Template for Front Hard cover Title page

Annex - A

EVIDENCE RETRIEVAL FOR EBM

MAAZ RAFIQ (FA21-MSCS-0003)

DEPARTMENT OF COMPUTER SCIENCE

FACULTY OF COMPUTING

MOHAMMAD ALI JINNAH UNIVERSITY

FALL 2023

EVIDENCE RETRIEVAL FOR EBM

SUBMITTED BY
MAAZ RAFIQ (FA21-MSCS-0003)

SUPERVISED BY
Dr. Shaukat Wasi

THESIS SUBMITTED TO THE [TYPE NAME OF THE FACULTY OF COMPUTING, MOHAMMAD ALI JINNAH UNIVERSITY, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

FALL 2023

CERTIFICATE OF APPROVAL

It is certified that the research work presented in this thesis, titled EVIDENCE RETRIEVAL FOR EBM was conducted by Maaz Rafiq under the supervision of Dr. Shaukat Wasi.

No part of this thesis has been submitted anywhere else for any other degree.

This thesis is submitted to the Department of Computer Science in partial fulfilment of the requirements for the degree of

Master of Sciences in Computer Science in FALL 2023
at the
Mohammad Ali Jinnah University
Karachi, Pakistan
January 25, 2023

Name of Candidate: _Maaz Rafiq_____ Signature: _____

Examination Committee Members:

a) Name of External Examiner: _____ Signature: _____
Designation, Affiliation:

b) Name of Internal Examiner: _____ Signature: _____
Designation, Affiliation:

c) Name of Supervisor: _____ Signature: _____
Designation, Affiliation:

_____
Name :
Dean, (Name of the Faculty) [for example (Faculty of

CERTIFICATE OF SUPERVISION

This is to certify that the thesis titled, "EVIDENCE RETRIEVAL FOR EBM", is submitted to the Department of Computer Science, Fall 2023 , by Maaz Rafiq for the award of the degree of Master of Science in the discipline of Computer Science. The thesis has been carried out under my supervision. I certify that the work submitted is original and not plagiarized from any other source, except as specified in the references. Neither the thesis nor the work contained therein has been previously submitted to any other institution for a degree.

DR. SHAUKAT WASI

MOHAMMAD ALI JINNAH UNIVERSITY
ORIGINAL LITERARY WORK DECLARATION

Faculty	*Faculty of Computing*	
Program	*Master of Sciences in Computer Science*	
Student Name: Maaz Rafiq		Reg. No: FA21-MSCS-0003
Email: fa21mscs0003@maju.edu.pk		Mobile No: 03357599280

Research Title:
__EVIDENCE RETRIEVAL FOR EBM_____

I do solemnly and sincerely declare that:

1. I am the author of this work.
2. This work is original.
3. Use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work; I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work.

4. I hereby assign all and every rights in the copyright to this work to Mohammad Ali Jinnah University (MAJU), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of MAJU having been first had and obtained.

5. I am fully aware that if in the course of making this work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by MAJU.

Student's Signature: _____ Place: Karachi-Pakistan.

Acknowledgements

I am thankful to Almighty for giving me the strength to work on a topic that I have never been through. I am deeply thankful to my supervisor, Dr. Shaukat Wasi, for his guidance, support, and knowledge sharing. I also extend a sincere and heartfelt obligation to all the medical experts without whom the completion of the project would not have been possible. It was hardly possible to proceed without them. I am thankful to my family for understanding my tough timelines and supporting me.

Abstract

Evidence-based medicine (EBM) is being used more frequently by medical professionals as a decision-making tool. Finding appropriate information manually is time-consuming and prone to errors because there is so much material online and it is expanding exponentially. The proposed method uses Information Retrieval (IR) with the Deep Learning (DL) model to implement quality retrieval of evidence. The formulated query given by the user will be searched in a database of medical data. The information will then be evaluated for quality and rated according to how relevant and strong the supporting evidence is. Our system's goal is to give medical professionals quick access to the most recent, evidence-based information about a patient's condition, so they can reduce the amount of time and effort they need to spend looking for relevant information while also improving the accuracy of what they find. By incorporating machine learning techniques into the EBM evidence retrieval process, we hope to help develop a decision support system for medical practitioners that is more effective and efficient.

Keywords: [Information Retrieval, Ranking, Machine Learning, Evidence Retrieval, Evidence-Based Medicine, Token Classification, EBM Implementation, Deep Learning, Information Retrieval]

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Evidence-Based Medicine	EBM
Critical Appraisal Exercise	CAE
BMJ Evidence-Based Medicine	BMJ EBM
Journal of Evidence-Based Medicine	JEBM
Low and Middle-Income Countries	LMICs
Population, Intervention, Comparator, and Outcome	PICO
Vector Space Model	VSM
Information Retrieval	IR
Natural Language Processing	NLP
Efficiently Learning an Encoder that Classifies Token Replacements Accurately	ELECTRA
Clinical Trials	CT
Randomized Clinical Trials	RCTs
Conditional Random Fields	CRFs
Convolutional Neural Networks	CNN
Neural Networks	NN
Bag Of Words	BOW
Patient, Intervention, Comparison, Outcome, System design	PICOS
Sample, Phenomenon of Interest, Design, Evaluation, Research type	SPIDER
Optical Character Recognition	OCR
Computer Vision	CV
Precision-Preferred Comprehensive Information Extraction System	PPCIES
National Council of Research	NCR
Named Entity Recognition	NER
Bidirectional Encoder Representations from Transformers	BERT

Machine Learning	ML
Inside, Outside, Begin, End, Single	IOBES
Begin, Inside	BI
Inside, Outside, Begin version-2	IOB2
Inside, Outside, End version-2	IOE2
Single, Begin, Middle, End	SBME
Begin, Inside, Last, Unit, Outside	BILUO
Inter Annotator Agreement	IAA

Evidence Retrieval for EBM

Maaz Rafiq

Mohammad Ali Jinnah University

Author Note

Department of Computer Science, 22-E, Block-6, P.E.C.H.S., Lal Kothi Stop,
Main Shahrah-e-Faisal, Karachi, 75400, Sindh, Pakistan.
Email ID: fa21mscs0003@maju.edu.pk

CHAPTER 1

INTRODUCTION AND MOTIVATION

Evidence-based medicine (EBM) is a way that doctors make decisions about your health by carefully considering the best and most reliable information from scientific research, combining it with their own experience, and considering what matters most to you as a patient (Rosenberg & Donald, 1995). It's like using a mix of the latest scientific findings, the doctor's know-how, and the patient's personal needs to ensure you get the best and most effective care.

To prescribe evidence-based medicine, medical practitioners need to follow the prescribed procedures of EBM, which require their valuable time and extra skill set (Suffian et al., 2018). There is a dire need to develop sophisticated EBM systems to help practitioners make effective decisions and prescriptions abruptly.

EBM involves using effective information search strategies to find reliable, current information from various sources and using extraction strategies to efficiently collect and analyze the retrieved information. This process includes a step called the Critical Appraisal Exercise (CAE) (Sarker et al., 2017). To achieve the CAE, we have to follow the following steps:

1. Identifying the specific issues or problems that a patient is experiencing and determining what evidence is necessary to address those issues.
2. Search the literature review.
3. Choosing the most reliable studies and using EBM guidelines to assess their validity.
4. Evaluating the quality of the evidence.
5. Extracting and synthesizing relevant evidence and using it to address the current problem or issue.

1.1 IMPORTANCE OF EBM

Evidence-Based Medicine (EBM) is important because it helps doctors make sure they're giving you the best possible care. Imagine that your health is like a puzzle, and EBM is like using the best pieces to solve it. It's like having a doctor who doesn't just guess what might work but looks at what science and research say works best. So, EBM helps doctors make smart choices based on what has been proven to be effective, keeping you safer and healthier. It's like using the best map to guide your journey to better health.

1.1.1 Background of EBM

1.1.1.1 BMJ Evidence-Based Medicine (BMJ EBM)

BMJ Evidence-Based Medicine (BMJ EBM) has been a leading force in the field of EBM since its launch in 1995. BMJ EBM is a pioneer in publishing "Clinical Evidence" summaries, concise overviews of the best evidence for key interventions, quickly becoming a trusted resource for clinicians at the point of care. BMJ EBM also features research and perspectives relevant to low- and middle-income countries, aiming to make EBM more accessible and applicable globally.

1.1.1.2 Journal of Evidence-Based Medicine (JEBM)

The Journal of Evidence-Based Medicine (JEBM) has been around since 2004, while not as long-standing but has carved its niche within the EBM landscape. JEBM stands out by concentrating on how Evidence-Based Medicine is used in China, making it different from Western-focused journals. Despite being new, JEBM has gained influence globally, with a rising impact factor and collaborations. It shares its research openly, reaching a broad audience and spreading knowledge in China and beyond.

1.1.1.3 World Congress on Evidence-Based Medicine (Barcelona, 2010)

The conference in Barcelona had a global focus on Evidence-Based Medicine (EBM), looking beyond developed countries and addressing challenges in low- and middle-income

countries (LMICs) with different healthcare systems and resource limitations. It emphasized bridging the gap between EBM research and practical application, particularly in LMICs, by discussing interventions and strategies.

1.1.2 EBM implementation approaches

1.1.2.1 Searching and Filtering Evidence

1.1.2.1.1 PICO Extraction

Automatic identification of key elements of research PICO (Population, Intervention, Comparator, Outcome) from texts like medical literature, blogs, articles, etc. makes it easier to find relevant studies.

1.1.2.1.2 Keyword Clustering

Grouping similar studies based on topic and keywords. This approach helps navigate the overwhelming medical research landscape.

1.1.2.1.3 Real-time Updates

Monitoring new publications to notify medical professionals of relevant findings, keeping them updated on the latest evidence.

1.1.2.2 Understanding and Analyzing Evidence

1.1.2.2.1 Evidence Synthesis

Automatically summarize large sets of research, providing concise overviews of the best available evidence.

1.1.2.2.2 Risk & Benefit Assessment

Analyzing data to predict potential outcomes of treatments based on patient characteristics and study findings.

1.1.2.2.3 Clinical Decision Support

Recommend evidence-based treatment options tailored to individual patients, considering their specific situation.

1.1.2.3 Applying and Monitoring Evidence

1.1.2.3.1 Clinical Documentation Analysis

Extracting information from patient records to assess adherence to EBM guidelines and identify areas for improvement.

1.1.2.3.2 Patient Education

Generating personalized reports Summarizing relevant evidence helps patients understand their treatment options.

1.1.2.3.3 Outcome Tracking and Feedback

Monitoring the effectiveness of treatments based on real-world data informs continuous improvement in healthcare practices.

1.2 EVIDENCE RETRIEVAL

Evidence retrieval in healthcare is like searching for the best information to help doctors make good decisions about your health. It's like looking for clues or answers in a vast library of medical knowledge. This way, your healthcare team can use the most reliable and up-to-date evidence to give you the best care possible.

1.3 PICO

PICO (Schiavenato & Chu, 2021) is a helpful tool that doctors use to make decisions about your health. It's like a formula that helps them ask the right questions.

- P stands for "patient" or "problem." It's about understanding who the patient is and what health issue they have.
- I stands for "Intervention": This is about the treatment or action the doctor is thinking of using.
- C stands for "Comparison": Sometimes, the doctor compares different treatments to see which one might be better.

- O stands for "Outcome": Doctors want to know what happens after the treatment, like if the patient gets better or if there are any side effects.

We use PICO to frame our questions because it helps us find the most relevant and useful information when searching for evidence. Breaking down the question into these parts guides us in searching for studies that match what we're specifically interested in, making our evidence-based decisions more precise and effective. It's like having a roadmap for finding the right answers in the vast world of medical information.

1.3.1 PICO Alternatives

Many alternatives to the PICO are available, among which two popular alternatives are Important named Sample, Phenomenon of Interest, Design, Evaluation, Research type (SPIDER) and Patient, Intervention, Comparison, Outcome, and System design (PICOS).

(Methley et al., 2014) performs analysis on these three tools. According to the study, SPIDER is focused only on highly relevant articles, potentially missing valuable research. It also reduces irrelevant papers but risks missing relevant ones. PICOS, on the other hand, offers decent sensitivity with some improvement in specificity but has high number of irrelevancies.

1.4 DRAWBACKS OF NON-EBM SYSTEMS

Non-EBM systems, or systems that don't use evidence-based medicine, have some drawbacks. One big issue is that decisions might be based more on tradition or what people have always done than on the most current and proven information. This can mean patients might not get the best and safest care. Also, without using evidence, there's a risk of relying too much on personal opinions or beliefs, which may not always lead to the most effective treatments. Evidence-based medicine helps make sure decisions are based on what has been proven to work, making healthcare more reliable and safer.

1.5 CHALLENGES IN EBM SYSTEMS

While EBM is valuable, it does face some challenges. One challenge is that finding good evidence takes time, and in real-life healthcare, decisions often need to be made quickly. Also, not all health issues have a lot of high-quality evidence available. Sometimes, doctors need to rely on their experience and judgment. Additionally, each patient is unique, and applying evidence to individual cases can be tricky. Despite these challenges, EBM remains crucial for making informed and effective healthcare decisions based on the best available evidence.

1.6 RESEARCH OBJECTIVES

This study proposes the development of an Evidence Retrieval system for EBM to contribute to the medical domain by using technology, such as improving retrieval of evidence and efficiently synthesizing evidence. Furthermore, to develop datasets of PICO and retrieval queries,

1.7 RESEARCH CONTRIBUTIONS

This research work has made the following contributions to the EBM community:

1. Proposing the framework for Evidence Retrieval
2. Token classification dataset of PICO labels in COVID Clinical Trials (CT) literature.
3. Information Retrieval dataset of COVID-19 CT literature.
4. Proposing algorithm for calculating the quality of paper.
5. Proposing an algorithm for calculating the quality of the author.

CHAPTER 2

LITERATURE REVIEW

In this extensive literature review, various types of research are explored, each providing insights into different aspects of evidence-based medicine and its practical application in the medical field. The focus of these studies is on incorporating modern artificial intelligence (AI) methods, deep learning algorithms, information extraction systems, and argument mining techniques. Their common objective is to transform the identification and analysis of crucial information in medical literature with the ultimate aim of enhancing medical decisions, treatment planning, and patient outcomes.

These studies significantly enhance evidence-based medicine by harnessing the capabilities of AI, deep learning, information extraction, and argument mining. They specifically address areas such as medical decision-making, information retrieval, and evidence synthesis, presenting compelling findings and valuable insights. The research demonstrates promising outcomes that have the potential to bring about substantial changes in healthcare practices through the application of these advanced methods.

It is important to note that while these studies show great promise, further research and development are necessary to address current limitations and enhance these methods for practical use in real-world healthcare applications.

2.1 A RELEVANCE & QUALITY-BASED RANKING ALGORITHM FOR EBM

In the vast domain of medical information, finding the most useful and reliable data can be a tough job for doctors who follow EBM. The usual ways of searching for information mainly focus on whether the text is relevant, but they often struggle to figure out if the evidence is really good. To tackle this problem, (Serrano-Guerrero et al., 2020) suggest a new ranking algorithm designed specifically for EBM searches.

The study introduces a two-step approach. First, it checks the content of documents to see if they are closely related to what the user is looking for. This involves using traditional methods to catch important medical concepts and terms. Second, the algorithm carefully evaluates the quality of evidence in each document. It looks at things like how well the study was done, the design of the study, and how strong the conclusions are. Metrics such as the type of study, the number of participants, and how the findings are reported are considered to make sure that only the most reliable and strong evidence is given top priority.

The authors bring these two aspects together to create a single ranking score. Documents that are both highly relevant and have strong evidence are placed at the top, while those lacking in either aspect are pushed down the list. This ensures that doctors not only get information directly related to their question but also receive the most trustworthy and clinically sound research.

To test the effectiveness of this algorithm, experiments were conducted using the Medline database, and the Cochrane Library was used as a benchmark. The results show that this new approach performs much better than traditional methods in finding high-quality and relevant EBM literature. This opens the door for exciting search tools that can help doctors make well-informed, evidence-based decisions more confidently and quickly.

2.2 COVID-19 IR WITH SEMANTIC SEARCH, QA, & SUMMARIZATION

The COVID-19 pandemic has led to an overwhelming increase in scientific publications, creating a strong need for quick and accurate information retrieval. (Esteva et al., 2021) tackle this challenge by introducing CO-Search, a new system designed to navigate the swiftly evolving COVID-19 literature landscape.

CO-Search goes beyond traditional keyword-based retrieval models by using deep learning. Its main feature is a semantic search approach, employing Siamese-BERT to understand the meaning of both queries and documents. This results in more relevant and nuanced results, going beyond simple keyword matching and capturing the underlying intent and context of information needs.

CO-Search offers a question-answering capability. By incorporating a multi-hop question-answering module, the system can directly address specific questions and extract precise answers from relevant documents. This eliminates the need for users to go through entire texts, providing a faster and more targeted information access experience.

To improve user experience further, CO-Search includes an abstractive summarization module. This creates concise summaries of retrieved documents, highlighting key points and allowing users to quickly understand essential information without delving into the full text. This is particularly valuable for busy researchers and clinicians who need to quickly synthesize large volumes of research.

The effectiveness of CO-Search was thoroughly evaluated using the TREC-COVID information retrieval challenge data. The system demonstrated significant improvements over traditional methods, showing superior performance in retrieving relevant and informative documents. This highlights the potential of CO-Search to empower researchers, clinicians, and anyone seeking reliable information about COVID-19, ultimately facilitating informed decision-making and accelerating scientific progress in this critical field.

2.3 ADVANCING EBM VIA TWO AUTOMATIC IDENTIFICATION OF PICO

Evidence-based medicine (EBM) relies on finding and evaluating relevant research, but dealing with complex medical literature can be challenging due to its sheer volume and intricate language. To tackle this, (Stylianou et al., 2020) propose EBM+, a new system designed to automatically extract essential information from medical studies, with a specific focus on the PICO framework - Population, Intervention, Comparison, and Outcome.

EBM+ works in two steps. First, it uses a PICO entity recognizer based on Convolutional neural networks (CNNs). These CNNs analyze text features to identify and classify entities representing populations, interventions, comparisons, and outcomes within the document. This helps streamline the initial process of gathering information by automatically extracting the core elements of the study.

EBM+ also introduces a PICO statement classifier. This classifier goes beyond just recognizing entities; it analyzes the extracted entities within the context of entire sentences. It identifies sentences that not only contain all PICO elements but also answer specific questions posed in PICO format. This advanced capability allows users to directly ask the system clinical questions and receive concise, PICO-structured answers extracted from relevant studies.

(Stylianou et al., 2020) demonstrate the effectiveness of EBM+ by evaluating its performance on a dataset of manually annotated medical abstracts. The system achieves top-notch accuracy in both PICO entity recognition and PICO statement classification, highlighting its potential to significantly enhance the efficiency and accuracy of EBM practice.

2.4 RELEVANCE-BASED AUTHOR RANKING ALGORITHM FOR PUBLICATION VENUES

Traditionally, metrics like the h-index for evaluating authors focus on their overall impact, often overlooking how relevant they are to specific publication venues. This can be a challenge for editors and conference organizers trying to identify scholars who are not only impactful globally but also specifically relevant to their communities. To address this issue, (Zhang et al., 2023) introduced RelRank, a new author ranking algorithm designed specifically for individual publication venues.

RelRank's key strength lies in its relevance-based co-authorship network. This network goes beyond traditional co-authorship analysis by including an author relevance factor. This factor measures the statistical connection between an author and a particular venue, considering factors like the proportion of the author's publications within that venue compared to others. This nuanced approach enables RelRank to capture how committed and dedicated an author is to a specific scientific community.

Using this enriched network, RelRank utilizes an improved version of PageRank. This modified algorithm includes the author relevance factor in the ranking process, ensuring that authors with stronger connections to the specific venue receive higher scores. This prioritizes scholars who are not only prolific but also deeply engaged with the research interests and communities of the venue.

RelRank's effectiveness was thoroughly evaluated on three datasets covering different research domains. In comparison to traditional author-level metrics, RelRank consistently showed better performance in identifying authors who were both highly cited and relevant to specific venues.

2.5 END-TO-END TRANSFORMERS FOR EBM

The field of medical research produces a huge amount of written information, providing opportunities and challenges for clinicians looking for the most relevant and trustworthy data. (Stylianou & Vlahavas, 2021) take a significant step in closing this gap by combining two important tasks: evidence-based medicine (EBM) and argument mining.

EBM focuses on finding and evaluating the best evidence to guide clinical decision-making. This often involves extracting key elements like population, intervention, comparator, and outcome (PICO) from medical studies. On the other hand, argument mining examines the structure of arguments in the text, identifying claims, premises, and their relationships.

TransforMED creatively brings together these two areas using advanced end-to-end transformer models designed specifically for medical literature. These models, known for their ability to understand complex contextual relationships within text, offer several advantages:

Joint PICO and Argument Extraction: TransforMED improves on traditional methods by predicting PICO entities and analyzing argumentative relationships simultaneously in a single model. This approach ensures a more comprehensive understanding of medical literature, helping users identify relevant evidence and understand the underlying rationale and potential counter-arguments.

Enhanced Performance: By combining EBM and argument mining tasks, TransforMED improves the accuracy of both models. Extracting PICO entities provides valuable context for argument analysis, while understanding argumentative structures can refine the identification of relevant evidence.

2.6 COMPREHENSIVE INFORMATION EXTRACTION SYSTEM IN TCM

Research in Traditional Chinese Medicine (TCM) is growing, but extracting important information from its clinical articles can be challenging. The complex language, varied formats, and non-standardized data create a barrier for researchers and practitioners trying to tap into the wisdom of TCM evidence. To tackle this challenge, (Xia et al., 2022) introduce the Precision-Preferred Comprehensive Information Extraction System (PPCIES), a revolutionary system for unlocking the secrets hidden in TCM clinical articles.

PPCIES is built on precision and comprehensiveness. Its main strength is a hybrid text processing approach, using Computer Vision (CV) and Optical Character Recognition (OCR) to handle the diverse layouts and scanned PDFs commonly found in TCM literature. This ensures accurate text extraction, regardless of the format, paving the way for further analysis.

PPCIES also employs rule-based extraction, a meticulous approach where hand-crafted rules tailored to TCM language and document structure guide information extraction. This ensures high precision, minimizing the risk of misinterpretations and inaccuracies. Unlike generic systems, PPCIES understands the nuances of TCM language and document structure, resulting in reliable and trustworthy information extraction.

However, PPCIES aims for more than just precision; it aims for comprehensiveness. It focuses on extracting a wide range of information relevant to TCM clinical trials. From patient demographics and interventions to comparisons, outcomes, and study design details, PPCIES covers 34 target fields across 14 crucial topics. This comprehensive approach allows researchers and practitioners to gain a holistic understanding of TCM research, facilitating informed decision-making and knowledge synthesis.

The potential of PPCIES is immense. By making key information easily accessible, it can empower researchers to conduct more efficient and insightful studies, accelerating the advancement of TCM knowledge.

2.7 ANNOTATED CORPUS OF CT PUBLICATIONS

Navigating the vast ocean of clinical trial publications can be daunting. Buried within abstracts and reports lie the crucial details of interventions, outcomes, and intricate relationships between them. Extracting this knowledge effectively requires tools that go beyond simple keyword matching. This is where (Sanchez-Graillet et al., 2022) enter with their annotated corpus and schema-based relational information extraction approach.

The study presents a meticulously curated dataset of 1,750 annotated clinical trial abstracts. Each abstract has been tagged with the key entities involved, ranging from medications and diseases to study participants and outcomes. But (Sanchez-Graillet et al., 2022) go the extra mile. They don't stop at identification; they unveil the connections.

Their work hinges on a predefined schema, the C-TrO ontology, which acts as a map capturing the intricate relationships between entities in clinical trials. This allows the authors to annotate not just the "who" and "what" but also the "how" and "why," revealing the complex interplay between interventions, diseases, and outcomes.

The subset of the corpus boasts sentiment labels, adding a nuanced layer of understanding. These labels identify positive, negative, or neutral opinions and conclusions expressed within the text, providing valuable insights into the authors' interpretations and potential biases.

The potential of this annotated trove is immense. It serves as a fertile training ground for developing powerful information extraction models. These models, fueled by the rich annotations, can extract not just isolated facts but also the meaningful connections and sentiment embedded within clinical trial publications.

2.8 TREC-COVID

The COVID-19 pandemic unleashed a torrent of scientific research, creating both a treasure trove of knowledge and a daunting information maze for researchers seeking answers. Amidst this deluge, (Voorhees et al., 2020) launched a groundbreaking initiative - TREC-COVID: a dynamic test collection specifically designed to guide the development of effective information retrieval (IR) systems in pandemic scenarios.

TREC-COVID's strength lies in its real-world focus. Forget generic queries; its topics stem from the actual information needs of researchers grappling with the pandemic. This ensures retrieved information directly addresses the pressing questions facing the scientific community. But pandemics evolve, and so does TREC-COVID. Unlike static test collections, it embraces a rolling release approach, continuously adding new topics as research priorities shift. This ensures constant relevance, keeping pace with the pandemic's unpredictable journey.

TREC-COVID goes beyond simple document retrieval. It presents a multifaceted IR challenge, encompassing tasks like question answering, topic search, and summarization. This comprehensive evaluation framework pushes the boundaries of IR, demanding systems that not only find relevant information but also synthesize and interpret it for researchers.

The test collection itself boasts several key components:

Diverse topics: From vaccine development to social impacts, TREC-COVID delves into various aspects of the pandemic, providing a holistic picture of information needs.

Vast document corpus: From research papers to news reports, the collection mirrors the real-world information landscape researchers navigate, ensuring systems are tested under realistic conditions.

2.9 SEMI-AUTOMATIC SLRs AND IE OF COVID-19 SCIENTIFIC EVIDENCE

The COVID-19 pandemic erupted like a volcanic fissure, spewing forth an overwhelming torrent of scientific research. Researchers found themselves adrift in a sea of information, their boats swamped by waves of data. It was a crisis of clarity, a battle against the very deluge of knowledge meant to save us.

COKE Project (Golinelli et al., 2022), a collaborative beacon of hope from the Italian National Council of Research (CNR) and the University of Bologna. It throws researchers a life raft, not of wood and rope, but of intelligent automation and insightful extraction. COKE offers a two-pronged approach to navigate this perilous research ocean. It acts as a skilled captain, charting a course through the churning waters with powerful language processing tools. Forget about fishing through murky keyword searches. COKE dives deeper, understanding the very essence of your research question through the PICO elements – Population, Intervention, Comparator, and Outcome. With this compass, it pinpoints studies directly relevant to your needs, saving you the Sisyphean task of endless reading.

COKE doesn't simply stop at identifying the right islands. It then maps the entire research landscape for you, grouping studies by their similarities and building a clear navigational chart. No more wading through endless, isolated articles. COKE presents you with a holistic picture of the relevant research terrain, allowing you to quickly grasp the lay of the scientific land. Once anchored in promising studies, COKE transforms into a meticulous treasure hunter, its machine learning and reading abilities unearthing the buried gems of knowledge within. It gathers crucial details like study design, interventions, outcomes, and findings, leaving no valuable stone unturned.

These gems are then strung together, forming interconnected knowledge graphs that unveil the hidden connections between different studies. Think of it as a luminous tapestry woven from threads of research, revealing the patterns and relationships that would otherwise

remain invisible. With this newfound clarity, researchers can finally see the full picture, gaining a deeper understanding of the scientific landscape and the potential answers hidden within. COKE's impact extends beyond the COVID-19. Its adaptable framework can be applied to diverse research domains, acting as a powerful tool to accelerate scientific discovery, improve evidence-based decision-making, and foster collaboration across disciplines. Imagine breakthroughs in medicine, technology, and countless other fields, all fueled by the streamlined navigation and insightful knowledge extraction that COKE offers.

The COKE Project stands as a testament to human ingenuity in the face of overwhelming information. It's a life raft in the research ocean, guiding us towards a future where knowledge is not a burden but a powerful tool for progress and understanding. As COKE continues to evolve, its ripples of impact will spread across the scientific landscape, ensuring that even amidst the stormiest seas of data, researchers can always find their way to shore.

2.10 SENT2SPAN: SPAN DETECTION FOR PICO EXTRACTION

Imagine a detective meticulously sifting through mountains of text, searching for the crucial clues to solve a medical mystery. That's essentially the task of PICO extraction in the realm of biomedical research, identifying the key elements Population, Intervention, Comparator, and Outcome that hold the secrets to understanding treatments and diseases. Traditionally, this "literary sleuthing" relied heavily on painstaking manual annotation of specific textual spans within research articles, a time-consuming and expensive endeavor. Enter Sent2Span (Liu et al., 2021), a groundbreaking approach that rewrites the rules of the game.

Sent2Span's brilliance lies in its reduced reliance on manual annotations, opting instead for sentence-level clues. Think of it as the detective focusing on specific paragraphs instead of poring over every word. These simpler annotations, indicating the presence or absence of PICO elements within a sentence, significantly ease the burden on human annotators. It employs a two-stage process to crack the PICO code 1) identifying candidate sentences; 2) Masked Span

Inference. This masked inference process effectively predicts the missing pieces of the PICO puzzle within each sentence, painting a clearer picture of the research at hand.

2.11 STEP-WISE MEDICAL NER IDENTIFICATION

Unearthing Medical Gems: Stepwise NER Unveils the PICO Treasure Trove

Imagine yourself a prospector (Zhang et al., 2020), panning through mountains of medical research papers, seeking nuggets of PICO gold: Population, Intervention, Comparison, and Outcome. These crucial elements hold the key to understanding treatments, diseases, and ultimately, improving healthcare. Researchers have relied on Named Entity Recognition (NER) tools to sift through text, haphazardly picking up these nuggets. But just like panning might miss buried treasures, traditional NER techniques often struggle with the complexities of medical language, leaving crucial PICO gems undiscovered.

This is where the "stepwise medical NER identification" method, introduced in a recent paper, offers a more refined approach. Think of it as a meticulously crafted sieve, designed to efficiently and accurately capture those gleaming PICO nuggets. The method works in three stages:

1. Classifies sentences into different PICO categories
2. NER techniques to identify and classify disease entities within each PICO-classified sentence
3. Mapping

This stepwise approach boasts several advantages like accuracy and etc.

2.12 PICO ENTITY EXTRACTION OF ANIMAL LITERATURE

Preclinical PICO (Wang et al., 2022), Unveiling Animal Lab Secrets with Bidirectional Encoder Representations from Transformers (BERT) Imagine delving into a maze of preclinical animal research papers, searching for the hidden gems of PICO - the Population, Intervention, Comparator, and Outcome that hold the keys to translating animal insights into clinical breakthroughs. Traditionally, this has been a laborious task, with clunky tools struggling to navigate the complex language and varied designs of animal studies.

BERT, a powerful language model, is here to illuminate the PICO path in the preclinical research labyrinth! The study introduces a novel approach that leverages BERT's prowess to efficiently and accurately unearth these crucial elements. Think of BERT as a skilled guide, trained on a vast library of biomedical texts. It understands the intricacies of scientific language and the unique features of preclinical research reports. With this expertise, BERT tackles the PICO puzzle in three key steps 1) Contextual Comprehension; 2) Entity Spotlight; 3) PICO Classification. BERT-powered approach offers several advantages over traditional methods.

CHAPTER 3

METHODOLOGY

3.1 CORPUS GENERATION

A Gold standard dataset plays a vital role in any domain for Evidence extraction research. Good evidence-extraction research depends on the availability of benchmark datasets. All supervised models require labeled training data to learn patterns. Hence, it is necessary to put labeled data in bulk quantity for training. However, COVID-19, despite being an impactful domain, has no PICO corpora publicly available for evidence extraction.

Keeping in view the importance, we first developed a gold standard PICO dataset for CT or Randomized Clinical Trials (RCT) in COVID-19 literature. By studying literature related to the PICO dataset on medical literature, we concluded that we should work on the abstract of the literature.

3.1.1 Data Collection

The corpus was originally curated by the Semantic Scholar team at the Allen Institute for AI (Wang et al., 2020). The corpus contains literature regarding COVID-19 until 2020. The corpus was curated iteratively in 5 rounds, starting from 2020-03-13 to 2022-06-02. We have worked on the final release of this data, which contains around 20,000 pieces of literature consisting of full-text and abridged literature and multiple subdomains of COVID-19, along with metadata like doi, authors, journal, published year, etc. The corpus also contains the embeddings of each abstract; however, the method of generating embeddings is not defined.

3.1.2 Data Scrapping

We also perform scrapping to collect some metadata related to the corpus for evaluating the quality of the study. The scrapped sources are 1) Scimago Journal Ranking (SJR) and 2) Google Scholar.

3.1.2.1 Metadata for Literature

To evaluate the quality of the literature, we required the following parameters:

1. Total citations
2. SJR rank
3. SJR quartile

In the above parameters, citations are scrapped from Google Scholar, and other parameters are scrapped from SJR.

3.1.2.2 Metadata for author

For the evaluation of the author, we required the following parameters:

1. Total citations
2. H-index

Both parameters are scraped from Google Scholar

3.1.3 Data Pre-Processing

As we mentioned the diversity of corpus in terms of subdomains in Section 3.2, we perform multilevel preprocessing. All strategies and their execution were performed with the collaboration of the domain expert team to minimize errors.

3.1.3.1 Preprocessing on subdomain

The scope of this step is to include only those samples in our dataset that are related to the CT of COVID-19. After brainstorming with the domain expert team, we perform this process by searching for a special Bag Of Words (BOW) in the full text of the literature. After this process, the domain experts team skims the whole filtered dataset to detect any false positives. This process reduces the dataset size from around 20,000 to 7,000.

3.1.3.2 Preprocessing on Abstracts

The goal of this step was to exclude all the literature that either doesn't have an abstract or has an abstract other than text format. The result of this step was the reduction of around 1000 samples from the dataset. This step was initiated for two reasons:

1. The document retrieval and evidence extraction mechanisms were based on the abstract of the study.
2. We found some studies that have abstracts in the form of podcasts and pictorials.

3.1.3.3 Preprocessing on PICO

This step aims to remove abstracts that don't have a single label for PICO. This step was to cover the scenario that the study is related to clinical trials, but it does not perform any clinical trials but surveys the previous studies or any article, etc. This step would help us exclude around 2000 samples from our dataset

3.1.3.4 Preprocessing on Digital Object Identifier (DOI)

In this step, we filtered out the literature studies that had invalid Digital Object Identifier (DOI). The need for this step under our scope of work is that we are scraping the metadata for evaluating the quality of literature and author based on literature DOI.

3.1.3.5 Preprocessing the Author

In author preprocessing, we exclude literature whose author's data is not available on Google Scholar. Filtering out this type of literature is necessary because no author data will result in an empty list after the calculation of author quality.

3.1.4 Data Annotations

3.1.4.1 Token classification tagging scheme

Feasible tagging of training data is crucial requirement of most Machine Learning (ML) algorithms (Keretna et al., 2015) which significantly affects the results of the algorithm. Tagging schemes are used for the identification of the boundaries of named entities, which

consist of multiple tokens. Such entities contain a beginning token and an ending token. The most widely used schemes are IOBES (Inside, Outside, Begin, End, Single), BI (Begin, Inside) IOB2 (Inside, Outside, Begin version-2), IOE2 (Inside, Outside, End version-2), SBME (Single, Begin, Middle, End) and BILUO (Begin, Inside, Last, Unit, Outside) (Keretna et al., 2015).

From the literature, we found that BI technique was good for identification of PICO boundaries.

Table 1

BI scheme labels

B-P	Beginning of Patient
I-P	Inside of Patient
B-I	Beginning of Intervention
I-I	Inside of Patient
B-C	Beginning of Comparision
I-C	Inside of Comparision
B-O	Beginning of Outcome
I-O	Inside of Outcome
O	Tokens other than PICO

3.1.4.2 Inter Annotator Agreement (IAA)

For the sake of the authenticity of our dataset, the same tokens were annotated by two different groups of annotators. Inter-annotator agreement is the degree of agreement between more than one annottor about how they agree on a certain category/class. It measures how the annotators understand a certain category (Cohn et al., 2008).

3.1.4.2.1 Cohen's Kappa coefficient

The most widely used statistic to measure the reliability of categorical data annotated by two different annotators is Cohen's Kappa coefficient, which measures agreement reliability according to the formula:

$$Kappa(k) = \frac{P_o - P_e}{1 - P_e}$$

k: possibility of the agreement occurring by chance

P_o: relative observed agreement among raters

P_e: hypothetical probability of chance agreement (Cohn et al., 2008)

Table 2

Kappa Score

S.No	Tag	Kappa Coefficient Score
1	PICO	0.98
2	BI	0.98

Based on kappa score obtained of *0.98*, our two annotators are in almost perfect agreement for both PICO and BI labels assigned to a token.

3.2 QUERY GENERATION

As the needs of our underlying problem were mentioned earlier in *3.1.3*, we filtered the original dataset according to those needs. This filtration process raised another problem with queries for retrieving the filtered dataset. Because the queries mentioned in the original study are unrelated to clinical trials, we created our own queries. This process was carefully done and evaluated by a team of medical experts.

3.2.1 Topic Modeling

To generate the queries, it is necessary to explore the corpus in terms of topics and sub-topics and map each sample under the topic or sub-topic. To achieve this, we use BERTopic architecture, which is very popular for topic modeling.

We use BERTopic in the default configuration, except for a few. We changed the embedding model and used two BERT-based pre-trained models alternatively: 1) PubMed BERT; and 2) Clinical BERT. Both models were trained on medical literature, including clinical trials, but pre-trained on COVID-19 data was not sure as it was not mentioned by the publisher explicitly. The major difference between the two models was the output embedding size, which was 786 and 1024, respectively. We also change the topic representation model and perform multiple experiments in combination with preceding models in the architecture.

The process of topic modeling was iterative, with the combination of the *Identification of Hierarchies* process, which will be discussed in the next section. The reason to perform topic modeling iteratively and with the combination of *Identification of Hierarchies* is that topic modeling work is an unsupervised approach, which means high chances of errors. This approach also works as an assistant for the domain expert teams. We apply Topic modeling to the whole data at once but select only those topics for hierarchy identification that have high confidence scores.

The below two figures show the intertopic distance between topics extracted by two different embedding models Clinical BERT and PubMed BERT. The circle shows the topic, and the circle within the circle shows hierarchical topics extracted by the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering algorithm (Campello et al., 2013).

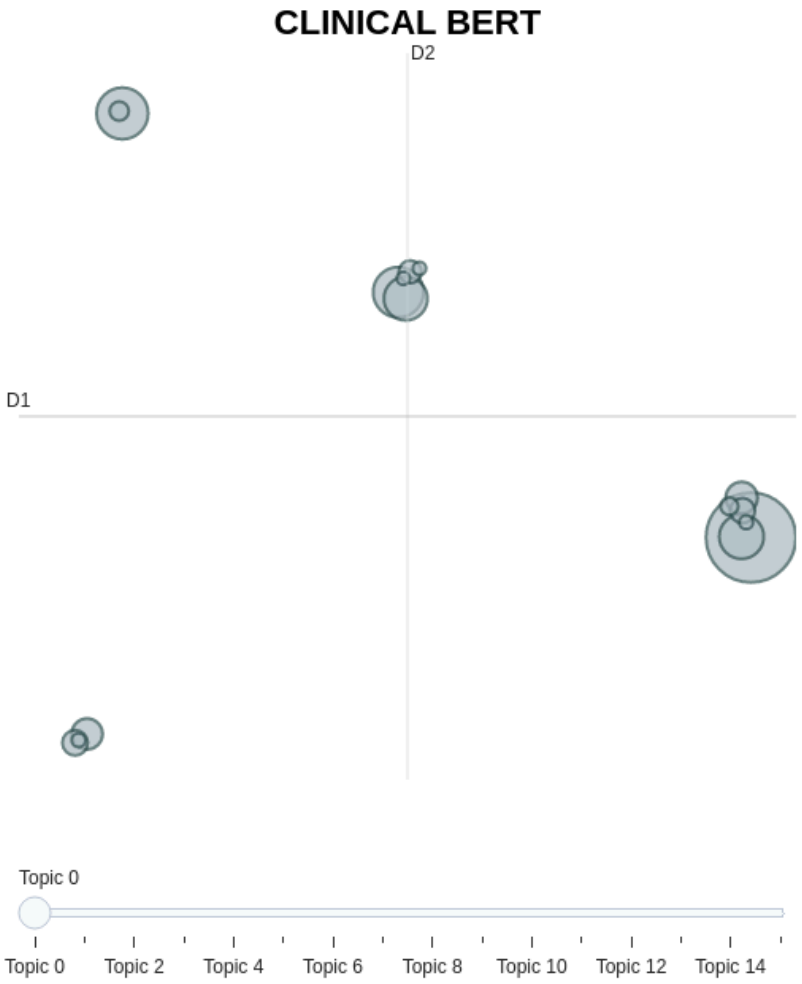

Figure 1. Intertopic Distance Map (Clinical BERT)

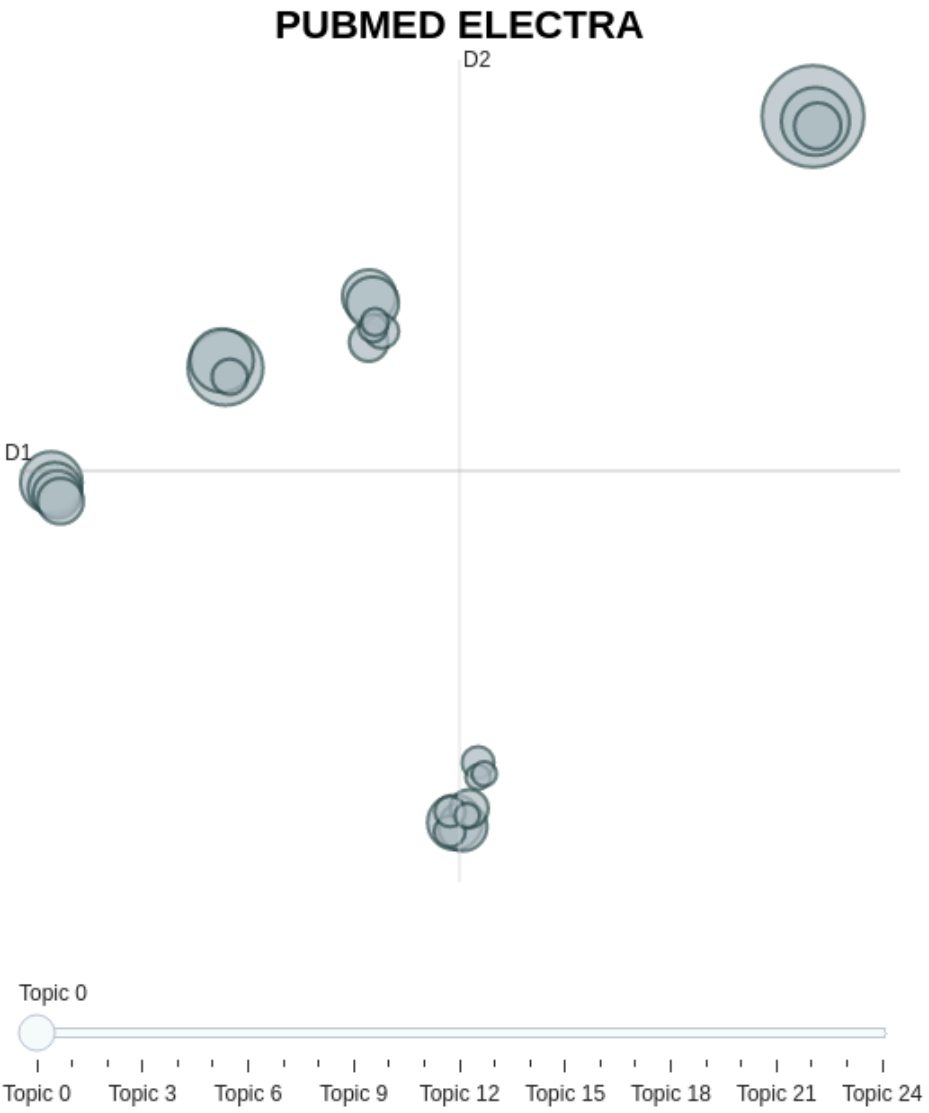

Figure 2. Intertopic Distance Map (PubMed BERT)

3.2.2 Identification of Hierarchies

Identification of hierarchies takes place as a subsequent step after topic modeling, as mentioned in *Section 3.1.4*. Domain experts take topics with marked literature and evaluate both topic names and their relevant literature. This process was done by two sub-teams of medical domains in parallel to avoid biases. The process of *Topic Modeling* with the *Identification of*

Hierarchies continued until every sample was marked under at least one topic. The process was very time-consuming because the topic representation done by BERTopic was not good and was changed by domain experts on every topic. Also, the clustering of the same literature was not up to mark.

Immunostimulatory	Vaccination	Nasal Administration		
		Subcutaneous Application		
		Oral		
		Virus-like particles (VLPs)		
		Edible		
		Carbohydrates		
		Plant-Based Vaccines		
	Others			
	Antimicrobial	Antimicrobial	Aminoglycosides	
			Prophylaxis	
			Macrolides	
		Antimalarial		
		Antiretroviral Drugs		
		Antiviral		
	Anticancer			
	antagomiR			
	Antibodies (ABs)	Monoclonal Antibodies (mAB)		
		broadly neutralizing antibodies (bnAbs)		
	Protease Inhibitor			
	Probiotics			
	Anti-inflammatory	Statins		
		Microsomal prostaglandin E2 synthase-1 (mPGES-1)		
		Corticosteriods		
	Anti-hypertensives	ARBs		
		ACE Inhibitors		
	Herpatofauna			
	Biotechnological Substances (BSs)			
	Marine-derived bacteria			
	Traditional Chinese Medicine (TCM)			
	Oncolytic Virotherapy			
	Phytopharmaceutical			
Ventilation	spontaneous breathing trial (SBT)			
	extracorporeal membrane oxygenation (ECMO) airway pressure release ventilation (APRV) Standard Oxygen Therapy neurally adjusted ventilatory assist (NAVA)			
	Lung Protective ventilation			
	Noninvasive Ventillation	High Flow Nasal Canula (HFNC)		
		Positive Pressure Ventilation		
Respiratory Viral infections	Corona	SARS		
		COVID-19	COVID-19	COVID-19
		MERS		
	Parainfluenza Virus			
	Human Metapneumovirus (HMPV)			
	Bocavirus			
	Picornaviruses	EnteroViruses	EV71	
		RhinoVirus		
	Pneumonia	Healthcare acquired pneumonia		
		Pneumocystis jirovecii pneumonia (PJP)		
		Pneumococcal Pneumonia		
		Community acquired pneumonia (CAP)		
		Hospital acquired Pneumonia		
		Ventilator associated pneumonia (VAP)		

Figure 3. Hierarchial Topics

The above *figure* is the sample topic hierarchy of our dataset. The hierarchy sequence is row wise from left to right. After the hierarchical mapping of topics in the abstracts, we collaborated with the field experts to carefully draft and evaluate the queries. We then drafted a

set of queries against each sub-topic, based on the abstracts. Half of the queries were drafted by one group of field experts and evaluated by another group, and vice versa. Set of formulated queries is presented in the *table*.

Table 3

Generated Queries

1. Efficacy of medical masks and respirators in protecting against respiratory infections in Healthcare Workers.
2. Studies to identify different aetiologies of Acute Otitis Media (OAM) in children.
3. Infections, CAR T cell therapy, intensive cytotoxic therapy in infants, and reduced-intensity therapy with decitabine.
4. Efficacy of nasal administration of mucosal Vaccine.
5. Efficacy of VLPs-based Vaccines on cancer, allergic and infectious diseases.
6. Changes in mortality and morbidity caused by different antibiotics on multidrug-resistant organisms.
7. Are natural products derived from Marine-sponges effective in treating neoplasm, infectious or autoimmune diseases?
8. Active ingredients of Traditional Chinese Medicine (TCM) to be used in medicine.
9. Utilization of Extracorporeal Membrane Oxygenation (ECMO) for Acute Respiratory Failure.
10. Effect of NonInvasive Ventilation (NIV) in patients with Acute Respiratory Failure or Acute Respiratory Distress Syndrome (ARDS).
11. Prevalence of different Respiratory virus infection (RVI) in hospitalized patients.

3.3 PROPOSED FRAMEWORK

From the literature review, we concluded that the implementation of EBM in the domain of COVID-19 does not exist. Another problem with current EBM implementations is that a lot of studies have done CAE steps individually, like implementation of retrieving relevant studies, evaluation of literature quality, extraction of evidence, etc., but not the implementation of an

end-to-end pipeline. A pipeline that takes queries from practitioners and provides the extraction of evidence from literature studies as a result, followed by the process of retrieving relevant studies and the evaluation of the literature quality. Keeping the discussed issue in mind, working on this framework is appropriate for our research work.

Our proposed framework derives from the methodology of (Serrano-Guerrero et al., 2020) with major changes and additional work. The additional work after this methodology is to incorporate the evidence-extraction process. The changes made in methodology will be discussed in the upcoming sections.

Our framework consists of 1) *relevance ranking*; 2) *quality evaluation*; 3) *fusion*; and 4) *evidence-synthesis* components. The workflow of the framework is depicted in *figure* with a description of each component in upcoming sections.

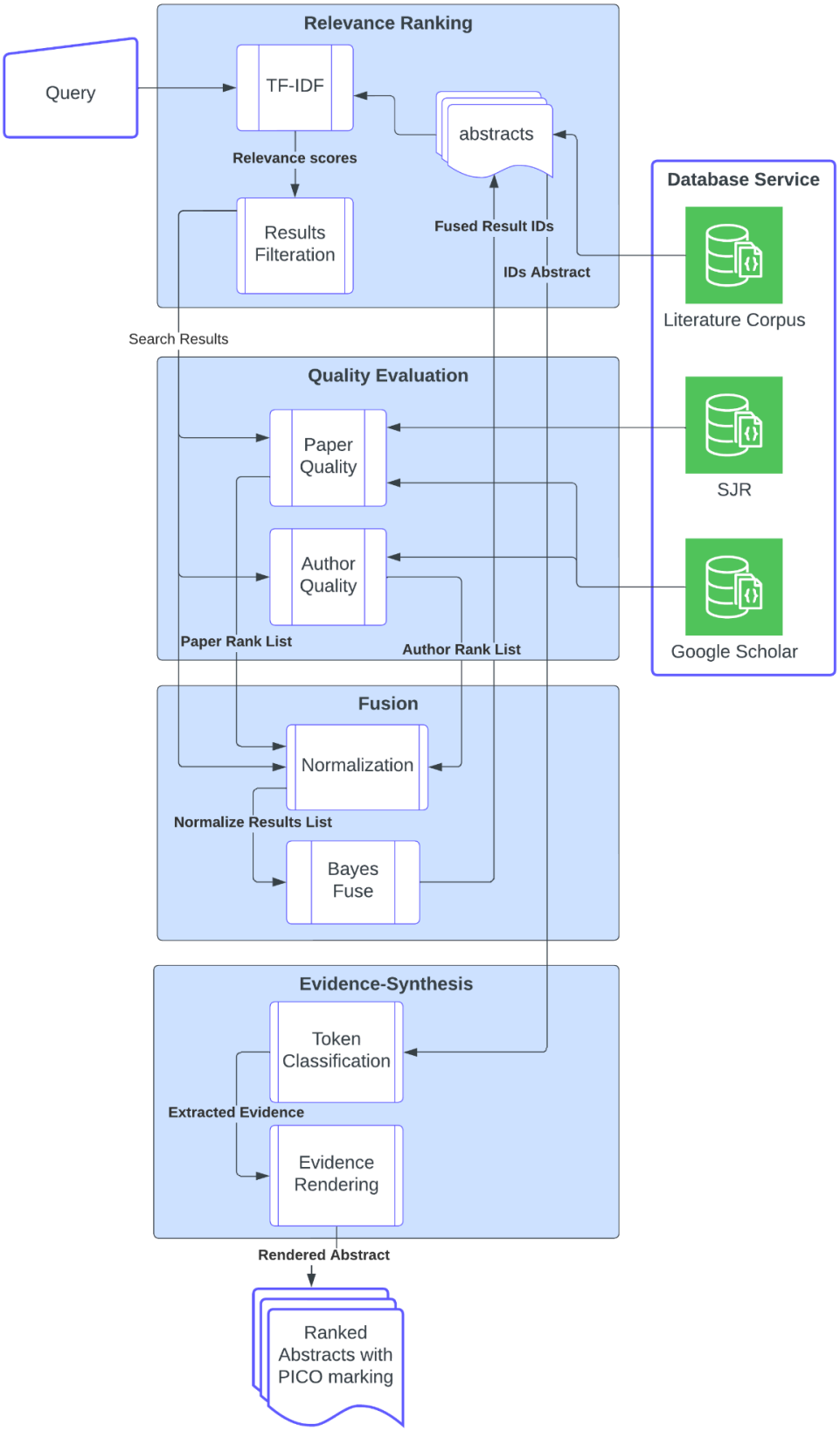

Figure 4. Proposed Framework

3.3.1 Relevance Ranking

Searching for documents that follow the principles of EBM is very important and complex. Consequently, we followed the idea of the mentioned work, which consists of two parts.

The first part uses textual content to refer to the Information Retrieval (IR) process. The significance of literature hinges on the system's proficiency in accurately representing, storing and retrieving literature in response to a practitioner’s query. To achieve this, the study used the Vector Space Model (VSM), which establishes the ranking based on the frequency and significance of words. Applying VSM provides a list of literature along with the relevancy score against a query. This relevancy score helps us in the *Fusion* process, which will be discussed in later sections.

The second part refers to the filtration of retrieved results. It is considered that documents should be filtered based on clustering techniques using the lingo clustering algorithm (Osiński et al., 2004). Lingo presents clusters with suitable topics. The study used this technique for two reasons 1) Reduce the retrieved results; 2) Present clusters with topics. The assumption behind assigning topics to clusters was that when practitioners entered the query and got the results in the form of clusters with topics on them, they could easily select the cluster whose topic seemed similar to what they searched for. But for the sake of automating the system, for now, they are selecting the cluster that has the maximum number of documents. The approach is quite good to cover this scenario, but we saw some problems with it from an implementation point of view:

1. Selecting a cluster with a maximum number of documents cannot justify the relevance of the cluster to the query and leads to the failure of the framework.

2. Assigning a topic to a cluster to get similarity with query open is a new paradigm of research that includes the techniques and evaluation of the topic assigned to the cluster based on the documents and on what conditions or definitions the topic will be assumed to be similar to what the query practitioner searches for because it is a subjective question.

Based on the above problems, we decided to use the thresholding method to filter the relevant results, as thresholding is the standard for reducing the size based on the similarity scores.

3.3.2 Quality Evaluation

When searching for the best literature in the field of EBM, retrieval of documents relevant to the query is not enough. Here come the principles of EBM, which state that the quality of the study should be evaluated. Here, quality refers to the two different aspects of the study of clinical trials. One aspect is the evaluation of the trials, which can be evaluated based on medical aspects and guidelines. The other aspect of evaluation is the validity of work, which is evaluated by checking the worth of the platform (Journal, Conference, etc.) and the worth of the author publishing the literature. If we talk about searching the literature for clinical trials or, in general, for any context, we first validate the work by checking the citations, publisher authors, etc. This is because it is a quick way to validate, and it saves us a lot of effort at the initial stage. This scenario directed us to design an algorithm for evaluating literature and authors separately.

The base idea for both algorithms came from the study mentioned in *Section 3.3*.

Their algorithms focused on the impact factor and literature category (journals, reviews, biographies, etc). In our understanding, using these parameters can capture the worth of the publisher but not the literature or author independently.

3.3.2.1 Literature Quality

The algorithm evaluates literature based on citations, SJR quartile, and SJR rank, as mentioned in *Section 3.1.2.1*.

3.3.2.2 Author Quality

The algorithm evaluates literature based on the author's H-index and total citations, as mentioned in Section 3.1.2.2. A single author is considered against literature by the selection of the highest total citations.

```
function evaluate Quality(docList):
    metadata = db.findFromAbstract(docList)
    pprEval = evaluatePaper(metadata)
    authEval = evaluateAuthor(metadata)
    return pprEval, authEval

function evaluatePaper(docList):
    paperEval = {}
    for each doc in docList:
        paperEval[doc['abstract']] =doc['hIndex']+doc['citations']+doc['sjrRank']+{"Q1":3,"Q2":2,"Q3":1,"Q4":0}[doc['sjrQuartile']]
    return paperEval

function evaluateAuthor(docList):
    authorEval = {}
    for each doc in docList:
        authorEval[doc['abstract']] = doc['authorCitations'] + doc['authorHIndex']
    return authorEval
```

Figure 5. Paper and Author Quality Algorithm

The above figure shows the implementation of the paper and author quality algorithm. The *evaluateQuality* takes an abstract list along with the metadata. After the extraction of metadata, both functions calculate abstracts based on the parameters mentioned in the above sections and return separate lists.

3.3.3 Fusion

Once lists of literature have been retrieved and ranked according to the multi-stage criteria, we need to combine them into a final list that has the properties of all criteria. For the combination of the provided list, we perform the Borda-fuse method (Aslam & Montague, 2001).

3.3.4 Evidence Synthesization

With the ranked list fulfilling the criteria of CAE steps two and three, it is necessary to extract evidence from the literature in the list to fulfill step five, which is the systemization of evidence. Here we are considering marking the PICO as evidence in EBM mentioned earlier in *Section 1.3*.

For marking PICO, we are using token classification, which is a popular technique in Natural Language processing. We used transformer architecture (Vaswani et al., 2017) for our model training and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) (Clark et al., 2020) based learning method. Here, marking was performed on the abstract of the literature only. The below figure is the sample synthesized evidence after rendering.

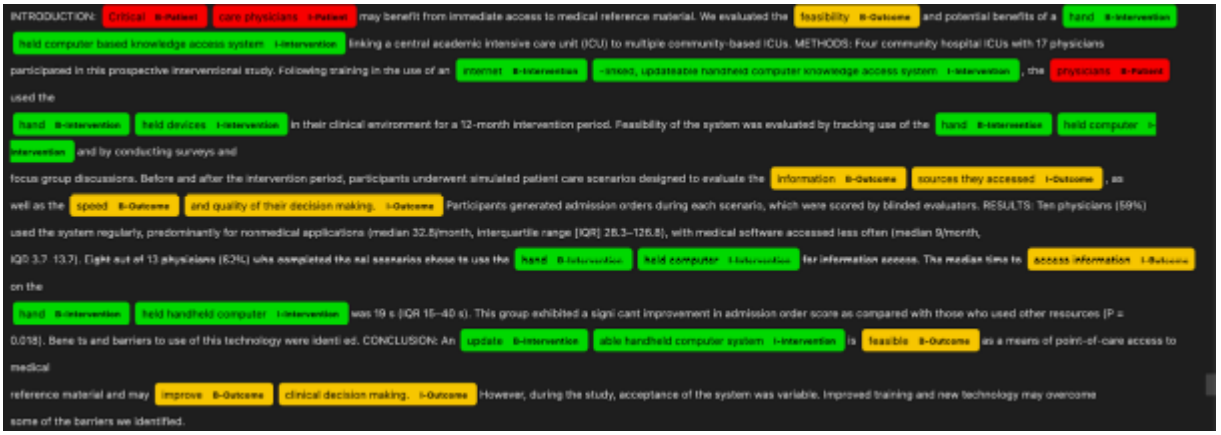

Figure 6. Synthesized Evidence

3.3.4.1 Transformers

Transformer is a method developed by Google for natural languages processing tasks, like translation and text understanding. It's unique because it uses a mechanism called self-attention to process input data in parallel rather than sequentially. It uses a self-attention mechanism, where the model can focus on different parts of the input text simultaneously. This allows it to process information in parallel, making it more efficient than previous models that processed data sequentially. The self-attention mechanism is what sets the Transformer apart. It allows the model to consider the importance of each word in the input text in relation to every other word. This parallel processing capability helps the transformer handle long-range dependencies and capture complex relationships within the data.

3.3.4.2 ELECTRA

ELECTRA is a method developed by Google for training language models. Instead of traditional methods where a model predicts each word in a sentence, ELECTRA introduces a unique twist: instead of masking or hiding certain words in a sentence, ELECTRA replaces some of them with incorrect words. These incorrect words are then tasked with being identified and corrected by the model during training. A discriminator model is used to distinguish between the original (correct) words and the replaced (incorrect) ones. The goal is for the model to get better at discriminating between real and fake words. Simultaneously, a generator model is trained to create convincing replacements for words. The generator aims to produce substitutions that are difficult for the discriminator to distinguish from the original words. What makes ELECTRA unique is that it is considered more computationally efficient than traditional masked language models, like BERT, as it only focuses on a subset of words for prediction. The discriminator's task provides a strong training signal as it has to differentiate between real and fake words, forcing the model to learn more intricate language patterns.

After the completion of the *evidence synthesizing* process, our framework returns the ranked list of abstracts that have been marked PICO if they exist. The ranked list with the marking of PICO ensures the implementation of evidence that is relevant and has up-to-the-mark quality, which is the requirement of CAE steps two, three, and five.

CHAPTER 4

EXPERIMENTS

The experiments in this study are based on the components of *relevance ranking* and *evidence synthesizing* the model presented in **FRAMEWORK FIGURE**. These components have models that were developed in recent research. The experiments were conducted using different pre-trained models with various configurations and approaches.

4.1 PICO EXTRACTION

4.1.1 Experiment 1

This experiment was performed on the (Nye et al., 2018) work to analyze the performance of their provided model and corpus.

The EBM-NLP corpus, comprising 5,000 abstracts extracted from Randomized Controlled Trials (RCTs) across diverse medical journals, has undergone comprehensive annotation at various levels. The annotation includes entity recognition, where the focus lies on identifying patients, interventions, comparisons, and outcomes within the abstracts. The corpus undergoes entity normalization, involving the mapping of recognized entities to established standard medical vocabularies. The corpus is subjected to relational analysis, aiming to identify and delineate relationships existing between the identified entities in the abstracts.

For identifying and classifying the text spans describing PICO, they used Conditional Random Fields (CRFs), attention-based Neural Networks (NN) with Bidirectional long short-term memory (BiLSTM) encoders, and their combined hybrid approach.

4.1.2 Experiment 2

This experiment was performed on the model provided by (Stylianou et al., 2020). The proposed model consists of an encoder based on BiLSTM using 2D Convolutional Neural Networks (CNNs) and a highway residual connection.

4.1.3 Experiment 3

This experiment was performed on the study of (Stylianou & Vlahavas, 2021). The study proposed a transformer-based approach for PICO classification. The model consists of three models for argument identification, relation classification, and PICO extraction.

4.1.4 Experiment 4.

This experiment was performed on the model provided by (Kanakarajan et al., 2021) study. The study trained the transformer model based on the ELECTRA method (Clark et al., 2020). This model was the latest, state-of-the-art best-performing among the above-experimented models. The source of the training dataset is not mentioned in the study.

In the above four experiments, the major problem was that they were trained on P, I, and O labels and did not deal with comparison separately. Another problem was that the label was based on only one or a maximum of two words. This creates a problem when the labels are not based on one or two words but on sentences or a sequence of words. This scenario happens frequently, as per the domain expert. They also treat the intervention and comparison labels together, which is technically not the same. Based on the above discussion, we decided to create our own data as mentioned in *Section 3.1,* and fine-tune it on (Kanakarajan et al., 2021) study that was experimented with in the last section.

4.1.4 Experiment 5

In this experiment, we use the (Kanakarajan et al., 2021) study model and fine-tune it on our created data, as mentioned in *Section 3.1*. Fine-tuning data gives us the edge in using features learned by the training of the base model. We perform fine-tuning by removing the last layer of the model, which has three neurons presenting P, I, and O, and adding our own layer of four neurons. During the training, the training process was performed only on the last layer while freezing other layers because of the limited data. This experiment gives the best accuracy among others.

4.2 RELEVANCE RETRIEVAL

4.2.1 ClinicalBERT

This series of experiments is based on a model named ClinicalBERT (Huang et al., 2019) created by a team of researchers at the University of Pennsylvania. This model is used for the extraction of features from literature and queries and is used with different approaches to documents. After extracting features or converting the inputs into embeddings, we calculate their cosine similarity as relevancy. The approaches to documents are:

1. title
2. abstract
3. title with abstract

4.2.2 PubMedBERT

This series of experiments is based on a model named PubMedBERT (Gu et al., 2022) created by Microsoft. This model was used for the conversion of inputs into embeddings and calculating similarity with the same document approaches as mentioned in *Section 4.2.1*.

4.2.3 TF-IDF

This series of experiments is based on a popular VSM model named Term Frequency-Inverse Document Frequency (TF-IDF) (Salton et al., 1975). This model is used for the conversion of inputs into vectors. The vector from this model is just the representation of word frequencies with their importance but not the semantic relation.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter contains the results of all the aforementioned experiments. We have also explained the results and discussed the shortcomings of the results.

5.1 RELEVANCE RANKING

To achieve the best retrieval results, we performed experiments on vectorization models and semantic models that are commonly used in IR. The VSM, despite being traditional, outperforms semantic models. **BELOW TABLE** depicts that TF-IDF results are far better than both semantic models.

Table 4

Comparison of Retrieval Models

Models	Avg Recall@5	Avg Precision@5
TF-IDF	0.38	0.53
ClinicalBERT	0.22	0.34
PubMedBERT	0.27	0.25

The below table further explains the precision and recall of each query when applying TF-IDF. We get the best results at k = 5.

Table 5

Evaluation of Query

Query	Precision	Recall
1	0.6	0.75
2	0.4	0.6
3	0.6	0.6
4	0.2	0.25
5	0.2	0.25
6	0.2	0.14
7	0.8	1.0
8	0.4	0.6
9	0.6	1.0
10	0.6	1.0
11	0.2	0.2

From above table, we can see that some queries have a recall score of 100 percent, but the precision is far lower. This is because the relevant number of relevant documents in the hierarchies is sufficiently small, thus auto-optimizing the value of *k*. Furthermore, some values of recall are very low. This is because the vectorization model works on frequencies of words, and the frequencies of occurrence of less relevant words may be higher than the relevant ones.

5.2 EVIDENCE EXTRACTION

As discussed in *Section 4.1.5* of Experiments, the token classification architecture of transformers based on the ELECTRA model after finetuning performed best among others.

Table shows the results of PICO extractions on both base and fine-tuned models.

Table 6

Comparison of PICO

Model	Accuracy	F1	Precision	Recall
Fine-Tune	93.5	36.8	36.6	37
Base Model	65.5	18.4	18.3	18.5

The overall result of the model shows that the finetune model outperforms the base model. If we see fine-tune results excluding accuracy, it is below fifty because our sequence of labels consists of different lengths of tokens, which creates an imbalance effect.

CHAPTER 6

CONCLUSION AND FUTURE WORK

We have created the EBM retrieval framework by working on the different aspects of EBM retrieval process. We have performed IR process by generating queries and implementing TF-IDF for relevance retrieval. We write algorithms for evaluating the *quality of paper* and *quality of author*. For *evidence synthesis*, we created corpus of token classifications on PICO labels. We also fine-tune BioELECTRA based model on created data.

The improvement of this work can be done by improving different parts of this framework. If hybrid approach of Bag Of Words (BOW) with semantics or any relational or context-capturing model is used in place of VSM, this will improve the relevance results by understanding the context and inclusion or exclusion of important words. Creating more data of PICO with reduced labels length as much as possible will improve the quality of features when train on transformer based approaches.

References

Aslam, J. A., & Montague, M. (2001). Models for metasearch. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 10.1145/383952.384007

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II* (pp. 160-172). Springer Berlin Heidelberg. 10.1007/978-3-642-37456-2_14

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv [cs.CL]*. http://arxiv.org/abs/2003.10555https://arxiv.org/abs/2003.10555

Cohn, T., Callison-Burch, C., & Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational linguistics (Association for Computational Linguistics)*, *34*(4), 597-614. 10.1162/coli.08-003-r1-07-044

Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2021). COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *npj digital medicine*, *4*(1). 10.1038/s41746-021-00437-0

Golinelli, D., Nuzzolese, A. G., Sanmarchi, F., Bulla, L., Mongiovì, M., Gangemi, A., & Rucci, P. (2022). Semi-automatic systematic literature reviews and information extraction of COVID-19 scientific evidence: Description and preliminary results of the COKE Project. *Information (Basel)*, *13*(3), 117. 10.3390/info13030117

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language

processing. *ACM transactions on computing for healthcare*, *3*(1), 1-23. 10.1145/3458754

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. 10.48550/ARXIV.1904.05342

Kanakarajan, K. R., Kundumani, B., & Sankarasubbu, M. (2021). BioELECTRA:Pretrained Biomedical text Encoder using Discriminators. *Proceedings of the 20th Workshop on Biomedical Language Processing*. 10.18653/v1/2021.bionlp-1.16

Keretna, S., Lim, C. P., Creighton, D., & Shaban, K. B. (2015). Enhancing medical named entity recognition with an extended segment representation technique. *Computer methods and programs in biomedicine*, *119*(2), 88-100. 10.1016/j.cmpb.2015.02.007

Liu, S., Sun, Y., Li, B., Wang, W., Bourgeois, F. T., & Dunn, A. G. (2021). Sent2Span: Span detection for PICO extraction in the biomedical text without span annotations. *Findings of the Association for Computational Linguistics: EMNLP 2021*. 10.18653/v1/2021.findings-emnlp.147

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services research*, *14*(1). 10.1186/s12913-014-0579-0

Nye, B., Jessy Li, J., Patel, R., Yang, Y., Marshall, I. J., Nenkova, A., & Wallace, B. C. (2018). A corpus with multi-level annotations of patients, interventions and Outcomes to support language processing for medical literature. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 197-207. 10.18653/v1/P18-1019

Osiński, S., Stefanowski, J., & Weiss, D. (2004). Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent Information Processing and Web*

Mining: Proceedings of the International IIS: IIPWM´04 Conference Held in Zakopane, Poland, May 17-20, 2004 (pp. 359-368). Springer. 10.1007/978-3-540-39985-8_37

Rosenberg, W., & Donald, A. (1995). Evidence based medicine: an approach to clinical problem-solving. *BMJ*, *310*(6987), 1122-1126. 10.1136/bmj.310.6987.1122

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620. 10.1145/361219.361220

Sanchez-Graillet, O., Witte, C., Grimm, F., & Cimiano, P. (2022). An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *Journal of biomedical semantics*. 10.1186/s13326-022-00271-7

Sarker, A., Molla, D., & Paris, C. (2017). Automated text summarisation and evidence-based medicine: A survey of two domains. *arXiv [cs.CL]*. http://arxiv.org/abs/1706.08162https://arxiv.org/abs/1706.08162

Schiavenato, M., & Chu, F. (2021). PICO: What it is and what it is not. *Nurse education in practice*, *56*(103194). 10.1016/j.nepr.2021.103194

Serrano-Guerrero, J., Romero, F. P., & Olivas, J. A. (2020). A relevance and quality-based ranking algorithm applied to evidence-based medicine. *Computer methods and programs in biomedicine*, *191*(105415). 10.1016/j.cmpb.2020.105415

Stylianou, N., Razis, G., Goulis, D. G., & Vlahavas, I. (2020). EBM+: Advancing Evidence-Based Medicine via two level automatic identification of Populations, Interventions, Outcomes in medical literature. *Artificial intelligence in medicine*, *108*(101949). 10.1016/j.artmed.2020.101949

Stylianou, N., & Vlahavas, I. (2021). TransforMED: End-to-End Transformers for Evidence-Based Medicine and Argument Mining in medical literature. *Journal of biomedical informatics*, *117*(103767). 10.1016/j.jbi.2021.103767

Suffian, M., Yaseen, M., & Wasi, S. (2018). Developing disease classification system based on keyword extraction and supervised learning. *International journal of advanced computer science and applications : IJACSA*, *9*(9). 10.14569/ijacsa.2018.090976

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. 10.48550/ARXIV.1706.03762

Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W. R., Lo, K., Roberts, K., Soboroff, I., & Wang, L. L. (2020). TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR forum*, *54*(1), 1-12. 10.1145/3451964.3451965

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., … Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. *arXiv [cs.DL]*.

Wang, Q., Liao, J., Lapata, M., & Macleod, M. (2022). PICO entity extraction for preclinical animal literature. *Systematic reviews*, *11*(1), 209. 10.1186/s13643-022-02074-4

Xia, Y., Cai, J., Li, Y., Dou, Z., Zhang, Y., Wu, L., Huang, Z., Xu, S., Sun, J., Liu, Y., Wu, D., & Han, D. (2022). A precision‐preferred comprehensive information extraction system for clinical articles in traditional Chinese Medicine. *International journal of intelligent systems*. 10.1002/int.22748

Zhang, T., Yu, Y., Mei, J., Tang, Z., Zhang, X., & Li, S. (2020). Unlocking the power of deep PICO extraction: Step-wise medical NER identification. *arXiv [cs.CL]*. 10.48550/arXiv.2005.06601

Zhang, Y., Wang, M., Zipperle, M., Abbasi, A., & Tani, M. (2023). RelRank: A relevance-based author ranking algorithm for individual publication venues. *Information processing & management*, *60*(1), 103156. 10.1016/j.ipm.2022.103156

PLAGIARISM VERIFICATION (for Final Submission)

(Note: *This form is filled at the final submission of Thesis/Research Project Reports for graduation.*)

Title of Research Thesis: _____

Total Pages: _____

Student Name : _____

Registration ID : _____

Supervisor: _____

_____

Program: _____

Faculty: _____

This is to report that the above document was scanned for similarity detection and has been submitted in turnitin repository. Process and outcome is given below:

Software used:	**Turnitin**	Date:	
Similarity Index:		Total word count:	
File Name:		Digital Receipt No.:	

_____

Signature and Stamp of the
Office of Research & Project