Automating Scientific Paper Screening with ChatGPT: An Evaluation of Efficiency and Accuracy

DANIEL BOTNARENCO*, YERAY BARRIOS FLEITAS, and FAIZAN AHMED, University of Twente, The Netherlands

Goals: This study aims to evaluate the performance of different language models, including BERT and GPT, in scientific paper screening. The primary research question is to assess their classification accuracy and language generation capabilities to gain insights into their potential and limitations. **Method:** The methodology involves evaluating the models for the specific task of scientific paper screening. The dataset comprises 6865 scientific papers with screening decisions provided as ground truth labels. Evaluation metrics such as accuracy and F1 scores are used, along with confusion matrices, to assess the models' classification performance. **Results:** The results show that the BERT model achieved the highest accuracy and F1 score among the tested models, while GPT-3 Turbo and 4 exhibited lower classification accuracy and F1 score performance. The processing speeds varied, with BERT benefiting from the CUDA framework. Each model provides, at best, twice as fast as a human coder processing speed of documents. **Implications:** The findings highlight the importance of prompt engineering and fine-tuning in improving language model performance for specific tasks. The study contributes to developing and understanding large language models in natural language processing tasks, facilitating their effective utilization in scientific paper screening tasks.

Additional Key Words and Phrases: ChatGPT, Large Language Models, Natural Language Processing, Contextual Information Extraction, Machine Learning

1 INTRODUCTION

In recent decades, scientific production has grown dramatically in practically all research fields. One of the main reasons is that research investment has increased, which has given more jobs to researchers, and at the same time, provided greater access to scientific databases such as WoS or Scopus[23]. All this results in a mass of scientific publications that often leads to infoxication, the inability to find what one is looking for due to the volume and dispersion of information. Therefore, literature reviews take on extraordinary value because they allow ordering knowledge so we can all progress.

A well-known problem when doing Systematic Literature Reviews is the significant amount of time necessary to invest in identifying and selecting papers. This issue is named the "screening problem". There exists a case in which a researcher has a large number of chosen scientific papers he would like to review to assess if they fit the research that is currently ongoing. Factors such as a large number of documents and the possibility of human error could lead to the researcher determining the wrong information from the papers. The research would automatically lose a significant percentage in its efficiency, as researchers would draw inaccurate conclusions. The paper of Kevin E.[3] serves as an excellent example that screening different scientific papers is one of the most time-consuming parts of the review process. Defining that the systematic literature review takes on average 33 days for thousands of articles. This study will investigate using state-of-the-art language models like ChatGPT and alternatives to expedite the process of categorizing scientific papers and extracting relevant information.

State-of-the-art language models like ChatGPT[20] are designed to process and analyze natural language data. These models employ natural language processing (NLP) and machine learning approaches to find patterns and relationships within the text, allowing them to hold high-level textual discussions.

Interesting research for a similar question, what is the efficiency of a large language model, came from China by Hangcheng et al.[11], which has researched the effectiveness of the Paragraph-BERT-CRF framework. Their test was defined by using different paragraphs of a scientific paper and retrieving contents. The results were impressive, 97% They argued that, because of the use of a pre-trained language model, there are ways to optimize the system (to gain a better performance). Although this paper shows the efficiency of a large language model in subtracting information from scientific papers, the new generation of state-of-the-art models like ChatGPT has made substantial improvements to their systems over the last few years. This research backs up the idea that a large language model could extract more contextual information from a scientific paper given more context.

According to a recent paper, ChatGPT and GPT-4 represent large language models (LLM) from the GPT series that can be fine-tuned on specific tasks and domains[16], including scientific papers. The authors noted that key innovations such as large-scale pre-training, instruction fine-tuning, and reinforcement learning from human feedback have contributed to the success of these models across diverse domains.

The solutions mentioned above are insufficient. This paper will research this problem in the context of the screening problem of scientific papers. The problem arises from the large number of scientific papers published; the load of a researcher increases from one day to another. This study addresses the growing need for automated tools to expedite the extraction of relevant information from the ever-growing body of scientific literature. Despite the complexity of these issues, the imperative for language model improvements is undeniable. Having now stated the problem statement, we can determine the following research question:

To what extent can state-of-the-art language models such as ChatGPT accurately identify and extract more contextual information from scientific papers, and what are the potential benefits and limitations of using language models for this task?

We can split this research question into multiple sub-questions.

TScIT 39, July 7, 2023, Enschede, The Netherlands

(1) What are the key features of state-of-the-art language models such as GPT-4, and how does it differ from earlier models?
(2) What are the benefits and limitations of using language models for research paper classification, and how do these compare to other classification methods, such as human coding?
(3) How accurate are language models in retrieving information from scientific papers, and how does this accuracy vary depending on the specific model used and the additional information available?
(4) What are the potential applications of using language models to retrieve information from scientific papers, and how might this impact the field of research?
(5) How can the accuracy of language model classification be improved, and what techniques or approaches are most effective in achieving this improvement?

This scientific paper aims to determine the efficiency of large language models in aiding researchers in their screening problem.

Firstly, it is essential to help the reader understand and to be on the same level when concluding. Therefore, chapter 2 has been created to create some common ground of understanding on this topic. This chapter overviews state-of-the-art language models such as ChatGPT and BERT. Discuss the potential benefits of using language models for extracting contextual information from scientific papers, such as improving information retrieval, enabling better data-driven decision-making, and accelerating scientific discovery. Discuss the potential limitations and challenges of using language models for this task.

Secondly, Chapter 3 discusses the Methodology used in this research. This chapter discusses the tests conducted to determine the efficiency of the large language models and the metrics used to evaluate said models, respectively. Our findings will be presented in Chapter 4, followed by a detailed discussion in Chapter 5 and conclusions in Chapter 6

2 BACKGROUND

The emergence of the latest iteration of Chat GPT, known as GPT-4, has garnered significant attention and acclaim. Notably, this advanced system has demonstrated remarkable capabilities on par with human abilities in various domains. Although fundamental inquiries regarding the potential replacement of humans by AI and the extent to which AI can enhance human performance remain open-ended, it is crucial to shift our attention to the evaluation of state-of-the-art large language models, such as GPT-4, in terms of their ability to attain a level of performance comparable to that of humans.

Despite GPT-4's comprehensive API, exploring other language models for diverse perspectives on performance and efficiency is vital. Since it would be great to implement a language model inside Excel, we would have to look for systems that have an API implemented. Systems that have an API similar to GPT-4 are: Cloud Natural Language API by [8] and Language Understanding (LUIS) API from [18]. After thorough research on this API, Microsoft will retire LUIS in 2025 and be replaced it with the improved and closer state-of-the-art system, Conversational language [19]. This new system takes a different approach than the previous model because it uses Conversational language understanding (CLU) technology developed by Microsoft. CLU is a cloud-based API service that applies machine-learning intelligence to build natural language understanding components. [2] Chat could also represent a decent alternative, but Bing Chat is built on top of GPT-4.

2.1 Paper screening

Scientific paper screening has evolved through various traditional methods and recent advancements. A notable shift in this process is the focus on abstract screening, a technique that reduces decision-making time and provides ample information about the study.

This paper provides guidelines for abstract screening in systematic reviews and meta-analyses, an essential aspect of conducting a high-quality and comprehensive review. The authors emphasize the importance of a disciplined and consistent approach to abstract screening, which can be time-consuming. The paper provides tips for abstract screening and highlights the need for guidelines to ensure a rigorous process.[22] Removing the other chapters and only reading the abstract helps tremendously in deciding. In addition to the guidelines for abstract screening, researchers have to be more explicit with information when writing the abstract such that there is no need for contributors to read the whole paper to understand what the content will be.

This paper provides an updated reporting guideline for systematic reviews that reflects advances in methods to identify, select, appraise, and synthesize studies. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement was first published in 2009 to address poor reporting of systematic reviews. The PRISMA 2020 statement comprises a checklist of 27 items recommended for reporting in systematic reviews and an "explanation and elaboration" paper providing additional reporting guidance for each item.[21]

2.2 NLP and screening papers

"Natural language processing" is a branch of artificial intelligence focused on how computers can process language as human beings do. The development of NLP technology was initiated with Alan Turing's 1950 paper [26], proposing the Turing test, a cornerstone for artificial intelligence algorithms. This step represents one of NLP research's (if not the one) most important aspects. Seventy years later, NLP is one of the hottest areas of AI thanks to many applications such as text generators, chatbots, and text-to-image programs that produce photo-realistic images of anything describable. NLP is also used in conversational agents like Siri and Alexa. Furthermore, Google uses NLP to improve search engine results and has added multiple functionalities to the system, as seen here [9]. Using NLP technology, we can summarize/extract information from a text/scientific paper. Previous research has demonstrated that NLP has a high chance of correctly summarizing abstracts, as shown by the previous research on this topic, where a developed NLP algorithm could retrieve keywords from an abstract of a scientific paper with decently high accuracy. We can also draw this conclusion from the paper of Basili et al., [1], where he performed extensive research on the efficiency of NLP classification of text. Depending on the different methods used, this process can be around 70% accurate. In

Table 1. Results of Paragraph-BERT-CRF framework

Test	Result
Precision	97.45%
Recall	97.44%
F1 Score	97.44%

contrast to this, this paper created by Dangovski et al.[5] at MIT shows that the developed neural network that can read scientific papers (in technical jargon) and summarize the chosen text in a couple of easy-to-understand sentences.

In G. Hangcheng's study[11], the efficiency of the BERT model was evaluated by identifying key elements in scientific paper abstracts such as purpose, method, results, and conclusion. The results of this study are available in Table1. Other papers discuss various algorithms to classify research papers [3] [14].

2.3 Literature gaps

Several recent studies have examined the accuracy and effectiveness of language models in processing scientific papers. For example, a team of scientists at MIT and elsewhere developed a neural network that can read scientific papers and render a plain-English summary in a sentence or two [5]. Another study used the GPT-3 language model to automatically generate summaries of scientific papers and discovered that the model could accurately capture essential information from the scientific papers [10].

This study investigates the performance of Large Language Models (LLMs) in biomedical tasks, but they have yet to be adequately investigated for more specific biomedical applications[4]. This study investigates the performance of LLMs such as the ChatGPT family of models (GPT-3.5s, GPT-4) in biomedical tasks beyond question-answering.

While state-of-the-art language models have shown promising results in processing information from scientific papers, limitations remain. For example, scientific papers often contain technical jargon that can be difficult for these systems to interpret. Additionally, language models may need help to capture the nuances and complexities of scientific arguments and hypotheses. There needs to be more literature researching how efficient are state-of-the-art language models in extracting information from a text or, in the case of the research problem, from a scientific paper.

RQ1: What are the key features of state-of-the-art language models such as GPT-4, and how does it differ from earlier models?

According to a research paper [29], GPT series models, including GPT-4, have exceptional natural language processing capabilities. These models have gained considerable attention due to their ability to generate human-like responses. However, the research also suggests that the overall ability of GPT series models on natural language understanding tasks does not increase gradually as the models evolve, especially with the introduction of the RLHF training strategy. While this strategy enhances the models' ability to generate human-like responses, it also compromises their ability to solve some tasks. The paper evaluates the performance of six GPT series models across nine different natural language understanding tasks. The research suggests that there is still room for improvement in model robustness.

RQ2: What are the benefits and limitations of using language mod- els for research paper classification and how these compare to other classification methods, such as human coding?

Language models have become increasingly popular for research paper classification due to their ability to understand and generate text. These models can recognize, summarize, translate, predict, and generate text and other content, making them useful for various applications in healthcare, software development, and other fields [25]. One of the benefits of using language models for research paper classification is their ability to process large amounts of data quickly and accurately. They can also identify patterns and relationships in the data that may not be immediately apparent to human coders. However, there are also limitations to using language models for research paper classification. One of the main limitations is that language models may only sometimes be able to accurately interpret the context of the text they are analyzing, leading to errors in classification. Language models may also need help capturing nuances such as technical language. When comparing language models to other classification methods, such as human coding, it is vital to consider the strengths and weaknesses of each approach. Human coding can be more accurate in cases where the text context is complex or ambiguous. However, it can also be more time-consuming and expensive than using language models [24]. Ultimately, the best approach will depend on the specific research question and the resources available for classification.

RQ4: What are the potential applications of using language models to retrieve information from scientific papers, and how might this impact the field of research?

Language models have vast potential in retrieving information from scientific papers, with applications like knowledge retrieval and clinical decision support in medicine [25]. Researchers can train large language models in biology to understand proteins, molecules, DNA, and RNA. Additionally, scaling and maintaining large language models can take time and effort. Overall, the potential applications of using language models to retrieve information from scientific papers are vast and could significantly impact the field of research. However, it is essential to consider these models' strengths and limitations carefully and ensure they are used ethically and responsibly.

Overall, using state-of-the-art language models in processing scientific papers is an active area of research. There is ongoing work to determine and improve the accuracy and effectiveness of these models.

3 METHODOLOGY

The constant improvement of this system facilitates the use of the GPT API. The decision to utilize GPT-4 in this study stems from its superior performance in natural language processing. To fine-tune the model using prompt engineering techniques[15], aligning the model with the specific task of scientific paper screening. This task-oriented fine-tuning is anticipated to produce more accurate

predictions, which is the primary research question this study seeks to answer.

The dataset used in this study is an extensive collection of 6865 scientific papers, each filled with rich information, from the paper ID to categories of science. It also includes the screening results, which detail each paper's inclusion/exclusion decision. These decisions are the ground truth labels we will use to train and evaluate the model. The dataset and the code used in this paper can be found in the refferences[6].

This study falls under the umbrella of quantitative research. We aim to analyze GPT-4's effectiveness in paper classification by comparing its predictions with the ground truth labels from our dataset. To evaluate the model's performance, we focus on two main aspects: its classification performance (measured through accuracy and F1 score) and its language generation capabilities (measured through fluency and coherence) see Section3.2. Our key evaluation metric will be the confusion matrix, which provides a comprehensive view of the model's classification performance. By assessing these aspects, we hope to comprehensively understand GPT-4's potential and limitations in scientific paper screening.

3.1 Data Processing

The provided corpus for this research consists of a public database containing scientific papers and meta-data regarding its context. The dataset is divided into two parts: Paper Data and Screening Results. The former provides the raw text and metadata for each paper, while the latter contains the screening decisions made by human experts. For a better overview of the dataset check Appendix A.3.

Creating a label specifically for each entry is necessary for the evaluation methods. We will follow the Decision column in this paper and label it accordingly. (0 - Included, 1 - Excluded, 2 - Not Sure) The categories for classification are imbalanced. To address this issue extra safety steps have been taken to address this issue. Even if a model predicts only the majority class for all instances, it can achieve high accuracy due to many negatives[12]. Therefore, accuracy should not be the sole evaluation metric. To combat this, the best metric for this issue we will use is the F1 score because metrics under the ROC curve (AUC-ROC) are more informative than accuracy[27]. To conduct the evaluation based on the paper's abstract, it is necessary to make sure that the paper's abstract is existent. Therefore, scientific papers without an abstract are removed from the dataset.

3.2 Evaluation

When defining evaluation criteria, it is essential to consider various aspects of the language model's performance. For example, evaluating the model's language generation capability may involve assessing factors such as fluency, coherence, grammaticality, and the ability to produce contextually appropriate responses.

Additionally, evaluating the model's understanding and comprehension can involve assessing its ability to accurately answer questions, provide relevant information, and demonstrate a grasp of context and nuances. By defining clear and robust evaluation criteria, we can effectively assess the strengths and limitations of large language models and facilitate their continual improvement and development.

Table 2. Results Comparison of the models based on Abstract

Model	Accuracy (%)	F1 Score (%)	N0 Samples
bert-base-uncased	83.00	83.00	6863
GPT 3.5 Turbo	74.22	75.69	5524
GPT-4	62.24	61.98	510
GPT-4 (modified)	60.77	61.68	1104

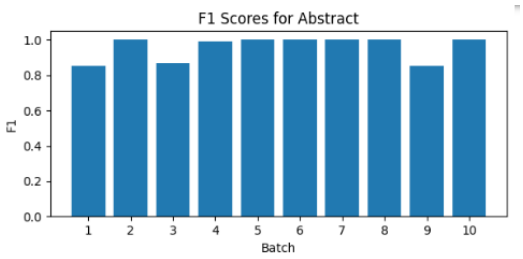

Fig. 1. F1 Scores of BERT trained with different batches.

3.2.1 Accuracy. Accuracy is a straightforward metric that measures the overall correctness of predictions. It is calculated as the ratio of correct predictions to the total number of predictions. Accuracy does not consider the specific types of errors made by the model. Accuracy can be misleading when the dataset is imbalanced, meaning some classes have significantly more samples than others.

3.2.2 F1 Score. The F1 score is a metric that considers both precision and recall, two critical measures in binary or multiclass classification tasks. Precision is the ratio of accurate positive predictions to the total predicted positive instances. It measures how many of the predicted positive cases are positive. The recall is the ratio of accurate positive predictions to the total actual positive instances. It measures how many of the actual positive instances are correctly predicted. The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances precision and recall, making it useful when both measures are important. The F1 score ranges from 0 to 1, with a value of 1 indicating perfect precision and recall and a value of 0 indicating poor performance.

4 RESULTS

One of the alternatives discussed before in the document refers to the use of BERT to solve the classification problem. The problem with using BERT is that it offers already pre-trained models that make it hard to use them for different tasks. There are models for specific tasks such as translating, classifying, summarizing, question answering and many others. For testing a classifying model has been used bert-base-uncased. The model has not been fined-tuned and therefore uses a predefined model that is available on Huggingface. Huggingface is a community-made website where people and companies can share their created models. After the epoch, we can determine different evaluations for the model such as accuracy and F1 score. See Table 2 for the results.

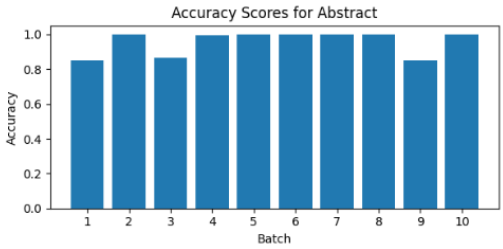

Fig. 2. Accuracy of BERT trained with different batches.

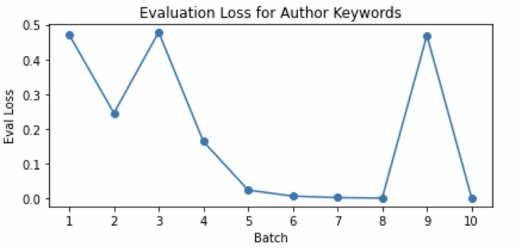

Fig. 3. Accuracy of BERT trained with different batches.

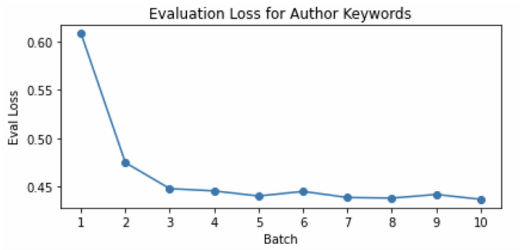

Fig. 4. Evaluation Loss of BERT trained with different batches.

CUDA is a platform developed by NVIDIA, which allows software developers to leverage the parallel processing capabilities of NVIDIA GPUs. This results in a significant speedup compared to CPU-based computations. This framework improves the performance of the algorithm in all aspects that comprise it: training and evaluation.

For the purpose of discussion, the Evaluation Loss results are shown in Figure 3. Results of the BERT model for pre-defined batches can be seen here 1 and 2. These figures show the F1 Score and the Accuracy of the model for a given batch. Evaluation for the model has been done for more variables such as Author Keywords and Title which can be seen here in the Appendix A.1. Tests of the models with random dataset batches; the results are available in Appendix A.2 (Figures 12, 10, 11) Results of the evaluation loss of the model after training can be seen in Figure 4. Comparing these results to the ones found in the Evaluation Loss on the predefined batches (See Figure 3).

Utilizing the GPT-API framework is easier compared to using BERT models. The fine-tuning of the GPT models is more inclined to the use of given prompts (Prompt Engineering). Compared to

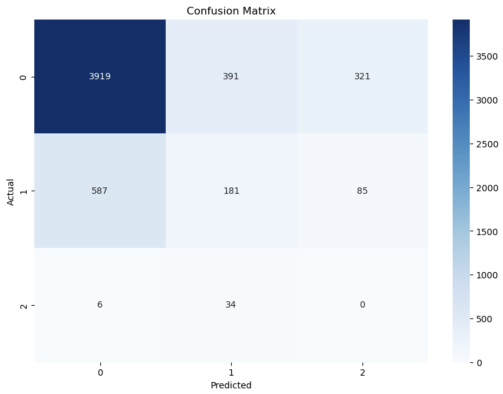

Fig. 5. Confusion Matrix GPT-3.5-turbo.

BERT models where the fine-tuning process is made out of training the model for a specific task.

After extensive modifications of the prompt to feed the GPT-API, we have selected a prompt that can answer the classification problem with a number. This process also helps lower the cost of the use of the API by decreasing the number of tokens used.

The test for the confusion matrix has been done on a sample of 5524 elements. Figure 5 shows the results of the test. The Matrix shows that the most picked answer was to Include the scientific paper, by making use of the darker blue colour of the entry such as a heat map. The second most predicted answer was to Include the papers when the actual answer was to Reject them. A more in-depth discussion can be found in Section 5.

Sadly, we did not get access to the GPT-4 API key to easily determine the results of the model. Therefore, the only option left was to buy the ChatGPT subscription which allows users to use the GPT-4 through the user interface on the website. This means that to conclude a result of the GPT-4 the only option left was to manually give the prompt and the necessary information in the user interface and manually input the results in an Excel spreadsheet. This requires more time to evaluate GPT-4 model. Results of the GPT-4 model can be seen in Table 2 and the confusion matrix can be seen in Figure 6.

The speed of using these large language models is faster than the processing speed of a human. Recorded when classifying the data, we use the time it takes for a human to complete the task as ground truth in this paper. Table 3 shows the time it takes to process a document based on Title or Abstract. The speed of BERT is only improved drastically when using the CUDA framework. In comparison, the GPT-API performs faster out of the box, but the model is more dependent on the usage of the API, which can make the responses of the model take longer than expected.

5 DISCUSSION

Based on the results in Table 2, BERT LLM demonstrates promising performance in classifying scientific papers. However, our evaluation reveals inconsistencies across batches, likely due to predefined

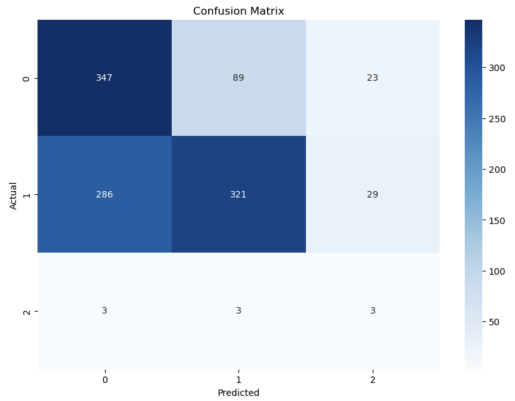

Fig. 6. Confusion Matrix GPT-4.

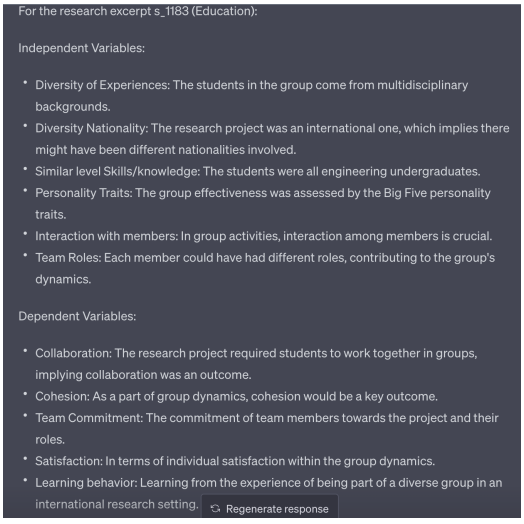

Fig. 7. Response of GPT-4 determining the Dependent and Independent variables.

Table 3. Comparison of Processing Speeds

Model	Task	Time (seconds)
Human	Abstract	20
	Title	8
BERT	Both	0.63
GPT 3.5 Turbo & 4	Both	0.26 - 30

batch configuration. Randomized batches were also evaluated, providing additional insight into the model's performance (Appendix A.2).

Despite not applying fine-tuning, the model exhibits consistent F1 Scores and accuracy across specific batches (Figures 1 and 2). We observed a drop in accuracy to 83.3% for batches 1,3 and 9, whereas other batches reached 100%. This may suggest a need for fine-tuning and model training on task-specific data to improve performance further.

These results might behave like this because of multiple reasons. Because the model is already pre-trained, it already has some general knowledge about answering similar questions. Training the model at least once with the data to be used to feed the model is still necessary. Therefore, it has a simple idea of the data and how it should work. Another reason could be the need to train the model longer. This task is vital for any NLP-specific tasks.

Following the evaluation of the BERT model, an evaluation of the GPT model has to be conducted. Firstly, a discussion on the GPT 3.5 turbo will be done. Compared to the previous models of GPT, this model performs better and faster in every task. Therefore, the other models have yet to be tested. Figure 5 shows the confusion matrix of the classification task defined before. This matrix showcases the most common answer for each label with the actual value and the predicted value in a heat-like map. Figure 5 shows the most common answer as "Included." On the other hand, there is a slight disagreement between the correct and predicted answers, as seen in Table 2.

Continuing with the OpenAI models, we will discuss GPT-4. Conducting the same tests for both models greatly shows us the difference. From Figure 6, we can see the confusion matrix and, in Table 2, the results of the GPT-4 model. The results show decreased accuracy and F1 score when using the newer model. The confusion matrix of the newer model shows a more detailed view of the predicted results. We observe that the model is more in line with the ground truth. This fact can be seen from the third most chosen element, predicted 1 - actual 1. Comparing this element to the third most chosen element of the GPT-3.5 model predicted 1 - actual 0. As stated, the labels used for this classification problem were: 0 - Included, 1 - Excluded and 2 - Not Sure. There was a discrepancy between the choice of the algorithm to Exclude papers that were Excluded.

Given the high expectations of OpenAI models, the modest results warrant a critical analysis.

The first point to argue would be the correctness of the provided prompt, in other words, the context. The algorithm might overlook some details when making decisions because of the need for more information in the prompt. Although, with the modified prompt, the model can classify better when papers should be Rejected when they were Rejected. Even with a more potent prompt, the accuracy did not improve but decreased. This discussion about the prompt refers to the previously mentioned Prompt Engineering [15].

Another functionality tested in this paper was to see if the database could be expanded with more useful information, such as Independent and Dependent variables. To do it is more complicated because it's hard to place the answers of ChatGPT in a data frame. This problem might be solved by working on a very elaborate prompt. However, we can check whether ChatGPT can provide this information for the user.

The dataset provided was updated with more information about around 60& of the total number of papers. This dataset tracks the papers and assigns each paper what independent and dependent variables are tracked in the paper in addition to the context the paper is situated. The variables were divided into specific parts to have a more harmonious dataset.

Considering this paper with id s_1183 where it has been assigned the Context of the paper: Education, Independent variable: Diversity Nationality and Personality Traits and the Dependent variable: Not specified. Comparing the results of the screening of this paper to the response of the GPT-4 Model in Figure 7, we can determine that the model is not far from the truth. The Model determines the context of the paper and Independent variables correctly, but it seems that the model has found more variables that fit the paper. The researchers screened this paper and decided that the dependent variable is not specified, but the model can determine 5 dependent variables.

It's necessary to have more comparison points to come up with a conclusion to determine if the model can help researchers expand their database. Therefore, another example is provided in Appendix A.1 for the paper with id s_0459. This paper has been classified with the following information: the context is Education, and the Independent Variables are: Personality Traits and Team Trust. Dependent Variables: Satisfaction. These results cannot be evaluated because of how the dataset is created and how the model answers. Therefore we have checked the outputs with the ground truth and decided that GPT-4 can improve the dataset with more information about the scientific paper.

RQ3: How accurate are language models in retrieving information from scientific papers, and how does this accuracy vary depending on the specific model used and the additional information available?

As presented in this research, the accuracy of language models in retrieving information from scientific papers ranges between 60 and 80%. It's important to note that this accuracy largely depends on the prompt given to the large language model, particularly for models like those from OpenAI that necessitate detailed prompts. Giving the model more contextual information about a specific scientific paper increases the chances of a correct decision. It also increases the price to be paid when using this service.

RQ4: How can the accuracy of language model classification be improved, and what techniques or approaches are most effective in achieving this improvement?

The accuracy of language models in classifying scientific papers can be improved by providing the model with a more detailed prompt and more contextual information about the specific query. Techniques such as Chain-of-Thought Prompting [28] and Few-Shot Learning [17], as suggested by the study on Prompt Engineering [15], offer effective strategies for achieving a more productive prompt.

Ethical issues that might arise by using large language models are the creation of bias in the algorithms and the struggle for recognition.

Bias is one of the biggest problems arising from using artificial intelligence algorithms. To lower the bias's impact when using the algorithms, they are a couple of key points where researchers and developers must be cautious. These key points are data collection and training of data. As discussed previously, if the data collection is done poorly and the person in charge is biased, the algorithm can implicitly inherit these ideas. As for the training of data, if the data used to train is biased, the algorithm can learn these traits to be biased. To solve this problem, researchers and developers must be careful when developing artificial intelligence algorithms. To reduce the impact of bias, ethical thinking must be at the project's forefront. This paper by [13] analyses artificial intelligence ethics. Researchers can use this overview of ethical artificial intelligence to be used in the project lowering the impact of any unethical bias to be added to the algorithm.

The issue regarding the struggle for recognition arises for a researcher when using a machine learning algorithm to evaluate or answer a question. A great paper showcasing how the Internet has affected us is "Social Implications of the Internet" by [7] The struggle for acceptance in the scientific world is ongoing. Historically, acknowledgement was frequently linked to one's standing in the academic hierarchy, with older researchers gaining more recognition than novice researchers. However, technology has begun to disrupt this hierarchy. Using a new machine learning algorithm can be seen by other researchers as the inability of the person to come up with their conclusion. Therefore, they would only accept the researcher's findings if he determined these results.

6 CONCLUSIONS

In summary, our evaluation of BERT and GPT-based models for scientific paper screening demonstrated a commendable accuracy, ranging from 60% to 80%. This opens up opportunities for employing such large language models (LLMs) to facilitate the initial paper screening process, freeing up researchers' time for more in-depth analysis.

Our findings resonate with the existing literature, corroborating the effectiveness of LLMs in complex tasks requiring high-level text comprehension and decision-making. Moreover, our research showcases that the performance of these models is heavily influenced by the quality of the prompts, underlining the relevance of Prompt Engineering for LLMs.

Despite its strengths, our approach has its limitations. For instance, while using pre-trained BERT models through Huggingface proved beneficial, it restricted the ability to fine-tune the models for our specific task.

Further, the research shed light on the importance of having a well-crafted prompt when using OpenAI's GPT models. Notably, our initial attempts using GPT-3.5 Turbo and GPT-4 revealed mediocre results, underlining the crucial role of Prompt Engineering. Our explorations with Chain-of-Thought Prompting and Few-Shot learning techniques showed their potential to improve LLMs' performance.

Turning to the ethical implications of using LLMs in scientific paper screening, there are potential concerns about privacy, bias, and transparency. It is crucial to ensure that these AI models do not perpetuate existing biases in the literature and can be held accountable for their decisions. Therefore, ongoing ethical and responsible AI discussions should accompany any conversation about employing these models for screening scientific literature.

This research makes a substantial contribution to the NLP field by demonstrating the potential utility of LLMs for research methodologies, specifically for scientific paper screening. It provides a foundation upon which further work can be built to optimize the application of these models in various research contexts.

Future research can further delve into improving the accuracy and efficiency of these models by exploring alternate training techniques, different architectures, or hybrid models. Additionally, integrating LLMs into other areas of scientific research can open up exciting avenues. For instance, they could be used for automated literature reviews, fact-checking in scientific discourse, or generating hypotheses for further investigation.

Overall, our findings provide a positive picture for using LLMs in scientific research, with the potential to revolutionize many parts of the research process. The study's findings highlight the importance of continued exploration and optimization of these models, bringing us closer to the era of AI-driven research approaches.

REFERENCES

[1] Roberto Basili, Alessandro Moschitti, and Maria Teresa Pazienza. 2006. EXTENSIVE EVALUATION OF EFFICIENT NLP-DRIVEN TEXT CLASSIFICATION. *Applied Artificial Intelligence* 20, 6 (2006), 457–491. https://doi.org/10.1080/08839510600753725 arXiv:https://doi.org/10.1080/08839510600753725

[2] Bing. 2023. Bing Chat. https://www.bing.com/new

[3] Kevin E. K. Chai, Robin L. J. Lines, Daniel F. Gucciardi, and Leo Ng. 2021. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews* 10, 1 (2021), 93. https://doi.org/10.1186/s13643-021-01635-3

[4] Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K. Savova, and Danielle S. Bitterman. 2023. Evaluation of ChatGPT Family of Models for Biomedical Reasoning and Classification. arXiv:2304.02496 [cs.CL]

[5] Rumen Dangovski, Li Jing, Preslav Nakov, Mićo Tatalović, and Marin Soljačić. 2019. Rotational Unit of Memory: A Novel Representation Unit for RNNs with Scalable Applications. *Transactions of the Association for Computational Linguistics* 7 (04 2019), 121–138. https://doi.org/10.1162/tacl_a_00258 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00258/1923613/tacl_a_00258.pdf

[6] Daniel, Botnarenco. 2023. Large Language Models Evaluation. https://github.com/Dani1232312/gpt-api-evaluation

[7] Paul DiMaggio, Eszter Hargittai, W. Russell Neuman, and John P. Robinson. 2001. Social Implications of the Internet. *Annual Review of Sociology* 27 (2001), 307–336. http://www.jstor.org/stable/2678624

[8] Google. 2016. Cloud Natural Language. https://cloud.google.com/natural-language

[9] Google. 2021. MUM: A new AI milestone for understanding information. https://blog.google/products/search/introducing-mum/

[10] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. arXiv:2209.12356 [cs.CL]

[11] G. Hangcheng, H. Yanqing, L. Tian, W. Zhenfeng, and D. Cheng. 2022. Identifying Moves from Scientific Abstracts Based on Paragraph-BERT-CRF. *Data Analysis and Knowledge Discovery* 6, 2-3 (2022), 298–307. https://doi.org/10.11925/infotech.2096-3467.2021.0973 cited By 0.

[12] Steven Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael Riegler, Pål Halvorsen, and Sravanthi Parasa. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports* 12, 1 (2022), 1–9.

[13] Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao. 2022. An Overview of Artificial Intelligence Ethics. (07 2022), 1–21. https://doi.org/10.1109/TAI.2022.3194503

[14] Sang-Woon Kim and Joon-Min Gil. 2019. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences* 9, 1 (2019), 30. https://doi.org/10.1186/s13673-019-0192-7

[15] Qintong Li, Zhiyong Wu, Lingpeng Kong, and Wei Bi. 2022. Explanation Regeneration via Information Bottleneck. arXiv:2212.09603 [cs.CL]

[16] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. arXiv:2304.01852 [cs.CL]

[17] Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 2824–2835. https://doi.org/10.18653/v1/2022.findings-acl.222

[18] Microsoft Azure. 2019. Language Understanding (LUIS). https://learn.microsoft.com/en-GB/azure/cognitive-services/luis/what-is-luis

[19] Microsoft Azure. 2023. Conversational language understanding. https://azure.microsoft.com/en-us/products/cognitive-services/conversational-language-understanding/#Getstarted

[20] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[21] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021). https://doi.org/10.1136/bmj.n71 arXiv:https://www.bmj.com/content/372/bmj.n71.full.pdf

[22] Joshua Polanin, Therese Pigott, Dorothy Espelage, and Jennifer Grotpeter. 2019. Best Practice Guidelines for Abstract Screening Large-Evidence Systematic Reviews and Meta-Analyses. *Research Synthesis Methods* 10 (05 2019). https://doi.org/10.1002/jrsm.1354

[23] Raminta Pranckutė. 2021. Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications* 9, 1 (2021). https://doi.org/10.3390/publications9010012

[24] Parisa Safikhani, Hayastan Avetisyan, Dennis Föste-Eggers, and David Broneske. 2023. Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence* 3 (02 2023), 6. https://doi.org/10.1007/s44163-023-00050-y

[25] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. arXiv:2212.13138 [cs.CL]

[26] A. M. TURING. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* LIX, 236 (10 1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433 arXiv:https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf

[27] Muhammad Umer, Zainab Imtiaz, Dr. Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. 2020. Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access* PP (08 2020), 1–1. https://doi.org/10.1109/ACCESS.2020.3019735

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

[29] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. arXiv:2303.10420 [cs.CL]

A APPENDIX

A.1 Appendix A.1

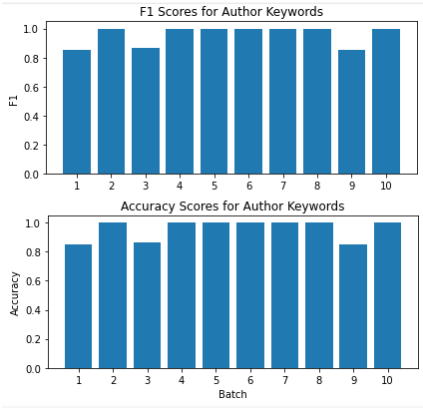

Fig. 8. Scores of Author Keyboard using pre-defined datasets.

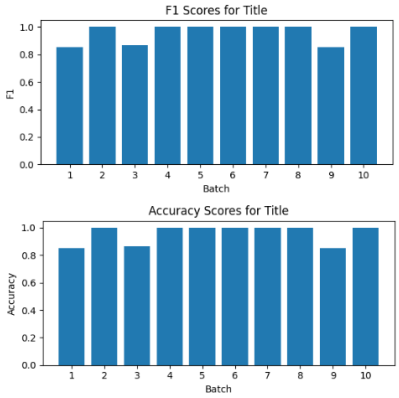

Fig. 9. Scores of Title using pre-defined datasets.

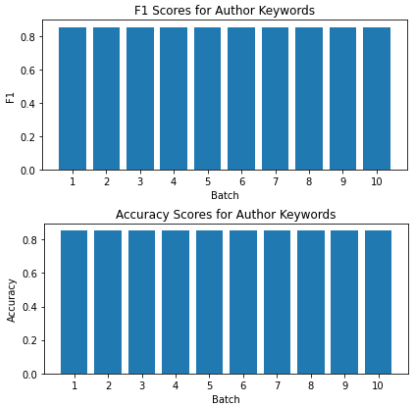

Fig. 10. Scores of Keywords.

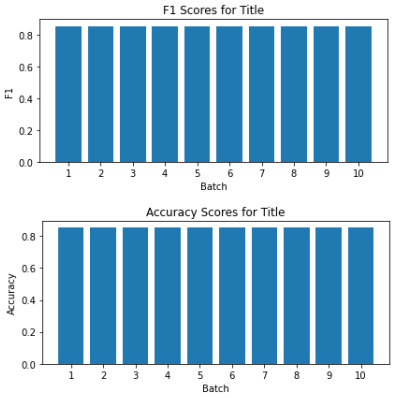

Fig. 11. Scores of Title.

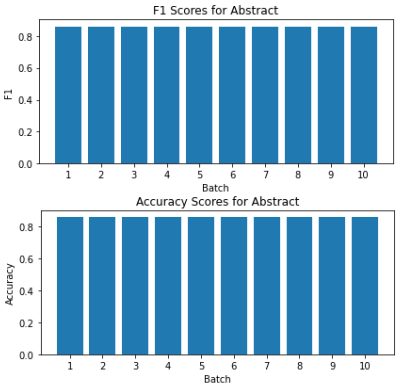

Fig. 12. Scores of Abstract.

A.2 Appendix A.2

For the research excerpt s_0459 (Industry/Business):

Independent Variables:

- Trust: The study investigates the role of trust in virtual teams.
- Conflicts: The challenges faced by virtual teams could be seen as conflicts.
- Medium of communication (virtual/f2f): The study is focused on virtual teams, so the medium of communication is virtual.
- Personality Traits: The study takes into consideration the Openness to Experience personality trait.

Dependent Variables:

- Productivity: Virtual team performance can be seen as a measure of productivity.
- Team Performance: The study is focused on investigating the members' ratings of virtual team performance.
- Trust: Trust also appears to be an outcome of certain conditions in this study.
- Satisfaction: Members' preference for virtual teams can be seen as a measure of their satisfaction.

Fig. 13. Response of GPT-4 in determining more information.

A.3 Appendix A.3

	Paper Data									
	ID_paper	Document Type	Year	Authors	Title	Abstract	Author Keywords	Index Keywords	Source title	Nº cites
3	s_1489	ARTICLE	2018	ZHANG JS;TAN	"THIS IS WHY WE PLAY": CHARACTERIZING ONLINE FAN COMMUNITIES OF TI	PROFESSIONAL SPORTS CON:	NBA;ONLINE FA	PROFESSIONAL ASPECTS;SPOR	PROCEEDINGS OF THE ACM O	5
4	s_1777	CONFERENCE PAPER	2017	SARIGIANNI C;\	"WE NEED TO TALK!" - PROJECT TEAMS DEALING WITH LOW CONNECTIVITY	THE CONTINUOUS TECHNOLO	COMMUNICATI(	COMMUNICATION;INFORMATION	PROCEEDINGS OF THE 25TH E	1
5	w_0669	ARTICLE	2020	BARNSCHEN S	``BEYOND SIMPLE EXPLANATIONS'' THE IMPACT OF DEMOGRAPHIC DIVERSI	THIS STUDY ANALYZED THE IN	DIVERSITY;TEA	SEX-ROLE STEREOTYPES;WORI	ZEITSCHRIFT FUR ARBEITS-UN	0
6	w_2763	ARTICLE	2012	FRANSEN K;VA	``YES, WE CAN!'': PERCEPTIONS OF COLLECTIVE EFFICACY SOURCES IN VO	COLLECTIVE EFFICACY CAN B	TEAM CONFIDE	TEAM PERFORMANCE;SELF-EFF	JOURNAL OF SPORTS SCIENC	39
7	s_5289	CONFERENCE PAPER	1992	KIRK SJ	`VENTURE': A GAMING/SIMULATION TO IMPROVE GROUP PERFORMANCE IN	THIS PAPER ADDRESSES HOW		COST ACCOUNTING;GAME THE(	SAVE PROCEEDINGS (SOCIET)	0
8	s_4622	BOOK CHAPTER	2005	SALAS E;STAGI	25 YEARS OF TEAM EFFECTIVENESS IN ORGANIZATIONS: RESEARCH THEME		CLASSICAL DE(		INTERNATIONAL REVIEW OF IN	157
9	s_3412	ARTICLE	2011	MONTOYA MM;	3D COLLABORATIVE VIRTUAL ENVIRONMENTS: EXPLORING THE LINK BETWE	INCREASINGLY, ORGANIZATIOI	3D COLLABORA		DECISION SCIENCES	76
10	s_1162	CONFERENCE PAPER	2019	AVILA DT;VAN F	46TH SEFI ANNUAL CONFERENCE: ARE TEAM MEMBER CONTRIBUTIONS CO	SOFTWARE ENGINEERING IS A	COHESION;MEI	EDUCATION COMPUTING;ENGIN	PROCEEDINGS OF THE 46TH S	0
11	s_2340	ARTICLE	2015	MCMAHON PE	A "THINKING FRAMEWORK" TO POWER SOFTWARE DEVELOPMENT TEAM PE	ESSENCE IS A NEW OBJECT M		HUMAN RESOURCE MANAGEME	CROSSTALK	0
12	s_2341	CONFERENCE PAPER	2015	FREIRE AS;DA	A BAYESIAN NETWORK MODEL TO ASSESS AGILE TEAMS' TEAMWORK QUAL	THE AGILE MANIFESTO STATE:		BAYESIAN NETWORKS;DISTRIBU	PROCEEDINGS - 29TH BRAZILI	6
13	s_0785	ARTICLE	2020	TURESKY EF;S	A CALL TO ACTION FOR VIRTUAL TEAM LEADERS: PRACTITIONER PERSPEC	PURPOSE: THE PURPOSE OF T	COACHING ANI		ORGANIZATION MANAGEMENT	10
14	s_1780	CONFERENCE PAPER	2017	STEINER MW;S	A CASE STUDY APPROACH FOR UNDERSTANDING THE IMPACT OF TEAM SEI	AN IMPORTANT INGREDIENT F	CAPSTONE DE:	ENGINEERING EDUCATION;CAP:	ASEE ANNUAL CONFERENCE A	2
15	s_0786	ARTICLE	2020	PAOLETTI J;BIS	A CHECKLIST TO DIAGNOSE TEAMWORK IN ENGINEERING EDUCATION	TEAMWORK IS INCREASINGLY	CHECKLIST;PBI	ENGINEERING EDUCATION;HUM	INTERNATIONAL JOURNAL OF	5
16	s_3681	CONFERENCE PAPER	2010	QIU RG	A COLLABORATIVE MODEL OF ENGINEERING EDUCATION FOR COMPLEX GL	ALTHOUGH TEAM-BASED PRO	COLLABORATI(	AD-HOC COLLABORATION;BEHA	PROCEEDINGS - FRONTIERS II	4
17	w_0342	ARTICLE	2021	RUEDA-MEDIN/	A COMBINATION OF SELF-DEBRIEFING AND INSTRUCTOR-LED DEBRIEFING I	BACKGROUND DEBRIEFING IS	DEBRIEFING;HI	SIMULATION;SKILLS	NURSE EDUCATOR	2
18	s_0406	ARTICLE	2021	SHARMA P;KAF	A COMPARATIVE STUDY OF MEN AND WOMEN ENTREPRENEURSHIP - WITH :	ENTREPRENEURSHIP LIES MC	COMPARATIVE		INTERNATIONAL JOURNAL OF	0
19	s_2880	CONFERENCE PAPER	2013	BASTIDA R;GUI	A COMPARATIVE STUDY OF THE EFFECT OF BLOGS AND EMAIL ON VIRTUAL	VIRTUAL TEAMS (VTS) OFFER	BLOGS;E-MAIL;	BLOGS;ELECTRONIC MAIL;SOCI	PROCEEDINGS - PACIFIC ASIA	4
20	w_2575	ARTICLE	2013	GREFFRATH G;	A COMPARISON BETWEEN CENTRE-BASED AND EXPEDITION-BASED (WILDE	THE PURPOSE OF THIS STUDY	GROUP EFFEC	EDUCATION;THERAPY;MODEL	SOUTH AFRICAN JOURNAL FO	5
21	s_3937	CONFERENCE PAPER	2009	EMIL WO;RENÉ	A COMPARISON BETWEEN TWO LEADERSHIP MODELS FOR SECURITY CHEC	THE IMPACT OF LEADERSHIP :	AIRPORT SECU	AIRPORT SECURITY;AVIATION SI	PROCEEDINGS - INTERNATION	2
22	s_2059	CONFERENCE PAPER	2016	TAUSCH S;TA S	A COMPARISON OF COOPERATIVE AND COMPETITIVE VISUALIZATIONS FOR	WE PRESENT A STUDY THAT II	COMPETITION;	COMPETITION;FEEDBACK;HUMA	CONFERENCE ON HUMAN FAC	8
23	w_3646	REVIEW	2006	BROPHY DR	A COMPARISON OF INDIVIDUAL AND GROUP EFFORTS TO CREATIVELY SOLV	UNTIL RECENTLY, MOST EMPIF		CONSENSUAL ASSESSMENT TE	CREATIVITY RESEARCH JOURI	32
24	w_2358	ARTICLE	2014	LAFORGE-MAC	A COMPARISON OF THE SELF-EFFICACY-PERFORMANCE RELATIONSHIP BET	ALTHOUGH STUDIES HAVE SH		COLLECTIVE EFFICACY;PATH-AN	JOURNAL OF APPLIED SPORT	3
25	w_0670	ARTICLE	2020	HOOGEBOOM I	A COMPLEX ADAPTIVE SYSTEMS APPROACH TO REAL-LIFE TEAM INTERACT	INTERACTION DYNAMICS ARE	TEAM INTERAC	GROUP DECISION-MAKING;INTE	GROUP \& ORGANIZATION MAI	16

	Paper Data			Screening Results						
	Nª Ref	Science Category	WoS category	Basic data?	Title	Abstract	Is about humans?	Quality keywords ?	Decision	ML
3	90			Yes	Not sure	Yes	Yes	Yes	Included	Included
4	59			Yes	Not sure	Yes	Yes	Yes	Included	Included
5	103	PSYCHOLOGY	PSYCHOLOGY, APPLIED	Yes	Yes	Yes	Yes	Yes	Included	Included
6	40	SPORT SCIENCES	SPORT SCIENCES	Yes	Not sure	Yes	Yes	Yes	Included	Included
7	0			Yes	Not sure	Yes	Yes	Yes	Included	Included
8	169			Yes	Yes	Special	Yes	Yes	Included	Included
9	89			Yes	Yes	Yes	Yes	Yes	Included	Included
10	16			Yes	Yes	Yes	Yes	Yes	Included	Included
11	16			Yes	Yes		Yes	Yes	Included	Included
12	31			Yes	Not sure	Yes	Yes	Yes	Included	Included
13	49			Yes	Not sure	Yes	Yes	Yes	Included	Included
14	25			Yes	Yes	Yes	Yes	Yes	Included	Included
15	67			Yes	Not sure	Yes	Yes	Yes	Included	Included
16	16			Yes	Not sure	Yes	Yes	Yes	Included	Included
17	25	NURSING	NURSING	Yes	Not sure	Yes	Yes	Yes	Included	Included
18	27			Yes	Not sure	Yes	Yes	Yes	Included	Included
19	55			Yes	Yes	Yes	Yes	Yes	Included	Included
20	54	SOCIAL SCIENCES - OTHER TOPICS	AL SCIENCES, INTERDISCIPLII	Yes	Not sure	Yes	Yes	Yes	Included	Included
21	21			Yes	Not sure	Yes	Yes	Yes	Included	Included
22	30			Yes	Not sure	Yes	Yes	Yes	Included	Included
23	148	PSYCHOLOGY	UCATIONAL; PSYCHOLOGY, M	Yes	Not sure	Yes	Yes	Yes	Included	Included
24	39	SCIENCES - OTHER TOPICS; PSYCHOLOGY; SPORT SC	URISM; PSYCHOLOGY, APPLIE	Yes	Not sure	Yes	Yes	Yes	Included	Included
25	116	PSYCHOLOGY; BUSINESS \& ECONOMICS	:HOLOGY, APPLIED; MANAGEN	Yes	Yes	Yes	Yes	Yes	Included	Included

Fig. 14. Overlook of the dataset.