Studies in English Language Teaching
ISSN 2372-9740 (Print) ISSN 2329-311X (Online)
Vol. 13, No. 1, 2025
www.scholink.org/ojs/index.php/selt

Original Paper

GapFinder: Exploring Research Gaps with Artificial Intelligence

Vilker Zucolotto Pessin[1], Kyria Rebeca Finardi[2] & Celso Alberto Saibel Santos[3]

[1,2] Post Graduate Program of Education, Federal University of Espírito Santo, Vitória, Brazil

[3] Post Graduate Program of Computer Science, Federal University of Espírito Santo, Vitória, Brazil

Correspondence: Vilker Zucolotto Pessin, Post Graduate Program of Education, Federal University of Espírito Santo, Av. Fernando Ferrari, 514. Vitória, Espírito Santo, CEP 29075-910, Brazil. Tel: +55 27998651194. E-mail: vilker.pessin@gmail.com

Abstract

"GapFinder" is an Artificial Intelligence (AI) solution developed to address an important challenge in the field of scientific/academic research, namely: the efficient mapping of research gaps based on existing scientific/academic literature. The exponential increase in all areas of academic publications stresses the relevance of this technological tool to help researchers identify gaps that are still unexplored in academic texts because they are undetected by traditional science mapping techniques to reveal fields for research and innovation. GapFinder applies Natural Language Processing (NLP) algorithms to identify potential research gaps in scientific documents using mining and extracting information techniques from unstructured texts in Portable Document Format (PDF). This AI solution fosters a more comprehensive understanding of the existing literature, emphasizing areas that need further investigation. The present study describes the development of GapFinder, from the conception of the idea to the practical implementation and its availability for access. The methodology used to process and analyze scientific documents in PDF format is described in the paper and followed by a simulation of GapFinder to demonstrate how it can facilitate the work of researchers from the perspective of corpus processing. The study concludes with the importance of innovation in scientific research through the implementation of technologies and methods that can act as catalysts for innovation in science, from the perspective of identifying gaps and new research strategies and also points out some implications for English language teachers and researchers.

Keywords

artificial intelligence, natural language processing, science mapping, language

1. Introduction

Artificial Intelligence (AI) establishes a new technological standard (Dwivedi et al., 2021), with significant impacts on automating processes that simulate and mimic human intelligence. Through machine learning and deep learning algorithms (Soori et al., 2023; Wu & Liao, 2023), various aspects of human behavior such as language, decision-making, and the ability to adapt and learn over time can be emulated by machines.

Stuart Russell and Peter Norvig (2020) propose four fundamental approaches to base AI: "acting humanly," "thinking humanly," "acting rationally," and "thinking rationally." Acting humanly means machines emulate human behaviors, such as expressing emotions and communicating, while thinking humanly involves duplicating the cognitive processes, like reasoning and learning that form the basis of human behavior. Acting rationally entails carrying out actions to secure the best possible outcome, guided by rational principles. Finally, thinking rationally involves applying logical reasoning to model thought processes (Pietikäinen et al., 2021).

These notions lay the groundwork for Natural Language Processing (NLP), an essential aspect of artificial intelligence that connects the intricate and nuanced nature of human communication requiring machines to grasp the essence of natural language to execute intelligent tasks. These tasks encompass speech recognition and synthesis, processing of written language, and even the generation of text, illustrating the critical role that NLP plays in bridging human and machine communication (Russell & Norvig, 2020).

The evolution of AI has been facilitated by the systematic advancement of some technologies (Ofosu-Ampong, 2024), including big data (Allam & Dhunny, 2019; Duan et al., 2019), advancements in machine learning algorithms and models (Soori et al., 2023; Wu & Liao, 2023), increased computational power and reduced costs through cloud storage (Amron et al., 2017), software frameworks and open-source libraries (Gudivada & Arbabifard, 2018). Application Programming Interfaces (API) (Xu et al. 2021) have been used within a global landscape of development and knowledge exchange (Pessin et al., 2022, 2023), creating a fertile environment for innovations.

The volume of publications on AI has significantly increased across various fields, including Education and research, leading to the emergence of current trends and vital discussions on issues that range from science mapping , systematic review of scientific literature (de la Torre-López et al., 2023), AI assistants, forecasting, intelligent tutoring systems to assessment, to name but a few (Crompton & Burke, 2023; Duan et al., 2019).

Traditional science mapping techniques often fall short in efficiently exploring unstructured texts, which are rich sources of information yet complex to analyze due to their varied formats and lack of standardization (Aria & Cuccurullo, 2017; Donthu et al., 2021; Li et al., 2021; Pessin et al., 2022).

A particularly noteworthy challenge in the evolution of science is the identification of research gaps (Chand, 2023; Chen et al., 2020). Often referred to as grey areas of knowledge, these gaps can remain opaque, hidden, or lost in vast amounts of unstructured published data (Abram et al., 2020). Depending on the origin, language, or channels that scientific output is published, it can also be invisible (Finardi et al., 2022, 2023).

A research gap can be seen as an excellent opportunity for scholars and teachers who incorporate research in their teaching. It represents an area where questions have not yet been asked or adequately answered, where curiosity and investigation can thrive. By identifying these gaps, researchers can significantly contribute to their fields of study, breaking new ground, expanding frontiers, and generating innovation.

The identification of gaps has direct effects and practical implications. For instance, in the global scientific context, efficiently discovering gaps can prevent duplicating efforts, directing resources and energy towards relevant and under-explored issues (Chand, 2023). In applied research, it can lead to the development of innovative technologies, solutions to pressing problems, and advances in various fields.

However, identifying research gaps is not a trivial task. Beyond considering the vast range of scientific publications in PDF format, gap identification requires not just a comprehensive understanding of the current state-of-the-art (Donthu et al., 2021; Hinojo-Lucena et al., 2019; Pessin et al., 2023), but also the ability to detect still relevant and unresolved issues and limitations.

The widespread availability of scientific literature in PDF format creates fertile ground for developing innovative tools in research. Using NLP techniques, valuable insights can be systematically and accurately extracted from scientific texts, such as research gaps.

Thus, tools like GapFinder emerge as significant resources to support researchers' work identifying unexplored areas of knowledge. By automating the detection of gaps in existing literature (de la Torre-López et al., 2023), GapFinder not only aids the literature review process but also provides meaningful insights to drive scientific innovation and discovery.

2. Method

The methodology for applying GapFinder in the process of identifying research gaps is structured in several stages, each playing a crucial role in the analysis and interpretation of scientific data.

Data Collection - The first stage consists of selecting recognized scientific databases acknowledged by the academic community. These platforms are supposed to be chosen for their quality and comprehensiveness, providing access to a wide range of relevant publications in study fields. The search for articles should be conducted using pertinent keywords related to the research topic, allowing for the capture of significant texts associated with the work to be undertaken.

Science Mapping - After collecting the data, a scientific mapping analysis is conducted using tools such as SMART Bibliometrics. The researcher should carry out this analysis to identify trends, prominent authors, and research opportunities, offering a comprehensive view of the state of the art in the areas of interest. At this stage, the researcher can select articles based on thematic relevance and scientific robustness, while also considering metrics such as citation frequency and publication impact factor.

Processing PDF Documents - With the key articles selected, the researcher may utilize GapFinder to identify potential research gaps within the PDF texts. GapFinder extracts information from unstructured texts, allowing the researcher to quickly identify limitations and underexplored areas in text paragraphs. After extracting the relevant paragraphs, GapFinder applies the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis technique to classify the emotional tone of the texts (Almansor et al., 2021). This analysis assigns scores that range from negative to positive, helping researchers discern between gaps that require immediate attention and areas that may only need refinement. Generally, negative sentiments are supposed to be associated with unresolved challenges or gaps in research.

Interpretation and Contextualization - The interpretation of results is a critical step where the researcher must contextualize the identified gaps considering the sentiment analysis. The researcher’s experience is essential, as they must apply a critical perspective to the extracted data, identifying nuances and contexts that may not be evident from automated analysis alone. This combination of insights generated by GapFinder and the researcher’s expertise facilitates the identification of answers to the gaps, promoting a more directed and productive investigation. It enables the detection of emerging trends and critical gaps in the knowledge base, which traditional tools might overlook. By improving these techniques, GapFinder helps to propel scientific discovery forward, offering researchers powerful insights and guiding more focused and productive research initiatives, bridging the gap between science and innovation (Sanabria-Z et al., 2024).

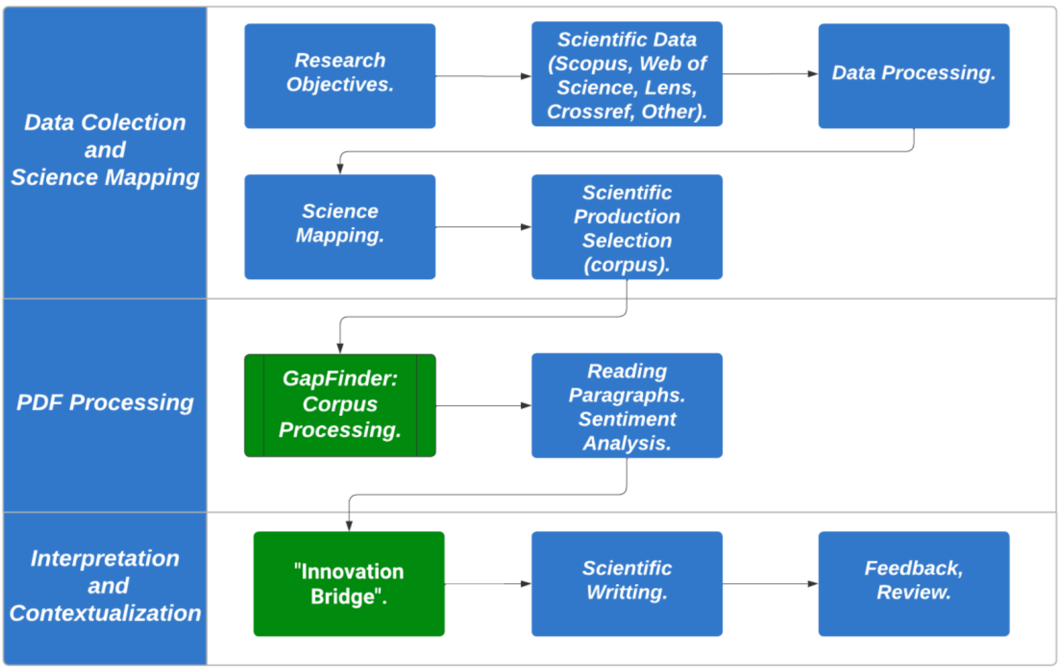

Figure 1. Process Flow through Scientific Production Using GapFinder

3. Result

This Initially, recognized academic databases such as Scopus, Web of Science, Lens and CrossRef were selected. These platforms were chosen for their comprehensive coverage and data quality, as they are endorsed by the scientific community to offer a wide spectrum of scientific publications from various disciplines.

The search for scientific papers was conducted using the keywords "artificial intelligence," "natural language processing," "education," and "research." This selection and combination of terms aimed to capture the intersection between AI, NLP, and their application in the fields of education and research, which are considered fundamental for the developmental objectives of GapFinder.

From the data gathered in the mentioned academic databases, a detailed scientific mapping analysis was conducted using SMART Bibliometrics (InsightSci, 2024) to reveal the state of the art and identify trends, gaps, and research opportunities. This scientific mapping facilitated an understanding of the current dynamics at the intersection of AI, NLP, and education and research, highlighting important documents for the future development of GapFinder and its practical application.

Particular attention was given to papers discussing the development of technological solutions in AI and NLP within the educational context, focusing on the challenges associated with scientific research in this domain. The selection of papers was meticulously carried out, considering not only thematic relevance and scientific robustness but also the alignment of studies with the specific goals and inherent challenges of GapFinder.

For the evaluation process of GapFinder, 10 papers were randomly selected from this research's corpus, all recently published, to assess the processing capabilities. These papers were specifically chosen because their titles include the terms "Artificial Intelligence" or "Natural Language Processing," ensuring that they are relevant to the system's focus areas.

During the GapFinder evaluation, the papers were uploaded to the system, which is then processed to generate a report (Fig. 2). This report consolidated data such as the PDF file name, DOI, author, title and keywords. Additionally, it processed the corpus by extracting the page number and paragraphs that contained indications of research gaps, and it created an 'Insight' column using AI technology.

The processing of these papers resulted in 40 paragraphs indicating possible gaps and limitations in the selected corpus analyzed. These paragraphs highlight numerous questions and themes that the scientific community has not yet fully explored, providing a fertile basis for future research and scholarly discussions. This identification provided a comprehensive view of the less explored areas or those presenting persistent challenges within the theme of "Artificial Intelligence".

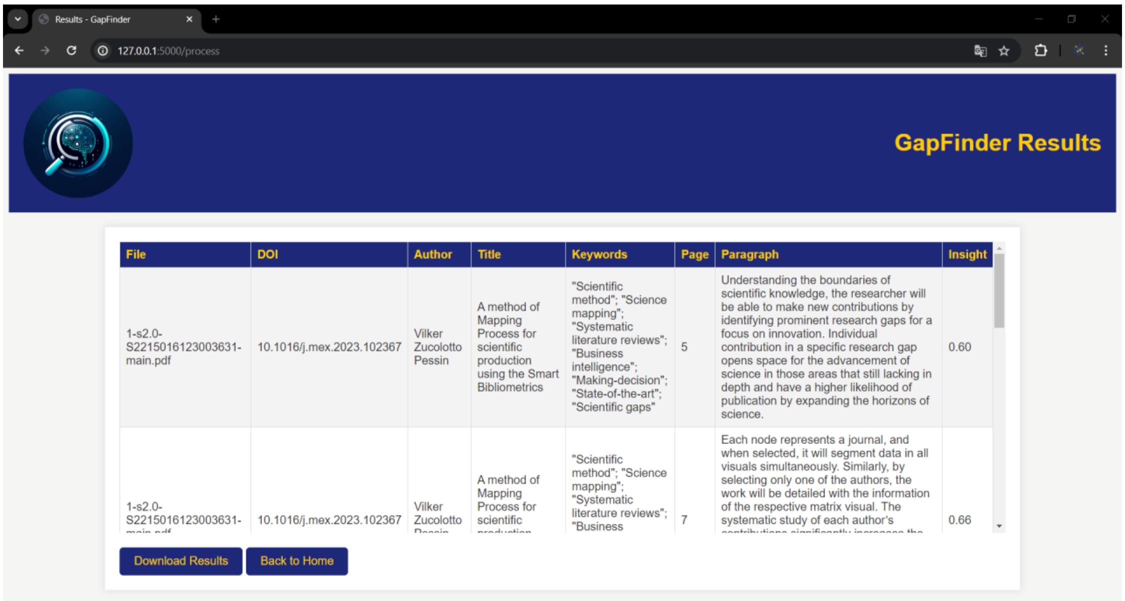

Figure 2. Data Processed by GapFinder

For this purpose, the analytical methodology involved a quantitative document processing analysis and a qualitative analysis (Abram et al., 2020) by employing a word cloud generation technique. This approach revealed not only the predominant themes and recurring concerns among the studies but also highlighted emerging trends and specific topics requiring further investigation.

The analysis of the data processed by GapFinder in a set of 10 documents shows significant insights considering the sample size. Among these scientific papers related to the theme of “Artificial Intelligence,” 8 contained at least one indication of a research gap in the corpus, showing that the tool was able to identify at least one opportunity for future research in 80% of the papers analyzed.

Additionally, a horizontal bar graph (see Fig. 2) was created from the data contained in the "Insights" column, which categorizes sentiments as strongly negative (-1.0 to -0.5), mildly negative (-0.5 to -0.1), neutral (-0.1 to +0.1), mildly positive (+0.1 to +0.5), and strongly positive (+0.5 to +1.0).

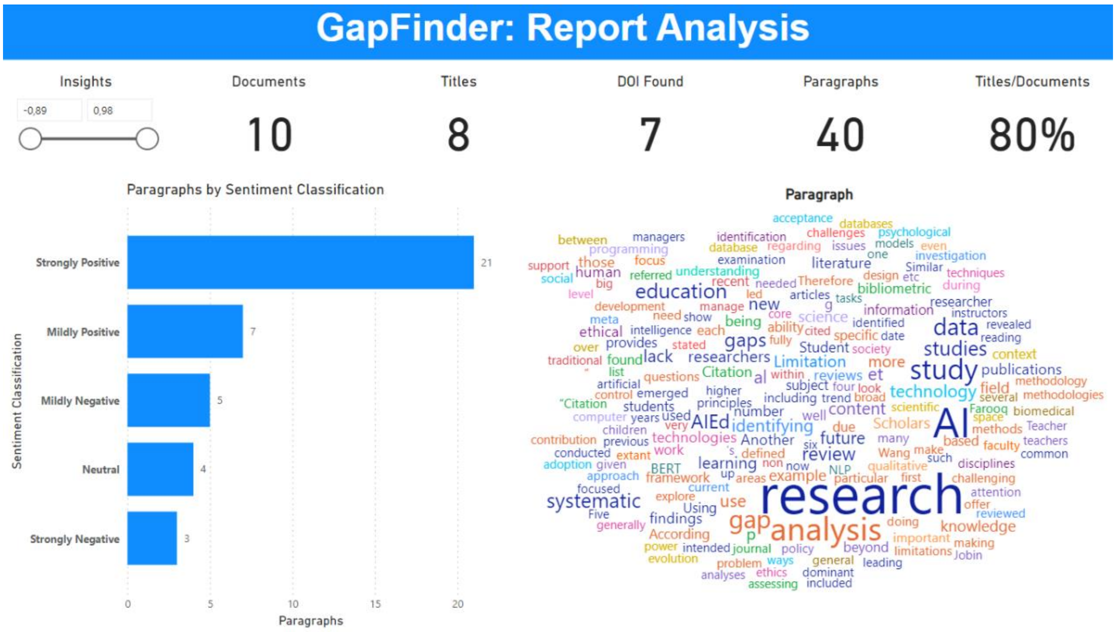

Figure 3. Analysis of Data Exported from GapFinder

Regarding the metadata shown in Fig. 3, the DOI information was successfully extracted from 7 of the processed documents. The success rate of metadata extraction depends on its availability in the PDF files, as some information may not be extractable if absent from the document.

The identification of a high percentage of titles with research gaps underscores the need for further investigation in these areas, potentially guiding the direction of future academic work and contributing to the advancement of scientific knowledge.

Beyond the results from document processing, a word cloud (Fig. 3) was generated to explore relationships and trends within the analyzed corpus. The word cloud, derived from the paragraphs extracted, reveals key themes and areas of interest. Notably, the keywords extracted, such as "AI", "research", "Gap", and "data", “future” pointing to significant and emerging areas of interest that represent fertile grounds for exploring new research frontiers. The word cloud effectively highlights these focus areas within the analyzed dataset, providing a valuable tool for outlining the themes addressed.

In conclusion, the outcomes of this study using GapFinder not only demonstrate its potential as a research tool but also emphasize the importance of continuing to explore and question starting by a gap.

4. Discussion

The application of NLP to PDF texts represents a significant advancement in the ability to identify, extract, analyze, and interpret data from a wide range of digital scientific documents (Olivetti et al., 2020). PDF documents, widely used in scientific contexts, often contain valuable information that, until recently, remained inaccessible for automated analysis due to technical limitations associated with processing text in this specific format. However, advancements in NLP algorithms have overcome these challenges, opening new avenues for efficient extraction of unstructured data (Pichiyan et al., 2023).

Python programming language (Python Software Foundation, 2023) facilitates NLP transformations with its simplicity and effectiveness, along with an extensive library of resources dedicated to NLP and PDF file manipulation. Python stands out in the NLP community due to its vast array of open-source libraries, such as Natural Language Toolkit (NLTK) (NLTK, 2023), SpaCy (Explosion, 2015), and PyPDF2 (Fenniak, 2023), among others, offering robust functionalities for text extraction, semantic analysis, and language processing. These libraries not only simplify the implementation of complex NLP algorithms but have also become a popular choice among researchers and developers at the forefront of textual data extraction.

Moreover, applying NLP to analyze PDF texts paves the way for significant advancements in systematic literature review and delving into the state of the art in various knowledge fields (Jahan & Oussalah, 2023; Moon et al., 2022), proving the ability to quickly identify relevant information in a large corpus of documents. This is particularly relevant in fast-evolving fields, where updating literature reviews and rapidly identifying research gaps can accelerate scientific progress.

These capabilities not only improve the efficiency and scope of scientific outcomes but also increase access to text analysis techniques, allowing a broader range of users to leverage information and data available in digital scientific documents. As NLP techniques continue to evolve, their application is expected to widen further, facilitating automation for productivity gains in areas requiring analysis of large textual volumes.

The development of the MVP (Minimum Viable Product) for GapFinder included a combination of advanced NLP techniques, the integration with other programming languages, coupled with a robust development and hosting architecture. This integrated approach enabled the delivery of a version that processes PDF texts efficiently, displaying results accessibly and supporting simultaneous user's access.

The GapFinder application was developed using multiple programming languages, each of which served a specific purpose. Python was used for processing PDF documents, extracting text, and integrating with AI algorithms. This choice ensured both deep and comprehensive text analysis. The results are displayed in HTML to ensure accessible presentation enabling users to interact with the system via a web interface.

At the heart of GapFinder is its NLP algorithm. Upon uploading scientific documents, the application is designed to process unstructured text to locate and extract paragraphs indicating grey areas in science or underexplored fields.

The core functionality of the tool is its potential to identify expressions and phrases within the text that typically indicate study limitations, underexplored areas, and gaps in the existing literature. This is achieved through NLP functions that search for specific linguistic patterns and semantic expressions.

After identifying these indicative or suggestive expressions, the application proceeds to extract the paragraphs associated with semantic cues, enabling data structuring for analysis. From each processed document, if any indication of a gap is found, the application extracts the file name, author's name, Digital Object Identifier (DOI), the paragraph containing research gap indicators, the page number where it was located, and generates an “Insight” column containing the processed data from the paragraphs using AI features for sentiment analysis. This extraction allows users not only to get an overview of the limitations found in the research corpus but also direct access to the sources where these gaps were identified.

The 'Insight' column employs the Valence Aware Dictionary and Sentiment Reasoner (VADER) analysis tool to assess the emotional tone of the text. VADER is particularly attuned to the nuances of sentiment in text, providing scores that range from strongly negative to strongly positive. These scores are categorized as follows: strongly negative (-1.0 to -0.5), mildly negative (-0.5 to -0.1), neutral (-0.1 to +0.1), mildly positive (+0.1 to +0.5), and strongly positive (+0.5 to +1.0) (Almansor et al., 2021).

Sentiment analysis can be applied across a variety of research contexts using the VADER technique, offering valuable insights to researchers. This tool is adept at gauging the emotional undertones of textual data, making it a versatile asset in diverse fields such as marketing, political science, social media analytics, and consumer behavior studies. For example, in marketing, VADER can analyze customer reviews to identify underlying sentiments that may affect brand perception. In social media, it can assess public opinion trends from posts. Each application allows researchers to unearth nuanced emotional responses that traditional data analysis might overlook, providing a deeper understanding of the subject matter and informing more strategic decisions in many scenarios (Chiny et al., n.d.; Ibrahim et al., 2022; Koukaras et al., 2022; Mammadova et al., 2023; Nkongolo Wa Nkongolo, 2023).

The application of VADER to paragraphs indicating potential research gaps significantly enhances the precision with which researchers can discern true gaps from areas merely requiring refinement. By assigning sentiment scores that range across a spectrum from strongly negative to strongly positive, VADER effectively reveals the emotional tone embedded in academic texts.

For instance, paragraphs characterized by negative sentiment scores might explicitly point to shortcomings or deficiencies in current research, signaling clear gaps that require immediate academic attention. These could be critical issues that have been consistently overlooked or emerging problems that have yet to be addressed adequately.

Conversely, paragraphs with positive sentiment scores often reflect satisfaction with the current state of research or highlight achievements and areas where significant progress has been made. These insights are crucial as they help differentiate between fields that are well-trodden and those that still hold uncharted territories. In this way, positive sentiments can help identify research areas that might benefit from further enhancement rather than the discovery of new gaps.

Moreover, VADER’s ability to detect neutral sentiment is equally valuable. Neutral scores typically indicate expository or background information that, while not directly highlighting gaps, provides essential context that aids in the deeper understanding of the field. This context is vital for researchers to make informed decisions about where to direct their research efforts.

Here's an important challenge: how can researchers effectively identify research gaps using GapFinder? Through the nuanced application of VADER within GapFinder, researchers can navigate the complex landscape of scientific literature more effectively, enhancing their ability to identify under-researched areas. This targeted approach not only saves time but also ensures that efforts are concentrated on the most promising avenues for impactful research and innovation. The integration of advanced sentiment analysis techniques into GapFinder significantly enhances its capability to identify sentiment in scientific literature.

In practical terms, researchers can find a gap by analyzing the output from GapFinder, which highlights sections of the literature that reveal limitations or unresolved questions. The sentiment scores provided by VADER assist researchers in discerning which gaps require immediate attention and which areas may benefit from further exploration. However, it is crucial for researchers to apply their expertise and experience during this process, as their critical judgment is necessary to interpret the findings effectively. By combining the insights generated by GapFinder with their own knowledge, researchers can more accurately identify and prioritize research gaps, ultimately driving innovation and advancing their fields of study.

4.1 Limitations

An important feature of GapFinder is its adaptable design, which anticipates updates and the incorporation of continuous improvements. One of the planned updates involves integrating new AI algorithms and machine learning resources with large-scale language models through APIs. This evolution will enable the application to further refine its gap identification capabilities using robust models trained on extensive datasets. Additionally, the GapFinder repository is available on GitHub under the MIT license, allowing developers to contribute to the enhancement of the solution. Interested contributors can access the repository at https://github.com/vilkerpessin, fostering collaboration and innovation in the ongoing development of this valuable research tool.

This enhancement will not only improve the accuracy of the tool but also allow for customization and adaptation to researchers' specific needs over time. Notwithstanding these improvements, this implementation will incur associated costs.

A current limitation of the system is the incorporation of specific GPT models tailored for some domains, aiming for increased precision, addressing this will inevitably involve additional costs. Enhancements will include more advanced computing resources and supplementary AI tools to refine these models for greater accuracy and domain-specific applicability. These upgrades are essential for expanding the system's capabilities, but they also mean that budgetary considerations will need to be carefully managed to support the increased expenditure on sophisticated AI technologies.

Another important aspect is the processing capacity of the utilized equipment and available subscriptions. The development of a more robust application is anticipated as new resources are integrated to handle large volumes of unstructured data (Abram et al., 2020), creating specific models from machine learning with specific data sets, including developing specific language models for different fields.

Costs involved and budget constraints remain significant, as the application requires investments to cover hosting costs and subscriptions to keep the application available. Deploying an application in the cloud can introduce certain budgetary constraints, particularly when it comes to functionalities that consume significant resources. For instance, enabling the download of detailed reports or allowing extensive processing by artificial intelligence can incur costs per user, which could be substantial depending on the scale of usage. These activities demand considerable computational power and data transfer, leading to higher operational costs.

As a result, some features, such as downloading comprehensive analytical reports or extensive AI-driven processing, might not be implemented initially due to financial limitations. This decision is made to ensure that the core functionalities of the application remain sustainable and operational within the available budget. However, we are actively seeking additional funding sources and exploring cost-effective technological solutions to overcome these financial barriers.

The future of AI, from the perspective of NLP, also encompasses significant challenges, such as improving the understanding of ambiguous or subtle natural language, reducing bias in language models, enhancing translation, summarization, creative writing, and creating more efficient and explainable systems. Solutions such as VADER can significantly aid in the disambiguation of language when it involves sentiments and other solutions will possibly be created contributing to mitigate some of these challenges.

Considering the potential of Gapfinder and the panorama laid out so far and notwithstanding the challenges and costs involved in using and enhancing this tool, we conclude that the benefits associated with this solution justifies further research and investment. In that sense, we are confident that the present study offers a relevant contribution towards that end.

5. Conclusion

The increasing complexity of academic research necessitates innovative solutions that leverage artificial intelligence to automate the discovery of research gaps. Such tools are essential for enabling researchers to navigate the vast and ever-expanding landscape of scientific literature more effectively. By employing advanced technologies like natural language processing, researchers and teachers can systematically analyze unstructured texts, revealing underexplored areas that traditional methods often overlook.

The importance of automating gap identification lies in its potential to enhance the efficiency and effectiveness of research endeavors. By streamlining the process of literature review, AI-driven solutions allow scholars to focus their efforts on the most pressing questions within their fields. This targeted approach not only minimizes the risk of duplicating existing research but also directs resources toward areas that promise significant contributions to knowledge and innovation. This is especially relevant in the education of future teachers and researchers.

Moreover, the ability to quickly identify and categorize research gaps empowers researchers and teachers to be more strategic in their inquiries, fostering a culture of innovation within academia. As scholars gain insights into the nuances of existing literature, they can better position their work to address critical challenges and explore new frontiers. Ultimately, the integration of AI in research methodologies into academic and teaching activities represents a transformative shift that enhances the capacity for discovery and drives meaningful advancements across various disciplines.

Acknowledgements

Kyria Rebeca Finardi would like to thank Cnpq and Fapes for support for research in the form of PQ scholarship (Cnpq) and 28/2022 (Fapes Universal).

References

Abram, M. D., Mancini, K. T., & Parker, R. D. (2020). Methods to Integrate Natural Language Processing Into Qualitative Research. *International Journal of Qualitative Methods*, *19*. https://doi.org/10.1177/1609406920984608

Allam, Z., & Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities*, *89*(January), 80-91. https://doi.org/10.1016/j.cities.2019.01.032

Almansor, E. H., Hussain, F. K., & Hussain, O. K. (2021). Supervised ensemble sentiment-based framework to measure chatbot quality of services. *Computing*, *103*(3), 491-507. https://doi.org/10.1007/s00607-020-00863-0

Amron, M. T., Ibrahim, R., & Chuprat, S. (2017). A Review on Cloud Computing Acceptance Factors. *Procedia Computer Science*, *124*, 639-646. https://doi.org/10.1016/J.PROCS.2017.12.200

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Chand, R. (2023). Framework for Identifying Research Gaps for Future Academic Research. *IRA International Journal of Education and Multidisciplinary Studies*, *19*(2), 160. https://doi.org/10.21013/jems.v19.n2.p12

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. In *Computers and Education: Artificial Intelligence* (Vol. 1). Elsevier B.V. https://doi.org/10.1016/j.caeai.2020.100002

Chiny, M., Chihab, M., Chihab, Y., & Bencharef, O. (n.d.). LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 7). www.ijacsa.thesai.org

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1).

https://doi.org/10.1186/s41239-023-00392-8

de la Torre-López, J., Ramírez, A., & Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, *105*(10), 2171-2194. https://doi.org/10.1007/s00607-023-01181-x

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*(March), 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data - evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63-71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Explosion. (2015). *spaCy 101: Everything you need to know*. Https://Spacy.Io. https://spacy.io

Fenniak, M. (2023). *Welcome to PyPDF2 — PyPDF2 documentation*. Pypdf2. https://pypdf2.readthedocs.io/en/3.0.0/

Finardi, K., França, C., & Guimarães, F. F. (2022). Ecology of knowledges and languages in Latin American academic production. *Ensaio: Avaliação e Políticas Públicas Em Educação*, *30*(116), 764-787. https://doi.org/10.1590/s0104-40362022003003538

Finardi, K., França, C., & Guimarães, F. F. (2023). Knowledge production and internationalization of Higher Education in the Global South. *Diálogos Latinoamericanos*, *32*(1), 1-19. https://doi.org/10.7146/dl.v32i1.127278

Gudivada, V. N., & Arbabifard, K. (2018). Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP. *Handbook of Statistics*, *38*, 31-50. https://doi.org/10.1016/BS.HOST.2018.07.007

Hinojo-Lucena, F. J., Aznar-Díaz, I., Cáceres-Reche, M. P., & Romero-Rodríguez, J. M. (2019). Artificial intelligence in higher education: A bibliometric study on its impact in the scientific literature. *Education Sciences*, *9*(1). https://doi.org/10.3390/educsci9010051

Ibrahim, A. F., Hassaballah, M., Ali, A. A., Nam, Y., & Ibrahim, I. A. (2022). COVID19 outbreak: A hierarchical framework for user sentiment analysis. *Computers, Materials and Continua*, *70*(2), 2507-2524. https://doi.org/10.32604/cmc.2022.018131

InsightSci. (2024, January 1). *Insights for Science*. Insights for Science. https://insightsci.com/

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. In *Neurocomputing* (Vol. 546). Elsevier B.V.

https://doi.org/10.1016/j.neucom.2023.126232

Koukaras, P., Nousi, C., & Tjortjis, C. (2022). Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. *Telecom*, *3*(2), 358-378. https://doi.org/10.3390/telecom3020019

Li, J., Goerlandt, F., & Reniers, G. (2021). An overview of scientometric mapping for the safety science community: Methods, tools, and framework. *Safety Science*, *134*(November 2020), 105093. https://doi.org/10.1016/j.ssci.2020.105093

Mammadova, M., Jabrayilova, Z., & Shikhaliyeva, N. (2023). DEVELOPMENT OF DECISION-MAKING TECHNIQUE BASED ON SENTIMENT ANALYSIS OF CROWDSOURCING DATA IN MEDICAL SOCIAL MEDIA RESOURCES. *Eastern-European Journal of Enterprise Technologies*, *5*(3(125)), 75-85. https://doi.org/10.15587/1729-4061.2023.289989

Moon, S., Lee, G., & Chi, S. (2022). Automated system for construction specification review using natural language processing. *Advanced Engineering Informatics*, *51*. https://doi.org/10.1016/j.aei.2021.101495

Nkongolo Wa Nkongolo, M. (2023). News Classification and Categorization with Smart Function Sentiment Analysis. *International Journal of Intelligent Systems*, *2023*. https://doi.org/10.1155/2023/1784394

NLTK. (2023, January 2). *Natural Language Toolkit — NLTK 3.4.4 documentation*. https://www.nltk.org/

Ofosu-Ampong, K. (2024). Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions. *Telematics and Informatics Reports*, *14*, 100127. https://doi.org/10.1016/J.TELER.2024.100127

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y. J., & Hiszpanski, A. M. (2020). Data-driven materials research enabled by natural language processing and information extraction. In *Applied Physics Reviews* (Vol. 7, Issue 4). American Institute of Physics Inc. https://doi.org/10.1063/5.0021106

Pessin, V. Z. (2024, April 23). *Validation of GapFinder Tool: Data Repository Submission*. Mendeley Data. https://doi.org/10.17632/wnw8j9xky2.1

Pessin, V. Z., Santos, C. A. S., Yamane, L. H., Siman, R. R., Baldan, R., & Júnior, V. L. (2023). A method of Mapping Process for scientific production using the Smart Bibliometrics. *MethodsX*, 102367. https://doi.org/10.1016/J.MEX.2023.102367

Pessin, V. Z., Yamane, L. H., & Siman, R. R. (2022). Smart bibliometrics: an integrated method of science mapping and bibliometric analysis. *Scientometrics*, *127*(6), 3695-3718. https://doi.org/10.1007/s11192-022-04406-6

Pichiyan, V., Muthulingam, S., G, S., Nalajala, S., Ch, A., & Das, M. N. (2023). Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis. *Procedia Computer Science*, *230*, 193-202. https://doi.org/10.1016/j.procs.2023.12.074

Pietikäinen, Matti, Silvén, & Olli. (2021). *CHALLENGES OF ARTIFICIAL INTELLIGENCE-FROM MACHINE LEARNING AND COMPUTER VISION TO EMOTIONAL INTELLIGENCE* (Vol. 1). http://urn.fi/urn:isbn:9789526231990

Python Software Foundation. (2023, January 1). *Welcome to Python*. Https://Www.Python.Org/. https://www.python.org/

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: a modern approach* (4th ed.). Pearson Education Limited.

Sanabria-Z, J., Cruz-Sandoval, M., Moreno-Romo, A., Bosch-Gómez, S., & Ramírez-Montoya, M. S. (2024). Research foresight in bridging open science and open innovation: Overview based on the complex thinking paradigm. *International Journal of Innovation Studies*, *8*(1), 59-75. https://doi.org/10.1016/j.ijis.2023.08.002

Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. In *Cognitive Robotics* (Vol. 3, pp. 54-70). KeAi Communications Co. https://doi.org/10.1016/j.cogr.2023.04.001

Wu, L., & Liao, X. (2023). Intelligent Machine Evolutionary Algorithm Learning Based on Artificial Intelligence. *Procedia Computer Science*, *228*, 1016-1022. https://doi.org/10.1016/j.procs.2023.11.133

Xu, Y., Wu, Y., Gao, H., Song, S., Yin, Y., & Xiao, X. (2021). Collaborative APIs recommendation for Artificial Intelligence of Things with information fusion. *Future Generation Computer Systems*, *125*, 471-479. https://doi.org/10.1016/j.future.2021.07.004