overlapping characteristics with other categories, which makes it difficult to accurately classify this pattern, making the task more complex for the model. The analyzes referring to the “Software” and “Technical Terms” standards are not present in the graph because they were analyzed using the BOW technique, which is a comparison model, which goes through the text word by word until finding the word of interest. The techniques featured dynamic analysis functionality, allowing percentages, currently used as a reference to find values in article analysis, to be updated as new patterns are included. This versatility could allow for broad scientific analysis beyond health issues related to COVID-19. Table 2 shows comparative analyzes between the models used.

### Table 2

*Comparison between the models used*

<table>
<thead>
<tr>
<th>MODELS</th>
<th>Naive Bayes - NB</th>
<th>Cosine Similarity</th>
<th>Bag of Words (BOW)</th>
<th>Words Artificial neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADVANTAGE</td>
<td>Similarity to human learning</td>
<td>Simple implementation</td>
<td>Simplicity</td>
<td>More accurate</td>
</tr>
<tr>
<td></td>
<td>Requires processing time</td>
<td>less Possibility adjustment</td>
<td>of Ease of implementation and making improvements better</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relatively fast training when compared to Semantic analysis neural networks</td>
<td>Processing speed</td>
<td>Able to recognize small features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Less accurate than Artificial neural networks</td>
<td>Does not analyze</td>
<td>Requires more time for training</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Requires a very large Difficult to understand database for training</td>
<td>May present results</td>
<td>Requires more false processing time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Typing errors with the more complex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Implementation is</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Revista de Gestão e Secretariado – GeSec, V. 15, N. 7, P. 01-22, 2024
São José dos Pinhais, Paraná, Brasil.
Table 2 demonstrates that, in addition to the accuracy data, comparative analyzes were carried out between the models used, to assist in analyzing the feasibility of using the models. However, these only present qualitative information that could support the choice of methods in future analyzes and were not used at this time with parameters for inclusion and/or exclusion of the techniques used.

1,770 scientific articles were registered and analyzed from January 2020 to December 2023, which were randomly collected in the Scopus database. In Fig.7 it is possible to see that the number of articles analyzed per year was not exact and uniform, since the objective was to collect them randomly.

**Figure 7**

*Number of articles analyzed by PRAC per year*

![Bar chart showing the number of articles analyzed by PRAC per year from 2020 to 2023.]

Despite not having determined the exact number of articles per year, uniformity was sought in the analysis of their average, as a minimum of 30 and a maximum of 45 articles were recorded in each of the journals, data shown in Fig. 8.
Processing of computer algorithms for traceability identification in scientific research

Figure 8

Average number of articles analyzed

The average RR of the journals was between 85% and 99%, while the number of articles analyzed was between 30 and 45. The techniques presented a dynamic analysis functionality, allowing the percentages, currently used as a reference to find values in article analysis, to be updated as new patterns are included. Such versatility could allow for broad scientific analysis beyond health issues related to COVID-19.

We did not specify desirable values for this measure, since the objective is not to categorize them as good or bad, but only identify the individual values relating to traceability and scientific reproducibility that were found. However, one of PRAC’s features is to predict the individual average RR value per journal and the average RR value of all journals that were analyzed, as represented in Fig. 9.
Figure 9

Journals with the highest percentage averages

The ten journals that had the highest averages are in descending order: Nature (99%), Science Advances (98.7%), Critical Care (98.4%), BMC Medicine (98.3%), Journal of Infectious Diseases (97.8%), Plos One (97.6%), Trials (96.5%), BMC Health Services Research (96.4%), and Frontiers in Immunology (96.3%). The analyzes carried out demonstrated that the use of NLP techniques used as a traceability and reproducibility analysis tool was effective and contributed to generating the discussion that science needs evaluations that prioritize criteria linked to scientific quality. This fact implies the general interest of researchers and society, since the knowledge produced through science is of collective interest. From the conceptual elaboration of the topic to the mechanism for making the data available, this work presented unprecedented results, whether in the absence of the approach in the literature or in the evaluation of the information made available in published scientific articles. The approach based on the combination of an artificial intelligence tool, with the use of NLP and Neural Networks techniques, demonstrated applicability and versatility in use by different research segments and by different scientific journals, as criteria for evaluating their publications.

Although the results were satisfactory and the AI model fulfilled its intended objective - identifying traceability and reproducibility patterns - in scientific articles, the analyzes demonstrated that both the performance of the tool and the NLP techniques used have room for improvement and that they still There are important challenges to be overcome to achieve a comprehensive understanding of natural language and a more efficient use of its techniques.
and analyzes as a whole. One of the opportunities for improvement concerns the way in which the techniques were validated, as each standard required different parameters for their identification and, therefore, it was necessary to use different models to obtain more reliable results during the analysis of the articles, an understandable limitation when it comes to identification of nine patterns at the same time, which requires the tool to be more comprehensive in the techniques used. Another point of attention to be considered concerns the need to use many phrases to validate each of the techniques used and their meanings, for the correct identification of them in the articles analyzed; This can be considered a restrictive condition, as it requires significant time and careful individual assessment. In this work, however, we believe we have overcome this point with the creation of the dataset based on peer-reviewed scientific articles, to guarantee precision and assertiveness, but which required detailed work in organizing the information. Still considering the opportunities for improvement, both the use of neural networks, with their effectiveness in various tasks, and traditional algorithms such as Naive Bayes with their simplicity, BOW and Cosine Similarity, played important roles in solving problems. Through PLN it was possible to prove the effectiveness of the techniques used for the proposed purpose given the complexity of the data involved since it was not just about identifying words in an article but about initially validating the meaning of these words, facilitating the understanding of the sentences in the article. text presented.

Conclusion

This work was able to develop and present an Artificial Intelligence model applied to scientific research, based on computational modeling, to:

a) provide four compatible NLP techniques for identifying traceability and reproducibility standards in scientific articles, regardless of the research area;

b) quantify the individual percentage of traceability and reproducibility in scientific articles and additionally;

c) present a virtual database system that can be used to verify the traceability percentage determined by the PLN tool in scientific articles.
Processing of computer algorithms for traceability identification in scientific research

References


Received: 06.07.2024
Accepted: 06.28.2024