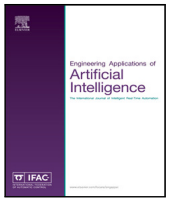
Engineering Applications of Artificial Intelligence 136 (2024) 109028

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Recommendation system of scientific articles from discharge summaries

Adrián Alonso Barriuso [a,b], Alberto Fernández-Isabel [a,*], Isaac Martín de Diego [a], Alfonso Ardoiz [b], J.F. J. Viseu Pinheiro [c]

[a] Rey Juan Carlos University, Data Science Laboratory[1], c/ Tulipán, s/n, 28933, Móstoles, Spain
[b] Dezzai[2], 6 Pollensa Street, ECU Bldg. 2nd floor, Las Rozas, Madrid 28290, Spain
[c] Instituto de Investigación Biomédica de Salamanca[3], Hospital Virgen de la Vega, 10ª floor, Paseo de San Vicente, 58-182, 37007, Salamanca, Spain

ARTICLE INFO

Keywords:
Medical text processing
Diagnosis retrieval
Computational semantics
Recommendation system
Scientific relevance

ABSTRACT

Medical professionals are often overwhelmed by the amount of patients they have to care for, leaving little time available to keep up to date in their respective specialities. They usually find it challenging to keep up with the vast amount of medical literature and identify the most relevant articles for their practice, especially those related to their patient's specific conditions. Therefore, a system that proactively supports healthcare professionals in selecting relevant articles related to the characteristics of the patients is crucial. This paper presents Medical Expert Linguist for Evaluating Nosology and Diagnosis Information (MELENDI) to tackle this issue. It is a recommendation system that effectively and efficiently recommends pertinent medical articles to healthcare professionals based on their patients' diagnoses. It combines a semantic similarity model generated using the content of discharge summaries, with a relevance estimator produced by analysing scientific publications. To test the system, 1,000,000 abstracts were obtained from PubMed and 10 discharge reports from 'Medical Information Mart for Intensive Care (MIMIC-III) were used. A group of 5 medical specialists has been involved in the system's evaluation. These evaluations demonstrated good overall performance, supporting the implementation of the system in a real-world environment, such as a hospital information system.

1. Introduction

Healthcare professionals must care for patients and keep updated in their specialities to ensure the latest evidence-based treatments.

The huge and growing amount of scientific literature in journals, congresses, clinical trials, and guidelines that usually appear, make this task more difficult (Johnson et al., 2018). This fact is aggravated in the case of the latest releases of drugs, medical advances, and techniques, and it requires a great deal of time to discriminate the most relevant approaches related to their field.

The lack of good-evidence updates for the healthcare givers has a potential impact on the quality of treatment and the possibility of saving lives. The situation was aggravated during and after the COVID-19 outbreak, coupling an explosion of scientific literature on the disease and the virus (Gianola et al., 2020) with extreme congestion of healthcare services (Olivas-Martínez et al., 2021).

The severity of the virus and the need to deal with a global pandemic quickly has led to a deterioration in the quality of research publications (Raynaud et al., 2021). Therefore, a system that supports healthcare professionals to keep up to date in their fields in the shortest possible time, offering relevant publications based on the characteristics of their patients, has become a key issue. Notice that these features are typically the nosology information (*i.e.* classification of diseases) and the diagnosis made (*i.e.* discharge summaries), as these items usually include the relevant details of potential diseases and disorders.

This paper introduces the MELENDI framework. It is a novel system that works automatically by extracting diagnoses from EHR, and by searching for relevant articles that combine the characteristics of those diagnoses, and proactively recommending them to the clinician.

MELENDI is a system specifically designed to be extremely fast, making answers in user time, and having a low cost regarding the computer resources consumption. Moreover, it could be embedded in a Hospital Information System (HIS) or be directly consumed by the clinician through a web application. This fact promotes easy access to healthcare professionals for consulting relevant articles related to the characteristics of patients.

* Corresponding author.
E-mail addresses: adrian.barriuso@urjc.es, a.alonso@dezzai.com (A. Alonso Barriuso), alberto.fernandez.isabel@urjc.es (A. Fernández-Isabel), isaac.martin@urjc.es (I. Martín de Diego), alfonso.ardoiz@dezzai.com (A. Ardoiz), jfviseu@saludcastillayleon.es (J.F.J. Viseu Pinheiro).

[1] www.datasciencelab.es.
[2] https://dezzai.com.
[3] https://ibsal.es.

https://doi.org/10.1016/j.engappai.2024.109028
Received 6 March 2024; Received in revised form 24 June 2024; Accepted 19 July 2024
Available online 26 July 2024

Regarding the system's architecture, it presents five modules that tackle specific tasks such as processing articles, detecting the diagnoses from discharge summaries, producing similarities between articles, calculating the relevance of articles, and generating a ranking of results.

The system has been adapted to the format of MIMIC-III EHRs (Johnson et al., 2016). Thus, it can detect and automatically extract main diagnoses from an established standard approach in the healthcare domain by combining regular expressions with a medical concept detection model (Kraljevic et al., 2021).

Several experiments have been carried out to evaluate the viability of the system. A group of experts in the healthcare domain from Biomedical Research Institute of Salamanca (IBSAL) have been involved in this task. MELENDI has been embedded into a real-world scenario provided by a HIS specifically designed for testing purposes. First, the achieved tests evaluate the most relevant modules of the system (*i.e.* those which calculate similarities between articles and detect the main diagnoses in the textual content of discharge summaries) and the different decisions in their design. Later, the system's complete performance and the results' quality are considered.

The rest of the article is structured as follows. Section 2 details the contributions of the proposal. Section 3 describes the fundamentals on which the system is based. Section 4 presents the details of the system and its components, and Section 5 focuses on the experiments carried out to validate the system. Finally, Section 6 concludes and proposes possible future guidelines.

2. Research contributions

The contributions of the proposal are summarised in the next points:

- The main innovation of MELENDI lies in its combination of a semantic similarity model with an automatic relevance estimator for scientific articles. This integration enables the system to make recommendations that are both academically relevant and contextually pertinent to the diagnosis at hand. This ensures that healthcare professionals receive information directly applicable to their clinical context.
- The modular nature of the proposed architecture allows for the decoupling of this implementation, making it adaptable to other data sources as well as manual user queries.
- The proposed solution prioritises a balance between relevance and semantic similarity; however, this trade-off can be adjusted to prioritise one over the other based on the specific needs and application domain.
- The system automatically recommends articles based on patient diagnoses, eliminating the need for user intervention. This obviates the need for specific technical knowledge inherent to search engines and information retrieval systems, thereby saving medical experts time in staying updated within their respective specialities.
- MELENDI has been designed to be efficient in terms of execution speed and computational cost, facilitating its potential implementation in a HIS.

3. Related work

This section introduces the foundations of the MELENDI framework. It covers the different perspectives related to the different modules of the system and how they are related to the healthcare domain.

First, Information Retrieval (IR) approaches are addressed by delving into semantic models that provide document similarity. These models are revisited considering general purposes and later specifically in the healthcare domain. Second, a comprehensive review based on Artificial Intelligence (AI) in medicine is tackled. There, the most typical approaches and topics in the healthcare area are considered. Finally, more specific approaches focused on the automatic diagnosis processing task are revisited and discussed.

3.1. Information retrieval

IR is one of the most important processes in computer science (Chowdhury, 2010). It consists of recovering the information of interest from a set of documents (*i.e.* a corpus). These documents usually contain textual data, which is unstructured or semi-structured information, but they can also contain structured information as videos and images.

In the case of textual content, IR tasks can be organised into processing text approaches and non-processing text approaches (Mala and Lobiyal, 2016). In the first case, the extraction of keywords (relevant words) from the text is the most common operation. This bag of words is standardised by transforming the words to their lexical root (*i.e.* steaming or lemmatisation). This task allows the computer to find similar words in the different texts and to establish similarities according to the matches. Typical works use algorithms like TF-IDF, cosine similarity, latent semantics, and combinations of them (Passalis and Tefas, 2018). In the second case, the whole structure of the textual content is considered. Thus, algorithms based on distributional semantics are the most typical in this context. These algorithms are focused on vectorising words, which allows for organising the textual content into several dimensions according to the co-occurrence of words. Typical approaches are those that implement recurrent neural networks and models based on transformers. Instances of them are Long-Short Term Memory (LSTM) networks and architectures based on Bidirectional Encoder Representations from Transformers (BERT) (Jiang et al.).

Recommendation systems are one of the most typical software architectures that make use of IR. These systems usually provide advice according to a set of keywords provided by users or elaborate a profile following their previously expressed preferences.

Regarding the healthcare domain, IR techniques are included in recommendation systems related to documents that extract similarities between symptoms, patients, and diseases (Stark et al., 2019). Thus, it is a key issue for the healthcare expert to find possible solutions already documented for a specific problem. Therefore, the development of these systems focused on decision-making support has been an important enhancement (De Croon et al., 2021).

Typical approaches that address this fact present two different perspectives: patients and physicians. In the first case, patients use a recommendation system to select the most interesting medical centre or doctor. Instances of these works are (Martinez et al., 2014) and Waqar et al. (2019). In the second case, healthcare experts use recommendations provided by these systems for finding new drugs, or medical treatments and procedures. Instances of these approaches are: (Katzman et al., 2018) and Zhang et al. (2015).

For the case of MELENDI, it is a recommendation system focused on decision-making support that provides the most relevant scientific documents according to the automatic processing of the discharge summaries. Thus, it can provide the most similar documents to specific diseases or syndromes easing the research work of physicians.

3.2. Artificial Intelligence in medicine

AI is a field of computer science focused on emulating the mental processes of intelligent individuals to solve specific problems through the usage of machines (Hunt, 2014).

This field can be organised into four main perspectives: case-based reasoning systems, expert systems, Bayesian networks, and behavioural-based systems.

Delving into the perspectives, case-based reasoning systems solve problems using the acquired experience through the study of previous cases (Kolodner, 2014). Expert systems can gather specific knowledge from humans with specific expertise, and then apply it to solve related problems (Liebowitz, 2019). In the case of Bayesian networks, they build graphs that use probabilities to transit between the states (edges). These probabilities are used to infer possible future situations according

to a current state (Marcot and Penman, 2019). Finally, behavioural-based systems consist of structures that establish relationships and interactions between the elements to solve complex problems (Riedl, 2019). Here, Multi-Agent Systems are the most typical approach. These intelligent systems cooperate, compete, and interact between themselves and with the environment to solve simple problems that can be joined to elaborate a more complex solution.

Regarding the medical domain, intelligent systems usually consider both patient (Ploug and Holm, 2020) and physician perspectives (Scheepers-Hoeks et al., 2013). The development of complex and intelligent systems focused on the healthcare domain is a widely spread issue.

Case-based reasoning systems in the healthcare domain have been widely used to detect possible diseases and infer the evolution of patients (Bentaiba-Lagrid et al., 2020). This fact has its foundations in the common progress of diseases in the majority of patients. Thus, considering the initial steps, the next ones can be inferred with a certain degree of certainty (Duan and Jiao, 2021).

Expert systems are well-known approaches focused mainly on the prediction and diagnosis of diseases using the expertise of physicians (Singla et al., 2014). Thus, they accumulate expert knowledge to infer possible solutions according to a set of symptoms. These symptoms are usually included manually by the users in the system as input.

Bayesian networks are approaches mostly oriented to patients. They allow discovering patterns in treatments and also in the diagnosis of diseases (Akila and Balaganesh, 2021). Thus, they ease the generation of ontologies and relationships between drugs, symptoms, and also secondary effects.

Behavioural-based systems are present in multiple perspectives in healthcare (Isern and Moreno, 2016). They can be useful for patients, where they can analyse multiple variables of a person through sensors, and for healthcare experts, due to they can simulate complex reactions to drugs. Moreover, they are relevant in the simulation of healthcare infrastructures and also biomedical experiments.

Finally, it is relevant to highlight systems that can provide support to healthcare professionals during the process of complex procedures. Though they do not usually include complex AI algorithms, they present robotic elements to simplify the tasks during operations and invasive treatments (Sun et al., 2020).

In the case of MELENDI, it is a decision-support system based on the expert knowledge provided through discharge summaries of several patients. The system processes this textual content semantically, and then it can use that knowledge to recommend the fittest scientific articles to healthcare professionals. This fact simplifies their work and eases to find new findings related to some drugs and treatments for diseases of interest.

3.3. Automatic diagnostic processing

There are specific systems adapted to detect and do effective diagnoses using data coming from previous patients. These systems are usually expert systems, that are trained following a set of rules or labels which are used to elaborate a classification. This classification consists of a positive flag or a negative flag, where the positive confirms the detection of a possible disease, and the negative discards the possibility with a certain degree of assurance. All these systems are also called in medicine: smart healthcare systems (Mansour et al., 2021). Notice that these systems have as a common point the explainability of the decisions made (Khodabandehloo et al., 2021). It is basic for healthcare experts, so they must explain to the patients the possible detected diseases. Moreover, it helps during the confirmatory process, as a human being must agree with the detection made by the system.

Delving into the rules-based systems, they follow the different indications (bottom-top perspective) establishing the possible positive according to the rules accomplished. These systems act as healthcare professionals, following the different points of interest and symptoms to produce the final diagnosis. Therefore, they are very similar to human decision-making procedures. Typical instances of these systems are those based on fuzzy logic (Mousavi et al., 2021), decision trees, and criteria graphs (Alves et al., 2021).

Rules-based systems have been also used in the literature to establish relationships between a diagnosis, drugs, and possible adverse reactions. Thus, the textual content is analysed to produce models usually based on graphs that can reflect and simplify these interactions and problems (Tan et al., 2022). This procedure eases the selection of the fittest treatment for healthcare professionals.

Alternatively, case-based reasoning systems are a specific part of the rules-based systems. These systems use previously studied cases to infer knowledge and make assumptions over a set of symptoms (Duan and Jiao, 2021). They also use some rules to discriminate between the cases, filtering to select the fittest ones.

In the case of label-based systems, they use the data to detect hidden patterns that are not usually detected by healthcare professionals (top-bottom perspective). These systems are based on Machine Learning (ML) techniques and models. Then, these models are trained using the supervised learning perspective to tackle the issue. Thus, they adapt their parameters according to the input data and the desired label, getting a specific configuration that detects the predefined disease in the data. Typical instances of these systems use well-known ML models such as Convolutional Neural Networks (Saha et al., 2021), Random Forests, and Support Vector Machines (Rauber et al., 2021).

In recent times, the emergence of Large Language Models (LLMs) has led to numerous applications in the healthcare field with mixed results (Thirunavukarasu et al., 2023). In the area of automatic medical entity detection, some approaches based on prompt engineering are beginning to be used, even without peer review, such as (Hu et al., 2024).

Special mention in this category to artificial vision systems. They are systems that use ML models specifically adapted to detect visual patterns. Therefore, they are useful for evaluating clinical images of patients, where possible diseases, problems, and symptoms are reflected (Castiglioni et al., 2021). They can focus on detailed regions and analyse the pixels better than a human professional. This capability allows using these systems as decision support systems.

In the case of the MELENDI framework, it has a semantic rule-based engine that uses textual content from discharge summaries to detect possible diseases or syndromes. It does not include a previously trained ML model, but it uses semantic information to infer the knowledge. Therefore, it could be considered a hybrid between a rule-based system and a case-based system, where the patterns to find information are previously known, and they can be used to achieve relevant classifications of diseases (*i.e.* nosological-related tasks) and related recommendations.

4. Proposed framework

The MELENDI framework is a complete system developed to save time for healthcare professionals. Its main functionality consists of making scientific article recommendations according to the diagnoses detected in specific discharge summaries.

To achieve this task, the system uses the most relevant information gathered from EHRs of several discharge summaries of patients. This information becomes the knowledge of the system. To obtain this knowledge, the information is processed through semantic techniques to detect diagnoses. These diagnoses are part of textual content that presents diseases, nosological entities, syndromes, and also any pathological or health condition of the patients. Unified Medical Language System (UMLS) (Bodenreider, 2004) provides support to accomplish the detection and extraction of these medical concepts. Then, the obtained knowledge is used by the system to search scientific articles related to these input concepts according to semantic similarity. Finally, these

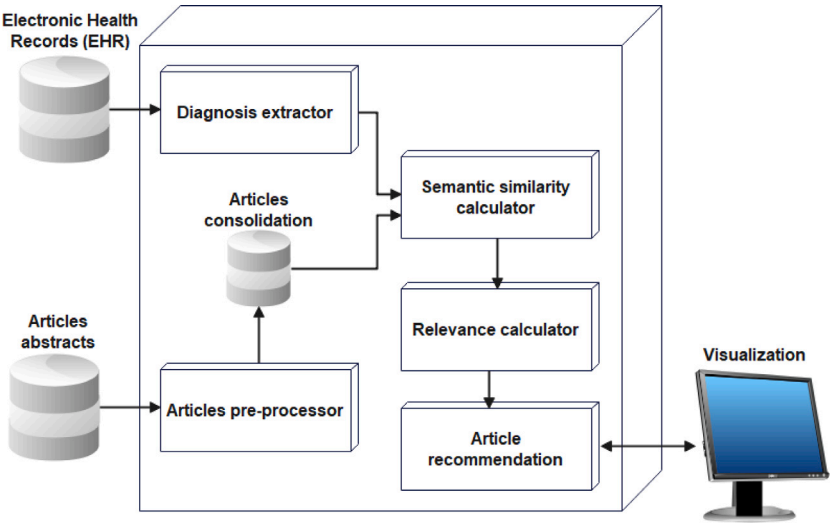


Fig. 1. Overview of the MELENDI framework architecture.

articles are organised through a set of specific features extracted from the metadata that measure their relevance in the healthcare domain.

Regarding the architecture of the system, it is organised into five main modules: *Articles pre-processor* module, *Diagnosis extractor* module, *Semantic similarity calculator* module, *Relevance calculator* module, and *Article recommendation* module. They are completed by two databases that consolidate the information gathered for scientific articles, and a *Visualisation* module to interact with users (see Fig. 1).

The first module is in charge of analysing abstracts from scientific publications related to the healthcare domain. The second module addresses the diagnosis detection and information-gathering tasks from the provided discharge summaries. The third module estimates the similarity between the input concepts (*i.e.* the concepts detected in the diagnoses and scientific articles). The fourth module calculates the relevance of the most similar articles obtained, and the fifth module ranks these articles according to similarity and relevance values.

The databases are formed by a set of *Articles abstracts* and its *Information consolidation*. The first comprises a collection of abstracts to analyse through semantic procedures to generate the corresponding sentence embeddings. These embeddings are stored in the second database for being used by the system in response to the requests made by users.

4.1. Articles pre-processor module

The system uses this module to process several abstracts from different scientific articles and to obtain sentence embeddings from them. This process is achieved following a background process one time per month. It allows updating the information about the state of the healthcare domain considering the last scientific novelties (see Fig. 2).

The source of information used to select and filter the scientific articles is Pubmed (White, 2020) since it is an open-access search engine. On the other hand, *msmarco-distilbert-base-v4* model (Reimers and Gurevych, 2019) has been the semantic model used to produce the embeddings.

The system also obtains the metadata of each article. This metadata is formed by the DOI, the year of publication, and the authors. This information is used later by the *Relevance calculation* module to estimate the reputation of the articles. Notice that the gathering process of this metadata is also achieved during the background process, while the estimation of the reputation is carried out in real-time answering to a request made by users.

Regarding the architecture of the module, it presents two components to address its two main tasks: the *Text content gatherer* and the *Embedding generator*. The first component extracts the texts and the metadata, while the second generates the associated embeddings and stores them in the article embeddings consolidation database.

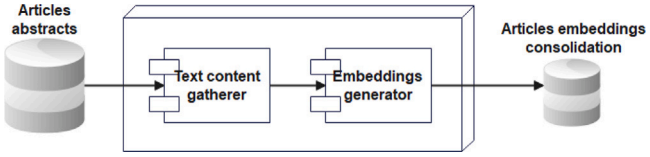


Fig. 2. Internal architecture of the *Articles pre-processor* module.

4.2. Diagnosis extractor module

This module has been specifically designed to address the challenges inherent in EHRs, which are characterised by their unstructured and intricate composition. Therefore, the system specifically targets the discharge summaries of the MIMIC-III dataset, a renowned and widely utilised medical corpus in healthcare analytics.

Using a hybrid methodology, the module integrates linguistic cues with LLMs to automate the segmentation of EHRs data into distinct components such as Main Diagnosis, Dates and Place, and Medications. This process, essential for managing the EHR's complexity, is commonly known as section identification problem (Pomares-Quimbaya et al., 2019).

This approach is performed through three sequential steps performed by three sub-modules (see Fig. 3 for further details): *text data pre-processing*, *primary diagnosis extraction*, and the *diseases entities recognition*.

During the *pre-processing* step, the raw EHR text is systematically transformed to ensure consistency and enhance the performance of the subsequent tasks. This is achieved through a series of operations. Initially, the text is stripped of extraneous formatting and special characters that can interfere with the process (*e.g.* \t, which typically represents tab spaces and other non-printable characters). In addition, functions are implemented to remove digits, punctuation, and excess whitespace unlikely to contribute to the diagnostic process. Finally, the textual content is split into paragraphs.

In the second step, *extracting the primary diagnosis*, the module employs a rudimentary classification mechanism based on keyword recognition. This algorithm ingests individual paragraphs and discriminates the content by leveraging a tailored keyword list. The "favourable" keywords encompass terms frequently encountered in diagnostic narratives

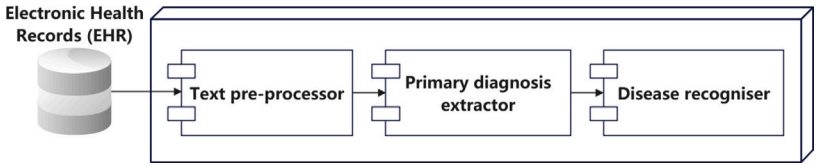


Fig. 3. Internal architecture of the *Diagnosis extractor* module.

such as *diagnosis*, *diagnosed*, *symptoms*, and *patient presents*. Conversely, "non-favourable" keywords, which typically signal the absence of diagnostic information, include terms like *procedures*, *dates*, *medications*, and *history of*. The compilation of this keyword list is derived from empirical heuristics, suggesting potential refinement in subsequent iterations to enhance accuracy.

However, the initial keyword-based heuristic occasionally fails to select any text segment as a likely main diagnosis. In those cases, an auxiliary Generative AI model is incorporated, operating on the entirety of the pre-processed EHR text to deduce the principal diagnosis when the primary method proves insufficient. This backup model, specifically instructed to identify diagnostic information (as detailed in Annex), builds upon the capabilities of Generative Pre-trained Transformer (GPT)-4 (Sanderson, 2023), thus leveraging its advanced language understanding to ensure robust extraction accuracy. These instructions have been provided through a customised few-shot prompt:

Diagnosis extractor few-shot prompt

You are a medical healthcare expert. Review the provided EHR text and extract the primary diagnosis. Present the primary diagnosis clearly and concisely. Use the following cases as examples:

1. EHR: "Finally, the patient was diagnosed with retinopathy secondary to Type II diabetes mellitus" OUTPUT: "Primary Diagnosis: retinopathy secondary to Type II diabetes mellitus"

2. EHR: "The patient complains of blurry vision and ocular pain. Past medical history includes multiple sclerosis. Current fundoscopy shows evidence of optic neuritis". OUTPUT: "Primary Diagnosis: Optic Neuritis secondary to Multiple Sclerosis"

3. EHR: "The individual has reported severe chest pains with shortness of breath and has a notable history of hypertension. Recent EKG findings indicate myocardial infarction". OUTPUT: "Primary Diagnosis: Acute Myocardial Infarction"

4. EHR: "Subject exhibits persistent cough, weight loss, and night sweats. Tuberculin skin test is positive, and chest X-ray reveals upper lobe infiltrates". OUTPUT: "Primary Diagnosis: Pulmonary Tuberculosis"

5. EHR: "Encounter for post-operative complications. Symptoms include abdominal pain, fever, and leukocytosis. The patient underwent a colectomy six days ago. The CT scan shows evidence of intra-abdominal abscess". OUTPUT: "Primary Diagnosis: Post-colectomy Intra-abdominal Abscess

In the third step, the *diseases entities recognition* leverages the capabilities of MedCAT, a sophisticated Named Entity Recognition (NER) tool designed for medical contexts (Kraljevic et al., 2021). MedCAT's NER+L functionality is instrumental in enhancing the quality of entity extraction by automatically associating detected medical entities with their corresponding UMLS metadata. This mapping process is the key component of this system as it serves as a filter and allows compiling only the entities whose UMLS *semantic type* is "T047" (disease or syndrome), which compose the diagnosis. Notice that this operation minimises possible errors committed by the NER tool.

Finally, the module produces a list of the desired entities found in the main diagnosis section, whose embeddings are calculated to serve as input for the *Semantic similarity calculator* module.

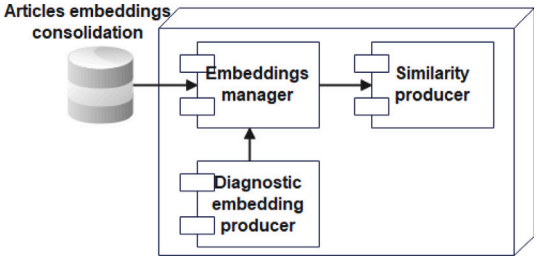


Fig. 4. Internal architecture of the *Semantic similarity calculator* module.

4.3. Semantic similarity calculator module

This module implements a semantic search engine based on cosine similarity that adheres to a symmetric semantic search model (see Eq. (1)). This model enables the discovery of textual content similar to a query (*i.e.* the detected diagnostic), effectively identifying synonyms and near-meaning words due to the contextual representation provided by a BERT-like architecture.

$$\cos(\mathbf{t},\mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \quad (1)$$

In this sense, notice that while an asymmetric search-oriented model might initially seem more appropriate for this use case, given that the abstracts are usually larger than the diagnoses, empirical evaluations and the advantages of a bounded metric led to select the symmetric semantic search model with cosine similarity.

Regarding the architecture of the module, it comprehends three components (see Fig. 4) the *Diagnosis embedding producer*, the *Embeddings manager*, and the *Similarity producer*.

The *Diagnosis embedding producer* component generates the embedding for the detected diagnosis. Thus, the text is converted into numeric information.

The *Embeddings manager* component joins the predefined embeddings of abstracts with the embedding developed with the diagnostic. This creates a new set of embeddings, adapting the original ones to the new inclusion.

The *Similarity producer* component estimates the similarity between the embeddings according to the new one. Then, it returns the n most similar abstracts, concluding the tasks of the module.

4.4. Relevance calculator module

This module implements the functionality of an enhanced release of a system previously developed called *Webelance* (Fernández-Isabel et al., 2020). It is a framework focused on the evaluation of the relevance of articles with content based on the healthcare domain.

Regarding the architecture of the module, it includes three main components to estimate the relevance of medical articles: the *Text processor*, the *Reputation calculator*, and the *Relevance calculator*. It is completed with the corresponding pre-trained relevance lexicon and neural network (see Fig. 5).

The *Text processor* component captures the information from abstracts applying text mining techniques. This information is mainly

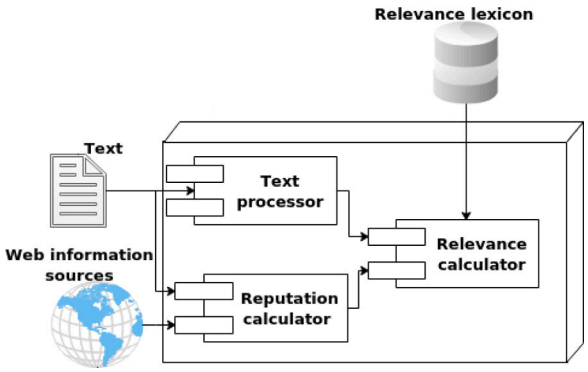


Fig. 5. Internal architecture of the *Relevance calculator* module.

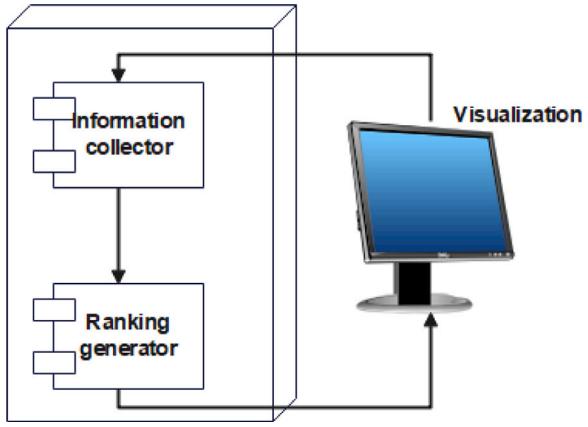


Fig. 6. Internal architecture of the *Articles recommendation* module.

formed by substantives in the form of uni-grams. These words are matched with the pre-trained lexicon to obtain the influence of the manuscript in the domain. Notice that in this release the neural network used in the original *Webelance* framework has not been considered.

The *Reputation calculator* component estimates the reputation of articles through a set of parameters provided by Web information sources (in this case, the Semantic Scholar API (Kinney et al., 2023)). The reputation is calculated using the equations provided by the *Webelance* system upgrading the equation of the reputation of authors as follows:

$$rep_i = \omega_1 \cdot inf\_citation\_count + \omega_2 \cdot h-index + \omega_3 \cdot seniority + \omega_4 \cdot papers\,, \quad (2)$$

where $\sum_{i=1}^{4} \omega_i = 1$. The $inf\_citation\_count$ parameter corresponds to the high influential cites count of the author, while the $h-index$ parameter provides a measure of an author's professional quality based on the number of citations received. The $seniority$ parameter represents the elapsed years between the first and the last scientific publication of the author, and the $papers$ parameter considers the number of publications of the author.

The *Relevance calculator* component addresses the calculation of the relevance of the manuscript by considering the outcomes from the preceding components. Thus, the influence of the article and the estimated reputation are mixed here to produce the desired result.

4.5. Articles recommendation

This module produces the final result of the system. It ranks the set of filtered manuscripts according to their estimated similarity with a previously selected diagnosis and their associated relevance.

Regarding the architecture, it presents two components: the *Information collector* and the *Ranking generator* (see Fig. 6). The first gathers the relevance and the similarity values previously calculated, while the second generates the final recommendation.

Delving into the *Information collector* component, it organised the n most similar articles according to the estimated similarity with the selecting diagnosis using the corresponding modules of the system. For each one of these manuscripts, their relevance is calculated. Notice that the range of values is the same for the similarity measure and the relevance, being in both cases 0 the worst scenario and 1 the optimal scenario.

The *Ranking generator* component uses the Euclidean distance between each pair: similarity-relevance to reach a compromise between the similarity value and the relevance value. Thus, the distance to the optimal point $[similary = 1, relevance = 1]$ is the value used to rank the articles. Notice that only i articles (being $i <= n$) are used to make the final recommendation. This value i is addressed through a configurable parameter of the system.

5. Experiments

This section presents the different experiments achieved to evaluate the viability of the proposal. It demonstrates the efficacy of the MELENDI system through a series of comparative analyses:

- Diagnosis Extractor Module: The performance of different NER tools is compared, with MedCAT emerging as the most effective for detecting medical entities.
- Semantic Similarity Models: Two models are evaluated for their ability to match diagnoses with scientific manuscripts. The cosine similarity model is chosen for its interpretability and bounded results.
- Real-World Scenario: The system's recommendations are compared with the selections made by medical experts. The system successfully recommended relevant articles in 9 out of 10 cases, aligning closely with expert choices.

These analyses validate the system's ability to provide accurate and relevant scientific article recommendations based on patient diagnoses.

Notice that the three experiments have involved a set of 5 experts in the medical domain from IBSAL. They reached a consensus and voting the results of the evaluations. They evaluated the elicitation of diagnoses, the relevance of the articles returned by the semantic similarity model according to those diagnoses, and the impact of combining semantic similarity with the intrinsic relevance of each returned article.

5.1. Evaluation of the diagnosis extractor module

The *Diagnosis extractor* module is one of the relevant modules of MELENDI. It is in charge of finding the diagnosis chunks in texts and then, extracting the relevant Named Entities related to the medical domain. This latter is prone to be achieved through different tools, which leads to considering some experiments to select the most appropriate.

In the current times, several NER tools especially focused on the medical domain have been developed. Traditional specialist systems such as BioBERT (Lee et al., 2020), MedCAT (Kraljevic et al., 2021), and Scispacy (Neumann et al., 2019) have established strong foundations in the field. While Generative AI, evidenced by emerging research on prompting techniques for LLM, promises to further revolutionise NER tasks, it must also align with the rigorous demands of the medical domain Hu et al. (2024).

To ensure the relevance and accuracy of NER within the medical context, it is imperative to integrate tools that interface with the UMLS. The linkage to UMLS is vital as systems like MedCAT and Scispacy can verify the medical specificity of detected Named Entities. This integration is particularly pertinent as MELENDI includes these terms to produce word embeddings that will be used to find related scientific manuscripts of the healthcare domain. In this case, MedCAT and Scispacy are the ones that accomplish that fact.

Table 1
Comparison in number of entities and diseases detected in 500 EHRs between MedCAT, SciSpacy, BioBERT and Exact-Matching.

NER System	Entities detected	Diseases detected
MedCAT	322,114	15,664
Scispacy	197,435	12,807
BioBERT	10,910	10,910
Exact-Matching	65,475	7,493
GPT4-Turbo	12,415	2,736
GPT4	8,196	3,780
GPT3.5	6,343	2,702

Therefore, a complete experiment has been designed to compare the performance of these two tools concerning other similar ones. This allows situating these approaches in the state-of-the-art according to their ability to detect Named Entities related to the medical domain (*i.e.* mainly diseases, drugs, and treatments). The prompting techniques used in this experiment are a variation of Hu et al. (2024) and are displayed in Annex.

In a first step, healthcare experts have analysed 500 random samples of discharge summaries from the MIMIC-III dataset (Johnson et al., 2016). These experts had as a main task to detect the diagnosis in the texts. Later, in a second step, the system used the different NER tools to evaluate the same samples to measure the performance compared to human beings.

Throughout the experiment, it was necessary to reassess multiple factors to determine the most suitable model for the framework. In the case of BERN2 (Sung et al., 2022), it proved to be a challenging option due to its high resource requirements. However, it demanded over 70GB of disk space, 63.5 GB of RAM, and 5.05 GB of GPU, making the deployment difficult in a practical setting. As most of the possible medical systems would struggle to allocate the necessary resources, the prospect of integrating this model into MELENDI was discarded. Furthermore, in the case of the Med-Flair model (ElDin et al., 2021), it was inaccessible and untraceable, also resulting in its exclusion from the evaluation.

Results yield that MedCAT is the best tool to achieve the NER task. Scispacy using the *en_core_scibert* model, is significantly worse at detecting Named Entities related to the medical domain. The others present similar results in particular when referring to disease or syndromes, which is a good indicator of the performance of MedCAT (see Table 1). As it can be seen, Scispacy, a fine-tuned BioBERT for disease recognition, and exact matching (look for an exact match between the UMLS descriptions and the texts from the EHR) detects significantly fewer entities, and diseases or syndromes.

Interestingly, the performance of LLMs represented by various versions of GPT—suggests a considerable gap in comparison to dedicated NER systems. This underscores the limitations present within current prompting methods when utilised for NER tasks. While Generative AI possesses significant potential for a multitude of applications, its current instantiation via simple prompting cannot rival the accuracy and depth of specialised NER models in the medical domain.

In addressing the issue of section identification within the EHRs, results demonstrated that the system successfully isolated the primary diagnosis in 75% of the documents. Incorporating the Generative AI system, specifically tailored for the task of section identification, it accurately extracted and reconstructed the main diagnosis in 96.3% of cases.

This proficiency can be ascribed to the extensive linguistic knowledge of the LLM, which it leverages to infer and compile the diagnosis from disparate sections without the necessity for explicit domain-specific understanding. This fact illustrates a fairly positive performance of the system.

5.2. Evaluation of the semantic similarity calculator module

The *Semantic similarity calculator* module is the other most relevant module of MELENDI. It is in charge of measuring the similarity between documents and the Named Entities present in the detected diagnosis. Following the state-of-the-art approaches, two similarity measures are the most typical ones: dot product and cosine similarity.

The dot product is usually used in asymmetric semantic search. Currently, the most interesting approaches in this field, both theoretically and potentially, lie in models that have been fitted using this similarity measure. On the other hand, the cosine similarity is usually oriented to symmetric semantic search. This fact should not fit so well with the proposed case in the system, as the abstracts of scientific manuscripts are usually more extensive than the set of concepts detected in the diagnosis chunks.

An experiment to confirm and validate this assumption has been proposed. To achieve this, two pre-trained models have been selected to represent each one of the two possibilities. In the case of the dot product, *msmarco-distilbert-base-tas-b* (Hofstätter et al., 2021) has been the candidate. For the case of the cosine similarity, *msmarco-distilbert-base-v4*(Reimers and Gurevych, 2019) has been picked. Both models have been selected due to the good balance between computational efficiency and good performance in general-purpose semantic similarity tests. A model trained for semantic similarity in the medical domain should be considered in further research.

Notice that one of the characteristics of the dot product is that the result is not bounded, making it difficult to interpret, combine and compare with other metrics. Therefore, the evaluation of the experiment has been designed to compare the performance of both models with 10 random diagnoses. This process consists of evaluating a set of 1,000 scientific manuscripts obtained by applying the dot product and cosine similarity respectively. Then, the 3 most similar scientific manuscripts to the diagnoses and, in addition, the 3 less similar scientific manuscripts (*i.e.* namely positions 998, 999, and 1000) are selected. This filter obtains a total of 120 manuscripts (*i.e.* 60 for each measure) that are evaluated by the previously selected 5 medical experts.

Thus, for each of the 10 diagnoses, the expert received 12 manuscripts, 6 related to the cosine similarity (m_i^c, $i = 1 \dots 6$), and 6 related to the dot product (m_i^d, $i = 1 \dots 6$). Notice that some of the manuscripts can be included in both resulting sets (*i.e.* a manuscript can be selected using the dot product and also the cosine similarity). Thus, m_i^c and m_i^d ($i = 1 \dots 3$) correspond to the most similar scientific manuscripts to the diagnoses using the cosine and the dot product measures, respectively. In addition, m_i^c and m_i^d ($i = 4 \dots 6$) correspond to the less similar scientific manuscripts to the diagnoses using the cosine and the dot product measures, respectively.

Next, the experts were asked to select the 3 manuscripts that best represent the corresponding diagnosis. Finally, the opinions of the experts and the outcome of the system using both measures are compared. This allows selecting the measure that fits better according to the human validation.

The result of the experiment can be seen in Tables 2 and 3. The first table illustrates the assigned ID of each diagnosis and the Named Entities detected in the diagnosis chunks. The second provides detailed information comparing the choices of the measures and the global opinion of human experts. There, each diagnosis also uses an id, having the 6 manuscripts recommended using the dot product, and the other 6 recommended using the cosine similarity. The manuscripts are ordered from left to right in decreasing order according to their respective scores. Thus, the manuscripts corresponding to the first 3 cells have the first, second, and third highest similarity scores respectively, while the last 3 correspond to positions 997, 998, and 999 respectively. This arrangement is used in both cases: dot product and cosine similarity.

The manuscripts with a high similarity that were selected by the experts appear in green colour. However, the manuscripts with a low

Table 2
Isolated diagnoses from ten different EHRs used to evaluate entire system.

EHR	Diagnosis
1	coronary artery disease, Diabetes Mellitus Type 2, hypertension.
2	COPD exacerbation.
3	Acute pancreatitis, Hypertriglicidemia, Metastatic Breast Cancer.
4	Pneumonia, asthma, diabetes.
5	Gastric perforation from marginal, ulcer.
6	Acute on Chronic Diastolic Congestive Heart Failure, Atrial fibrillation with rapid ventricular response.
7	Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure.
8	Amyotrophic Lateral Sclerosis.
9	Chronic Type A Aortic Dissection, s/p Repair ascending dissection reanastamosis of grafts , Coronary Artery Disease, Atrial Fibrillation, s/p Coronary Artery Bypass Grafting and Maze Procedure on, Hypertension, Dyslipidemia, h/o of transient ischemic attack.
10	Hyponatremia secondary to Fanconis syndrome versus SIADH.

Table 3
Comparative between the opinion of experts and the results provided by the dot product and cosine similarity.

Diagnosis	Manuscripts using DOT						Manuscripts using COS					
	m_1^d	m_2^d	m_3^d	m_4^d	m_5^d	m_6^d	m_1^c	m_2^c	m_3^c	m_4^c	m_5^c	m_6^c
1	OK	OK										KO
2	OK		OK				OK	OK	OK			
3			OK	KO			OK					
4	OK							OK	OK			
5	OK		OK				OK	OK				
6	OK		OK				OK		OK			
7		OK					OK	OK				
8	OK	OK		KO			OK					
9	OK	OK		KO			OK					
10		OK					OK					KO

similarity that were selected by the experts appear in red. The rest of them appear with no colour, indicating no choice made by experts. The expert selected as relevant the manuscript with the highest score 7 times in the case of the dot product and 8 times in the case of the cosine similarity. It can be seen that the cosine similarity approach has fewer low-similarity manuscripts selected by experts. This leads to including the measure in the module of the system. This issue is relevant because this measure provides a better interpretability and it is easier to combine it with the relevance measure, as both have boundaries between 0 and 1.

Finally, it is important to indicate the high correlation between the selections made by experts and the manuscripts provided by both measures. Therefore, it can be concluded that the module works with high quality and the results are truthful.

5.3. Evaluation of the system in a real-world scenario

This experiment shows the viability of the proposal in a real-world scenario. To achieve it, a complete set-up of MELENDI is considered. Word embeddings of $1,000,000$ abstracts coming from scientific manuscripts randomly gathered from PubMed have been used as a database. Moreover, the parameters related to relevance (see Section 4.4) have been fixed. In this case, the ω_n parameters take the following values: $\omega_1 = 0.1$, $\omega_2 = 0.2$, $\omega_3 = 0.3$, $\omega_4 = 0.4$. These weights are related to the importance given to the different elements that the *Relevance calculator* module uses to produce its outcome. Following (Fernández-Isabel et al., 2020), the rest of the parameters related to the *Webelance* system maintain their standard configuration.

Once the system is ready to work, the 10 EHRs and their evaluations made by the experts used in the previous experiment have been selected to be analysed here (see Table 2).

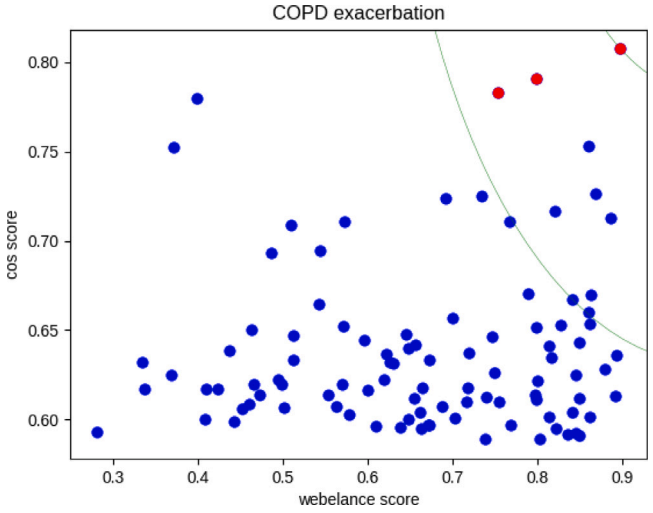


Fig. 7. Representation of articles distributed according to their relevance and cosine similarity to the diagnosis *COPD exacerbation*.

Then, in this experiment, the relevance and the similarity are jointly evaluated. For this purpose, the top 100 manuscripts with the highest cosine similarity for each of the 10 diagnoses and their respective relevance were calculated. The top 10 manuscripts nearest to the ideal point ([relevance, cosine similarity] = $[1, 1]$) are obtained for each diagnosis.

Next, a comparison between the relevance of the 10 selected manuscripts and the opinion of the group of experts is established.

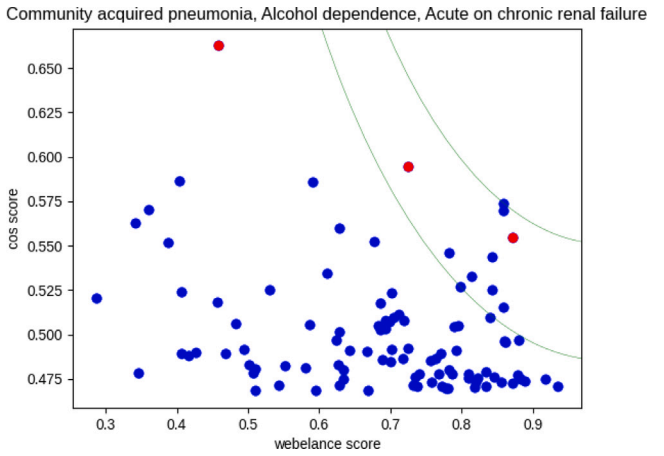


Fig. 8. Representation of articles distributed according to their relevance and cosine similarity to the diagnosis *Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure*.

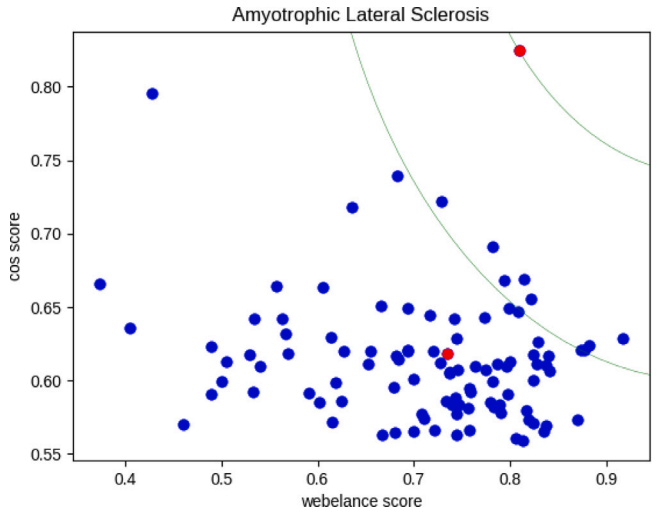


Fig. 9. Representation of articles distributed according to their relevance and cosine similarity to the diagnosis *Amyotrophic Lateral Sclerosis*.

In Fig. 7, the top 100 manuscripts for the diagnosis *COPD exacerbation* with id = 2 are shown. The selected top 10 are marked within the two green curves. In addition, those articles selected as the most relevant by the experts in the previous experiment (3 as a maximum) are represented in red. In this particular case, the 3 articles fall within the top 10 recommended by the system.

On the other hand, in Fig. 8, corresponding to the diagnosis *Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure* with id = 7, 2 (out of 3) papers are within the top 10. The other paper falls out of the top 10 but is within the top 100 due to its high cosine score. Notice that it is the paper with the highest similarity. However, its relatively low relevance score penalises it.

Finally, the results for the diagnosis *Amyotrophic Lateral Sclerosis* with id = 8 are represented in Fig. 9. In this case, only 1 of the 3 papers selected by the experts is within the top 10 of the recommended papers. Notice that this one appears in the first position. However, the second paper has a relatively high cosine similarity and relevance score locating it within the top 50. The third is outside the top 100 manuscripts with a low cosine score of 0.414.

The results of the 10 diagnoses have been compiled in Table 4. There, it can be seen the relative position for each of the 3 papers selected by the experts compared with the system recommendation.

Table 4
Comparison between the evaluations of the experts and the recommendations made by MELENDI. Each row shows the three papers selected by experts for each diagnosis with the position of the article in the ranking proposed by the system (cosine similarity value).

Diagnosis	Paper 1	Paper 2	Paper 3
1	1	4	>100
2	1	2	6
3	>50	>100	>100
4	1	>10	>10
5	1	>10	>10
6	1	2	>100
7	3	4	>10
8	1	>10	>100
9	2	3	>100
10	2	>10	>100

Delving into the results presented in Table 4, it can be considered that a paper far away than the 50th position could be a possible error of the system. Thus, the experiment counts on 23 hits and 7 misses under this criteria. Comparing the results of the system with the choice of the experts can be stated that MELENDI obtains an accuracy of 0.77.

Notice that to reinforce these results, the Friedman test and subsequent Nemenyi test (Holander and Wolfe, 1973) have been used to accentuate the uniqueness of the proposed method. In this case, no significant differences appeared between the MELENDI and the experts in the choice of the most important paper in the proposed diagnoses.

After concluding the experiment, the following conclusions can be drawn:

- In 9 of the 10 diagnoses the system has recommended at least 1 of the 3 papers selected by the experts, always within the top 3 nearest to the ideal point.
- In 6 of the 10 diagnoses the paper with the best score of the system was selected by the human experts.
- In 5 of 10 diagnoses there are at least 2 papers selected by the human experts within the top 5 selected by the system.
- In 6 of 10, human experts have selected at least one paper outside the top 100 recommended by the system.

In conclusion, the system can properly provide knowledge in a similar way to experts in the medical domain. On the other hand, the assumption of the importance of relevance is accomplished, as the system includes some manuscripts in the recommendation according to this feature. This issue is interesting, as the relevance is automatically calculated, and it is a difficult task to achieve for human beings, as they do not usually know the quality of the authors or the importance of the journal where the manuscript was published. Thus, the system allows users to automatically update the knowledge about a diagnosis (and also its treatments) being able to stop worrying about checking the quality indices of the publications where it is addressed.

6. Conclusions

This paper introduces MELENDI, a scientific paper recommendation system for healthcare professionals that allows them to keep up to date with advances related to the diagnoses of their patients.

The main novelty of the proposal is based on the hybridisation of a semantic similarity model with an automatic scientific article relevance estimator to produce the final ranking of scientific articles. In this way, the recommendation achieves a trade-off between the relationship between the recommended paper with its corresponding diagnosis and the relevance of the paper itself. This trade-off ensures that the recommended articles are strictly related to the patients and relevant in their subject matter and origin.

Experiments have been carried out with two semantic similarity models, one tuned for the scalar product and the other for the cosine distance. These experiments, carried out by a group of medical

specialists from IBSAL have shown insignificant differences between both models, selecting the model adjusted for cosine similarity for simplicity and interpretability of the results. On the other hand, the top 100 articles with the best cosine similarity concerning 10 diagnoses extracted from 10 EHRs of MIMIC-III were calculated to evaluate the effectiveness of this recommendation. The obtained results indicate that in 9 of the 10 diagnoses, the system has recommended at least one of the 3 most relevant papers selected by the IBSAL experts, all of them within the 3 nearest to the ideal point of similarity and relevance. It demonstrates the good performance of the recommendation system and supports its possible implementation in a real environment.

Despite the good performance found, this work has had different limitations. Firstly, the diagnostic detection system has been adapted to the typology and casuistry of MIMIC-III discharge reports, which limits its implementation with other types of EHRs. For this reason, developing a system that fully automates diagnostic detection independent of the typology of the EHRs will be contemplated for future work, subject to the availability of labelled data for this purpose. Moreover, the semantic similarity model is domain-general, so fine-tuning the model for the medical domain and the inclusion of LLMs to produce responses in natural language would positively impact its performance. This fine-tuning process will require the system to be in production and collect user feedback through an interactive interface. In the case of LLMs, they should have only a supportive role in the system due to the problems with hallucinations in a relevant domain such as healthcare. Finally, the trade-off between relevance and similarity could also be adjusted according to the relative importance given by the user to each feature, resulting in a system adapted to the user's particularities.

CRediT authorship contribution statement

Adrián Alonso Barriuso: Writing – original draft, Resources, Investigation. **Alberto Fernández-Isabel:** Writing – original draft, Supervision, Methodology, Investigation. **Isaac Martín de Diego:** Validation, Supervision, Conceptualization. **Alfonso Ardoiz:** Resources, Methodology, Formal analysis. **J.F. J. Viseu Pinheiro:** Validation, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Research supported by grants from the Spanish Ministry of Science and Innovation, under the Knowledge Generation Projects program: XMIDAS (PID2021-122640OB-100) and from Madrid Autonomous Community, Spain (ND2019/TIC-17169); medical corpus provided by Dezzai; and donation of the Titan V GPU by NVIDIA Corporation.

Annex

NER Experiment:

```
### Task
Your task is to work as a Medical NER system to generate an HTML
a version of an input text, marking up specific entities related to
healthcare.
The entities to be identified are: "medical entity" and "disease or
syndrome". Use HTML <span> tags to highlight these entities.
Each <span> should have a class attribute indicating the type of the
entity.

### Entity Markup Guide
Use <span class="medical entity"> to denote any medical entity.
Use <span class="disease or syndrome"> to denote specific entities
related to diseases or syndromes.
Leave the text as it is if no such entities are found.

### Entity Definitions
Disease or Syndrome is defined as a disorder of structure or
function in a human, animal, or plant, especially one that has
a known cause and a distinctive group of symptoms, signs,
or anatomical changes.

### Annotation Guidelines
Only complete noun phrases (NPs) and adjective phrases (APs) should be
marked. Terms that fit concept semantic rules, but that are only used
as modifiers in a noun phrase should not be marked.
Include all modifiers with concepts when they appear in the same
phrase except for assertion modifiers.
You can include up to one prepositional phrase (PP) following a
markable concept if the PP does not contain a markable concept and
either indicates an organ/body part or can be rearranged to
eliminate the PP (we later call this the PP test).
Include articles and possessives.
Conjunctions and other syntax that denote lists should be included
if they occur within the modifiers or are connected by a common set
of modifiers. If the portions of the list are otherwise
independent, they should not be included.
Similarly, when concepts are mentioned in more than one way in the
same noun phrase (such as the definition of an acronym or where a
generic and a brand name of a drug are used together), the concepts
should be marked together.
Concepts should be mentioned in relation to the patient or someone
else in the note. Section headers that provide formatting, but
that are not specific to a person are not marked.

### Examples
Example Input1: He had been diagnosed with osteoarthritis of the
knees and had undergone arthroscopy years prior to admission.
Example Output1: He had been diagnosed with <span class="disease
or syndrome">osteoarthritis of the knees</span> and had
undergone <span class="medical entity"> arthroscopy</span >
years prior to admission.

Example Input2: After the patient was seen in the office on
August 10, she persisted with high fevers and was admitted
on August 11 to Cottonwood Hospital.
Example Output2: After the patient was seen in the office on
August 10 , she persisted with <span class="medical entity">
high fevers</span> and was admitted on August 11 to Cottonwood
Hospital.

Example Input3: HISTORY OF PRESENT ILLNESS: The patient
is an 85 - year - old male who was brought in by EMS with
a complaint of a decreased level of consciousness.
Example Output3: HISTORY OF PRESENT ILLNESS: The patient
is an 85 - year - old male who was brought in by EMS with
a complaint of <span class="medical entity"> a decreased
level of consciousness</span >.

Example Input4: Her lisinopril was increased to 40 mg daily.
Example Output4: Her <span class="medical entity">
lisinopril</span> was increased to 40 mg daily.

### Final Instructions
Take a deep breath and work on this problem step-by-step.
You will be rewarded if the quality of your answer is top-notch.
```

References

Akila, D., Balaganesh, D., 2021. Semantic web-based critical healthcare system using Bayesian networks. Mater. Today Proc.

Alves, M.A., Castro, G.Z., Oliveira, B.A.S., Ferreira, L.A., Ramírez, J.A., Silva, R., Guimarães, F.G., 2021. Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. Comput. Biol. Med. 132, 104335.

Bentaiba-Lagrid, M.B., Bouzar-Benlabiod, L., Rubin, S.H., Bouabana-Tebibel, T., Hanini, M.R., 2020. A case-based reasoning system for supervised classification problems in the medical field. Expert Syst. Appl. 150, 113335.

Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32 (suppl_1), D267–D270.

Castiglioni, I., Ippolito, D., Interlenghi, M., Monti, C.B., Salvatore, C., Schiaffino, S., Polidori, A., Gandola, D., Messa, C., Sardanelli, F., 2021. Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: a first experience from Lombardy, Italy. Eur. Radiol. Exp. 5 (1), 1–10.

Chowdhury, G.G., 2010. Introduction to Modern Information Retrieval. Facet publishing.

De Croon, R., Van Houdt, L., Htun, N.N., Štiglic, G., Abeele, V.V., Verbert, K., et al., 2021. Health recommender systems: systematic review. J. Med. Internet Res. 23 (6), e18035.

Duan, J., Jiao, F., 2021. Novel case-based reasoning system for public health emergencies. Risk Manag. Healthc. Policy 14, 541.

ElDin, H.G., AbdulRazek, M., Abdelshafi, M., Sahlol, A.T., 2021. Med-Flair: medical named entity recognition for diseases and medications based on Flair embedding. Procedia Comput. Sci. 189, 67–75.

Fernández-Isabel, A., Barriuso, A.A., Cabezas, J., de Diego, I.M., Pinheiro, J.V., 2020. Knowledge-based framework for estimating the relevance of scientific articles. Expert Syst. Appl. 161, 113692.

Gianola, S., Jesus, T.S., Bargeri, S., Castellini, G., 2020. Characteristics of academic publications, preprints, and registered clinical trials on the COVID-19 pandemic. PLoS One 15 (10), e0240123.

Hofstätter, S., Lin, S.-C., Yang, J.-H., Lin, J., Hanbury, A., 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. arXiv preprint arXiv:2104.06967.

Holander, M., Wolfe, D.A., 1973. Nonparametric statistical methods, vol. 497, John Wilew and Sons Inc. Publications, New York.

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V.K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., et al., 2024. Improving large language models for clinical named entity recognition via prompt engineering. J. Am. Med. Inform. Assoc. 1–10.

Hunt, E.B., 2014. Artificial Intelligence. Academic Press.

Isern, D., Moreno, A., 2016. A systematic literature review of agents applied in healthcare. J. Med. Syst. 40 (2), 1–14.

Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., Zhao, L., Cross-lingual information retrieval with BERT. In: LREC 2020 Language Resources and Evaluation Conference 11–16 May 2020. p. 26.

Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. Scient. Data 3 (1), 1–9.

Johnson, R., Watkinson, A., Mabe, M., 2018. The STM report. In: An Overview of Scientific and Scholarly Publishing, fifth ed. p. 94.

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol. 18 (1), 1–12.

Khodabandehloo, E., Riboni, D., Alimohammadi, A., 2021. HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline. Future Gener. Comput. Syst. 116, 168–189.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., et al., 2023. The semantic scholar open data platform. arXiv preprint arXiv:2301.10140.

Kolodner, J., 2014. Case-Based Reasoning. Morgan Kaufmann.

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A.A., Roberts, A., et al., 2021. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. Artif. Intell. Med. 117, 102083.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.

Liebowitz, J., 2019. The Handbook of Applied Expert Systems. cRc Press.

Mala, V., Lobiyal, D., 2016. Semantic and keyword based web techniques in information retrieval. In: 2016 International Conference on Computing, Communication and Automation. ICCCA, IEEE, pp. 23–26.

Mansour, R.F., El Amraoui, A., Nouaouri, I., Díaz, V.G., Gupta, D., Kumar, S., 2021. Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems. IEEE Access 9, 45137–45146.

Marcot, B.G., Penman, T.D., 2019. Advances in Bayesian network modelling: Integration of modelling technologies. Environ. Model. Softw. 111, 386–393.

Martinez, M.L., Vazquez, D.A., Maya, D., Olvera, X., Guzman, G., Torres, M., Quintero, R., Moreno, M., 2014. Geospatial recommender system for the location of health services. In: 2014 14th International Conference on Computational Science and Its Applications. IEEE, pp. 200–203.

Mousavi, S.M., Abdullah, S., Niaki, S.T.A., Banihashemi, S., 2021. An intelligent hybrid classification algorithm integrating fuzzy rule-based extraction and harmony search optimization: Medical diagnosis applications. Knowl.-Based Syst. 220, 106943.

Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669.

Olivas-Martínez, A., Cárdenas-Fragoso, J.L., Jiménez, J.V., Lozano-Cruz, O.A., Ortiz-Brizuela, E., Tovar-Méndez, V.H., Medrano-Borromeo, C., Martínez-Valenzuela, A., Román-Montes, C.M., Martínez-Guerra, B., et al., 2021. In-hospital mortality from severe COVID-19 in a tertiary care center in Mexico city; causes of death, risk factors and the impact of hospital saturation. PloS One 16 (2), e0245772.

Passalis, N., Tefas, A., 2018. Learning bag-of-embedded-words representations for textual information retrieval. Pattern Recognit. 81, 254–267.

Ploug, T., Holm, S., 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. Artif. Intell. Med. 107, 101901.

Pomares-Quimbaya, A., Kreuzthaler, M., Schulz, S., 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. BMC Med. Res. Methodol. 19, 1–20.

Rauber, T.W., da Silva Loca, A.L., de Assis Boldt, F., Rodrigues, A.L., Varejão, F.M., 2021. An experimental methodology to evaluate machine learning methods for fault diagnosis based on vibration signals. Expert Syst. Appl. 167, 114022.

Raynaud, M., Zhang, H., Louis, K., Goutaudier, V., Wang, J., Dubourg, Q., Wei, Y., Demir, Z., Debiais, C., Aubert, O., et al., 2021. COVID-19-related medical research: a meta-research and critical appraisal. BMC Med. Res. Methodol. 21 (1), 1–11.

Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Riedl, M.O., 2019. Human-centered artificial intelligence and machine learning. Hum. Behav. Emerg. Technol. 1 (1), 33–36.

Saha, P., Sadi, M.S., Islam, M.M., 2021. EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers. Inform. Med. Unlocked 22, 100505.

Sanderson, K., 2023. GPT-4 is here: what scientists think. Nature 615 (7954), 773.

Scheepers-Hoeks, A.-M.J., Grouls, R.J., Neef, C., Ackerman, E.W., Korsten, E.H., 2013. Physicians' responses to clinical decision support on an intensive care unit—comparison of four different alerting methods. Artif. Intell. Med. 59 (1), 33–38.

Singla, J., Grover, D., Bhandari, A., 2014. Medical expert systems for diagnosis of various diseases. Int. J. Comput. Appl. 93 (7).

Stark, B., Knahl, C., Aydin, M., Elish, K., 2019. A literature review on medicine recommender systems. Int. J. Adv. Comput. Sci. Appl. 10 (8).

Sun, Y., Wang, L., Jiang, Z., Li, B., Hu, Y., Tian, W., 2020. State recognition of decompressive laminectomy with multiple information in robot-assisted surgery. Artif. Intell. Med. 102, 101763.

Sung, M., Jeong, M., Choi, Y., Kim, D., Lee, J., Kang, J., 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics 38 (20), 4837–4839.

Tan, H.X., Teo, C.H.D., Ang, P.S., Loke, W.P.C., Tham, M.Y., Tan, S.H., Soh, B.L.S., Foo, P.Q.B., Ling, Z.J., Yip, W.L.J., et al., 2022. Combining machine learning with a rule-based algorithm to detect and identify related entities of documented adverse drug reactions on hospital discharge summaries. Drug Safety 45 (8), 853–862.

Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. Nat. Med. 29 (8), 1930–1940.

Waqar, M., Majeed, N., Dawood, H., Daud, A., Aljohani, N.R., 2019. An adaptive doctor-recommender system. Behav. Inf. Technol. 38 (9), 959–973.

White, J., 2020. PubMed 2.0. Med. Ref. Serv. Q. 39 (4), 382–387.

Zhang, Q., Zhang, G., Lu, J., Wu, D., 2015. A framework of hybrid recommender system for personalized clinical prescription. In: 2015 10Th International Conference on Intelligent Systems and Knowledge Engineering. ISKE, IEEE, pp. 189–195.