PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

INTERACTIVE AND EXPLAINABLE MACHINE LEARNING TO IMPROVE EFFICIENCY IN MEDICAL DOCUMENT SCREENING

ANDRÉS FRANCISCO CARVALLO DE FERARI

Thesis submitted to the Office of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Doctor in Engineering Sciences

Advisor:
DENIS A. PARRA.

Santiago of Chile, September 2022

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

INTERACTIVE AND EXPLAINABLE MACHINE LEARNING MODEL TO IMPROVE EFFICIENCY IN MEDICAL DOCUMENT SCREENING

ANDRÉS FRANCISCO CARVALLO DE FERARI

Members of the Committee:

DENIS PARRA

MARCELO MENDOZA

HANS LÖBEL

GABRIEL RADA

EDUARDO VEAS

HÉCTOR JORQUERA

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Santiago of Chile, September 2022

To my family.

ACKNOWLEDGEMENTS

First and foremost, I am grateful to my wife and parents for their unwavering support throughout this challenging journey. I also appreciate the IALab members for maintaining active collaboration despite physical distance due to the pandemic, and I thank Ivania Donoso and Hernán Valdivieso for their assistance with this thesis.

My heartfelt thanks go to the professors and research collaborators who offered inspiration, corrected errors, and helped me make the most of my hard work and persistence. I am also thankful to committee members Marcelo Mendoza, Gabriel Rada, and Hans Lobel for their valuable feedback, which enabled me to enhance my work and explore new approaches to the primary issue. I extend my gratitude to Gabriel for his support from the Epistemonikos Foundation and its collaborators.

Lastly, I wish to express my appreciation to my advisor Denis Parra for believing in my abilities and for his patience throughout the entire process, including the numerous revisions of the initial journal article drafts, this thesis, and related publications.

This research received funding from the following sources: the Millenium Institute Fundamentals on Data, the CONICYT FONDECYT Regular Project (grant number 1191791) and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Document screening is a fundamental task within Evidence-based Medicine (EBM) that seeks to validate scientific evidence to support medical decisions. This thesis proposes an active learning-based setting for document screening in EBM to reduce the number of documents that physicians need to label for answering clinical questions. Moreover, given the context of the COVID-19 pandemic, the number of indexed documents increased exponentially, so there is a need to sample articles to fine-tune the model aiming to improve its performance using a small proportion of the total examples. Through a user study, we evaluate whether visualizing the attention of a transformer-based model as highlighted words in the abstract is perceived as helpful for users on document classification and if there is a preferred encoding to visualize these attentions. Concerning active learning, our results indicate that uncertainty sampling combined with a BioBERT document representation and a Random Forest outperforms other proposed approaches. Furthermore, for COVID-19 article classification, we obtained that the XLNET language model outperformed other state-of-the-art models. We showed that we could save more than 65% of experts' workload using an uncertainty-sampling strategy, measured as the number of documents needed to review manually. Results from the user study indicate that, in general, attention is not perceived as helpful. However, there is an interaction between the type of article and visual encoding in the perception of helpfulness of attention as an explanation. Moreover, we provide evidence that using attention as an explanation improves users' performance since users who use visualizations obtain an increase of 5.27% (pd accuracy) compared to users who do not use any visualization.

Keywords: **Natural Language Processing, Active Learning, XAI, Evidence-based medicine**.

RESUMEN

La revisión de documentos es fundamental en Medicina Basada en Evidencia (MBE) ya que busca validar evidencia científica para respaldar decisiones clínicas. Esta tesis propone una solución a la sobrecarga de información basada en active learning que busca reducir la cantidad de documentos que los médicos deben etiquetar para responder preguntas clínicas. Además, en el contexto de la pandemia COVID-19 la cantidad de articulos indexados creció exponencialmente, proponemos estrategias de sampleo de evidencia para hacer finetuning de un modelo con una pequeña proporcion de toda la evidencia existente. Finalmente, mediante un estudio de usuario evaluamos si las atenciones aprendidas por un modelo basado en transformer son percibidas como útiles y si existe alguna forma mejor para visualizarlas. Con respecto a Active Learning los resultados indican que el muestro basado en incerteza combinado con representación BioBERT y un Random Forest supera a otros enfoques propuestos. Respecto a la clasificación de artículos de COVID-19, obtuvimos que el modelo XLNET supera a otros modelos del estado del arte y demostramos que podemos ahorrar más del 65% de la carga de trabajo de los expertos utilizando una estrategia de muestreo basado incerteza. Finalmente, los resultados del estudio de usuario indican que, en general, las atenciones no son percibidas como utiles para los usuarios como una forma de explicación. Sin embargo, observamos un efecto de interaccion entre el encoding visual y el tipo de artículo con respecto a la percepción de utilidad de las atenciones. Además obtuvimos que los usuarios que visualizan las atenciones tienen una efectividad de un 5.27% mayor comparado a aquellos que no utilizan visualización.

Palabras Claves: **Active Learning, Inteligencia Artificial Explicable, Medicina basada en Evidencia, Modelos de Lenguaje**.

1. INTRODUCTION

Evidence-based Medicine (EBM) is a practice that provides scientific evidence to support medical decisions. This evidence is obtained from biomedical journals, usually accessible through the portal PubMed[1], a search engine, which provides free access to abstracts of biomedical research articles, as well as to the MEDLINE database. An existing problem is to find relevant documents given a clinical question or a query within a massive volume of information. As a consequence, the time required for search and screening of articles can take long, and sometimes it consumes a large part of a physician's workday (Miwa et al., 2014; Elliott et al., 2014). When people conduct this repetitive task, there is a good chance of overlooking relevant articles, which can have a negative impact on decisions such as the patient's treatment (Keselman & Smith, 2012).

Moreover, the publication of medical papers has grown exponentially in the last decade. Since 2005, PubMed has indexed more than 1 million articles per year, which means that the process of searching and manual screening of medical evidence will become increasingly more difficult for physicians without the support of information retrieval and machine learning algorithms. For this reason, some systems have emerged to support experts in the collection of evidence such as Embase[2], DARE[3] and Epistemonikos[4].

Furthermore, the rapid spread of COVID-19 since late 2019, pushed research related to this disease shown by more than 200,000 new articles indexed, with a peak of more than 23,000 new papers indexed per month[5]. Given this context, EBM discipline has turned

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://www.elsevier.com/solutions/embase-biomedical-research
[3] https://www.crd.york.ac.uk/CRDWeb/
[4] https://www.epistemonikos.org/en
[5] https://www.science.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat

essential, since new evidence needs to be classified the best way possible given the short period to decide how to approach this disease.

This thesis researches different methods to improve the efficiency and efficacy of document screening in EBM practice. In other words, we aim to reduce the effort made by physicians when they screen documents to find the evidence needed to support the answers to a medical question. Due to the context of massive indexing of evidence related to COVID-19 and the advances in recent years of language models, we researched the efficient classification of documents with a new topic where labeled data is scarce, and propose a solution to automate the process of selecting relevant evidence based on their study design. In addition, as the NLP area has evolved substantially during this last time, we evaluate different language models that better represent medical documents as input to automatized models to improve COVID-19 evidence classification depending on their methodology considered in production on physicians using an evidence-based medicine system. Finally, we evaluate through a user study if using models attention outputs as a visualization of highlighted words of the article's abstract improves user performance for the document screening task and if these explanations are perceived as helpful for the task.

1.1. Hypothesis and research questions

Given the problem of information overload that physicians have to deal with every day to screen novel evidence related to medical subjects, there are two open challenges related to this problem: (1) finding a way to select a proportion of documents to reduce their workload on the document screening task, (2) how to represent these texts best so that a computer correctly interprets them. Another aspect that we want to investigate is if providing explanations improves the user's performance on the document screening task.

Considering the recently described problems, we propose the following research questions for this thesis:

Questions related to offline experiments:

(i) Is there potential to improve the document screening task to answer clinical questions by using active learning strategies?
(ii) What is the most informative way to represent medical articles as vector representations to the model for improving its performance in the document screening task?

Question that involve human experts:

(i) How does the inclusion of explanations influence the decision and reduce the cognitive effort of health experts?
(ii) How do certain types of visual encoding influence health experts on choosing relevant evidence?

To answer the first question of alleviating the work of physicians in document screening, we hypothesize that an active learning approach where a proportion of documents is selected to be reviewed in a limited number of iterations may be the best approach to face the problem of information overload. Concerning the second question, since there has been a significant advance in natural language processing, more recent models based on transformers can generate a more informative representation for a computer.

Concerning questions that involve human experts, we believe that including an explainable framework will improve the performance of physicians in the document screening task, reduce their cognitive effort and make the predictions of the models more *"interpretable"*. Regarding the best encoding of how to visualize these explanations, we

hypothesize that background color is perceived as the preferred way to highlight words in the abstract.

1.2. Contributions

There are two problems we are considering: (1) find a way to reduce the workload in the document screening task and (2) generating interpretable predictions for non-expert users.

The main contributions of this work are the following:

(i) Improving the efficiency of medical experts in the labour of screening evidence relevant to medical treatments.
(ii) Using state-of-the-art language models representing medical documents to improve in the task of document screening.
(iii) Studying how different visualization encodings affect decisions on medical experts related to find relevant evidence to medical treatments.

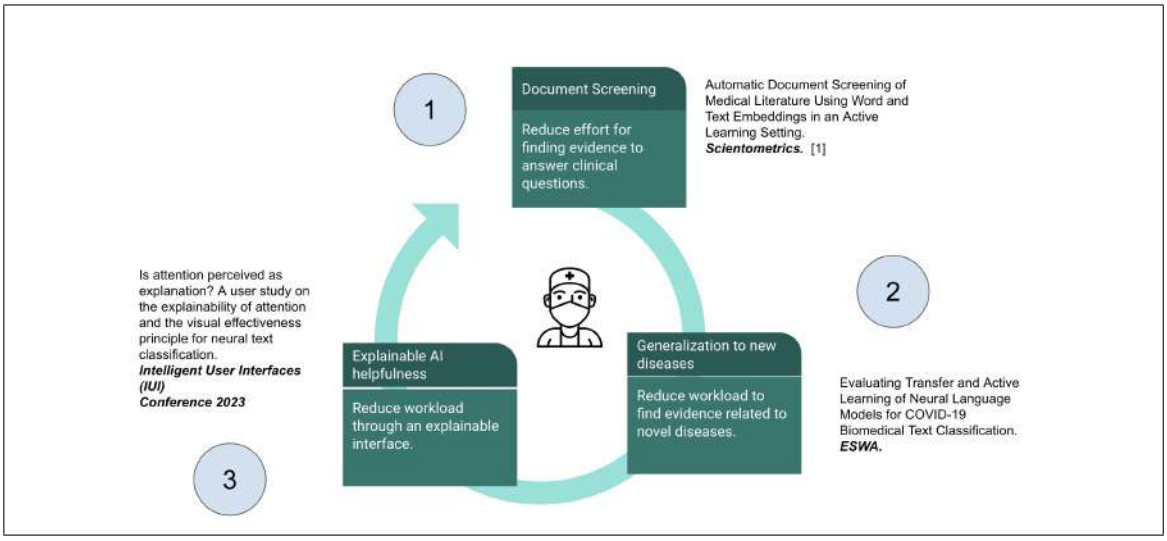


Figure 1.1. Overall diagram: This diagram shows how the thesis and its chapters are organized, with their corresponding publication.

As shown in figure 1.1, it can be seen that in Chapter 3, the problem we address is the information overload for the task of searching for relevant evidence to answer clinical questions. Then in Chapter 4, given that the COVID-19 pandemic appeared, which meant an increase in the quantity of evidence in a short time, we propose a classification system according to the type of study. In addition, we seek to sample a proportion of COVID-19 documents to improve the model's performance. Finally, in the last chapter, we studied how to reduce cognitive load and improve user performance through an explainable interface.

1.2.1. Automatic document screening

As previously discussed, we developed an active learning framework in which we combined active learning strategies, machine learning models, and ways of representing texts (Carvallo, Parra, Lobel, & Soto, 2020; Carvallo & Parra, 2019). We obtained that the best

way to represent text is BioBERT, a language model based on transformers given that was trained with medical texts has a semantic representation adapted to the medical domain. On the other hand, we obtained that the uncertainty sampling strategy, which consists of sampling examples in which a machine learning model is less sure of its prediction, is the one that yields the best results in terms of reducing workload for physicians. Finally, the best model introduced in the active learning loop is the Random Forest.

1.2.2. Evaluation of a biomedical language model in production

Given that natural language processing has evolved rapidly in recent years, and transformer models are the current state-of-the-art language models in several NLP tasks. We compared state-of-the-art language models based on transformers architecture for biomedical text classification in the context of evidence-based medicine. The objective is to distinguish robust types of studies from other studies to focus efforts on only one kind of evidence and reduce the daily workload on physicians.

In addition, another critical factor to consider was the COVID-19 pandemic, which, given its impact on society in 2020, generated an exponential increase in evidence related to this disease. Given this context, we had to develop a model that could ease physicians' daily workload and generalize to new diseases. To overcome this problem, we propose a language model to alleviate this problem and take it to production in an evidence-based medicine system to evaluate its performance with real users.

In an article under review (Expert Systems with Applications Journal), we obtained that the XLNET model is the best model to represent medical documents among other state-of-the-art models. Furthermore, when taking XLNET production in the Epistemonikos evidence-based medicine system with real users, we obtained that by selecting

to label documents where the model was unsure of its prediction, we were able to reduce more than 65% of the daily workload by physicians.

1.2.3. User study on Explainable Artificial Intelligence

This last section addresses the research questions related to the interpretability of automatic classification models; we validated them through a user study on the Epistemonikos evidence-based medicine platform.

We want to validate two significant aspects, (1) to investigate whether using the model's attention as an explanation is perceived as helpful for users. (2) to study if there is an interaction effect between the visual encoding and the type of article being reviewed on the perceived helpfulness of the model's attention as explanations. Moreover, we also studied if explanations reduce cognitive overload on the document screening task and if the model's predicted probability is relevant information for making a classification decision.

We obtained that:

(i) Using model's attention as an explanation is not perceived as helpful by users for document classification task.

(ii) Although attention as an explanation is not perceived as helpful there is an interaction effect between the visual encoding and the type of article being reviewed in the perception of usefulness of explanations as highlighted words.

(iii) When users give a high score on perceived helpfulness of highlighted words as an explanation, the model placed more attention on article specific words such as ”meta-review” or specific treatments.

(iv) When comparing the performance of users when using visualization of highlighted words in the article there is an increase in performance.

(v) The model's predicted probability is perceived as helpful for users in the task of document screening.

1.3. Related work

Before viewing each contribution in detail, we must contextualize on what has already been proposed to solve this problem, for each of the sub-tasks we are trying to solve: *document screening*, *language models for document representation* and *explainable artificial intelligence* focused on text-based applications.

1.3.1. Document screening in the medical domain

The task of finding relevant documents related to a medical question through citation screening has been studied and it is known as the *total recall problem*: given a medical topic or question, find all the documents that are relevant about a particular topic. Recently, the CLEF eHealth task 2 Kanoulas et al. (2017, 2018, 2019) is a challenge that calls for solving the problem of prioritizing which documents to screen to reduce work overload for experts. They provide a public dataset with medical topics and a set of candidate documents; participants have to rank documents by relevance for every specific medical subject in the minimum of iterations to make more efficient the document screening process (Grossman et al., 2016).

In the literature, the approaches for solving this problem are based on three general lines: **information retrieval**, **machine learning methods**, and **natural language processing**. The latter is used to support the first two.

In the **information retrieval** area, there have been many attempts to solve the problem using techniques such as relevance feedback (Donoso-Guzmán & Parra, 2018), query

expansion (G. E. Lee & Sun, 2018), ranking and inference based on external knowledge (Goodwin & Harabagiu, 2018).

From the **machine learning** community, the approaches usually focus on semi-automating the screening process of medical articles, which is still conducted or validated by physicians.

There have been efforts to solve this problem by using automatic classification (Bekhuis et al., 2014; Choi et al., 2012; Adeva et al., 2014; Mo et al., 2015; Wallace et al., 2012). In these previous works, authors compared classifiers such as Naive Bayes, K-NN, and SVM, using different ways to represent text, such as word embeddings and bag-of-clinical terms from titles and abstracts. There is also literature indicating the use of active learning (Hashimoto et al., 2016; Figueroa et al., 2012; Wallace et al., 2010; Miwa et al., 2014) for medical topic detection and clinical text classification. Moreover, a few deep learning models have been proposed for the classification of relevant evidence and categorization of documents in medical questions (Del Fiol et al., 2018; Hughes et al., 2017). The majority of work done has used datasets of up to 50 medical topics/questions and 200,000 documents. The Epistemonikos dataset includes 948 medical questions and $370,000$ potential documents, allowing models to generalize and to improve their performance compared to the state of the art.

Moreover, for both machine learning and information retrieval approaches, there is an increasing use of more powerful Natural Language Processing techniques mainly derived from deep learning models (Peters et al., 2018; Devlin et al., 2018; Howard & Ruder, 2018).

1.3.2. Biomedical text classification

The *Biomedical text classification* task's primary assignment is to classify a full article or its segments into one of several predefined categories, based on the manuscript's content. In P. Lewis et al. (2020), several language models pre-trained on the medical domain are compared in two biomedical tasks: sequence labeling and classification. Results show that language models based on the Transformer architecture and pre-trained on biomedical data, outperform other traditional language models. The classification tasks showed in this work included identification of cancer concepts, chemical-protein interactions, gene-disease interactions, drug interactions, and clinical events within a medical document.

The approach presented by Yao et al. (2019) combined rule-based features and knowledge-guided deep learning for the task of disease classification by training a convolutional neural network with word embeddings, including additional information from unified medical language system (UMLS) for learning the embeddings. The proposed method outperformed state-of-the-art participants from the i2b2-2008 obesity challenge[6] that consists in identifying obesity information and co-morbidities in a document.

The work described in Y. Wang et al. (2019) proposed using weak-supervised learning and an embeddings representation of documents to reduce the human effort of labeling large amounts of data. They offered a rule-based NLP algorithm to generate labels combined with BioW2Vec (Pyysalo et al., 2013) pre-trained word embeddings. They compared this approach with other machine learning models, such as Support Vector Machines, Multilayer Perceptron, Random Forest, and Convolutional Neural Networks. The task they tried to solve was smoking status classification and proximal femur fracture classification. They showed that convolutional neural networks capture additional features

[6] https://www.i2b2.org/NLP/Obesity/

from weak supervision compared to other machine learning models and achieved better performance.

Concerning Deep Learning architectures, Gargiulo et al. (2019) used a Hierarchical Deep Learning architecture to identify MeSH terms in PubMed articles. Since most of the time, this problem can be interpreted as a multi-class and multi-label classification problem since MeSH terms are hierarchical. In the same spirit, Du et al. (2019) used a deep learning architecture for multi-label classification of medical texts. They evaluated their model in the Hallmarks of Cancer classification dataset and on the Chemical exposure assessments dataset, where the main task is to extract chemical entities. They combined the model predicted confidence scores and contextual information from the target document extracted from ElMo model representation. They concluded that their proposed method required less human effort for feature engineering as traditional machine learning models and is highly efficient for large datasets.

Recently, Mujtaba et al. (2019) presented a survey of clinical text classification and showed that in most of the cases, proposed methods use content and concept-based features as input for machine learning models, and that most of the datasets and tasks consisted in identifying medical concepts in clinical texts and classification of clinical reports. Moreover, Nadif & Role (2021) surveyed several approaches solving the task of biomedical classification and found that self-supervised learning, where labels do not have to be manually created by humans, though automatically derived from relations found in the input texts, allowed for the effective word embedding representation of biomedical articles.

Some approaches have used machine learning models to extract relevant evidence arguments from medical articles (Šuster et al., 2021; Nye et al., 2020; Schmidt et al., 2021;

Mayer et al., 2018). In the same line, some works seek to assign categories to entities inside the text related to PICO tags, namely Patient, Intervention, Comparison, and Outcome (Demner-Fushman & Lin, 2007; Kim et al., 2011).

To the best of our knowledge, there are no studies that, given the article's content, use a model to classify what type of evidence it is depending on its methodology or study design, which is the task we are trying to solve in this work. Another advantage of our approach is that we seek to find a model that adapts to clasiffy documents concerned on new diseases not seen during training, which have not been studied for the task of Biomedical text classification.

1.3.3. Evidence based medicine systems

Regarding relevant literature for evidence-based medicine systems, we review the most used *evidence-based medicine systems*, particularly during the COVID-19 pandemic. PROSPERO[7], is a database available since 2011 where physicians can share their systematic reviews. Another online platform that emerged during the pandemic is Australia COVID-19 Clinical Evidence Taskforce[8], based at Cochrane Australia, which counts on physicians specialized in tagging COVID-19 clinical trials to make the construction of systematic reviews easier for future researchers.

Similarly, the COVID-NMA initiative[9] is an online database that contains all registered trials and is updated in real-time, considering their quality and results. Following with other available tools for finding in an easier way evidence related to COVID 19 is COVID-19 Evidence Network to support Decision making[10] (COVID-END), a network of

[7]https://www.crd.york.ac.uk/prospero/
[8]https://covid19evidence.net.au/
[9]https://covid-nma.com/
[10]https://www.mcmasterforum.org/networks/covid-end

organizations that coordinate through an online platform for synthesizing and curation of COVID-19 related evidence.

Another relevant initiative during the pandemic is RECOVERY-trial[11], which large-enrollment clinical trial of possible treatments for people in the United Kingdom admitted to hospitals with suspected or confirmed COVID-19 infections. This initiative consists of more than 39,000 physicians identifying treatments that may be beneficial for people hospitalized with this disease.

In the same line of RECOVERY-trial, SOLIDARITY[12] is another platform to find clinical trials related to COVID-19 treatments. Furthermore, this initiative counts on more than 10,000 patients enrolled in more than 500 hospitals in over 30 countries, consolidating one of the most extensive randomized trials on COVID-19.

Regarding sources of extraction of evidence not only related to COVID-19, there are various alternatives. Some of them are the Cochrane Database of Systematic Reviews (CDSR)[13], JBI Database of Systematic Reviews and Implementation Reports[14], EPPI-Centre evidence library[15], and WHO Institutional Repository for Information Sharing (IRIS)[16].

For this work, we collaborate with Epistemonikos since this foundation owns the LOVE[17] platform, one of the largest tagged databases of medical evidence related to both COVID-19 (more than 410,000 articles) and other diseases. This evidence database also

[11] https://www.recoverytrial.net/

[12] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments

[13] https://www.cochranelibrary.com/

[14] https://journals.lww.com/jbisrir/pages/default.aspx

[15] http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=62

[16] https://iris.wpro.who.int/

[17] https://app.iloveevidence.com/topics

has the advantage of having physician-validated tags depending on the type of article and study design. Additionally, as suggested above, it extracts evidence from most of the sources mentioned and is updated weekly.

1.3.4. Transfer learning in the biomedical domain

There have been efforts to improve language models in the biomedical domain by using transfer learning strategies, where a pre-trained model is fine-tuned with a labeled dataset.

Giorgi & Bader (2018) proposed that finetuning a LSTM with noisy and automatically generated data for Biomedical Entity Recognition (BNER) task on different entity classes, allows the model to improve the performance, compared to training only with the original data.

Peng et al. (2019) compared BERT and ElMo models' transfer learning adaptation performance on biomedical text-related tasks, namely sentence similarity, named entity recognition, and document classification on 1,580 Pubmed abstracts annotated with ten currently known hallmarks of cancer. Results showed that transformer-based methods, such as BERT, benefits more than ElMo by using transfer learning on most of the tasks.

Sachan et al. (2018) used unlabeled data to improve the performance of Biomedical NER models. They trained a bidirectional language model (BLM) on unlabeled data and transferred their weights as parameter initialization of another BLM. They evaluated the model's performance on four Biomedical NER datasets, obtaining that the fine-tuned model outperformed models trained from scratch with the original training dataset. Along the same line, other authors proposed using transfer learning on the biomedical NER task,

finding that additional knowledge obtained from other sources of data results in an improvement of models' performance (Francis et al., 2019; Mehmood et al., 2020; Khan et al., 2019).

Nourani & Reshadat (2020) proposed using transfer learning for the task of association between genes and diseases. They combined a Convolutional Neural Network and an Attention-BiLSTM and trained their model for the association between genes and diseases. They then used the learned weights as initialization for training on the extraction of phenotype task, obtaining an improvement of the performance, compared to only training with the original dataset for phenotype extraction.

Although favorable results have been obtained in the medical field using transfer learning techniques, most of them applied these techniques on the biomedical named entity recognition, sentence similarity, and document classification tasks. To the best of our knowledge, transfer learning strategies have not been applied to solve classifying evidence according to the type of document or study methodology in more extensive datasets. Besides, different strategies for efficient sampling of few examples have not been compared to require the minimum effort for physicians for curating new documents.

1.3.5. Explainable AI (XAI) for text applications

Recently, NLP tasks such as text summarization, question answering, text classification, and sentiment analysis, among others, have improved their performance thanks to transformer-based architectures. Something distinctive about these architectures is that they learn through attention mechanisms, where their attention output may be helpful as an explanation.

Some works have open-source tools so that users can explore the attentions of their NLP models or tasks. Examples of this are the work done by Vig (2019), Alammar (2021),

and Geva et al. (2022). Both tools allow visualizing attention words from a text and how they change along with the different layers and attention heads.

Moreover, works have analyzed attentions obtained by a transformer-based model and tried to explain their behavior, as in the work done by Vig & Belinkov (2019). In the same spirit, Wu et al. (2020) analyzed attentions learned by transformer models for sentiment analysis, and Dalvi et al. (2021) studied the latent concepts learned by the BERT language model.

Other works seek to demonstrate theoretically if attentions are helpful as an explanation. The first approach proposed by Jain & Wallace (2019) involves extensive experiments across various NLP tasks that aim to assess how attention weights provide meaningful explanations for predictions, finding that they vastly do not. The work that refuted those mentioned earlier was done by Wiegreffe & Pinter (2019). They demonstrated through a rigorous experimental design, obtaining that model attentions were helpful as a means of explanation.

Although several works study whether attentions are helpful as explanations for NLP tasks, both through experimental studies or by providing open-source tools, in this thesis, we seek to demonstrate through a user study if the attentions are helpful for the document classification task in the domain of evidence-based medicine.

1.4. Outline

The remainder of this thesis is distributed as follows: Chapter 2 presents all the preliminaries and key concepts to understand this work. Chapter 3 then shows our proposed solution to address the problem of document screening. In Chapter 4 we show the evaluation of a language model in production for identifying relevant evidence based on the type

of study. Chapter 5 describes the user study that evaluates whether model's attentions as explanations are perceived as helpful for physicians to screen documents. Finally, Chapter 6 presents discussion, conclusions and future work.

1.5. Publications

Several results and discussions presented in this work have been previously published or under review in Scientific Journals. In this section, we list all the publications used in this thesis:

- Automatic document screening of medical literature using word and text embeddings in an active learning setting, Scientometrics. ISSN: 1588-2861. DOI: https://doi.org/10.1007/s11192-020-03648-6. Published.
- Evaluating Transfer and Active Learning of Neural Language Models for COVID-19 Biomedical Text Classification. Andrés Carvallo, Denis Parra, Hans Lobel. Expert Systems with Applications ISSN: 0957-4174. Under Review.
- Is attention perceived as explanation? A user study on the explainability of attention and the visual effectiveness principle for neural text classification. Andrés Carvallo, Denis Parra, Ivania Donoso, Hernán Valdivieso, Peter Brusilovsky, Katrien verbert. Intelligent User Interfaces 2023 Conference. Under Review.
- Comparing Word Embeddings for Document Screening based on Active Learning. Andrés Carvallo and Denis Parra. BIRNDL Workshop at the SIGIR 2019 Conference.
- Analyzing the design space for visualizing neural attention in text classification. D Parra, H Valdivieso, A Carvallo, G Rada, K Verbert, T Schreck. 2nd workshop on visualization for AI Explainability.

- Neural language models for text classification in evidence-based medicine. Andres Carvallo, Denis Parra, Gabriel Rada, Daniel Perez, Juan Ignacio Vasquez, Camilo Vergara. LATINXAI Workshop at Neurips 2020 Conference.

2. PRELIMINARIES

In this section, we give additional information on the necessary context and concepts needed to understand the following chapters. In the first part, we provide a general definition of evidence-based medicine and the problem this medicine area seeks to resolve. We then contextualize language models that are essential, since we must represent medical texts as vectors as an input to the computer. In the following sections, we talk about explainable artificial intelligence and how to view explanations given by automatic models to facilitate the understanding of the platform users. Finally, we refer to important concepts to carry out a user study, which seeks to evaluate whether the explanations make any difference in the document screening task.

2.1. Evidence-based medicine

Evidence-based medicine (EBM) is a medical practice that aims to find all the available evidence to support medical decisions. Nowadays, this evidence is obtained from research published in biomedical journals, usually accessible through online databases like PubMed (Lindsey & Olin, 2013) and EMBASE (Lefebvre et al., 2008), which provide free access to articles' abstracts and, in some cases, to full articles.

There are two problems related to the area of EBM that can be solved with the help of an automated model: *document screening* and *biomedical text classification*. For the first problem, document screening, the objective is to find all the relevant evidence related to a specific medical question. In the information retrieval area, it is called the total recall problem, where the purpose is to retrieve all possibly relevant documents given a query. For example, given the question: what are the adverse effects of the COVID-19 vaccine? In the case of the total recall problem, the objective is to retrieve all the relevant evidence related to that question in the first positions.

In the case of biomedical text classification, in EBM, one of the main objectives is to distinguish relevant evidence depending on the study methodology. Therefore it is approached as a classification problem depending on its content, where the potential document categories are:

- **Primary study RCT:** studies that use a methodology where subjects are randomly assigned to one of two groups: the experimental and control. The experimental group receives the intervention that is being tested, and the other group receives an alternative treatment, which in most cases is a placebo. Considered articles are those that report a randomized trial, also including trial registries and protocols.
- **Primary study non-RCT:** case studies that do not use an RCT methodology and show isolated results on particular cases without a robust study design methodology. Moreover, they are primary studies that do not fulfill a randomized trial. A primary study is an umbrella term that includes any study design, qualitative or quantitative, where new data is collected from individuals, populations, or any experimental subject. Other criteria of inclusion are (a) pre-clinical research in humans and (b) modeling studies.
- **Systematic review:** a type of article that uses an explicit methodology to summarize, identify and appraise all the evidence related to a specific medical issue. In most cases, they are composed of RCT study design due to their robustness as relevant evidence. A systematic review seeks to answer a research question, employs a comprehensive and reproducible search strategy, identifies all relevant studies, and can take years to be completed with the work of several collaborators. To consider a systematic review for inclusion it fulfills the following criteria: (a) provides a description of at least one eligibility criterion, (b) its main objective is to synthesize primary studies (other syntheses might be used as an

additional source for studies) and (c) reports an explicit method that includes searching in at least one electronic database.

- **Broad synthesis:** a type of article that summarizes relevant evidence related to a medical issue but is not as extensive as a systematic review. Furthermore, these types of articles synthesize systematic reviews and, sometimes, primary studies. Broad synthesis considered for inclusion (a) reports an explicit method that includes searching in at least one electronic database, and (b) its main objective is to synthesize systematic reviews.
- **Excluded:** documents that are excluded or not considered as relevant evidence, as they do not belong to any of the other categories.

Regarding a conventional EBM practice, the substantial relevant evidence to support medical decisions, in most cases, corresponds to systematic reviews and primary studies RCT (Egger et al., 2008). However, in some cases, broad synthesis articles are also considered robust if they have RCT within their references (Mays et al., 2005; Dwan et al., 2008).

Both problems in this thesis: document screening, and biomedical text classification, are essential to improve EBM development. We address them using two different approaches. For the document screening problem, we propose an active learning strategy (Settles, 2012) that consists of selecting a proportion of documents related to a medical question to be labeled, thus saving the work of manually reviewing each one of the documents. Then for the biomedical text classification problem, we propose a classifier based on a state-of-the-art language model to automatically categorize the types of documents and reduce the work for manual categorization of evidence.

2.2. Language models

A language model is an algorithm that calculates the probability of occurrence of a target word given a series of observed events. One way to obtain a semantic representation of words in the form of vectors is by calculating their probability and using a vector learned by the language model to represent a word or text.

The first generation of models for document representation were based on the vector space model (Salton et al., 1975) using TF-IDF vectors, but more recent approaches have represented words with models such as word and text embeddings. The methodology to obtain these embeddings has evolved, starting with Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and then full text embedding representations such as ELMO (Peters et al., 2018), ULM-fit (Howard & Ruder, 2018) and BERT (Devlin et al., 2018). The latter representation is state of the art in the field of language models, and it is based on the so called transformer architecture for neural networks (Vaswani et al., 2017) which includes attention mechanisms. BERT, for instance, predicts hidden words previously masked, and it also learns to predict sentences: if the second sentence in a pair of sentences is its subsequent in the original document or not. It can also be adapted to tasks such as text classification in the medical domain. For instance, Lee et al. (2019), re-trained BERT focusing on the biomedical domain with more than one million PubMed articles, thus generating a version of BERT called BioBERT.

In the same spirit, the exponential growth of indexed medical articles in databases such as PubMed has made it possible to train language models in large medical corpora. Therefore, language models trained with medical document have emerged, such as BioBERT (J. Lee et al., 2019), BlueBERT (Peng et al., 2019), BioELMO (Jin et al., 2019), and BiomedicalW2VEC(Pyysalo et al., 2013).

Furthermore, there have been improvements over BERT since one of the limitations of this language model is that it cannot process sentences with a length over 512 words due to their computational complexity. During recent years, there have been efforts to solve this problem, particularly with a model called XLNET (Z. Yang et al., 2019), that allows training with longer texts than 512 tokens and use an autoregressive training strategy that efficiently evaluates all potential word combinations from a sentence.

In this thesis, since only BERT and BioBERT models were available for the first paper, we consider those language models to represent titles and abstracts from biomedical documents as input for active learning based on machine learning models for the document screening task. Then in the following chapter related to biomedical text classification for COVID-19 evidence, we used XLNET since it can deal with large texts.

2.3. Active Learning

The concept of active learning can be defined as choosing unlabeled examples that best improve your model to be labeled by an expert in a domain. This technique is essential when you have many data without labels, and there is a need to have labeled examples to automate a task with a machine learning model implied. Also, when you need domain experts that are costly and scarce, and it is essential to optimize their time to the maximum. Active Learning approaches can be divided into stream-based selective sampling (Cohn et al., 1996) and pool-based (D. D. Lewis & Gale, 1994). Moreover, the main difference between stream-based selective sampling and pool-based sampling is that the first makes an independent judgment on whether each sample in the data stream needs to query the labels of unlabeled samples, while the latter chooses a sample based on the evaluation and ranking of the entire dataset. This thesis focuses on pool-based active learning since medical documents are indexed to web portals, namely PubMed, massively and simultaneously.

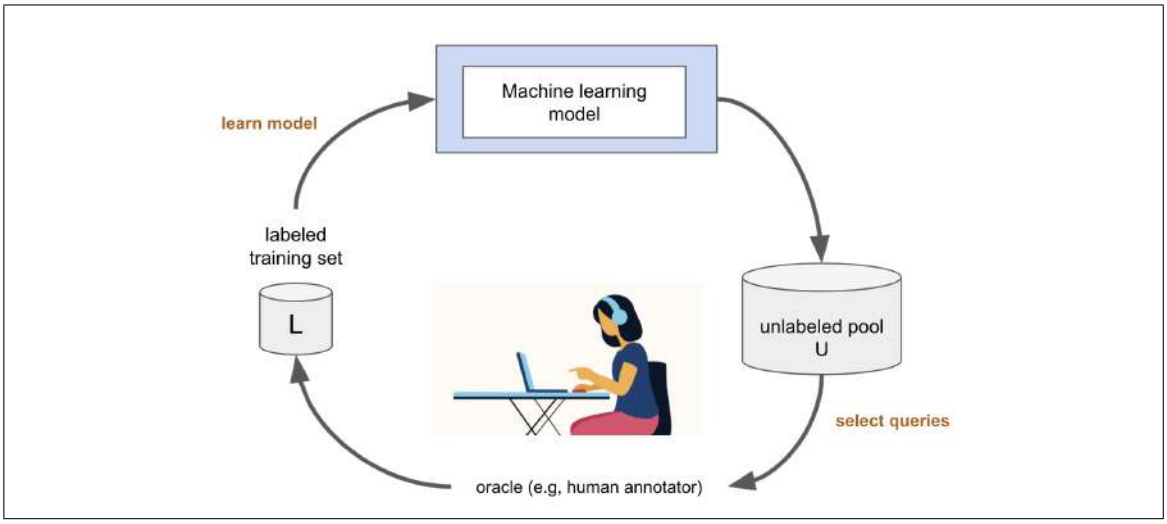


Figure 2.1. The pool-based active learning cycle.

Figure 2.1 shows the pool-based active learning cycle. A learner may begin with a small number of instances in the labeled training set L, request labels for one or more carefully selected instances, learn from the query results, and then leverage its new knowledge to choose which instances to query next. Once a query has been made, there are usually no additional assumptions on the part of the learning algorithm. The new labeled instance is added to the labeled set, and the learner proceeds from there in a standard supervised way.

The most used strategies for choosing unlabeled examples for being labeled by the oracle are: uncertainty sampling (D. D. Lewis & Gale, 1994), query by comitee (Seung et al., 1992), expected model change (Settles & Craven, 2008), variance reduction (Cohn et al., 1996) and error reduction (Roy & McCallum, 2001). This thesis uses uncertainty sampling since it has been demonstrated computationally efficient compared to other strategies (Schein & Ungar, 2007). In the field of evidence-based medicine, given that many medical documents are received without screening, it is necessary to efficiently choose those that

allow the model to be improved based on its potential label's level of certainty on their prediction.

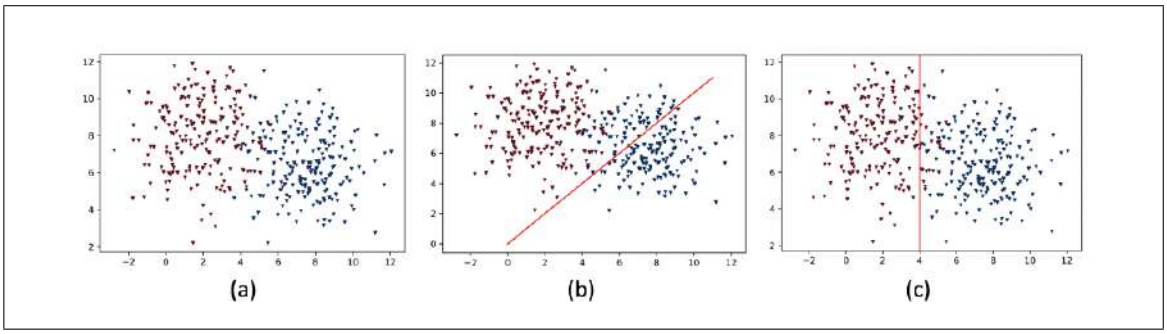


Figure 2.2. Example showing how uncertainty sampling pool-based active learning works with a dataset generated from 400 instances evenly sampled from two classes. (a) 2D graph showing the instances in the hyperplane, where the red points belong to one class and the green points to another. (b) Logistic regression trained with 30 labeled examples randomly sampled, where the line shows the decision boundary between the two classes (accuracy = 70%) (c) Logistic regression that changes its decision boundary by training using 30 chosen examples using uncertainty sampling (accuracy = 90%).

Figure 2.2 shows us with an illustrative example that we can save the number of documents that we can ask the oracle to label based on an uncertainty sampling strategy, reaching an accuracy of 90% with only 30 labeled examples out of a total of 400. In the example, we use a dataset generated from using isotropic gaussian blobs with a standard deviation of 1.8, each of them showing a different distribution for each class.

Figure 2.2(a) shows the 400 dataset points in a 2D hyper-plane where 200 belong to one class (red) and the other 200 to another (green). In Figure 2.2(b) We show a hypothetical case in which we do not have all the labels. In the first iteration, we randomly sample 30 dataset points to train a logistic regression and adjust the decision line. As a result, we obtain an accuracy of 70%. Finally, in Figure 2.2(c), we show how when using uncertainty sampling to choose 30 examples to train, by adjusting the decision line, the

model improves and obtains an accuracy of 90%. The results on this toy dataset show that using an active learning strategy over a random sampling allows obtaining a substantial improvement in the model's performance and with only a tiny proportion of the dataset to request labels from an oracle.

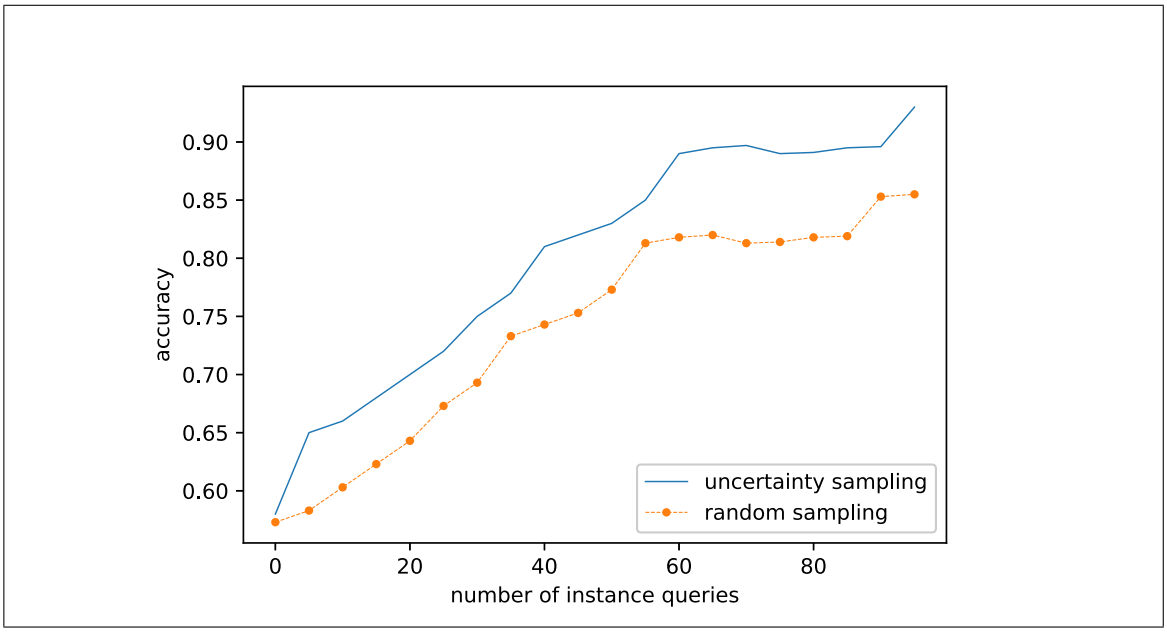


Figure 2.3. Learning curves for text classification: baseball vs. hockey. Curves plot classification accuracy as a function of the number of documents queried for two selection strategies: uncertainty sampling (active learning) and random sampling (passive learning). We can see that the active learning approach is superior here because its learning curve dominates that of random sampling.

In Figure 2.3 we show an example similar to the previous one where we compare random sampling with uncertainty sampling but in a real text classification dataset, where we train a model to classify whether the text corresponds to baseball or hockey. For this example, we use the dataset 20 Newsgroups corpus (Lang, 1995), which consists of 2,000 documents divided equally into the two classes. A common way to evaluate Active Learning strategies is to see how a performance metric (i.e., accuracy) evolves, as we ask the

oracle to tag us for more examples. Figure 2.3 shows the learning curve of two strategies, uncertainty sampling, and random sampling. The reported results correspond to a logistic regression averaged over ten folds using cross-validation. It can be seen that after labeling 30 examples, the accuracy of uncertainty sampling is 0.810, while the random sampling strategy reaches only 0.730. As we can see, the curve that an active learning strategy uses is always on the curve that uses random sampling. This example provides evidence that using an active learning strategy shows superiority over a baseline (in this case, random sampling) since it reaches a higher accuracy throughout the entire learning curve.

2.4. Explainable AI (XAI)

This section will develop the concept of explainable artificial intelligence (XAI). But first, we must clarify three key concepts: *interpretability*, *transparency*, and *explicability*. Regarding the first concept of *intepretability* according to general knowledge it is something capable of being understood, although in the area of artificial intelligence this concept is related to *transparency* and *explicability* (Lipton, 2018). *Transparency* is a property that allows the user of a system to understand how a system works to obtain a given result or prediction. *Explicability* is attributed to a system that can be explained, but unlike transparency, this explanation does not necessary reveal how the system works internally, it can be a post-hoc explanation of its behavior. In this thesis, we will focus on making an explainable model. Users visualize the importance of the model assigned to certain words to predict the type of study or its relevance to a medical question. We do not focus on transparency since we do not show the user the inner workings of the language model used for obtaining predictions.

Given this introduction of key concepts, we understand XAI as shown by Gunning & Aha (2019) as a system capable of explaining to users what it was built for, details on

their strengths and weakness and provide information on how may they work in the future. Beyond this concept, Gunning & Aha (2019) proposed six questions that a user may be able to answer given an explanation given by an AI system: a) *Why did you do that?*, b) *Why not something else?*, c) *When do you succeed?*, d) *When do you fail?*, e) *When can I trust you?* and f) *How do can I correct an error?*.

In the area of XAI there are two potential ways of giving an explanation: *global* and *local*. In the case of *global explanations*, they allow the user to examine the model from a broader perspective in order to understand the reasoning for obtaining their predictions and how the model works internally for every decision made. An example of global explanation can be found in the work done by Strobelt et al. (2017), that proposes a tool to understand the internal working of LSTM models during their training process.

On the other hand, *local explanations* allow the user to understand reasons for a particular prediction given by an AI model. These kinds of explanations are the ones that we are going to consider in the last chapter of the thesis where we evaluate if local explanations on texts for the task of document classification are useful for final users. Example of frameworks that make use of local explanations are LIME (Ribeiro et al., 2016), SHAP Lundberg & Lee (2017), ELI5[1] and Grad-CAMM++ Selvaraju et al. (2016).

Since we focus on local explanations for this thesis, there are three types: feature importance, nearest neighbors, and counterfactuals in this area. Concerning features importance, this type of explanation indicates to the user which input features are more critical for the model given a specific prediction; for example explanations obtained from attention mechanisms (Larochelle & Hinton, 2010), LIME, or SHAP. The nearest neighbors strategy gives an explanation as information on similar examples to the one that the user is

[1] `https://github.com/TeamHG-Memex/eli5`

reviewing. Finally, the counterfactual approach allows the user to modify specific parameters of the model to verify how predictions change and have an idea of how the model makes decisions; one example of this is the work done by Byrne (2019).

In this thesis, we focus on local explanations applying a feature importance explanation strategy since we receive a medical document as input. We output the importance given by the model to each word of the article to provide the physician with more information on which words are more important or less necessary for the model to classify given evidence as relevant.

2.5. Data visualization

The concept of visualization can be defined as a way of communicating a set of information through graphic representations (Ward et al., 2010). However, in this thesis, we will focus on Tamara Munzner's definition of visualization (Munzner, 2014) as a computer system capable of giving a visual representation to a dataset designed to support individuals to interpret the data.

Munzner also argues that design decisions do not change regardless of the domain in which data is being viewed. However, its interpretation can be framed within multiple areas of knowledge (i.e., psychology, design, or statistics). Given this, Munzner proposes an interpretable framework for various areas of expertise, which is the one that we will use in this thesis. This framework seeks to answer three fundamental questions that can be obtained from a data set: (1) what data will be visualized? (*what?*), (2) why does the user need to use the visualization? (*why?*) and (3) how will the visualization be designed? (*how?*). To better guide the design process, Munzner proposes a four-level nested framework: domain, data abstraction, visual encoding, and algorithm. It can be

seen in Figure 2.4 the dependency between these four levels and at what levels the three questions (*why?*, *what?* and *how?*) are answered.

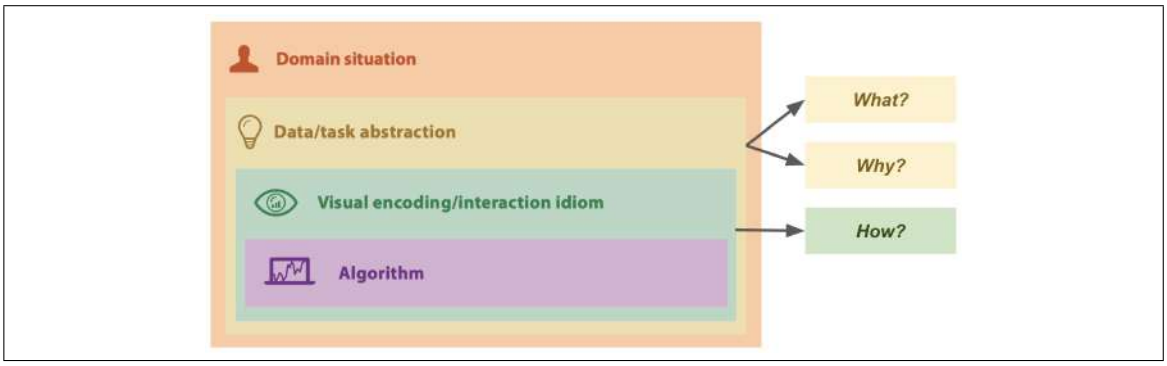


Figure 2.4. Nested visualization framework proposed by Tamara Munzner Munzner (2014) showing the dependency of the four abstraction levels and questions answered on each level.

Figure 2.4 shows the framework proposed by Munzner (2014) that will be used for our user study, each of the levels are described below:

- **Domain situation:** Defines the target user group, questions to answer with the visualization, and available data. The objective of this level is to understand the problem being solved by using visualization.
- **Data/task abstraction:** after defining the general problem in the previous level, this level seeks to answer two of the three fundamental questions: *what?* and *why?* In order to interpret these answers with the visualization, the data and the user's needs are mapped to a specific terminology to achieve this.
- **Visual encoding/interaction idiom:** this abstraction level seeks to answer the third fundamental question: *how?*. To answer this question, we need to define a way to create and manipulate the visualization to display the data as the user needs. There are two primary choices to make at this level: visual encoding and interaction idioms. On the one hand, visual encoding refers to mapping the data

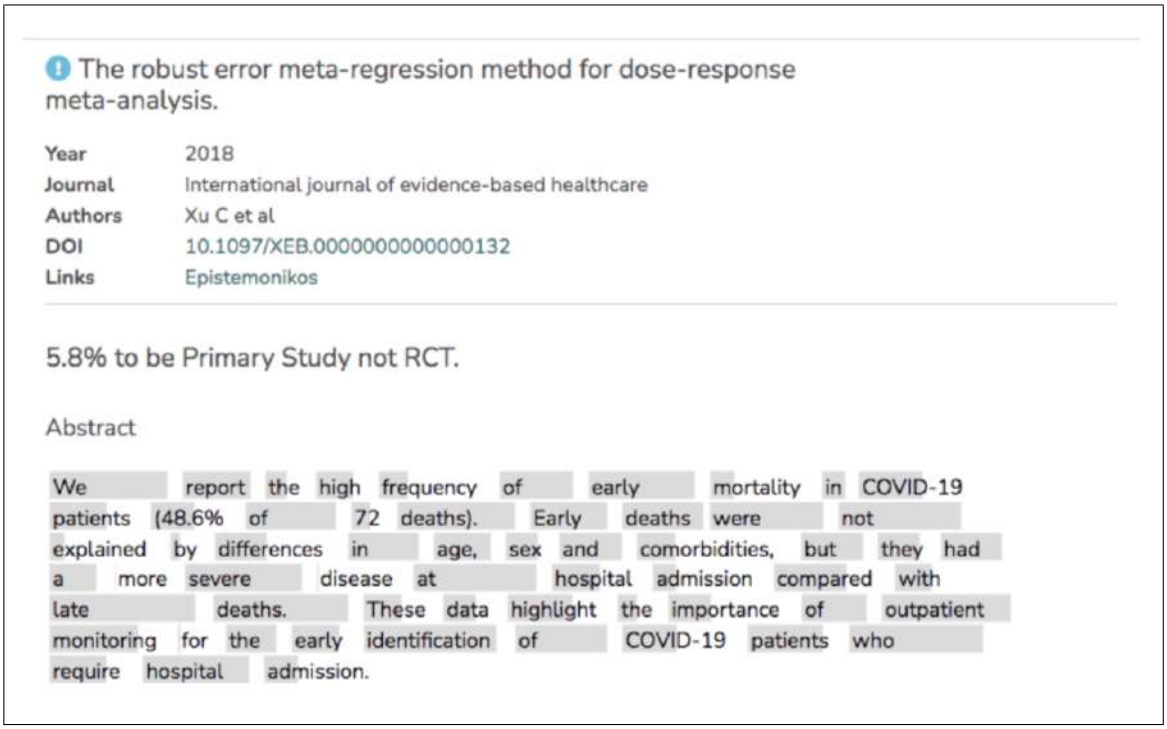


Figure 2.5. Screenshot of the proposed visualization, showing the document text and other metadata such as title, year, journal, authors, and predicted probability, marking the relevance of words depending on the bar length.

to some visual representation, in other words, how the data will appear in the visualization. On the other hand, interaction idioms determine how the user will control or modify the data through the visualization to understand it better.

- **Algorithm:** this last level ensures efficient data management, visual encoding, and interactions. If the algorithm is not efficient, this will result in poor user experience, which is seen as the core of the framework.

In this thesis, and specially on the chapter where physicians will interact with a visualization for the evidence classification task, we will use the Munzner framework. Specifically the answer to the three questions using our proposed visualization for the evidence-based medicine (see Figure 2.5) is:

- ***What?:*** the user sees the text of the document being reviewed along with other metadata such as title, authors, journal, and the probability given by the model for a given type of article.
- ***Why?:*** the user intends to use our visualization tool to efficiently classify medical documents as relevant or non-relevant depending on the type of study.
- ***How?:*** the visual encoding consists of marking relevant words on the text that the model considered as relevant for predicting a given label.

In the chapter designated to the user study, we will give more details on the visualization interface and the connection between our proposed solution with the Munzner framework.

2.6. User experience evaluation on a user interface

User experience (UX) is defined as a person's perceptions and responses that result from the use or anticipated use of a system (Hassenzahl, 2008; Law et al., 2009). Other interpretations supplement this formal definition: UX explores how a person feels about using a system, i.e., the experiential, affective, meaningful, and valuable aspects of its usage (Vermeeren et al., 2010). Therefore, UX is understood as dynamic, given the ever-changing internal and emotions of a person and differences in the circumstances during and after interactions (Law et al., 2009; Karapanos et al., 2009).

Therefore, UX should be seen as something evaluable after interacting with an object and before and during the interaction. While it is relevant to evaluate short-term experiences, given dynamic changes of user goals and needs related to contextual factors, it is also essential to know how experiences evolve and the reasons behind this behavior.

During recent years, there has been an interest both by the industry and the academic world in improving user experience (Vermeeren et al., 2010). However, the evaluation of user experience is not trivial since the alternatives of analysis can be broad (Vermeeren et al., 2010). There are multiple assessment dimensions: usability, utility, speed, availability, cognitive load, among others.

In this work, we will focus on evaluating accuracy (in terms of precision and recall), efficacy (time y clicks) y cognitive load (Vermeeren et al., 2010) measured using the NASA TLX test (Hart & Staveland, 1988), compared to the traditional interface that users are used to use. Another essential aspect that we must try to avoid in the evaluation is the bias of giving specific answers based on the context. To surpass this effect, also known as the learning effect, we will use the Latin Square methodology (Kirk, 2010).

The NASA TLX test is used to measure the cognitive load of a proposed solution in your system compared to using the traditional interface. In the case of this work, we will compare an interface that includes explanations that will be described in detail in the chapter where we describe this problem. The NASA TLX test is used to measure the cognitive load of a given task. This test consists of evaluating how much the assignment implies to the user in six dimensions: mental, physical, temporal, performance, effort, and frustration. Each dimension has a set of questions where the user has to give a score from 1 to 10. In most cases, the way to assess cognitive load in a set of study users is to have a control group using an unchanged system and a study group using a modified design. Finally, the control group's responses with the study subjects are compared to see if the changes affect the cognitive load.

Another important framework to be considered in a user study is the Latin Square strategy. This strategy explicitly evaluates the platform, avoiding biases that can confuse the results.

Subject	Consecutive tests (or study periods)			
	1	2	3	4
#1	A	B	C	D
#2	B	C	D	A
#3	C	D	A	B
#4	D	A	B	C

Figure 2.6. Example of the Latin Square design. Source: `https://paasp.net/within-subject-study-designs-latin-square/`

A possible unwanted effect in the study is that users ”learn” an interface (learning effect) by the order in which it is presented, which can bias the analysis of the results as an undesired effect. We will rotate the interfaces to show each user through a Latin Square design to minimize this effect. Each user is shown the interfaces in a different order for the various testing iterations, moving one space down at each step, thus avoiding the learning effect mentioned above. In the example shown in Figure 2.6, the user is presented with the interfaces in this order: interface A, then interface B, interface C and interface D. Then the following user will be shown B, C, D, and A first, and so on, however, the application of this methodology for our particular user study will be discussed in more detail in Chapter 5.

3. CHAPTER 3: AUTOMATIC DOCUMENT SCREENING OF MEDICAL EVIDENCE USING AN ACTIVE LEARNING SETTING

In this chapter we answer the first and second research questions, related to reducing the effort made by physicians at screening documents to find the evidence needed to support the answers of a medical question, and which is the best way to represent documents. Rather than building a classification model using a traditional machine learning, where a large dataset of labeled documents is used to train a model, we choose to experiment with an active learning approach Settles (2012). We use active learning due to its similarity with the actual task carried upon by physicians in EBM: label a few documents in several iterations, and get better at classifying more documents after each iteration. One of the main tasks of active learning is choosing the appropriate data points (documents) to be labeled by the experts in order the train the model with as few examples as possible. In order to evaluate our approach, we experiment with a large dataset of medical questions, unlike previous works that use smaller datasets G. E. Lee & Sun (2018).

We aim to answer the first two research questions:

- How can we improve the document screening task to answer clinical questions?
- What is the most reliable way to represent medical articles titles and abstracts for the document screening task?

In other words, we evaluate if a strategy based on state-of-the-art language models, such as BERT and BioBERT, in conjunction with an active learning approach, helps to improve the efficiency and efficacy of document screening in the medical domain?. Furthermore, we study if these approaches represent a considerable advantage compared with traditional word embedding language models (Word2Vec and GloVe) and TF-IDF representation.

3.1. Proposed Method

The process of finding documents that answer a clinical question requires first retrieving a set of candidate documents. Then, physicians perform the document screening where they select from the candidates abstracts and titles that are related to the medical question. This process may involve a large amount of time and cognitive effort from experts.

In this work, we propose the use of an active learning strategy to reduce the labeling effort from experts. Figure 3.1 illustrates the proposed approach.

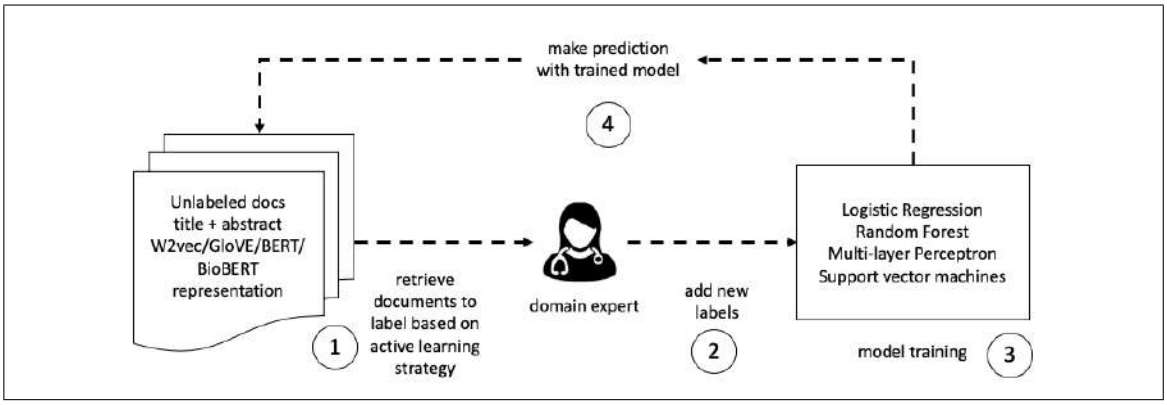


Figure 3.1. Illustration of the active learning approach. It starts with a set of candidate documents which based on an active learning strategy (uncertainty or random sampling) are retrieved to be labeled. Then the oracle (domain expert) adds new labels and the system uses the labels to train a machine learning model, and next it makes predictions with the latest model trained. Predictions are used to sample the new set of candidate documents.

3.1.1. Efficient labeling using active learning

Given a medical question q, a set of unlabeled candidate documents $C = \{c_i\}_{i=1}^{N}$, and a labeling oracle O, in our case a physician who knows if a document is relevant to q, the goal of the process is to train a classifier of relevant documents M^q, using as few labelings

from the oracle as possible. To achieve this, we iteratively select informative samples of documents to be labeled by the expert. Using these labelings, we progressively train the classifier, until we obtain a model with the desired performance, generating a sequence of classifiers $\{M_i\}_{i=1}^{k}$.

As the number of available oracle labelings is highly constrained, the critical aspect of the process is the selection of an appropriate sample to be presented to the oracle. To achieve this, we use an active learning approach Settles (2012), evaluating two different strategies for sampling, namely uncertainty sampling and random sampling. These strategies were selected based on their lower computational complexity compared to other methods such as error-based, gradient-based, and variable reduction Settles (2012). The first active learning strategy is uncertainty sampling, where one tries to select the sample that the classification model is most uncertain about. Then to estimate this uncertainty, the scheme selects the sample with the lowest classification confidence when assigned to its most likely label. Formally, given an initial model Θ, we select a new sample $\hat{x}$ based on the following equation:

$$\hat{x} = \text{argmax}_x \, 1 - P_\theta(\hat{y}|x),$$

where $\hat{y}$ is the class label with the highest posterior probability given the classification model θ. The second strategy considered for experiments is random sampling. In this scheme, active learning randomly chooses examples to be labeled and then trains the model θ with these new labels.

Based on the selected sampling strategy, we obtain a small set of unlabeled documents $X = \{x_i, \}_{i=1}^{n}$ from C, with $n << N$. Following this, we query the oracle O for a binary labeling $Y = \{y_i, \}_{i=1}^{n}$ of the n examples in X, where $y_i = 1$ identifies relevant documents. Finally, using X and Y, we train a classification model $M_i(X, Y)$, that is used to predict the labels for unobserved documents. We repeat this process to create updated versions

of the classification model M. In practice, for the initial model M_1, we start with five randomly sampled labeled documents for each medical question and train the first version of the classification model.

3.1.2. Document representation

In this work, we compare TF-IDF representation (bag-of-words document representation with TF-IDF weighting) with word embeddings such as Word2Vec Mikolov et al. (2013) and GloVe Pennington et al. (2014) , as well as with the state of the art text embedding BERT Devlin et al. (2018), and a fine-tuned BERT model called BioBERT J. Lee et al. (2019).

In order to train the word vectors, Word2Vec uses a feed-forward neural network for two possible tasks: given a sequence of words, predict the most probable next word (continuous bag of words) or given the word predict most probable context words (skip-gram). In this work, we use the Word2Vec skip-gram technique to obtain word embeddings, because it represents well even rare words (such as specific medical terms) compared to a continuous bag of words that presents higher accuracy for more frequent words Mikolov et al. (2013).

In the case of GloVe, word embeddings are obtained based on a probabilistic approach. In this neural language model, the objective is that the dot product of a vector of a target word with a matrix of vectors from words of their context is as close as possible to the original word co-occurrence matrix. After that, when the vectors are already optimized using ordinary least squares, these word embeddings are used as a way to represent words in a latent space.

Concerning text embeddings such as BERT or BioBERT, they use a transformer architecture Vaswani et al. (2017), an attention model that learns relations between words

and sentences. As the transformer has an encoding and decoding architecture, in this case, BERT uses only the encoder. This language model reads all the sequence at once through a *query, key, value* structure and a positional encoding, using an attention mechanism to solve two tasks. The first task is predicting a hidden word, and the second aims to capture the relation between sentences, in this case, titles and abstracts of medical documents.

Our main goal in this article is to evaluate differences among models based on word embeddings and text embeddings. When using word embedding models (Word2Vec and GloVe) to represent a document, as shown in Figure 3.2, we have to aggregate the obtained embeddings from each word of the title and abstract to represent the document as a vector, and eventually use it as input in a machine learning model.

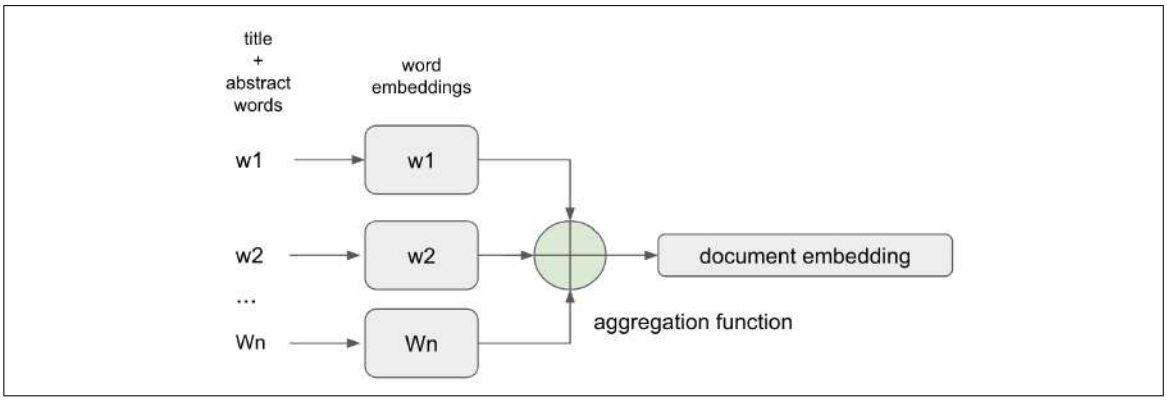


Figure 3.2. Using Word embedding model (Word2Vec and GloVe) to transform the title and abstract words of an article into a single document embedding.

On the other hand, text embedding models such as BERT or BioBERT, as shown in Figure 3.3, take as input the complete document (title and abstract tokens) and independent of its length, they output a fixed-sized embedding that represents the document. Concerning BioBERT, it is a fine-tuned version of BERT with more than one million full-text

documents from PubMed[1] and approximately 4.5 billion words. This model is adapted to the medical domain for tasks such as document classification.

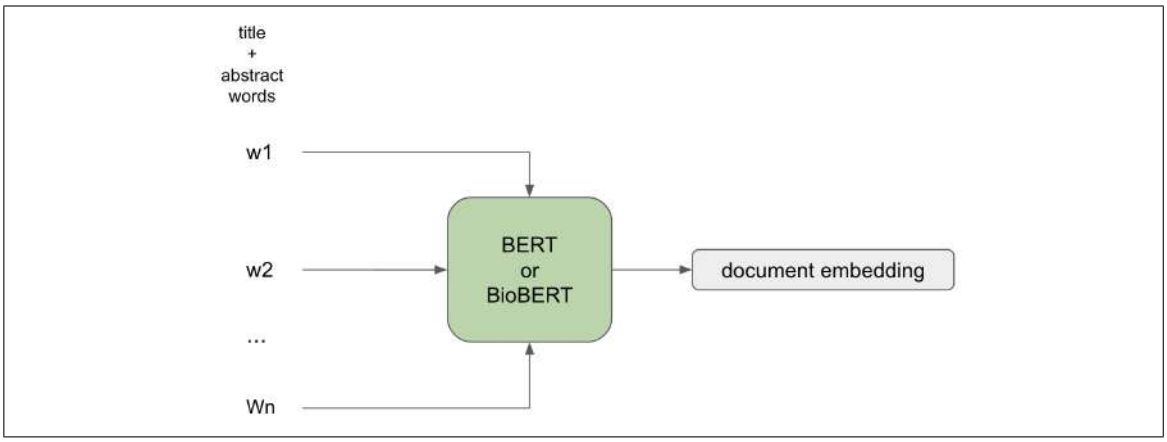


Figure 3.3. Using text embedding model (BERT and BioBERT) to transform the title and abstract of an article into a document embedding.

To generate the document representations that serve as input for the active learning procedure, we employ the concatenation of title and abstract. As shown by G. E. Lee & Sun (2018), the combined information from title and abstract is more informative than each one of them separately. Once concatenated, we lowercase the text and remove stopwords. The resulting text is then processed by the selected embedding technique. For Word2Vec and Glove, a 300-dimension embedding vector is generated for each word, and the final representation is generated by averaging these vectors, ending up with a document vector of 300 dimensions. In the case of BERT and BioBERT, the whole text is processed at once, generating a 768-dimension embedding vector as the final representation. Then for TF-IDF representation, we obtain a vector for each document and apply latent semantic indexing to the document-term matrix in order to reduce the dimensionality of each document to one hundred. If we do not perform this step, we might end up with document vectors of several thousands of dimensions (the size of the vocabulary),

[1]https://www.ncbi.nlm.nih.gov/pubmed/

what might increase the chance of facing the *curse of dimensionality* when building the machine learning classification models. We have chosen the above embedding dimensions because it has been shown in several experiments that GloVE Pennington et al. (2014) and Word2vec Mikolov et al. (2013) achieve their best performance with embeddings of size 300. Moreover, for BERT-base Devlin et al. (2018) and BioBERT J. Lee et al. (2019), which was the one used in this case, the ideal dimension is 768, because we used a pre-trained BERT language model, which size is of 768 per document embeddings. Since we are using the optimal size for each language model (rather than the same dimension to all of them), we are giving them an equal chance of performance based on that parameter.

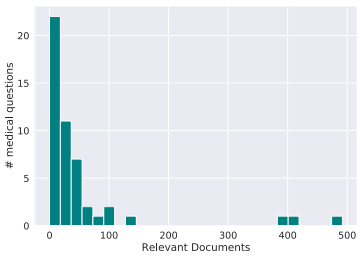
3.2. Dataset

To evaluate the proposed method, we use two datasets: CLEF eHealth[2] and Epistemonikos[3]. These datasets define a set of medical questions, where each is associated to a Systematic Review, which is a type of article that collects and synthesizes the relevant primary studies and trials related to a question. The information of each document in both datasets consists of the title, abstract, author, year and a label indicating if the document is relevant (or not) to the question or medical subject. For evaluation purposes, we split the documents related to each question (both relevant and not) into 70% for training and 30% for testing. We describe further characteristics of both datasets below.

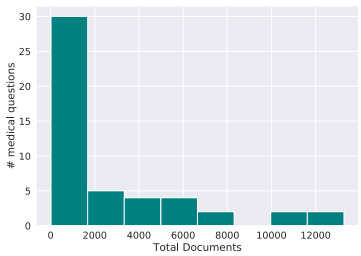
3.2.1. CLEF eHealth dataset

The CLEF eHealth dataset is conformed of 50 medical questions (ex. *which are the most effective treatments for the common cold?*) and 200,000 documents that were crawled

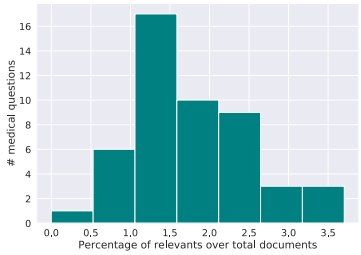
[2] https://sites.google.com/site/clefehealth2017
[3] https://www.epistemonikos.org/



(a) Distribution of relevant documents per question.



(b) Distribution of total documents per question.



(c) Distribution of the percentage of relevant documents per question.

Figure 3.4. CLEF eHealth dataset distribution of relevant and total documents per question.

from PubMed using each document id. Figure 3.4 presents the main characteristics of the distribution of the documents in the dataset:

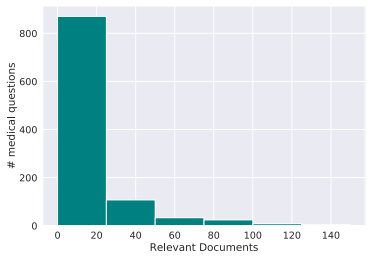
Figures 3.4(a) and 3.4(b) present the distribution of relevant documents and total documents per question in the CLEF eHealth dataset, respectively. On both, the y-axis represents the count of questions and the x-axis the number of documents. We can appreciate that most of the questions in CLEF eHealth have between 1 and 50 relevant documents, observing a long tail distribution. Regarding the total of documents, we can observe

something similar since most of the questions have between 1 and 1000 documents. For instance, Figure 3.4(a) indicates that about 25 medical questions have between 1 to 10 relevant documents. The long bar indicates that this is the most frequent case. Then Figure 3.4(b) indicates that about 30 medical questions have between 1 and 1,800 documents (including relevant and not-relevant ones) which experts have to screen in order to identify relevant documents. So, plots a) and b) differentiate because a) shows the distribution of only relevant documents per question and b) is the distribution of total documents including both relevant and not-relevant. Based on this, we argue that CLEF eHealth is a complex dataset because the proportion of relevant documents over the total number of documents is quite low. To assess this, Figure 3.4(c) presents the distribution of the proportion of relevant documents per question. It can be observed that most of the medical questions have a proportion between 1.5% and 2%, producing a highly unbalanced dataset.

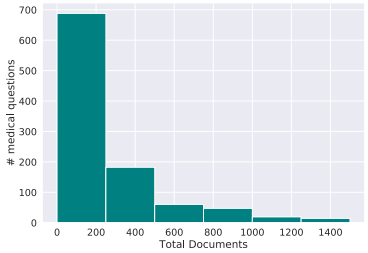
3.2.2. Epistemonikos dataset

The Epistemonikos Evidence Synthesis Project is a collaborative initiative established in 2012 to collect, organize, and to compare all relevant evidence for health decision-making, through a multilingual platform. The resulting Epistemonikos dataset is composed of 948 medical questions and 372,829 potential documents. The labels were previously curated by senior medical students, in which they had to select papers related to a set of medical questions. Figure 3.5 presents the main characteristics of the distribution of the documents in the dataset:

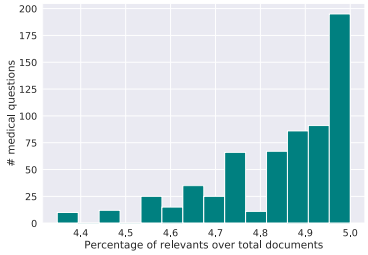
Figures 3.5(a) and 3.5(b) present the distribution of relevant documents and total documents per question in the Epistemonikos dataset, respectively. In Figure 3.5(a) and 3.5(b) we have the distribution of relevant documents and the total documents in the Epistemonikos dataset. On both, the y-axis represents the count of questions and the x-axis the number of documents. We can appreciate that most of the questions in Epistemonikos



(a) Distribution of relevant documents per question.



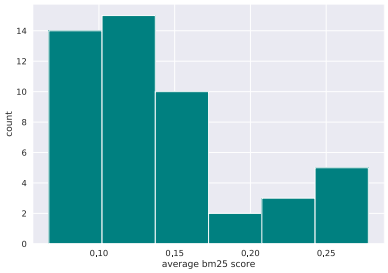
(b) Distribution of total documents per question.



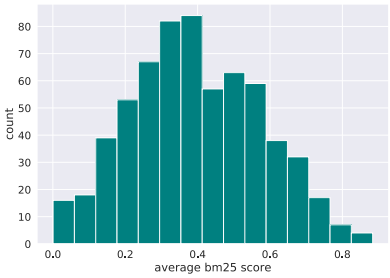
(c) Distribution of the percentage of relevant documents per question.

Figure 3.5. Relevant document distribution on the Epistemonikos dataset.

have between 1 and 20 relevant documents, observing a long tail distribution. Regarding the total of documents, we can observe something similar since most of the questions have between 1 and 200 documents. For clarification, we provide an example. Figure 3.5(a) indicates that about 820 medical questions have between 1 and 20 relevant documents, then Figure 3.5(b) indicates that about 690 medical questions have between 1 and 200 documents (including the relevant ones) where experts have to screen to identify relevant documents. So, plots a) and b) differentiate because a) shows the distribution of only relevant documents per question and b) is the distribution of total documents including



(a) Distribution of BM25 score on documents for CLEF eHealth queries



(b) Distribution of BM25 score on documents for Epistemonikos queries

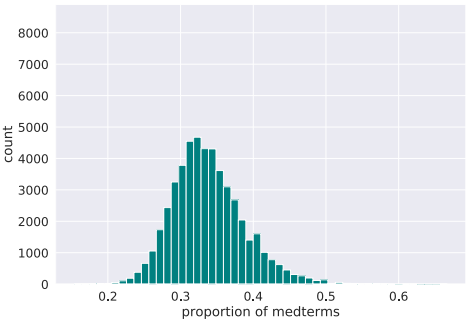
Figure 3.6. Epistemonikos and CLEF eHealth comparison of BM25 query similarity

both relevant and not-relevant. Then, the proportion between relevant and total documents in this dataset is, on average, a 4.61%, which makes it less complex compared to CLEF eHealth. From Figure 3.5(c) we can be observe that most of the medical questions have a proportion between 4.8% and 5%.

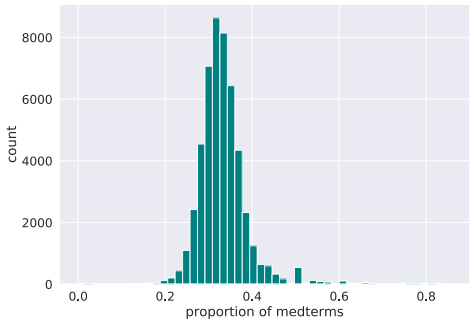
3.2.3. Epistemonikos and CLEF eHealth datasets complexity comparison

In this section, we compare the complexity of both datasets Epistemonikos and CLEF eHealth in terms of BM25 score similarity between medical questions and their respective medical documents (Figure 3.6). Also, we calculate the proportion of medical terms over total words on each title and abstract from each dataset documents (Figure 3.7).

It can be seen from Figure 3.6 that most CLEF eHealth documents have a BM25 score between 0.1 and 0.2 compared to that of Epistemonikos, which is between 0.35 and 0.40. That indicates that the level of specificity and complexity of the CLEF dataset is higher for this task given by a lower chance to discriminate relevant documents only by the co-occurrence of words from the query and documents.



(a) Distribution of medical terms proportion on CLEF eHealth documents



(b) Distribution of medical terms proportion on Epistemonikos documents

Figure 3.7. Epistemonikos and CLEF eHealth comparison of medical terms proportion on document titles and abstracts

If we observe from Figure 3.7, the density of medical terms per document in CLEF eHealth, we see that it is higher than in Epistemonikos. Thus showing that the CLEF eHealth dataset has a vocabulary more focused on the medical domain, making it more complicated in the document screening task for the model learned since there is a larger probability of words unobserved during training to be used in testing data. CLEF eHealth texts have more medical terms compared to the Epistemonikos dataset. For BERT or BioBERT, it is easier than for GloVe or Word2vec to create meaningful aggregated document representations for the task addressed in this article.

3.3. Experimental evaluation

In this section, we compared the performance of combinations of different active learning strategies and documents representations. Experiments were programmed in Python3 using libact Y.-Y. Yang et al. (2017), sci-kit-learn Pedregosa et al. (2011), pandas and

gensim libraries. For tf-idf representation we used sci-kit-learn Pedregosa et al. (2011) for feature extractor and reduced dimensionality with truncated SVD implementation.

In order to perform a large-scale evaluation, experiments are performed using a simulation of the active learning labeling process of documents for medical questions, using as the oracle the labels of the corresponding datasets.

Active learning setting: for each medical question, we hide the document labels and we leave only five random chosen documents with their respective labels to start building the model and then iterate with active learning to receive feedback from the oracle. For each prediction made by the machine learning model in each iteration, we sort the results depending on the predicted probability of being relevant for each model, so the evaluation metrics were calculated with the ranked list of potential candidates given by each strategy. We provide the oracle provide 10 documents for each iteration and complete the task using 10 iterations. Moreover we start with the same random sets of 5 documents for each ML algorithm using a 10-fold cross validation for testing each model. Another relevant setting is that, as we are runing offline experiments, we assume the oracle labels are always correct.

Relevance Feedback setting: we used two algorithms of relevance feedback Rocchio and BM25 as used by Donoso-Guzmán & Parra (2018) with the same meta parameters and setup but applied on this Epistemonikos dataset.

Classification models: we evaluate four different techniques for document relevance classification in our experiments: Multi-layer perceptron (MLP), Random Forest, Support Vector Machines (SVM), and Logistic Regression. These methods present a representative sample of machine learning techniques applied to text classification. The hyperparameters chosen for each method are, for Multilayer-Perceptron, three hidden layers of size 100, ReLU activation function, and Adam optimizer with a batch size of 32 (using a grid search

optimization to obtain the best combination of hyperparameters). Then for Random Forest 100 estimators and a Gini Index criterion. Concerning Support Vector Machines, we used a radial basis function kernel and linear kernels with a regularization constant set at 1.0 (using a grid search optimization to tune it for the best hyperparameters). And finally, for Logistic Regression, we used an ℓ_2 regularization and a maximum of 100 iterations until convergence.

Evaluation metrics: we used traditional information retrieval metrics such as *recall@k*, *precision@k* and mean average precision (*MAP*), similar to G. E. Lee & Sun (2018). Also, non traditional metrics are used, such as *LastRel%* and *work saved over sampling (WSS)*, that were used as two-task submission evaluation metrics for CLEF eHealth 17 Competition Goeuriot et al. (2017). *LastRel%* stands for last relevant percentage, which is the percentage of candidates documents that need to be screened and is essential because it indicates the number of documents needed to review to get all the relevant documents for that medical question. For example, if we have a list of 50,000 documents related to a medical question, where only 100 are relevant, the ideal would be that these 100 documents were in the top positions (last relevant in place 100) so that the expert did not have to review all 50,000 documents, indicating how efficient the proposed model is for solving this document screening task. Ideally, this metric should be as low as possible to avoid reviewing the entire list of articles until finding the last relevant document.

Justification of evaluation metrics: Recall@k indicates the ratio between between the retrieved relevant documents over the total relevant documents for a medical question. It is crucial because we do not want to miss any relevant document for a medical question. However, we need additional metrics because a naive optimization of recall and recall@k will make us find all the relevant documents (efficacy), but not in the most reasonable ranking (efficiency) to save physicians time. Then, precision@k calculates the proportion of relevant documents over k documents retrieved; it is still essential because we want

to retrieve the maximum quantity of relevant documents on each iteration of the active learning loop.

Mean Average Precision (MAP), computes precision at each recall positions (i.e., every position at which we find a relevant document) and averages over them. This metric penalizes a ranking that retrieves relevant documents in positions too far away from the top. Finally, LastRel% reflects the number of medical documents that need to be revised until finding the latest relevant document for a specific medical question; this indicates if effectively the model is saving work from physicians. The same for WSS that reflects the amount of labor saved for the task of labeling relevant documents.

Then *WSS* stands for work saved oversampling, which is a metric that shows how many candidate documents can be removed from manual screening. Evaluation metrics are related to the task of finding relevant documents for medical questions in the minimum possible iterations and on the first positions. Also, they allow evaluating if the proposed framework saves work to physicians for finding relevant documents without the need to review all available documents.

Including both metrics (LastRel% and WSS) facilitates the comparison between models to verify the amount of work saved from physicians thanks to the proposed framework, for the task of document screening related documents to medical questions.

3.4. Results

3.4.1. CLEF eHealth dataset results

For these experiments we evaluated the active learning framework combining document representation, active learning strategies and machine learning models for a small

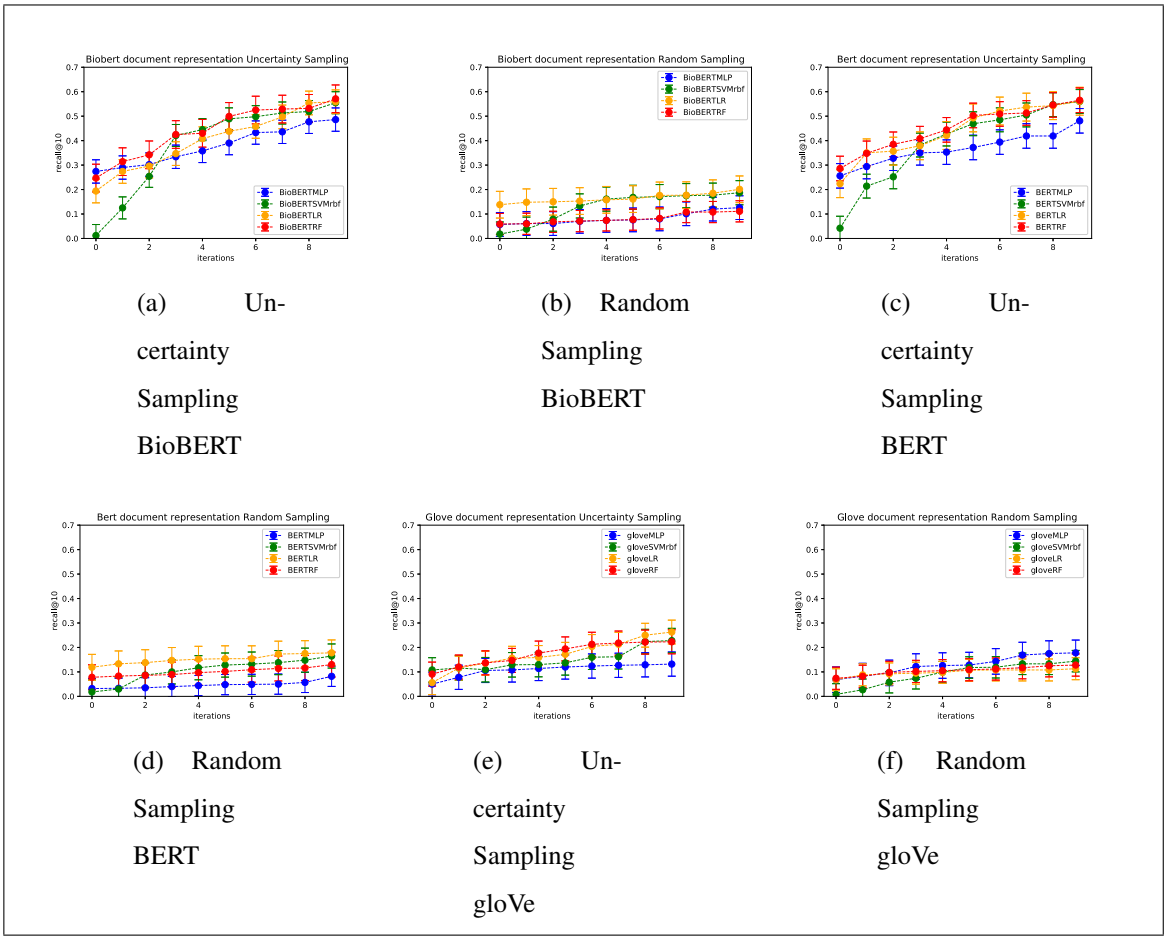
dataset (CLEF eHealth). The results shown are recall at three cut off levels (recall@10, recall@20, recall@30), precision at three cut off levels (10,20,30), mean average precision, Lastrel% and WSS after ten labeling iterations of ten documents each. In *Table 1*, we see the results on a small dataset (CLEF eHealth).

In Table 3.1, the first column indicates the dataset as well as the type of embeddings. The second column shows the active learning strategy (US vs. RS), as well as the learning model (MLP, RF, LR, SVM). Then the following nine columns show recall at three cut off levels (recall@10, recall@20, recall@30), precision at three cut off levels (precision@10, precision@20, precision@30), Mean average precision (MAP), Lastrel% and WSS. The larger these metrics, the better the model, except for Lastrel% (the smaller, the better). As shown in Table 3.1 for the CLEF eHealth dataset, the combination of random forest (RF) with an uncertainty sampling (US) strategy and BioBERT representation achieves the best performance in recall@k, and the best in precision@k. However, there are no significant differences with the results obtained using BERT with RF and BioBERT with SVM-linear. When comparing state-of-art representations (BERT, BioBERT) with word embeddings and TF-IDF representations, we noticed that although these representations do not report the best results, they are more consistent and robust to changes in the sampling strategy.

If we look at the latest relevant documents rather than the top-k, we see an interesting result. Concerning work saved over the document screening task, the RF model combined with a BioBERT representation, with an uncertainty sampling strategy, has the best performance, since the expert would have to review on average only a five percent (4.5%) of the list until finding the last relevant document. In contrast, with GloVe representation using random sampling, but with an SVM with RBF kernel as the learning method, the expert had to review an average of 96.1% of the full list.

3.4.1.1. CLEF eHealth model learning analysis

In this section, we present the results for the CLEF eHealth dataset of recall@10 after each iteration of documents for machine learning models (Multilayer-Perceptron, Random Forest, Support Vector Machines with linear kernel and Logistic Regression). We considered the best three document representations results obtained for the CLEF eHealth dataset (BioBERT, BERT, and GloVe). For each representation (*Figures 8-10*), we have a comparison of uncertainty based active learning with random sampling. On the x-axis, we have the number of iterations of ten documents that we ask the oracle to label, and on the y-axis is the metric of recall@10 on each iteration. For this particular task of document screening, on the iteration analysis, we focused on the recall@10, because our goal is not to leave out relevant documents for a medical question.



(a) Uncertainty Sampling BioBERT

(b) Random Sampling BioBERT

(c) Uncertainty Sampling BERT

(d) Random Sampling BERT

(e) Uncertainty Sampling gloVe

(f) Random Sampling gloVe

Figure 3.8. Comparison of Uncertainty and Random Sampling performance for CLEF eHealth dataset, iterations versus recall@10.

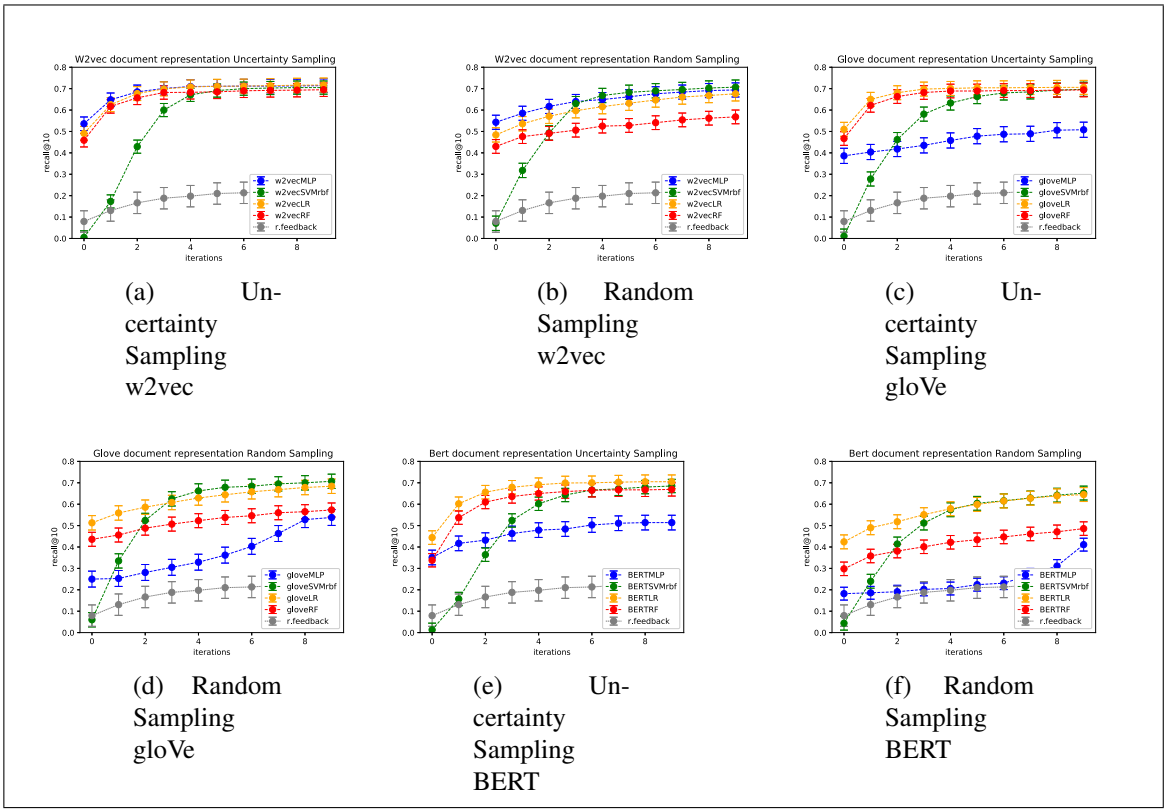
Figures 3.8 (a-f) show that BioBERT document representation for CLEF eHealth dataset gets higher levels of effectiveness at tenth iteration. Also, with BioBERT document representation, Logistic Regression and Random Forests gets better results in fewer iterations and are a clear winners over other models. Regardless of how we represent the documents and the machine learning model, the strategy of active learning based on uncertainty surpasses the baseline random sampling in all cases.

3.4.2. Epistemonikos dataset results

For these experiments, we evaluated our active learning framework on a large dataset of questions (Epistemonikos, 948 questions) combining document representations, active learning strategies, and machine learning models. Similar to CLEF eHealth, we used the same evaluation metrics to make them comparable. We also compared our results with traditional relevance feedback algorithms (using BM25 and Rocchio), using the same setting as Donoso-Guzmán & Parra (2018) but applied on this dataset.

Table 3.2 presents the results for the Epistemonikos dataset. The first column indicates the dataset as well as the type of embeddings. The second column shows the active learning strategy, as well as the learning model. Later, the following nine columns then show recall at three cut off levels (recall@10, recall@20, recall@30), precision at three cut off levels (precision@10, precision@20, precision@30), Mean average precision (MAP), Lastrel% and WSS.
As shown in Table 3.2 for the Epistemonikos dataset, it can be seen that the combination of an uncertainty sampling strategy with a logistic regression (LR) using a Word2vec representation of documents achieves the best results in terms of performance at recall@10. However, there is not a major improvement over GloVe representation using the same model and active learning strategy, and there are no significant differences compared to SVM and MLP. Concerning work saved oversampling (WSS), the LR model combined with a Word2vec representation has the best performance since the expert would have to review, on average, only 14.8% of the list until finding the last relevant document.

With this same Word2vec representation, we see an excellent performance of US-MLP in terms of recall@k and precision@k metrics, indicating that MLP performs well at ranking the top documents. Concerning a general comparison between active learning versus relevance feedback approaches (Rocchio and BM25, at the end of Table 3.2), regardless



(a) Uncertainty Sampling w2vec

(b) Random Sampling w2vec

(c) Uncertainty Sampling gloVe

(d) Random Sampling gloVe

(e) Uncertainty Sampling BERT

(f) Random Sampling BERT

Figure 3.9. Comparison of Uncertainty and Random Sampling performance for Epistemonikos dataset, iterations versus recall@10.

of the representation of documents, machine learning models, or active learning strategy, a definite improvement can be observed on all metrics.

3.4.2.1. Epistemonikos model learning analysis

In this section, we present the results for the Epistemonikos dataset of recall@10 after each iteration of the active learning process with ten documents per iteration. The machine learning models (Multilayer-Perceptron, Random Forest, Support Vector Machines with linear kernel and Logistic Regression) are compared on the best three document representations results obtained for this dataset (Word2Vec, GloVe, and BERT), all of them

compared to the baseline relevance feedback model (Rocchio). For each representation, we have a comparison of uncertainty based active learning with random sampling. On the x-axis, we have the number of iterations of ten documents that we ask the oracle to label, and on the y-axis is the metric of recall@10 on each iteration.

Figures 3.9 (a-f) show that for the Epistemonikos dataset, all methods converge more quickly with uncertainty sampling than with random sampling, which saves a considerable deal of effort to physicians for labeling. Moreover, Word2Vec embedding representation seems to speed up convergence compared to GloVe and BERT on several learning methods (notice the effect on SVM). However, there are no significant differences in Word2Vec with GloVe or BERT after ten iterations (also shown in Table 2). In all cases, the Logistic Regression (LR) and Random Forest (RF) models reports higher levels of recall than the other methods from iteration one and converges after the 3rd or 4th iteration, which is in deep contrast to SVM or MLP which converge only at the 7th or 4th iteration. This result provides essential evidence of the effort that could be saved to physicians as oracles, with only 40 documents labeled rather than 60 or 70 to achieve a similar level of classifier performance by using uncertainty sampling with logistic regression for the sampling strategy and learning algorithm, respectively. Finally, using an uncertainty-based sampling strategy, independent of model or representation of the document that we use, we outperform the relevance feedback baseline very quickly compared to random sampling.

3.5. Discussion

In this chapter, we supported results from previous studies in terms of showing that active learning with an uncertainty sampling (US) strategy yields good results for the task of biomedical document screening. Our main contribution was comparing the performance of different schemes to represent documents in an active learning setting, namely

TF-IDF, Word2vec, GloVe, BERT, and BioBERT. In our experiments with two datasets (CLEF eHealth and Epistemonikos), we found that an active learning strategy based on uncertainty sampling with either a BERT or BioBERT document representation, yields the best results. However, the conclusions are not completely clear in terms of the learning algorithm. In the Epistemonikos dataset, the US strategy combined with a logistic regression achieves better results in fewer iterations for retrieving documents to be labeled by an expert. Still, there are no significant differences with SVM or random forests, but LR is considerably faster for training models iteratively. In the CLEF eHealth 2017 dataset, we found that US with BioBERT document representation reaches the best performance with a random forest, leaving the logistic regression in third place after SVM. After additional analysis we found stronger similarities between the documents in train and test splits of the Epistemonikos datasets and larger differences between train and test document similiarities in the CLEF eHealth dataset, an element that might explain differences in performance of the top learning methods LR, and RF.

We have evidence to answer the first two research questions given the results obtained. In the case of the former, regarding an efficient way to alleviate the task of document screening, we show that using an active learning approach performs well and requires a low percentage of documents to be tagged. Moreover, regarding the second question, we obtained that BioBERT, a transformer-based model, allows us to get a better representation of medical documents compared to traditional word embeddings. However, for the next chapter, we will explore a new language model that came out after BERT, which is XLNET, which improves the representation since it allows to obtain a representation of the complete text without being limited to 512 tokens such as BERT.

In the next chapter, we will evaluate the performance of a language model on a real evidence-based medicine system. The main objective is to assess if sampling only documents where the model is more uncertain on their predictions reduces the daily workload on physicians in the task of document screening.

Table 3.1. Average results of recall@k (r@k), precision@k (pr@k), Mean Average Precision (MAP), Lastrel% and WSS performance measured in CLEF eHealth dataset using active learning strategies (US: uncertainty sampling, RS: random sampling) using a batch of 10 documents per feedback iteration for TF-IDF, Word2vec, GloVe, BERT-base and BioBERT-base representation. Results in **bold** font are the best for each metric, while the second and third best are underlined. Statistical significance is calculated with multiple t-tests using Bonferroni correction. The symbol $*$ indicates the statistically significant best result. No significant difference is shown with †.

Dataset	AL-Model	r@10	r@20	r@30	pr@10	pr@20	pr@30	MAP	LastRel%	WSS
CLEF eHealth	US-MLP	.081	.120	.163	.173	.151	.133	.128	85.9	.141
50 SRs	US-RF	.334	.392	.418	.367	.408	.320	.404	76.9	.231
TF-IDF	US-LR	.255	.320	.355	.414	.308	.241	.278	75.7	.243
	US-SVM (rbf)	.292	.335	.364	.471	.313	.241	.327	74.7	.206
	US-SVM (linear)	.268	.310	.331	.453	.313	.246	.315	77.6	.224
	RS-MLP	.034	.067	.097	.072	.076	.076	0.07	85.4	.145
	RS-RF	.167	.227	.295	.293	.221	.196	.211	76.1	.238
	RS-LR	.126	.189	.242	.238	.201	.187	.175	75.0	.249
	RS-SVM (rbf)	.116	.180	.224	.255	.207	.184	.178	78.3	.216
	RS-SVM (linear)	.144	.212	.280	.255	.221	.197	.205	73.9	.260
CLEF eHealth	US-MLP	.132	.200	.221	.233	.173	.133	.176	76.0	.240
50 SRs	US-RF	.223	.281	.313	.341	.219	.165	.266	83.0	.170
GloVe	US-LR	.263	.311	.330	.396	.256	.188	.290	64.8	.352
300 dim	US-SVM (rbf)	.228	.265	277	.341	.213	.148	.247	74.0	.260
	US-SVM (linear)	.239	.274	.289	.343	.221	.160	.260	76.0	.240
	RS-MLP	.122	.139	.198	.225	.155	.122	.154	78.8	.212
	RS-RF	.128	.183	.218	.203	.150	.122	.156	84.0	.160
	RS-LR	.113	.186	.247	.185	.156	.143	.161	74.9	.251
	RS-SVM (rbf)	.144	.227	.279	.226	.174	.142	.181	96.1	.039
	RS-SVM (linear)	.126	.181	.240	.187	.145	.119	.146	68.1	.318
CLEF eHealth	US-MLP	.215	.264	.278	.322	.220	.167	.252	66.8	.332
50 SRs	US-RF	.265	.308	.330	.382	.245	.184	.297	74.9	.251
Word2vec	US-LR	.228	.266	.278	.345	.215	.160	.252	68.0	.320
300 dim	US-SVM (rbf)	.237	.286	.308	.412	.278	.205	.293	71.6	.284
	US-SVM (linear)	.235	.272	.279	.394	.249	.177	.259	73.2	.268
	RS-MLP	.118	.173	.232	.179	.156	.134	.166	77.1	.229
	RS-RF	.144	.187	.256	.197	.137	.119	.170	83.0	.170
	RS-LR	.102	.167	.252	.137	.121	.121	.118	74.0	.260
	RS-SVM (rbf)	.121	.180	.238	.191	.156	.133	.160	85.3	.147
	RS-SVM (linear)	.164	.226	.249	.173	.145	.118	.162	69.0	.329
CLEF eHealth	US-MLP	.481	.663	.762	.802	.688	.597	.816	12.9	.871
50 SRs	US-RF	.565†	.727†	.804†	.833†	.695†	.597†	.893†	6.2	.938
BERT-base	US-LR	.561†	.721†	.800†	.837†	.693†	.591†	.852†	9.8	.902
768 dim	US-SVM (rbf)	.560†	.705	.783	.835†	.678	.579	.826	22.1	.779
	US-SVM (linear)	.570†	.736†	.813†	.841†	.706†	.601†	.876	13.4	.866
	RS-MLP	.082	.125	.174	.106	.099	.089	.108	80.6	.194
	RS-RF	.130	.165	.189	.141	.107	.080	.130	83.9	.161
	RS-LR	.178	.271	.320	.272	.219	.181	.214	73.1	.269
	RS-SVM (rbf)	.165	.232	.288	.194	.158	.137	.173	89.7	.103
	RS-SVM (linear)	.147	.212	.248	.183	.143	.128	.168	67.6	.323
CLEF eHealth	US-MLP	.486	.667	.758	.806	.697	.604	.840	12.0	.880
50 SRs	US-RF	**.571***	**.738***	**.819***	**.853***	**.715***	**.614***	**.910***	**4.5***	**.955***
BioBERT-base	US-LR	.559	.723	.805	.831	.696	.595	.855	9.5	.905
768 dim	US-SVM (rbf)	.555	.702	.781	.824	.677	.577	.822	18.9	.811
	US-SVM (linear)	.571†	.736†	.815†	.841†	.706	.603	.881†	12.2	.878
	RS-MLP	.126	.174	.225	.139	.113	.105	.140	81.1	.189
	RS-RF	.111	.142	.177	.191	.133	.114	.142	86.7	.133
	RS-LR	.201	.254	.290	.219	.165	.138	.216	70.5	.295
	RS-SVM (rbf)	.187	.248	.280	.232	.180	.146	.205	86.4	.136
	RS-SVM (linear)	.176	.243	.273	.216	.174	.140	.203	67.8	.321

Table 3.2. Average results of recall@k (r@k), precision@k (pr@k), Mean Average Precision (MAP), Lastrel% and WSS performance measured in Epistemonikos dataset using active learning strategies (US: uncertainty sampling, RS: random sampling), with batch size of 10 documents per feedback iteration for TF-IDF, Word2vec, GloVe, BERT-base and BioBERT-base representation. Results in **bold** font are the best for each metric, while the second and third best are underlined. Statistical significance by multiple t-tests using Bonferroni correction. The symbol $*$ indicates the statistically significant best result. Results with no significant difference with the best one are indicated with †.

Dataset	AL-Model	r@10	r@20	r@30	pr@10	pr@20	pr@30	MAP	LastRel%	WSS
Epistemonikos	US-MLP	.242	.347	.434	.294	.215	.173	.255	75.5	.245
948 SRs	US-RF	.516	.587	.630	.534	.347	.262	.518	68.4	.290
TF-IDF	US-LR	.507	.591	.633	.517	.333	.247	.477	66.7	.333
	US-SVM (rbf)	.441	.513	.552	.442	.281	.207	.416	68.7	.313
	US-SVM (linear)	.483	.556	.600	.491	.317	.235	.460	67.7	.323
	RS-MLP	.143	.227	.313	.110	.091	.083	.130	76.5	.234
	RS-RF	.380	.468	.527	.366	.246	.189	.345	70.5	.294
	RS-LR	.428	.531	.589	.399	.279	.218	.392	63.2	.367
	RS-SVM (rbf)	.428	.515	.569	.392	.265	.205	.391	64.1	.358
	RS-SVM (linear)	.433	.525	.582	.406	.278	.213	.402	64.2	.357
Epistemonikos	US-MLP	.508	.666	.744	.555	.421	.337	.591	32.2	.678
948 SRs	US-RF	.694	.832	.884	.696	.497	.385†	.765	23.5	.765
GloVe	US-LR	.706†	.844†	.898†	.689	.494	.385†	.768	15.3	.847
300 dim	US-SVM (rbf)	.697	.828	.877	.670	.470	.361	.744	17.9	.821
	US-SVM (linear)	.704†	.841†	.896	.693	.495	.384†	.772	16.1	.839
	RS-MLP	.538	.673	.737	.439	.319	.253	.492	60.2	.398
	RS-RF	.573	.694	.754	.483	.338	.265	.522	49.1	.509
	RS-LR	.684	.814	.866	.589	.419	.329	.668	33.3	.667
	RS-SVM (rbf)	.707	.830	.877	.616	.436	.340	.705	72.0	.280
	RS-SVM (linear)	.708	.832	.877	.619	.437	.338	.708	31.0	.690
Epistemonikos	US-MLP	.714†	**.854***	**.903***	**.707**	**.504***	**.392***	**.787***	16.1	.839
948 SRs	US-RF	.695	.832	.888	.703†	.501	.388†	.765	23.5	.765
Word2vec	US-LR	**.717***	.851†	.900†	.697	.492	.381	.768	**14.8***	**.852***
300 dim	US-SVM (rbf)	.705†	.844†	.898†	.698	.496	.385†	.769	16.3	.837
	US-SVM (linear)	.706†	.835	.889	.688	.489	.379	.763	18.1	.819
	RS-MLP	.694	.821	.872	.605	.431	.338	.692	33.0	.670
	RS-RF	.568	.699	.764	.487	.348	.275	.525	48.6	.514
	RS-LR	.676	.807	.864	.579	.416	.329	.658	35.0	.650
	RS-SVM (rbf)	.705	.832	.878	.619	.439	.342	.709	70.7	.293
	RS-SVM (linear)	.694	.817	.868	.604	.428	.334	.691	33.3	.667
Epistemonikos	US-MLP	.514	.685	.771	.577	.428	.335	.577	37.7	.623
948 SRs	US-RF	.669	.802	.856	.673	.473	.364	.718	32.5	.675
BERT-base	US-LR	.705†	.834	.883	.702†	.494	.381	.767	21.9	.781
768 dim	US-SVM (rbf)	.685	.814	.864	.680	.476	.361	.733	26.0	.74
	US-SVM (linear)	.692	.825	.876	.701†	.496	.380	.755	24.7	.753
	RS-MLP	.411	.542	.621	.326	.239	.194	.342	68.1	.319
	RS-RF	.486	.614	.684	.393	.282	.225	.410	62.4	.376
	RS-LR	.645	.767	.824	.566	.396	.310	.623	47.9	.521
	RS-SVM (rbf)	.652	.764	.821	.567	.393	.306	.626	73.1	.269
	RS-SVM (linear)	.647	.765	.815	.559	.389	.303	.628	29.9	.700
Epistemonikos	US-MLP	.450	.612	.695	.518	.389	.309	.513	42.5	.575
948 SRs	US-RF	.443	.587	.674	.422	.307	.246	.411	50.2	.498
BioBERT-base	US-LR	.673	.806	.868	.656	.463	.359	.712	23.2	.768
768 dim	US-SVM (rbf)	.664	.797	.853	.651	.456	.353	.695	26.5	.735
	US-SVM (linear)	.666	.794	.850	.641	.447	.343	.691	26.3	.737
	RS-MLP	.469	.603	.676	.393	.285	.228	.418	72.9	.271
	RS-RF	.557	.684	.750	.470	.335	.264	.503	57.4	.426
	RS-LR	.690	.812	.860	.604	.427	.333	.681	38.4	.616
	RS-SVM (rbf)	.681	.804	.852	.597	.422	.329	.674	76.8	.232
	RS-SVM (linear)	.683	.803	.848	.596	.418	.323	.671	22.6	.773
Epistemonikos	Rocchio	.261	.369	.432	.655	.419	.330	.631	26.31	.737
948 SRs	BM25	.131	.173	.209	.427	.295	.240	.254	67.24	.328

4. EVALUATING TRANSFER AND ACTIVE LEARNING OF NEURAL LANGUAGE MODELS FOR COVID-19 BIOMEDICAL TEXT CLASSIFICATION

In this chapter, we propose a different approach for the first and second research questions of this thesis. Moreover we seek to reduce the effort made by physicians at screening documents given the context of the COVID-19 pandemic and the advances in the natural language processing area. Mainly, we evaluate the models' capabilities to generalize to recent articles related to COVID-19.

The rapid spread of COVID-19 since late 2019 increased research related to this disease shown by more than 200,000 new articles indexed, with a peak of more than 23,000 new papers indexed per month[1]. Furthermore, to help researchers find relevant evidence and extract patterns in the content of the articles, research groups active in machine learning such as Google, Chan Zuckerberg Foundation, and the Allen institute collaborated to create the CORD-19 open source dataset(L. L. Wang et al., 2020), which consists of articles related to COVID-19.

Given this context, evidence-based medicine (EBM) discipline is now essential, since new evidence needs classification the best way possible, given the short time frame to decide how to approach this disease. One of the most critical tasks for the practice in EBM is classifying articles into types of studies, namely systematic review, randomized controlled trial, non-randomized controlled trial, broad synthesis, or excluded. This way, by having categorized evidence, the task of finding relevant evidence is more manageable for physician-researchers (Sackett, 1997).

Even though state-of-the-art methods for biomedical text classification has been succesfully used for identification of diseases (P. Lewis et al., 2020; Yao et al., 2019; Y. Wang et al., 2019), MeSh terms (Gargiulo et al., 2019), medical concepts (Du et al., 2019) and

[1]https://www.science.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat

PICO tags (Demner-Fushman & Lin, 2007; Kim et al., 2011), they have not been tested on new diseases not seen by the model during training, thus hindering generalization. Another problem is that most of these methods are fully automated without human-in-the-loop incremental feedback.

The lack of generalization is critical, especially with the appearance of new diseases, where there is no previous evidence, and it is necessary to react quickly to make decisions about public policies, treatments, and diagnoses, among others. Regarding the problem of lack of human-in-the-loop feedback, blindly trusting predictive models based only on offline performance metrics might not be optimal (W. Wang & Siau, 2018). Including expert physicians in the process also is essential, since they help with learning and validating predictions made by automated models.

To deal with the previously mentioned problems, we propose a neural language model-based classification scheme in the context of EBM to categorize COVID-19 related documents into one of the five potential types of articles introduced before. Specifically, we propose to finetune neural language models on a subset of the document corpus by sampling part of them, in order to verify which sampling strategy mostly improves the model's performance. In addition to evaluating the proposed scheme in the CORD-19 dataset, we further validate our approach by putting the finetuned model and sampling strategy in production on a real EBM system (Epistemonikos) to select COVID-19-related documents for being reviewed by volunteer physicians, thus providing an explicit human-in-the-loop validation. In summary, the main contributions on this chapter are the following:

(i) We evaluate, implement, and analyze language models for the medical documents classification task, obtaining that the XLNet model finetuned with an extensive EBM dataset can generalize for articles related to novel diseases and different type-of-article distributions.

(ii) Although the finetuned XLNet performs well on a dataset consisting only of COVID-19 articles, we show that results can be improved by labeling a small fraction of testing documents using an uncertainty sampling strategy.

(iii) We validate, with actual data from an EBM production environment, that the model's prediction uncertainty is a good proxy for identifying the fraction of documents to be manually labeled. In fact, this could reduce approximately 65% of expert labelers' daily workload, measured as the number of documents needed to be manually reviewed.

The remainder of this Chapter is distributed as follows: first, we describe the EBM system where experiments were carried out. The subsequent section shows the proposed method, which was evaluated offline and then in production. In the following section, we explain datasets used for experiments. Then, we analyze the results, while in the next section, we perform a user evaluation of our proposed method taken to production on an EBM system. Finally, we state discussions and conclusions.

4.1. Evidence based medicine interface

This section describes the Epistemonikos annotator interface used by volunteers in order to classify new documents and decide whether to include them in the Epistemonikos database.

The Epistemonikos annotator interface is shown in Figure 4.1. It can be seen that the collaborator can (A) search for other evidence to find additional information about the document being reviewed. Can review the (B) title of the article , (D) metadata (i.e., authors, journal, year, DOI and links to other sources), (E) abstract, and the (F) model prediction certainty for each possible category, i.e., Systematic Review, Primary Study RCT, Primary Study non-RCT and Broad Synthesis. Then, given all this information, the

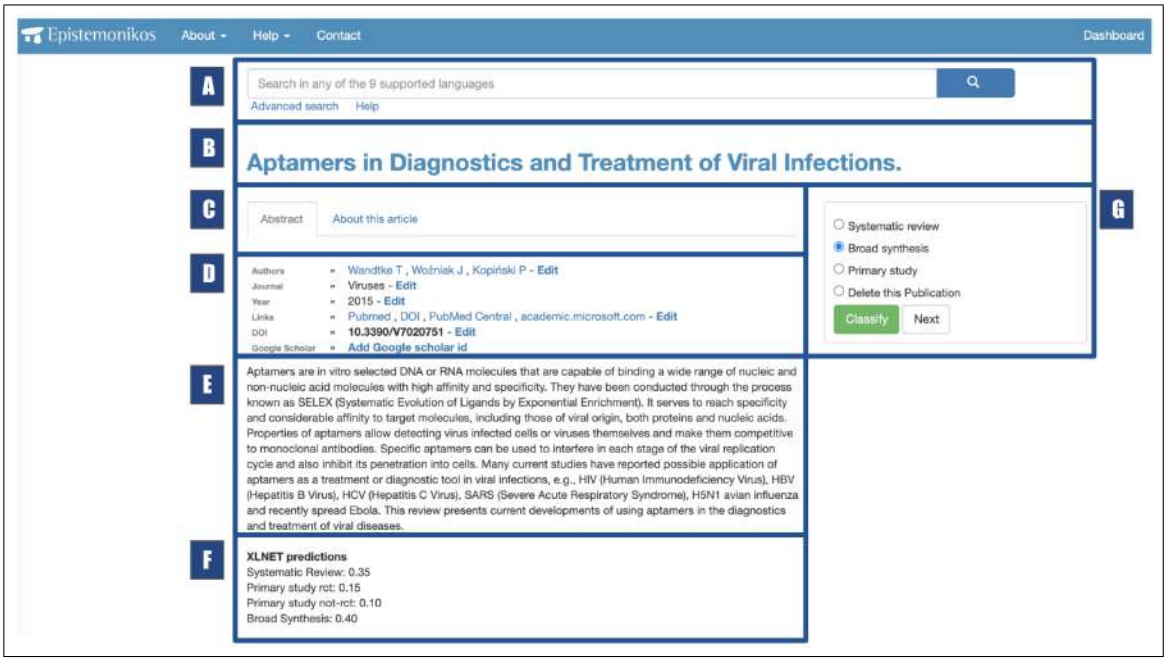


Figure 4.1. Epistemonikos annotator interface. (A) Search bar where physicians can retrieve articles. (B) Title of the current article being reviewed. (C) Menu to read the current article's abstract or add additional metadata in the "About this article" tab menu, i.e., study design. (D) Additional metadata of the current article, e.g., Authors, Journal, Year, External Links, DOI, and Google scholar link. (E) Current document abstract. (F) XLNet model prediction certainty for each potential class: Systematic Review, Primary Study RCT, Primary Study not-RCT, and Broad Synthesis. (G) Classification box where physicians, according to the given information, assign a label to the current document in Systematic Review, Broad Synthesis, Primary Study, or Exclude the publication. The user can also skip the current article by pushing the "next" button.

physician can (G) classify the given article in one of the potential classes. It is worth mentioning that in the (C) "about this article" tab, the user can add additional metadata to the article such as study design (randomized or non-randomized controlled trial), country, language, among others.

The resulting open-source interface for searching human validated evidence is shown in Figure 4.2. It can be seen that the final user has access to types of articles manually

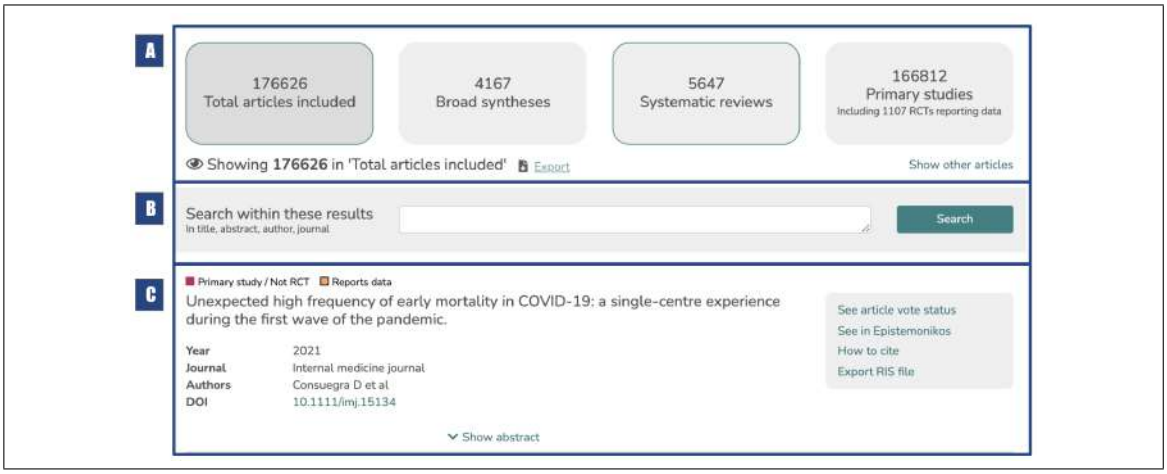


Figure 4.2. Epistemonikos platform after obtaining human validated articles with their related content. (A) Statistics of the number of articles of each category, namely Broad Synthesis, Systematic Reviews, Primary Studies RCT and Primary Studies non-RCT and Total Articles. Users can also click on the type of study to filter the results. (B) Search bar. (C) Search results include the type of study, article title, metadata, i.e., year, journal, authors, and DOI, article vote status, Epistemonikos link to the article, citation text, and possibility to export the bibliographic citation (RIS) file.

validated by physicians indexed at Epistemonikos (A), namely Broad Synthesis, Systematic Reviews, or Primary Studies, each one of them with their corresponding number of documents available in the database. Also, the user can click on each type of article to filter the results. Other alternatives are available to access additional information, such as further related articles and the article votation results from collaborators. Then for each of the search results, there is information on the type of article, title, abstract, and metadata (year, journal, authors, and article DOI).

4.2. Proposed method

In this chapter, we compare several recent document classification methods' performance, such as XLNet (Z. Yang et al., 2019), BioBERT (J. Lee et al., 2020) and BERT

(Devlin et al., 2018) language models. For each of these three methods, we compute embeddings of the documents and then perform classification using a fully connected layer.

We then experiment with several finetuning strategies to verify whether it is possible to improve the previous results by using a small proportion of COVID test set evidence.

Finally, based on the finetuning results, we use the best performing strategy in production in a real EBM system to evaluate if this strategy allows physicians to save a part of their daily workload, measured in the number of documents needed for being manually reviewed.

4.2.1. Medical documents categorization

In the context of EBM, documents are classified on the following categories:

- **Primary study RCT:** studies that use a methodology where subjects are randomly assigned to one of two groups: the experimental and control. The experimental group receives the intervention that is being tested, and the other group receives an alternative treatment, which in most cases is a placebo. Considered articles are those that report a randomized trial, also including trial registries and protocols.
- **Primary study non-RCT:** case studies that do not use an RCT methodology and show isolated results on particular cases without a robust study design methodology. Moreover, they are primary studies that do not fulfill a randomized trial. A primary study is an umbrella term that includes any study design, qualitative or quantitative, where new data is collected from individuals, populations, or any experimental subject. Other criteria of inclusion are (a) pre-clinical research in humans and (b) modeling studies.

- **Systematic review:** a type of article that uses an explicit methodology to summarize, identify and appraise all the evidence related to a specific medical issue. In most cases, they are composed of RCT study design due to their robustness as relevant evidence. A systematic review seeks to answer a research question, employs a comprehensive and reproducible search strategy, identifies all relevant studies, and can take years to be completed with the work of several collaborators. To consider a systematic review for inclusion it fulfills the following criteria: (a) provides a description of at least one eligibility criterion, (b) its main objective is to synthesize primary studies (other syntheses might be used as an additional source for studies) and (c) reports an explicit method that includes searching in at least one electronic database.
- **Broad synthesis:** a type of article that summarizes relevant evidence related to a medical issue but is not as extensive as a systematic review. Furthermore, these types of articles synthesize systematic reviews and, sometimes, primary studies. Broad synthesis considered for inclusion (a) reports an explicit method that includes searching in at least one electronic database, and (b) its main objective is to synthesize systematic reviews.
- **Excluded:** documents that, as they do not belong to any of the other categories, are excluded or not considered as relevant evidence.

Regarding a conventional EBM practice, the most relevant evidence to support medical decisions corresponds in most cases are systematic reviews and primary studies RCT (Egger et al., 2008). However, in some cases, broad synthesis articles are also considered robust if they have RCT within their references (Mays et al., 2005; Dwan et al., 2008).

4.2.2. Text representation and classification methods

In this section we describe architectures of state-of-the-art language models based on transformers: BERT, BioBERT and XLNet for the classification of EBM articles.

4.2.2.1. BERT-based models

As a baseline model, we use the BioBERT language model encoder, which is a transformer-based BERT model pre-trained on one million full PubMed articles to represent documents as a 768 dimension vector, and then use a fully connected classification layer with a softmax to make the corresponding prediction on the type of document.

We also use the traditional BERT-base language model encoder as a baseline, pre-trained on BooksCorpus (800M words) and English Wikipedia (2,500M words), corresponding to general domain corpora.

Both models are trained using a masked-training approach where random words are hidden, and the model has to predict the corresponding word. Also, another task learned by these models is to predict if two sentences are related. One limitation of this approach is that due to computational complexity, the length of the input tokenized document has to be lower than 512.

For training, we finetuned both the classification layer and the BERT and BioBERT encoders to improve models performance for this particular EBM article classification task. Then for testing, we used the resulting trained model to make the predictions on the CORD-19 dataset.

The architectures of the BERT and BioBERT language models and fully connected layer classifiers are shown in Figure 4.3. It can be seen that the model receives as input the document's words, then they pass through a BioBERT/BERT tokenizer that adds special

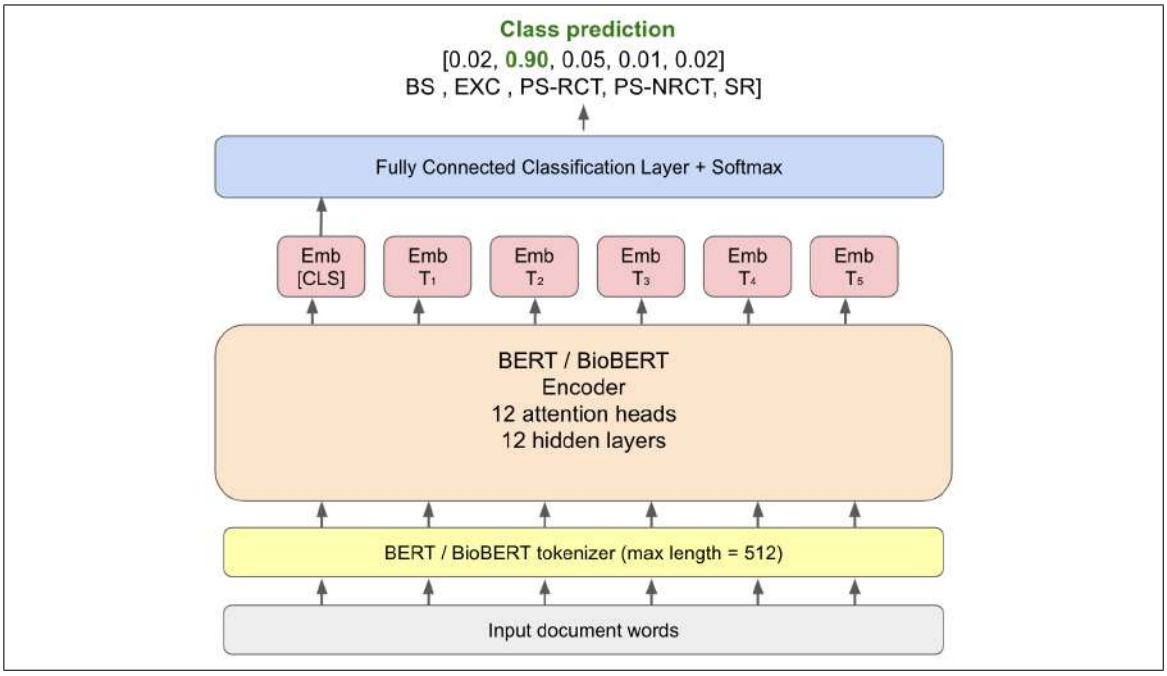


Figure 4.3. BERT and BioBERT language models classification architecture. The input of the model is the document's title and abstracts words. Then words are tokenized using the BERT or BioBERT model tokenizer with a maximum length of 512. After that, they are passed through the BERT/BioBERT encoder (12 attention heads and 12 hidden layers) and outputs an embedding for each token. Finally, the CLS token embedding is used as input of a fully connected layer using a softmax function to output the class prediction given by the maximum predicted probability. Where BS= Broad Synthesis, EXC= Excluded, PS-RCT: primary study RCT, PS-NRCT= primary study non-RCT and SR= Systematic Review.

tokens and split the document. Then they pass through a BioBERT/BERT encoder to obtain embeddings for each token. Finally, the *CLS* token embedding is used to represent the document as input for a Fully Connected classification layer, apply softmax and obtain the corresponding class prediction.

For the non-finetuned models we trained BERT and BioBERT linear layer for 5 epochs, with a learning rate of 0.0001, weight decay of 0.01 and Adam (Kingma & Ba, 2014) optimizer. We chose 5 epochs as the model converged on these number of epochs. The

architecture of BERT and BioBERT encoders consists of 12 attention heads and 12 hidden layers.

Concerning the finetuned model, we trained the linear layer and the BERT/BioBERT encoder weights for 20 epochs with a learning rate of 0.0001, a weight decay of 0.01 and we used Adam optimizer.

4.2.2.2. XLNet

For this approach we propose using the XLNet transformer-based language model encoder (Z. Yang et al., 2019) to represent documents as a 768-dimension vector, and then use a fully connected classification layer with a softmax to make the corresponding prediction on the type of document. The architecture of XLNet encoder consists in 16 attention heads and 24 hidden layers.

The XLNet model uses a permutation language model approach that is trained to predict one token given preceding context like a traditional language model, but instead of predicting the tokens in sequential order, it predicts tokens in some random order. This way the model is forced to model bidirectional dependencies. Another benefit of this model is that as it is based on a Transformer-XL architecture it allows as input a tokenized document of any given length.

For training, we finetuned both the classification layer and the XLNet weights to improve models performance for this particular classification task. For testing, we then used the resulting trained model to make the predictions.

For the non-finetuned model, we trained the linear layer for 5 epochs, with a learning rate of 0.0001, weight decay of 0.01 and Adam optimizer. We chose 5 epochs because the model converged on these number of epochs.

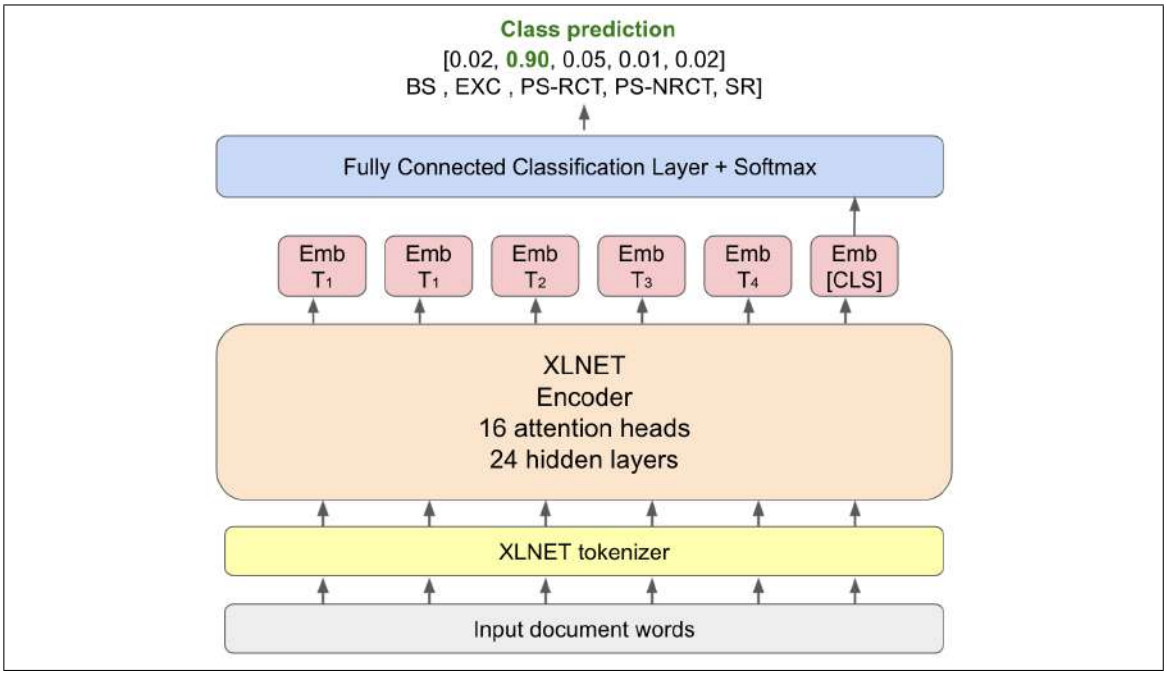


Figure 4.4. XLNet language model and classification architecture. The input of the model is the document's title and abstracts words. Then words are tokenized using the XLNET model tokenizer. After that, they are passed through the XLNET encoder (16 attention heads and 24 hidden layers) and outputs an embedding for each token. Finally, the CLS token embedding is used as input of a fully connected layer using a softmax function to output the class prediction given by the maximum predicted probability. Where BS= Broad Synthesis, EXC= Excluded, PS-RCT: primary study RCT, PS-NRCT= primary study non-RCT and SR= Systematic Review.

Concerning the finetuned model we trained the linear layer and the XLNet parameters for 20 epochs with a learning rate of 0.0001, a weight decay of 0.01 and Adam optimizer.

The architecture of this language model and fully connected layer classifier is shown in Figure 4.4. It can be seen that the model receives as input the document's words, then they pass through a tokenizer that adds special tokens and split the document. Then they pass through an encoder to obtain embeddings for each token. Finally, the *CLS* token

embedding is used to represent the document as input for a fully connected classification layer, apply softmax and obtain the corresponding class prediction.

This architecture is similar to BioBERT, although what changes is the tokenizer and the encoder. Another difference with BioBERT is that the CLS token is located at the end of the document.

4.2.3. Finetuning strategies

We hypothesize that the results obtained by the previously described models can be improved by finetuning the model with a small proportion of the test dataset (CORD-19).

We test three finetuning strategies and compare their results with the original models to verify if there is space for improvement. In order to carry out the experiments, we make a partition of the CORD-19 dataset in two test sets, the first one is for making the sampling strategies and training the models, and the second half is for making predictions shown in the results section.

The datasets used for sampling strategies and for testing the model's predictions consist both of 9,427 documents. The first one is distributed into 4,556 primary-not-rct , 2,839 excluded , 1,658 systematic reviews , 250 broad synthesis and 124 primary rct. Then the second dataset used for predictions consists of 4,548 primary-not-rct , 2,790 excluded, 1,720 systematic reviews , 241 broad synthesis and 114 primary rct.

For results consistency, we carry out ten iterations of bootstrap sampling for each of the sample strategies and calculate the average recall, precision, f1 score and weighted average.

The taken strategies were the following:

- **Without finetuning :** the model trained with the Epistemonikos dataset without finetuning on CORD-19 dataset.
- **Random sampling:** we perform a random sample of 200 documents from the CORD-19 test set and finetune the model with these documents.
- **Random sampling augmented:** we use the same random sampled documents from the random sampling strategy. However, we use data augmentation, obtaining 600 documents in total, where 200 are original from the test set, and the other 400 were synthetically augmented. We use a data augmentation strategy proposed by Ma (2019), where random words from the article are replaced with a probability of 0.5, with the closest BioBERT word embedding in the latent space. Furthermore, this finetuning strategy aims to increase the number of minority class documents to be balanced.
- **Uncertainty sampling:** We perform two iterations of a classification uncertainty sampling strategy based on the maximum predicted probability (Settles, 2009). To achieve this, we sample articles where the maximum probability predicted by the model was lower than 0.5, which are documents where the model is more uncertain of the possible class. We finetuned the model with these examples and then extracted a sample again, repeating this process for two iterations. It is worth mentioning that the number of articles sampled decreased after each iteration since the model increased the certainty of predicted classes. Given this, we sampled 200 documents on the first iteration and then 200 in the second iteration. Performance stayed the same or worsened for subsequent iterations, so they were not considered in the analysis.

4.3. Datasets

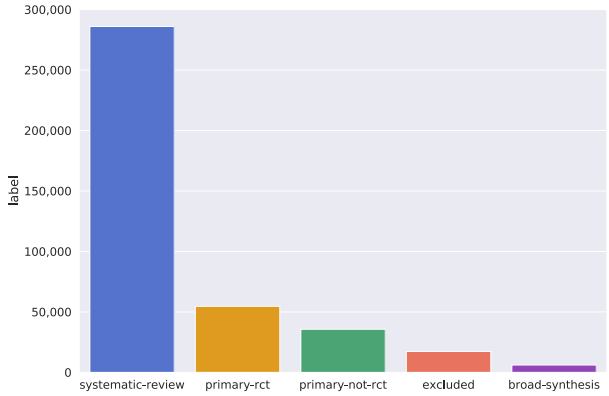
In this section, we describe the datasets used for training and testing our proposed method and baselines. Regarding the training set, we use a dataset provided by Epistemonikos that contains 399,737 medical articles distributed in systematic reviews, primary studies RCT, primary studies non-RCTs, broad synthesis, and excluded. These articles do not include any evidence related to COVID-19 since they were all published before November 2019. The content of these articles consists of their titles and abstracts.

For testing, we use CORD-19, a well-known COVID-19 publicly available dataset (L. L. Wang et al., 2020), which for this study was adapted by crossing CORD-19 Pubmed IDs with the Epistemonikos database to obtain the type of documents. The original CORD-19 has 59,000 articles, and by crossing both databases, we obtained 18,854 documents consisting of their title and abstract with their corresponding EBM labels.

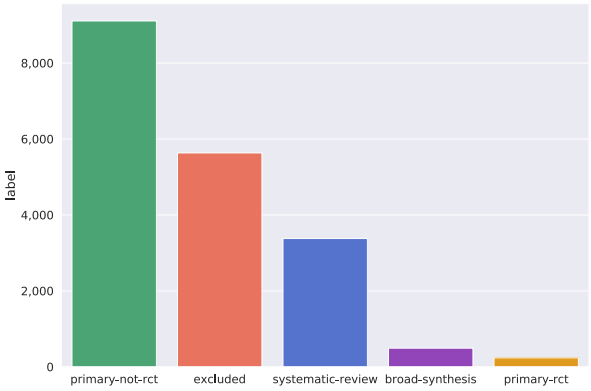
The following subsections give more details on these datasets concerning the distribution of types of documents, document length, and medical terms for both Epistemonikos and CORD-19 datasets.

The distribution of documents among different types of documents are shown in Figure 4.5. It can be seen that most of the articles from the train set (Epistemonikos) are systematic reviews (286,050), followed by primary studies RCT (54,623), primary studies non-RCT (35,644), excluded (17,324), and broad synthesis (6,096).

Then for the test set (CORD-19), we observe that the distribution is different from the train set. Since the primary study non-RCT (9,110) is the most frequent class, followed by excluded (5,634), systematic reviews (3,380), broad synthesis (492), and primary studies RCT (238).



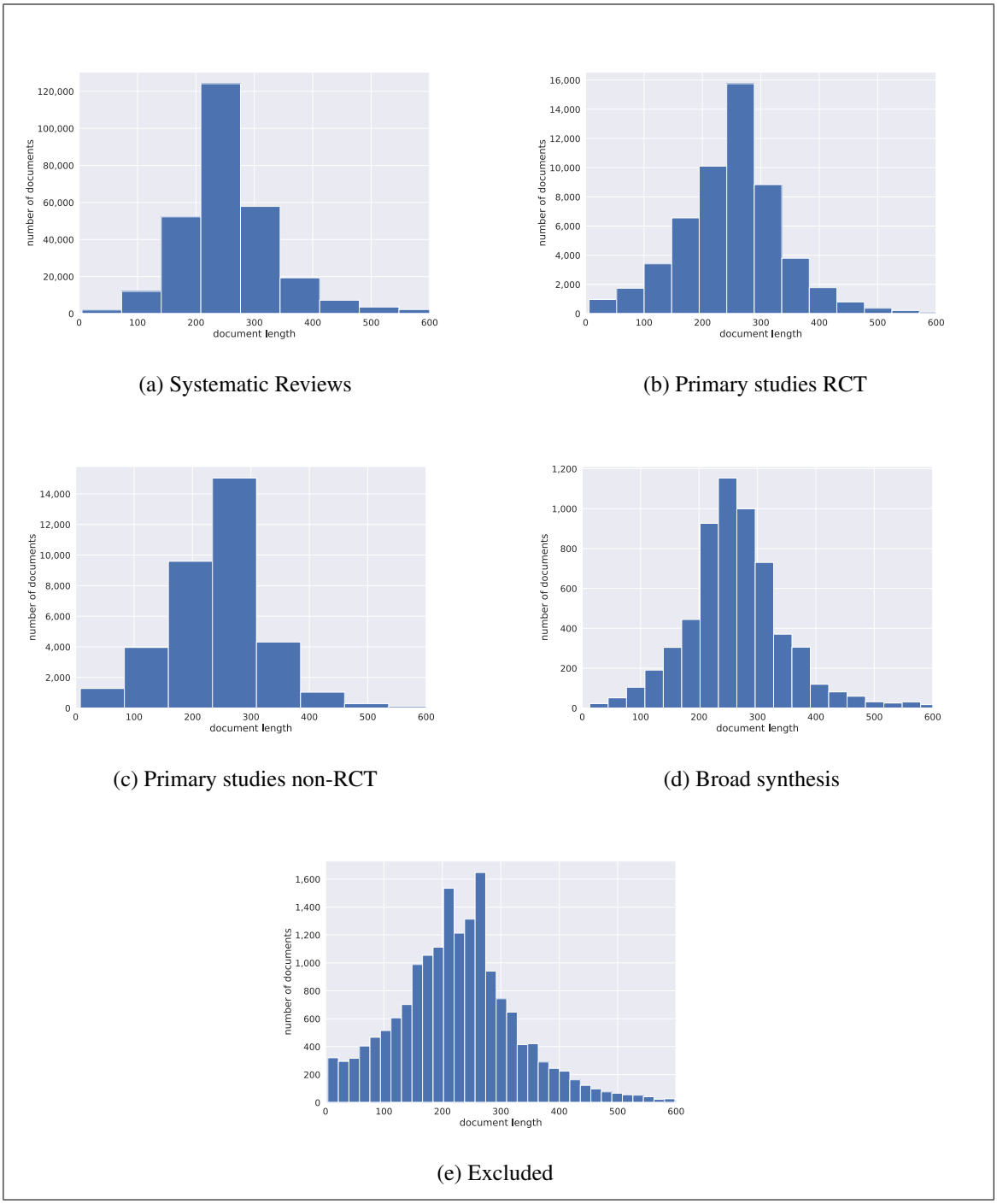
(a) Epistemonikos train set distribution type of articles.



(b) CORD-19 test set distribution type of articles.

Figure 4.5. Distribution of type of articles in the train set (Epistemonikos) and test set (CORD-19). (a) Distribution of types of articles in the train set (Epistemonikos), i.e., Systematic Review, Primary Study RCT, Primary Study non-RCT, Excluded and Broad Synthesis. (b) Distribution of types of articles in the test set (CORD-19), i.e., Systematic Review, Primary Study RCT, Primary Study non-RCT, Excluded and Broad Synthesis.

There is a considerable difference in the document distributions between both train and test sets. This may be produced because, as COVID-19 is a novel disease, at that moment, there was not enough evidence from primary studies to include in systematic reviews. Furthermore, this difference in document types distribution between the train and test sets implies an additional challenge in proposing a solution capable of generalizing new domains and distributions.



(a) Systematic Reviews

(b) Primary studies RCT

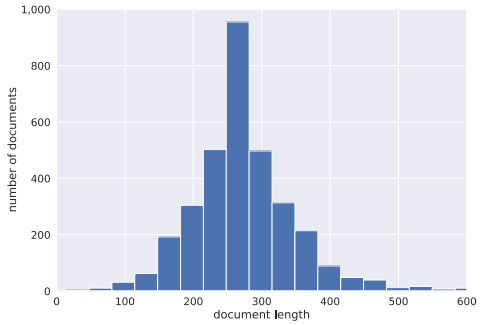
(c) Primary studies non-RCT

(d) Broad synthesis

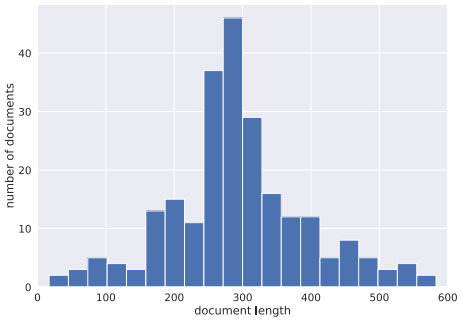
(e) Excluded

Figure 4.6. Document length distribution in the train set (Epistemonikos) among different types of documents where the x-axis indicates the document length and the y-axis shows the number of documents that have that length. (a) Distribution of document length of Systematic Review articles. (b) Distribution of document length of Primary studies RCT articles. (c) Distribution of document length of Primary studies non-RCT articles. (d) Distribution of document length of Broad Synthesis articles. (e) Distribution of document length of Excluded articles.

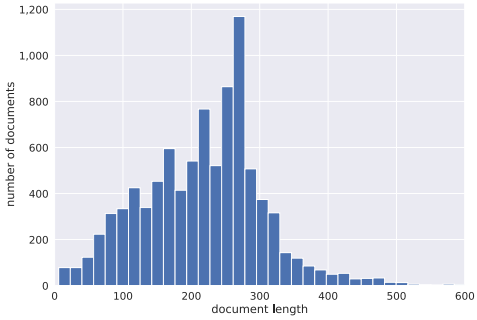
The distribution of the type of article length from the train set (Epistemonikos) is summarized in Figure 4.6. The *y-axis* shows the count of documents and the *x-axis* indicates document length. In the case of systematic reviews, about 120,000 documents have a length between 200 and 300 words. This distribution is similar to other types of articles since most documents have that number of words.



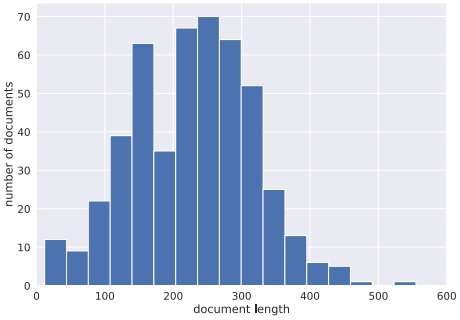
(a) Systematic Reviews



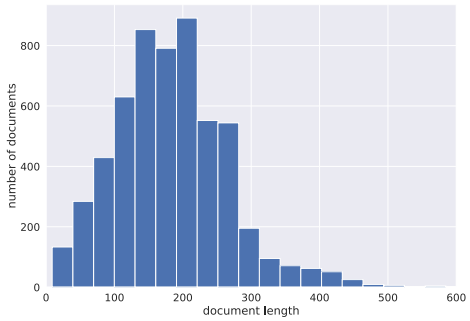
(b) Primary Studies RCT



(c) Primary Studies non-RCT



(d) Broad Synthesis



(e) Excluded

Figure 4.7. Document length distribution in the test set (CORD-19) among different types of documents where the x-axis indicates the document length and the y-axis shows the number of documents that have that length. (a) Distribution of document length of Systematic Review articles. (b) Distribution of document length of Primary studies RCT articles. (c) Distribution of document length of Primary studies non-RCT articles. (d) Distribution of document length of Broad Synthesis articles. (e) Distribution of document length of Excluded articles.

The distribution of type of article length from the test set (CORD-19) is presented in Figure 4.7. The *y-axis* shows the count of documents, and the *x-axis* indicates their document length as the number of words. In the case of systematic reviews, primary studies, and broad synthesis, we observe a similar distribution as the train set (Figure 4.6), where most of the documents have a length between 200 and 300 words. Although, for the excluded, most of the articles have less than 200 words.

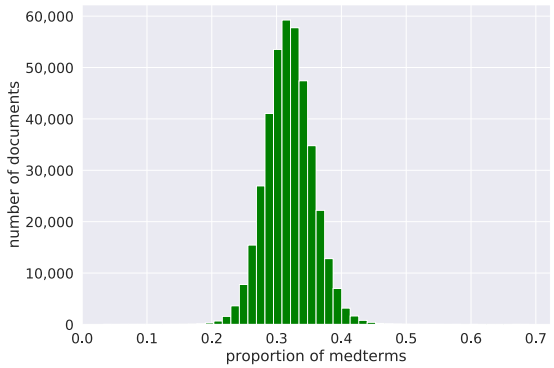
The distribution of the density of medical terms in train and test sets is shown in Figure 4.8. The *y-axis* shows the count of documents, and the *x-axis* the proportion of medical terms over the total words of the document. It can be seen that the proportion of medical terms of both train and test datasets show a similar distribution, where the majority of documents have a ratio between 0.28 and 0.33 medical terms.

4.4. Results

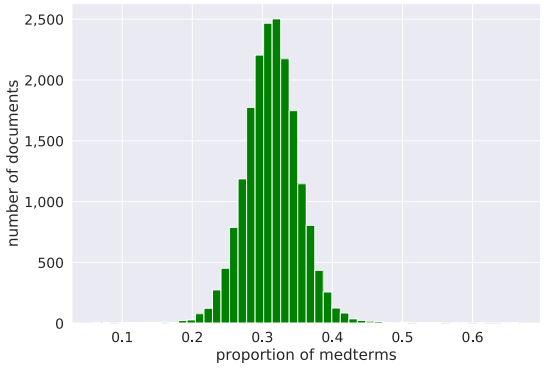
In this section we show and analyze the results of the text classification models on the CORD-19 dataset: i) using pretrained neural language models for embeddings followed by a linear classification layer trained in the Epistemonikos set, and ii) jointly active learning finetuning strategy on the language model and the classification layer end-to-end in the Epistemonikos set. In the following section we show the results of applying different finetuning strategies to the best performing model using data from CORD-19.

4.4.1. Classification results with Epistemonikos training data

For these experiments we compare the performance of XLNet, BioBERT and BERT trained on Epistemonikos dataset and tested on CORD-19. We measure their capability to classify types of documents in terms of precision, recall and f1-score for novel COVID-19 articles.



(a) Distribution of medical terms proportion on Epistemonikos train set documents



(b) Distribution of medical terms proportion on CORD-19 tests set documents

Figure 4.8. Distribution of medical terms proportion among the number of documents, where x-axis shows the proportion of medical terms and the y-axis indicated the number of documents that have that proportion. (a) Distribution of medical terms proportion on the train set (Epistemonikos). (b) Distribution of medical terms proportion on the test set (CORD-19).

The results on Table 4.1 were obtained by using pretrained language models followed by a classification layer trained with Epistemonikos data. In terms of weighted average,

Table 4.1. Results obtained for document classification using neural language models trained only on the linear layer on Epistemonikos dataset (N = 399,737) and tested on the CORD-19 dataset (N= 18,854). Classes are Broad Synthesis, Systematic Review, Primary Study RCT (Primary RCT), Primary Study non-RCT (Primary non-RCT), and Excluded (ex). In **bold** we show the model with best performance on that given type of document and metric.

Type of article	BERT			XLNet			BioBERT		
	Prec.	Rec.	F-1	Prec.	Rec.	F-1	Prec.	Rec.	F-1
Broad Synthesis	.00	.00	.00	.00	.00	.00	.00	.00	.00
Excluded	.57	.00	.00	**.77**	**.13**	**.22**	.73	.06	.12
Primary RCT	.07	.09	.08	.27	.61	.37	**.30**	**.68**	**.42**
Primary non-RCT	.91	.03	.07	.79	**.67**	**.72**	**.85**	.55	.67
Systematic Review	.19	**.99**	.31	**.34**	.98	**.51**	.28	**.99**	.44
Weighted Avg	.65	.20	.09	.68	.54	.51	.68	.47	.44

Table 4.2. Results obtained for document classification using models finetuned end-to-end on the Epistemonikos dataset (N = 399,737) and tested on the CORD-19 dataset (N = 18,854). Classes are Broad Synthesis, Systematic Review, Primary Study RCT (Primary RCT), Primary Study non-RCT (Primary non-RCT), and Excluded (ex). In **bold** we show the model with best performance on that given type of document and metric.

Type of article	BERT			XLNet			BioBERT		
	Prec.	Rec.	F-1	Prec.	Rec.	F-1	Prec.	Rec.	F-1
Broad Synthesis	.53	.37	.44	**.84**	**.77**	**.81**	.56	.69	.62
Excluded	.86	.83	.84	**.97**	**.96**	**.97**	.90	.62	.73
Primary RCT	.63	.84	.72	**.83**	**.89**	**.86**	.64	.80	.71
Primary non-RCT	.91	.93	.92	**.99**	**.99**	**.99**	.82	.96	.88
Systematic Review	.90	.93	.91	**.94**	**.97**	**.96**	**.94**	.92	.93
Weighted Avg	.88	.88	.88	**.97**	**.97**	**.97**	.85	.84	.84

XLNet outperformed both BioBERT and BERT models tested on the CORD-19 dataset (.51 f1-score). In terms of classification of randomized controlled trial articles, BioBERT performed better than the other models (.42 f1-score). Concerning systematic reviews, excluded and non-randomized controlled trials XLNet yielded better performance shown by a higher f1 score on all types of articles (.22, .72, and .51 f1-score, respectively). Finally,

concerning Broad Synthesis, none of the models trained only on the linear layer could identify these types of articles.

The results on Table 4.2 were obtained by finetuning the pretrained language models jointly with a linear classification layer using Epistemonikos data. In this setting, XLNet outperformed other neural language models on average and for all the types of articles. This indicates that although the XLNet model was trained with articles not related to COVID-19, testing on CORD-19 showed its capability to generalize to novel diseases.

By comparing models trained only on the linear layer (Table 4.1) with the models trained end to end (Table 4.2), we achieved a significant improvement in the classification of all of the classes, especially in identifying Broad Synthesis. In the discussion section we review reasons in the models, which explain these results.

In the next section, we will compare several sampling strategies to choose informative articles to finetune the XLNet model with the least number of examples.

4.4.2. Results of active learning finetuning strategies

Based on the finetuning results for XLNet, we argue that there is space for improvement by subsequent finetuning the model with a small proportion of new articles related to COVID-19. To assess this, we finetune the best performing XLNet model from the previous section using a sample of CORD-19 articles based on two active learning strategies (random sampling, and uncertainty sampling) as well as data augmentation.

The purpose of doing this is to see if updating the model by sampling a small portion of the COVID documents is enough to improve the model's performance already pre-trained with a large dataset.

Table 4.3. Results on transfer learning strategies on CORD-19 dataset using XL-Net model finetuned on Epistemonikos for broad synthesis (bs), excluded (exc), primary study RCT (ps-RCT), primary study with non-RCT (ps-nrct), systematic review (sr), and weighted average (avg) for each of the metrics. In **bold** we show the transfer learning strategy that outperforms the original model in that corresponding metric. In parenthesis, we show the standard deviation of ten bootstrap sampling iterations for each model. The dataset used for testing the model's predictions consists of 9,427 documents spread into 4,548 ps-nrct, 2,790 exc, 1,720 sr, 241 bs, and 114 ps-rct. The * indicates the most relevant types of articles in the field of EBM, which are sr and ps-rct.

Strategy	**Metrics**	**# docs**	**bs**	**exc**	**ps-rct***	**ps-nrct**	**sr***
XLNet	Prec.		.842	.969	.810	.990	.946
Finetuning	Rec.	-	.747	.968	.856	.987	.967
Epistemonikos	F1-score		.791	.969	.832	.988	.957
Random	Prec.		.778 (.007)	.893 (.025)	**.866** (.013)	.990 (.001)	**.956** (.004)
Sampling	Rec.	200	.568 (.054)	**.973** (.002)	**.857** (.016)	.958 (.007)	.937 (.012)
200 docs	F1-score		.655 (.037)	.931 (.013)	.861 (.003)	.974 (.008)	.946 (.004)
Random	Prec.		.773 (.012)	.868 (.018)	**.864** (.014)	**.992** (.001)	**.959** (.002)
Sampling	Rec.	400	.561 (.055)	**.976** (.001)	**.859** (.006)	.943 (.011)	.929 (.008)
400 docs	F1-score		.649 (.039)	.918 (.009)	**.862** (.004)	.966 (.005)	.943 (.002)
Data	Prec.		.704 (.005)	.958 (.015)	**.812** (.002)	.986 (.006)	.947 (.002)
Augmentation	Rec.	600	**.836** (.036)	.962 (.011)	.856 (.001)	.983 (.007)	.966 (.002)
	F1-score	(400 aug)	.764 (.023)	.960 (.005)	.833 (.001)	.985 (.003)	.956 (.001)
Uncertainty	Prec.		.795 (.016)	.889 (.014)	**.856** (.013)	**.991** (.001)	**.960** (.001)
Sampling	Rec.	200	.731 (.052)	**.975** (.001)	.822 (.012)	.946 (.009)	.937 (.003)
iteration 1	F1-score		.760 (.024)	.930 (.007)	.838 (.005)	.968 (.004)	.948 (.001)
Uncertainty	Prec.		.733 (.039)	.968 (.006)	**.862** (.004)	.984 (.003)	**.955** (.001)
Sampling	Rec.	400	**.803** (.014)	.952 (.006)	**.865** (.012)	.988 (.003)	**.972** (.004)
iteration 2	F1-score	(US1+iter2)	.766 (.015)	.960 (.003)	**.863** (.004)	.986 (.003)	**.963** (.001)

The results of different sampling strategies on the best XLNet model from the previous section are summarized in table 4.3. They show a comparison of the performance of test

set sampling strategies among classes used for finetuning the XLNet model in terms of precision, recall, and f1-score. The compared methods are random sampling, data augmentation, and uncertainty sampling on different numbers of iterations. We show the average of each bootstrap sampling for each class and finetuning strategy on the CORD-19 partition used for testing models predictions.

It can be seen that there is an improvement in terms of recall for broad synthesis articles by using uncertainty sampling (.803) and data augmentation (.863).

In the case of excluded documents, random sampling of 400 documents is the strategy that most increases the model's performance in identifying these types of articles. Something similar occurs with non-randomized primary studies since the random sampling of 400 articles is the strategy that most increases the precision of the model for distinguishing these types of articles.

For systematic reviews and randomized primary studies, which are the most robust evidence in EBM (Egger et al., 2008), the uncertainty sampling strategy with two iterations of 200 articles is the only one that increases the model performance in identifying both types of articles in terms of precision (ps-rct=.852, sr=.955), recall (ps-rct=.865, sr=.972) and f1-score (ps-rct=.863, sr=.963).

Although XLNET yields good generalization results without finetuning, the performance of predictions of SR and ps-RCT, which are the most relevant evidence in EBM, have space for improvement. This improvement is achieved by using the uncertainty sampling method for choosing a small proportion of COVID-19 documents, allowing the model to identify with more accuracy SR and ps-RCTs shown by an increase of 13.7% f1-score for ps-RCT and 3.7% f1-score for SR by using this sampling strategy.

4.5. User Evaluation

After the offline model evaluation, we pick the best performing model (XLNET) and test it in production with users of the Epistemonikos platform. Our main goal is saving time from physicians' work by asking them to label documents only if the model is not confident about the classification. With this user study we can quantify if this strategy works, and how much time we can save from the physicians' effort.

To achieve this we select the finetuned document classification model in production on a real EBM system (Epistemonikos) between September 2020 to May 2021. To select papers for labeling, we used the uncertainty sampling strategy on COVID-19-related documents for being reviewed by volunteer physicians. Every day, Epistemonikos volunteers are given documents already classified by the model, and if they find an error in the classification, they assign the correct label that corresponds to the articles; in another case, the document is classified in the same way as the model predicts.

Although the model was trained to classify into one of the five classes mentioned earlier, the class "excluded documents" was not included among the potential documents to classify, since they are categorized using other heuristics that do not depend entirely on the model's prediction.

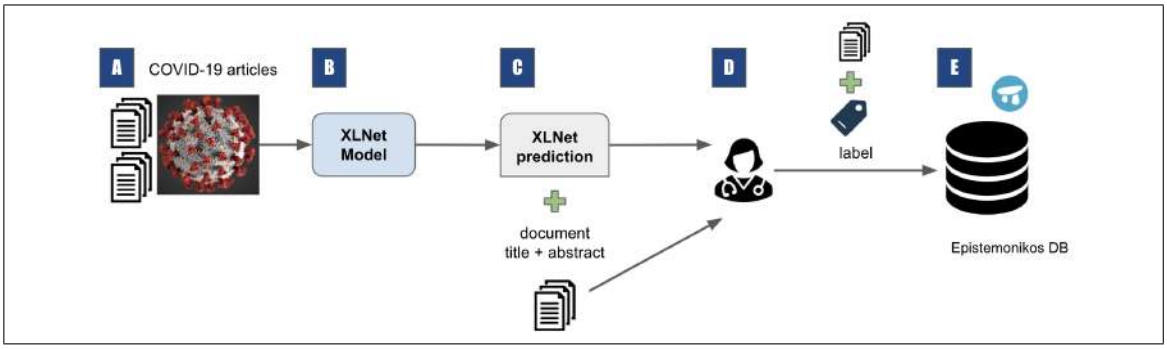


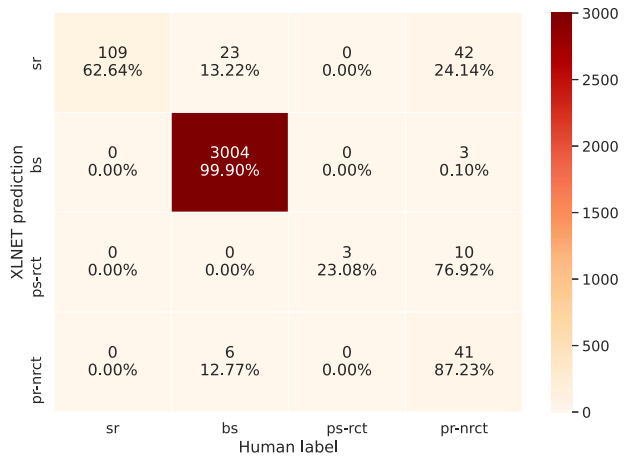
Figure 4.9. Epistemonikos document classification process. (A) Documents titles and abstracts related to COVID-19 disease are passed to an XLNet model. (B) XLNet model processes the documents and makes predictions. (C) XLnet prediction for each class and documents are given to physicians to review. (D) Physicians decide on the document label by considering XLNet prediction and the document's title and abstract. (E) Documents and human assigned labels are included in the Epistemonikos database avaliable for future EBM researchers.

A diagram of the human validation process of the XLNet model predictions and inclusion of labeled articles to the Epistemonikos database is shown in Figure 4.9. In this process, novel COVID-19 articles indexed to online databases are passed and processed through an XLNet model. After that, the model outputs a prediction certainty for each of the possible types of articles, namely Broad Synthesis, Systematic Review, Primary study non-RCT, Primary study RCT, or Excluded. Then physicians, given the document title and abstract and the XLNet predictions, assign a label to the document. Finally, the documents and their corresponding categories (not considering excluded ones) are included in the Epistemonikos database, available as open-source for EBM researchers.

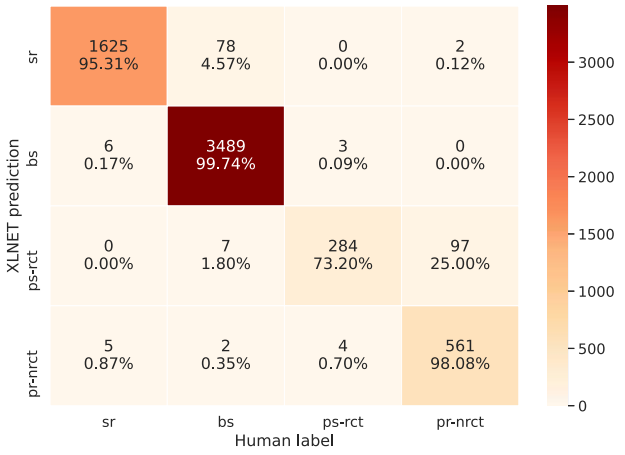
To evaluate the performance of the model on real users from an EBM platform, we divide articles into two groups, one where the model is more confident of its prediction and the other where it is uncertain.

We consider that the model is more confident on their prediction if the probability is higher than 0.5 on the class with maximum probability, and less confident in other case.

This analysis aims to assess whether we can trust the model's prediction certainty and assign experts to review only those in which the model is less confident on their possible category. After dividing these two groups, we obtained that on 6,163 (65%) of the articles, the model was confident on their predictions.



(a) XLNet uncertain predictions i.e., when the maximum probability class is less than 0.5



(b) XLNet certain predictions i.e., when the maximum probability class is more than 0.5

Figure 4.10. Confusion matrices based on XLNet predictions with Epistemonikos human labels when the model is more uncertain (maximum probability class higher than 0.5) and less certain (maximum probability class lower than 0.5).

Figures 4.10(a) and (b) show the results of confusion matrices on users' evaluation of XLNet predictions on cases where the model was more confident on their predictions, compared to cases where the model was uncertain. The results show that regardless of the model's level of certainty, both have high-grade efficiency (recall metric) in predicting broad synthesis, with 99.7% correct predictions when the model is certain and 99.9% when the model is uncertain.

Regarding systematic reviews articles, we have many more examples in which the model is more confident of its prediction, with 1,705 cases, compared to 174 in which it is uncertain. Furthermore, when the model is certain, the error rate of this class is 4.69%, compared to 37.36% when it is uncertain.

Concerning randomized controlled trials (ps-RCT), when the model is uncertain, the number of misclassified examples (10 cases, 76,62%) is higher than the ones correctly classified (3 cases, 23,08%). Then, when the model is sure of its predictions, the error rate of this type of article is 26.8%.

Finally, in the case of non-randomized controlled trials (ps-nrct), the error rate when the model is uncertain is 12.77%, which is considerably higher than the 1.92% error rate for this type of article when the model is certain on its predictions.

Table 4.4. Results obtained for document classification on Broad Synthesis, Systematic Review, Primary Study RCT, and Primary Study non RCT for the user evaluation on examples where XLNet was certain (max probability > 0.5) and uncertain (max probability < 0.5). In **bold** we show the model with the best performance on that given type of document and metric.

Type of article	Max Prob < 0.5 N = 3,214			Max prob > 0.5 N = 6,163		
	F-1	Prec.	Rec.	F-1	Prec.	Rec.
Broad Synthesis	.99	.99	.99	.99	.99	.98
Primary study RCT	.38	.23	.99	**.84**	**.73**	**.98**
Primary study non RCT	.57	.87	.43	**.91**	**.98**	**.85**
Systematic Review	.77	.63	.99	**.97**	**.95**	.99
Macro Avg	.68	.68	.85	**.93**	**.92**	**.95**
Weighted Avg	.97	.98	.97	.97	.97	.97

Table 4.4 presents the results obtained in terms of performance metrics when the model is certain and uncertain on their predictions. We can see that there is a considerable improvement in overall effectiveness if we compare the predictions when the model is certain compared to when it is uncertain. Evidence of this is given by an improvement of f1-score in the prediction of primary studies RCT (121%), non-RCT (59%), and systematic reviews (25%).

However, as analyzed in the confusion matrices, the performance for the prediction of broad synthesis is not affected by the model's certainty.

This analysis allowed us to provide strong evidence that models' certainty is a good proxy for choosing articles for being reviewed by experts. Moreover, to quantify the work saved for physicians, if we give experts only documents to review where the model is

uncertain of their predictions, we have to consider only 3,214 (35%), showing that using this strategy, we are saving approximately 65% of the workload required by physicians to review novel evidence expressed as the number of reviewed articles.

4.6. Discussion

In this chapter, we supported the results from previous studies showing that fine-tuning end to end on neural language models for a classification task considerably improves the model's performance. Although the best model, XLNet, yields good results, we performed different sampling strategies to finetune the model with a small proportion of examples from the test set, improving their performance in terms of precision and recall. We gave evidence that with a small number of documents from the test, by using an uncertainty sampling strategy, the model improved its classification performance on Systematic Reviews and Primary studies, which are the most important categories to identify in the context of EBM. Finally, we validated our results on users in a real EBM system, obtaining that their workload was reduced by approximately 65% measured as the number of documents needed for revision.

We summarize the main take-aways and some challenging aspects of our research on the following points:

- **Dataset complexity:** one of the main challenges with this data is that the test set contains diseases not seen by the model during training. Another difficulty of the dataset is that the distribution of articles differs among train and test sets. These differences were handled by proposing finetuning strategies that allow the model to better generalize for novel diseases by using the least effort regarding the number of documents needed for physician revision.

- **User evaluation workload savings:** results from user evaluation on a real EBM system allowed us to empirically demonstrate that by giving physicians only documents to review in which the model is less certain on their predictions, their workload was reduced approximately by 65%, measured as the number of documents needed for being manually reviewed.
- **Time complexity and performance trade-off:** Although the XLNet model was the model that surpasses other state-of-the-art neural language models, this model requires GPU usage, which means supplementary resources in a production environment. For training XLNET until convergence modifying the whole model weights on two *GeForce GTX 1080 Ti* GPUs took approximately 72 hours, and another issue is that without GPU, the model takes an average of 1.26 seconds for each document, which may be inefficient when needing the prediction of a vast number of documents, for example classifying 18,000 examples will take 6.3 hours. Compared with GPU, the time taken to classify each document using XLNET was an average of 0.092 seconds per document; i.e., classifying 18,000 documents takes only 26 minutes.
- **Finetuning strategies:** results showed that the model was able to generalize to new diseases since it was trained on non-COVID documents and tested on the CORD-19 dataset. However, while XLNet yields high-grade results on COVID documents, the fact of choosing documents for finetuning the model using an uncertainty sampling strategy allowed to improve its performance for EBM substantial evidence, namely systematic reviews, and randomized controlled trials. In general, uncertainty sampling helps XLNET improve systematic reviews and randomized controlled trials without affecting its performance in classifying other types of articles. Furthermore, random sampling has a better performance on most frequent articles, the performance on other classes is affected.

- **End-to-end XLNet:** results indicate that the XLNet model trained end to end surpasses other state-of-the-art models such as BERT and BioBERT. One of the main reasons for obtaining better performance than other transformer-based models is the capability of XLNet to process the entire content and not be limited to 512 tokens as BERT and BioBERT. Another explanation of XLNet's better results is its autoregressive training approach it considers all combinations of words that enable finding word relations that, in some cases, are not considered by BERT-based models that train using a mask-based method, where random words are hidden, and the task is to predict that word.

4.7. Conclusions

In this chapter, we performed a set of experiments for evidence classification tasks. We showed that finetuning an XLNet neural language model on an extensive EBM dataset significantly improves performance among other models such as BERT and BioBERT for classifying COVID-19 related articles. Although XLNet performed well on COVID-19 documents, we demonstrated that by using an uncertainty sampling strategy, the model improves its performance using a small proportion of COVID-19 articles, adapting this way to the information overflow of evidence related to this pandemic disease, especially for systematic reviews and randomized trials, which are considered as the most robust evidence in the EBM field. Finally, we validated our offline results in an official EBM system by using human-in-the-loop feedback from physicians using the platform. We found that we can save approximately 65% of experts' workload, measured as the number of documents needed to be manually labeled.

For future work, we will evaluate if by showing physicians, in a visualization, the words to which the XLNet model paid most attention, positively improves their performance on document classification and their confidence in automatic classification models in the context of EBM.

5. CHAPTER 5: USER STUDY FOR EVALUATION OF MODEL GENERATED EXPLANATIONS

In this chapter, we investigate whether including explanations in the predictions of an automated classification model is perceived as helpful for users and if they reduce the cognitive load for users on an Evidence-based medicine (EBM) platform doing a particular task. Furthermore, we also want to validate which visual encoding as highlighted words in the abstract is preferred as an explanation for document screening in the context of evidence-based medicine. Given that the document screening task receives texts as input, the explanations are shown to the users as highlighted words using different visualization alternatives as seen in Figure 5.2. We compare the helpfulness and reduction of cognitive load of three visual encodings and compare the results with a control group without visualization.

In recent years, there have been efforts to visually explain to users the predictions of an AI system such as SHAP (Lundberg & Lee, 2017) or LIME (Ribeiro et al., 2016) that highlight words that contribute more to a given prediction in the context of text classification tasks. Other works proposed using the attentions of a language model (Vaswani et al., 2017; Vig, 2019; Alammar, 2021; Geva et al., 2022), although neither of them was evaluated with final users.

Our work extends past research on visual explanations in two important directions. First, we seek to demonstrate if attention mechanisms on words bring helpful explanations for users. Second, we compare several ways of visualizing attention weights and evaluate which is more beneficial in terms of cognitive effort and performance.

In order to address these challenges, we built an explainable interface for document classification on Epistemonikos. This online web platform supports physicians in categorizing and searching for new evidence related to several medical issues (e.g., COVID19)

where collaborators that use this system curate new incoming evidence depending on the type of article.

Using this system, we conducted a controlled user study to investigate the effect of visual explanations on the article's content by comparing two interfaces, the traditional Epistemonikos system, against three different visualizations of words attended by a transformed-based model for predicting the type of article.

We used the XLNET language model, a state of the art transformer-based model that learns based on the attention mechanism, in this case on words from a document. We chose this model since it yields the best results from the previous chapter to classify documents by type of article depending on their content to carry out this user study. This model was set in production on an evidence-based medicine platform where users label new articles depending on the type of study: Primary Studies randomized trials (RCT), Primary Study non-randomized trials (non-RCT), Systematic Review, Broad Synthesis, or Excluded. Since the language model is based on the attention mechanism, the explanations allocate more importance to more relevant words to produce the final prediction. Furthermore, although the attention mechanism improves model performance there is no consensus that this mechanism is helpful as an explanation.

In this work, we validate if certain visual encodings on words are perceived as helpful compared to not using visualizations. Moreover, we propose four visual encodings: horizontal bar length, background color, brightness, and the control group without visualization.

In summary, the main contributions of this chapter are the following:

- Generally, words with large attention weights are not perceived as helpful explanations, but how extreme this perception is depends on the document type and the visual encoding.
- We empirically show that the information of the model's output prediction probability is considered the most helpful source for users to make a final decision.
- We show that the background color visual encoding for highlighting words is the one that most reduces the cognitive load. However, using no highlighted words reduces physical load and frustration.
- Background color visual encoding is perceived as more helpful for survey-type documents (i.e., SR and BS). However, brightness is perceived as more helpful for randomized and non-randomized trials (i.e., RCT and non-RCT).

In the context of this thesis, these contributions are related to the third and fourth research questions. One of them addresses if explanations are helpful for physicians for making decisions on biomedical text classification. Other one is related to physicians' preferred visual encoding to perform their day-to-day tasks. To the best of our knowledge on the use of attention mechanisms, there is no empirical evidence on users about the quality of the explanations or whether they reduce the cognitive load or the preferred visual encoding in the domain of evidence-based medicine. Although Jain & Wallace (2019) and Wiegreffe & Pinter (2019) concluded that attentions obtained from recurrent neural network models are not helpful as an explanation by using an experimental setting without human validation, in this work, we offer strong evidence through a user study that explanations allow to reduce cognitive effort and improve users performance, and we compare different visual encodings as ways of showing these explanations to users.

The remainder of this Chapter is organized as follows: first, we describe the proposed system used to evaluate explanations. We then describe the study design. In the following

sub section we explain how we measure cognitive load and usability. We analyse the results of our user study. Finally, we discuss the results and state our conclusions.

5.1. Proposed system

For this research, we proposed a user-controllable interface integrated into Epistemonikos[1], an evidence-based medicine system physicians use to curate scientific articles related to medical issues. The interface using an XLNET model in the backend for document classification was deployed into production by installing a Google Chrome extension which modifies the web page when the user uses Epistemonikos in the web browser. We chose the XLNET model to output the predictions and attention on each word using the last layer of the model according to work done by Carvallo, Parra, Rada, et al. (2020), where authors outperformed state-of-the-art language models on biomedical text classification for evidence-based medicine. The decision to use the Epistemonikos system to conduct the user study was that we wanted to have physicians collaborators evaluating the interface in real-time without interfering with their day-to-day work. It also allowed us the capability of tracking all users interactions.

The information shown by the AI method correspond to values between 0 and 1, one for each word in the document, representing the attention output obtained by the neural document classification model. With these data, based on Tamara Munzner visual design framework (Munzner, 2014), we identify these visual tasks the user is facing in the interface: (1) identify the words with the highest attention and (2) compare the attention between words in the document.

[1] https://www.epistemonikos.org/

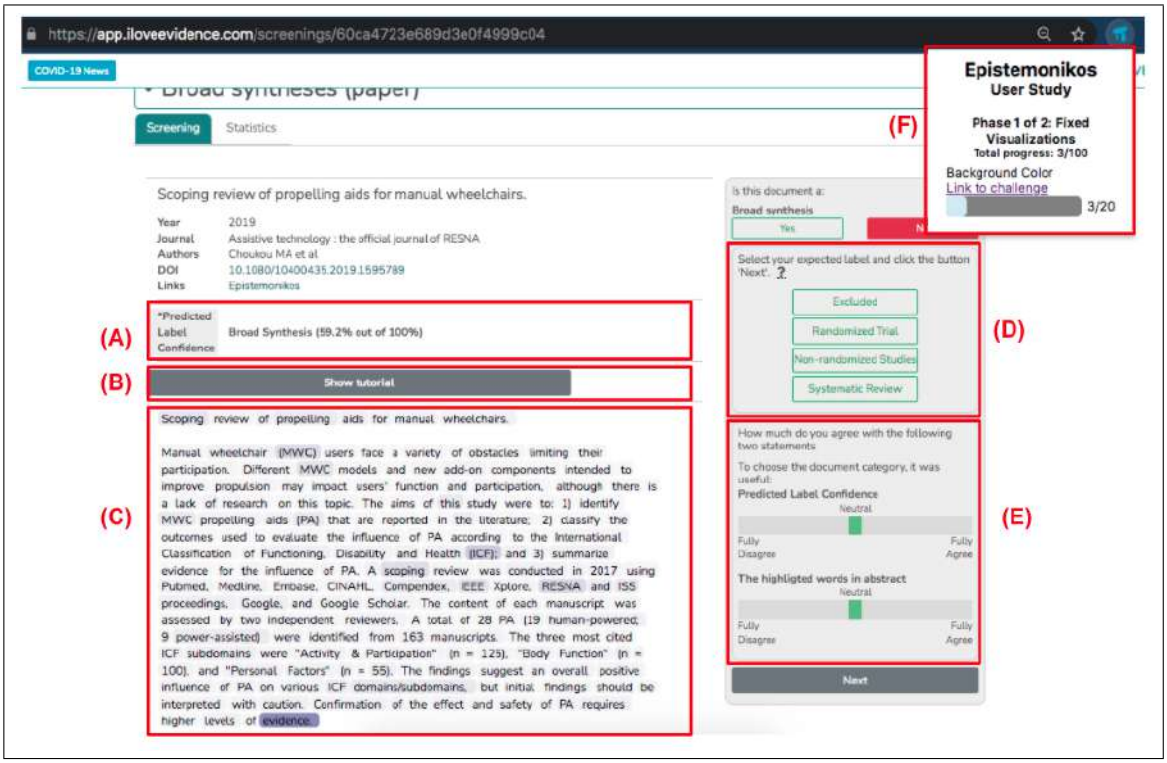


Figure 5.1. User evaluation system screenshot. (A) Document type predicted and model's predicted probability. (B) Tutorial button to give the user a tutorial on how to use the platform. (C) Abstract of the article with visualization on most relevant words considered by the model for classification, in case of no visualization the abstract has no highlighted words. (D) Classification choice where the user has to choose given the information which is the type of article. (E) Two questions where the user has to give a score between 1 and 5. The first question is how helpful was the model's predicted probability for assigning a label to the current article and the second question is how helpful are highlighted words for assigning a label to the current article. (F) Progress bar Google Chrome extension depending on how many documents are left to finish the current visualization setup and the whole user study.

The user-controllable interface (Fig 5.1) consists on three main components: (A) model type of article prediction and confidence, (B) word attention visualization, (C) category selection where the user can make a selection on the type of article according to the given information. Other additional features of the interface are a progress bar showing

the user the current progress in the study and also a tutorial that explains to new users how to use the interface.

Finally, after classifying each document the user has to answer two questions: (1) *How helpful (in a score from 1 to 5) was the model's predicted probability information to classify this document?* (2) *How helpful (in a score from 1 to 5) were the highlighted words in the abstract to classify this document?*

5.2. User study design

In this section we show how we design a user study which addresses the last two research questions of this thesis which are *Q3: How does the inclusion of explanations of automated decisions influence the decision of health experts?* and *Q4: How does certain types of visual encodings influence health experts to make the decision related to choosing a document as relevant evidence?* Additionally we study if the information of model's predicted probability is helpful for users for improving their performance in the task of document classification. Inorder to approach these questions we propose the following study design:

(i) Phase 1: where users have to interact with the interface for document screening using different visual encodings and labeling different types of articles. At the end of each evaluation users have to rate the helpfulness of current visual encoding for highlighted words in the abstract and rate the helpfulness of model's predicted probability information.

(ii) Phase 2: where users can choose one of the visual encoding or decide not to use any visualization to continue with the document classification process.

The three types of visual encodings along with the control group without visualization are further described as follows:

- **No visualization:** As seen in Figure 5.2 (A), in this case there is no visualization and it is used as a control variable.
- **Background color saturation:** As seen in Figure 5.2 (B), the XLNET attentions over words are encoded in the level of luminance of the words' background color.
- **Word Luminance:** As seen in Figure 5.2 (C), the XLNET attentions over words are encoded in the level of luminance of the words' font color.
- **Bar Length:** As seen in Figure 5.2 (D), the XLNET attentions over words are encoded in the length of a background bar.

Figure 5.2 show more details on each visual encoding proposed for the user study.

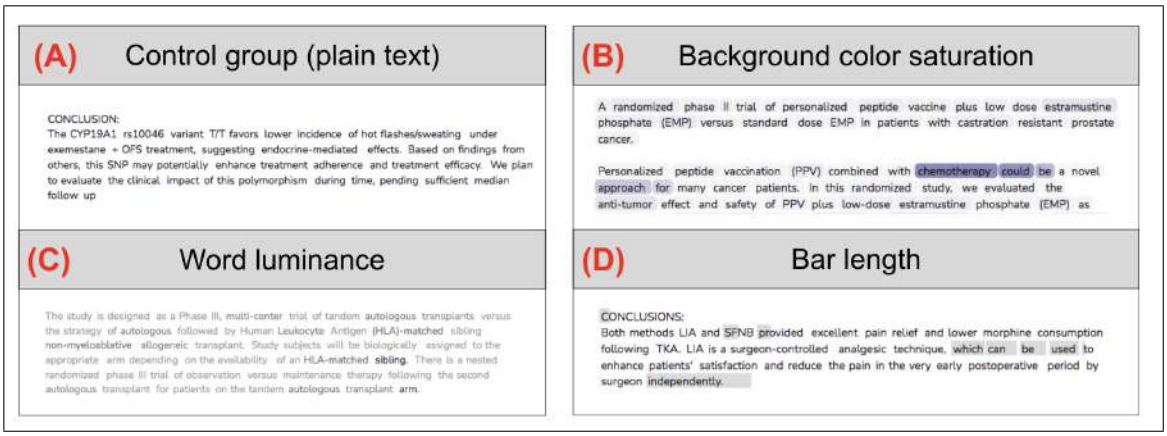


Figure 5.2. Screenshot of visual encodings. (A) Control group with no visualization. (B) Background color intensity where the more clue color intensity the higher the importance, (C) Word luminance where the more opacity the more important the word is, and (D) Bar length, where the larger the bar means the word is more relevant for the model prediction.

We seek to evaluate whether any of the three proposed visual encodings is the most useful for physicians to classify medical documents in the context of EBM, along with other analyses such as if one type of visual encoding is preferred on a particular type of article or if users prefer not using any visual encoding at all. We chose to compare these visual encodings shown in Figure 5.2 based on the work done by Felix et al. (2017) where they compared the effect on human performance of using no visual encoding compared to font channels and mark channels on visual tasks related to wordclouds.

Furthermore, the proposed visual encodings shown in Figure 5.2 indicate different levels of accuracy in visual perception for users. On the one hand, comparing the length of 2 bars (Figure 5.2 (D)) is especially precise for human evaluators, while when comparing two hues of colors (Figure 5.2 (B) and (C)), it is easy to know which one is darker, but it is difficult to indicate how much darker one color is than the other.

Another aspect we consider in designing our user study is preventing a learning effect, where users ”learn” an interface (learning effect) by the order in which they are presented, which can bias the results as an undesired effect. In order to have a robust user study that avoids this unwanted effect, we propose a two-phase design as shown in the following diagram:

As shown in Figure 5.3, we counted on users using the Google Chrome extension when using Epistemonikos Web interface, where the study had 2 phases. In the first phase, we assign a visualization encoding (horizontal bar length, background color, or brightness or non-visualization encoding, following a within-subject Latin Square methodology[2], divided into different challenges or groups of documents depending on the type of document. The objective of this methodology is to prevent a fixed order on the sequence of treatments (in this case, the visualizations). For our user study this means not always starting with

[2]https://paasp.net/within-subject-study-designs-latin-square/

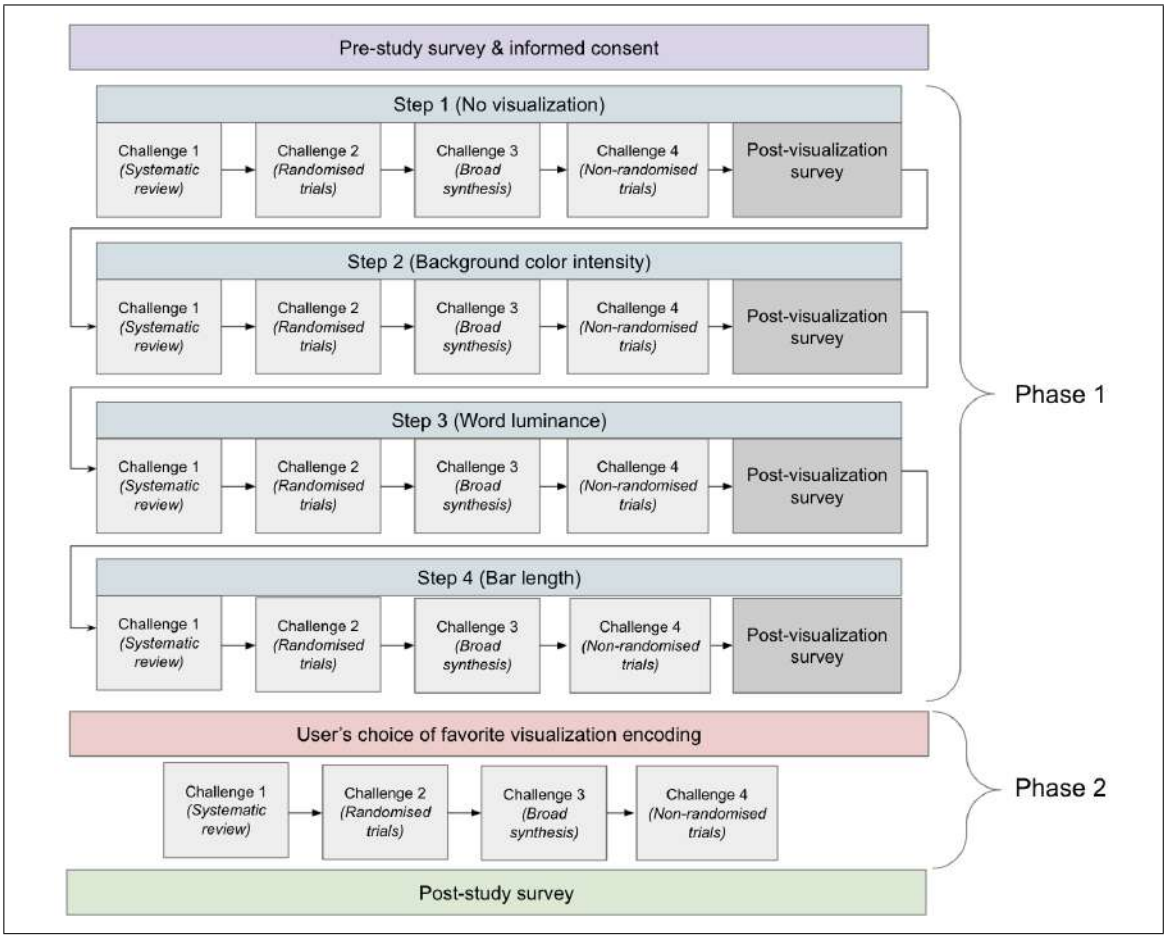


Figure 5.3. User study design. Before starting the study, the user accepts all the terms of the informed consent and a pre-study survey. In the first phase the user goes through each of the visual encodings. Each visual encoding consists of reviewing each document type also referred to as challenges. Subsequently, in the second phase, the user can choose their favorite visual encoding to review articles from each challenge. Finally, a survey is conducted at the end of the study on users' general perceptions.

the no visualization setting and the systematic review document; the order is modified to prevent the aforementioned "learning effect" (Challenge 1 in Figure 5.3).

To evaluate the cognitive load of using visualizations compared to not using them, we make a cognitive load survey at the end of each visualization encoding to assess changes in each visual setting.

Then to verify which is the most useful visual encoding, in a second phase, we give users the option to either turn off the display or choose one of the visual encodings shown above. Next, we do a final cognitive load survey of this last setting where users can select their favorite visualization to finish the user study. Finally, to end the study, we ask the users for a general evaluation of the proposed tool, independent of the type of visualization.

5.3. Experimental configuration

This section describes decisions made to map the output attention of the model to the visual encoding channels.

Figure 5.4 shows the model architecture to output the model's predicted probability along with the attention vector for each of the tokens. In the first stage, abstract words are split, and then transformed into XLNET special tokens that pass through the XLNET transformer encoder to obtain an embedding for each attention head and for each token. Finally, the *[CLS]* token embedding is used as input of a fully-connected layer using a softmax function to predict the type of article, namely Broad Synthesis, Excluded, Randomized Controlled Trial, Non-randomized Controlled Trial, or Systematic Review.

For the user study, we made two relevant decisions concerning the XLNET tokenizer, as in some cases, this tokenizer generates two or more tokens from a single word (for example, the word *Transformer* turns into two tokens *Trans* and *#Former*), we sum over the attention output of each token. Then, concerning the chosen layer to obtain the attention

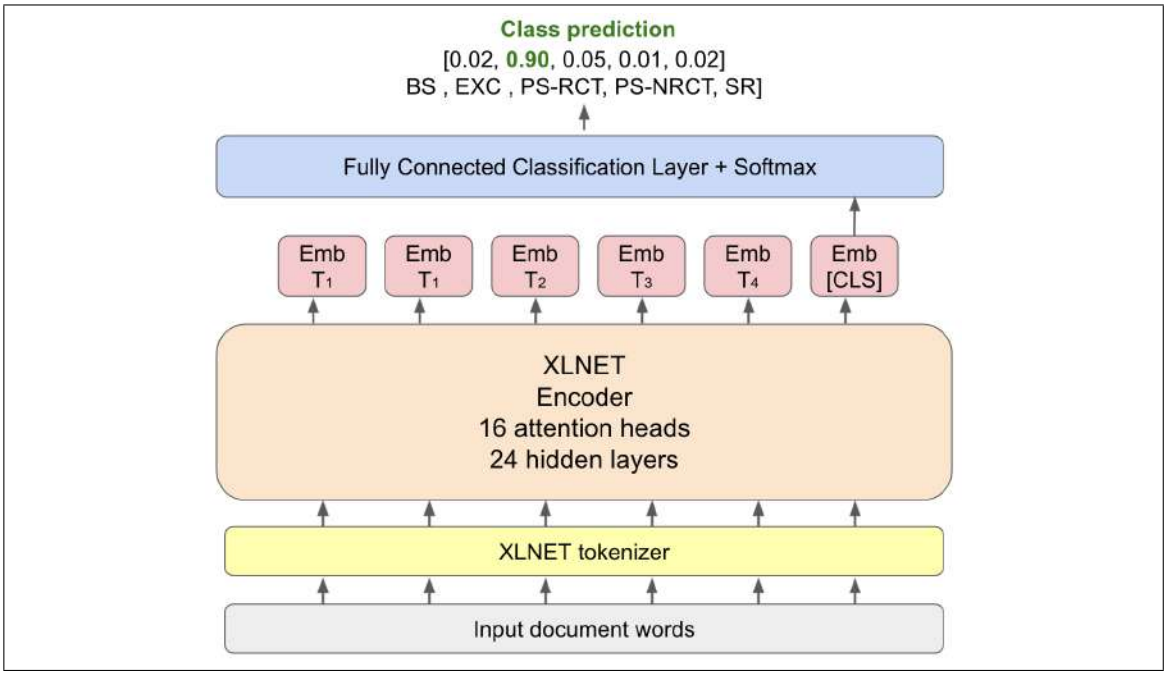


Figure 5.4. XLNet language model and classification architecture. The input of the model is the document's title and abstracts words. Then words are tokenized using the XLNET model tokenizer. After that, they are passed through the XLNET encoder (16 attention heads and 24 hidden layers) and outputs an embedding for each token. Finally, the CLS token embedding is used as input of a fully connected layer using a softmax function to output the class prediction given by the maximum predicted probability. Where BS= Broad Synthesis, EXC= Excluded, PS-RCT: primary study RCT, PS-NRCT= primary study non-RCT and SR= Systematic Review.

output, we chose the last layer. Concerning the aggregation of attention heads, we averaged the attention values of the 16 attention heads. We made these decisions based on the work by Clark et al. (2019) and Htut et al. (2019).

The formulas to pass from attention values to each of the visual encodings are described as follows:

- **Word luminance:** Given the raw attention for each word in a document abstract we applied a square root scaling to the data between 0 and the maximum attention value on the abstract based on the following equation:

$$\hat{a} = \sqrt{\frac{a}{\max(a_{doc})}}$$

where $\hat{a}$ is the new attention value for each word, a is the original attention value, $\max(a_{doc})$ is the maximum attention value obtained on the current document. Then that attention value was mapped using the the D3.js setting to luminance from light-grey (#bdbdbd") to black ("#000000"). It was necessary to apply square root since a linear scale with luminosity caused words with less attention to be not visible.

The D3.js code used for scaling the attentions values to luminance in the text is the following:

```
d3.scaleSqrt().domain([0,
    maxWeight]).range(["#bdbdbd", "#000000"])
```

- **Background color saturation:** Given the raw attention for each word in a document abstract we applied a linear scale to the data between 0 and the maximum attention value on the text based on the following equation:

$$\hat{a} = \frac{a}{\max(a_{doc})}$$

where $\hat{a}$ is the new value of the attention for each word, a is the original attention value, $\max(a_{doc})$ is the maximum attention value of the current document. Then that new attention value ($\hat{a}$) was mapped using the D3.js library setting to saturation from light-white (#f7f7f7) to a type of purple (#8a86be).

The D3.js code used for scaling the attentions values to background color saturation in the text was:

```
d3.scaleLinear().domain([0, d3.max(weights)])
    .range(["#f7f7f7", "#8a86be"])
```

- **Bar length:** Given the raw attention for each word in a document abstract we applied a linear scale to the data between 0 and the maximum attention value on the text based on the following equation:

$$\hat{a} = \frac{a}{\max(a_{doc})}$$

where $\hat{a}$ is the new value of the attention for each word, a is the original attention value, $\max(a_{doc})$ is the maximum attention value of the current document. Then that new attention value ($\hat{a}$) was mapped using D3.js library to a number between 0 and the length in pixels of the largest word from the text.

The D3.js code used for scaling the attentions values to bar length in the text was:

```
lengthBar = d3.scaleLinear().domain([0,
    d3.max(weights)]).range([0, maxLength])
```

5.4. Evaluation

We considered N=5 users for the evaluation, who labeled 200 records each, adding up to 1,000 labeled documents. The number of users is low since they were required to have a minimum level of expertise in evidence-based medicine and be familiar with the Epistemonikos platform. These documents are divided into the first phase, where they reviewed articles for each of the four challenges. Finally, for the second phase, users choose

their favorite visualization to review the four challenges. It should be noted that what we call a *challenge* in Epistemonikos is equivalent to reviewing one type of document: Systematic Review, Broad Synthesis, Randomized trials, and Non-randomized trials. In the same spirit, we seek to evaluate three general dimensions: (1) Utility of explanations and model's predicted probability, (2) Cognitive load, and (3) Preferred visual encoding.

In general, efficiency refers to whether the interface helps improve users' performance on performing a particular task, in this case, screening medical documents in order to label its type of article. It is measured with the time the user takes to classify the document given different visualization encodings.

Cognitive load measures how much effort requires using a given visual encoding or non visualization. This load is measured in four dimensions: mental demand, physical demand, time, performance, effort, and frustration. Furthermore, to compare the cognitive load of the interface among with the different visual encodings, we use the NASA TLX test (Cao et al., 2009) as shown in Table 5.1.

Cognitive effort dimension	Question
Mental demand	How mentally demanding was the task?
Physical demand	How physically demanding was the task?
Temporal demand	How hurried or rushed was the pace of the task?
Performance	How successful were you in performing the task? How satisfied were you with your performance?
Effort	How hard did you have to work to accomplish your level of perfomance?'
Frustration Level	How insecure, discouraged, irritated, stressed, and annoyed were you?

Table 5.1. User study NASA TLX cognitive effort survey for user studies. Users should assign a score from 1 to 100 depending on the level of intensity for each dimension, namely, mental demand, physical demand, temporal demand, performance, effort, and frustration level.

As seen in Table 5.1 to evaluate the cognitive load for each dimension, we ask a set of questions for each measurement at the end of each visual encoding setting. The user has to assign a score between 1 and 100 depending on how much effort is required.

Another aspect that we seek to evaluate is the perception of the helpfulness of explanations shown as visual encodings and the information on models' predicted probability.

We assess this in two ways: when reviewing each document, we ask the user if the visualization was perceived as helpful for the document screening and if the information on models' predicted probability assisted them in making a correct decision. Moreover, in a second phase, we want to evaluate if there is a preferred visual encoding when the user must choose their favorite encoding to finish the study. The objective of the second phase is to give additional evidence of the preferred visual encoding in terms of perception of helpfulness, if any.

Finally, to evaluate the general perception of the user about the interfaces presented in the user study, we carried out a post-study survey. The user post-study survey from the end of the study is described in Table 5.2. The purpose of this post-study survey is for users to evaluate the overall interface independent of the type of visual encodings.

5.5. User study results

In this section we analyse the results obtained from the user study in order to answer the last two research questions addressed in this thesis. We evaluate the proposed platform in three major aspects: perceived helpfulness of visualizations and model's predicted probability, preferred visual encoding and cognitive load.

N	Question
1	I understood why documents were automatically classified as a particular category.
2	The suggested classifications seemed accurate, given the content of the document.
3	I quickly felt familiar with the interface.
4	I felt the system was easy to use.
5	I didn't realize how time passed while using the classification interface.
6	The system made me think that it helped me understand automatic decisions for document classification.
7	I would use the system again to classify documents
8	I would recommend the system to a colleague.
9	I think the system requires other kinds of visualization to understand automated decisions

Table 5.2. User post-study survey. This survey consists of a multiple selection on strongly disagree, disagree, neutral, agree or strongly agree.

5.5.1. Visual explanations preception of helpfulness

The perception of helpfulness of attention outputs as visual explanations is obtained from answers given by users after classifying each document, where they were asked to assign a rating on a scale of 1 to 5 how much they agree that the visualizations were helpful to make a decision.

Figure 5.5 shows that 37% of users have a neutral perception (score 3) about helpfulness of visualizations. Moreover, in 51% of cases users disagree and partially disagree (score 1 and 2) with visual explanations. Finally, users who found that the visualization was helpful or very useful (score 4 and 5) in only 12% of the cases.

In order to better understand the poor perception of users concerning visualizing weights on attended words, we conducted an analysis by type of document and type of visualization.

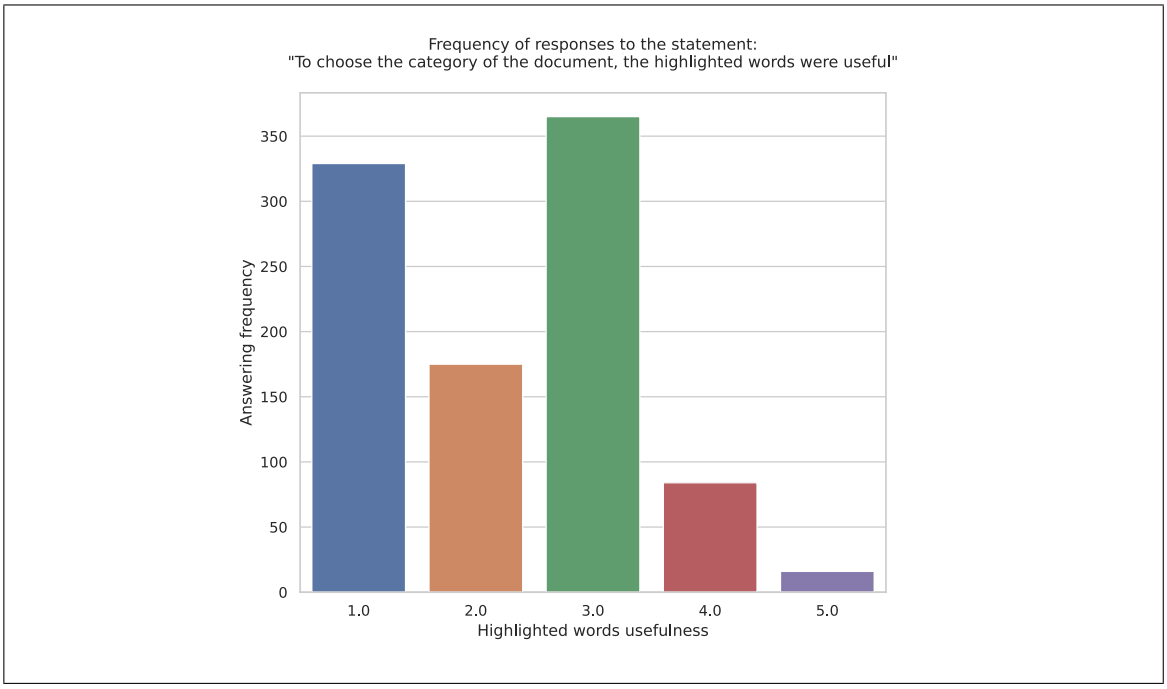


Figure 5.5. Visualizations' helpfulness. The x-axis indicates the user's agreement with the statement above the plot, where 1 completely disagrees, and that of 5 totally agree. The y-axis shows the frequency of answers for each rating provided on highlighted words utility.

Figure 5.6 shows the results of visual explanations as highlighted words for each type of document and visual encoding combination. It can be seen that, in general, regardless of the type of document and visual encoding, there is a prevalence of the lowest scores to measure the helpfulness of visual explanations. However, most users gave a neutral score to background color on systematic reviews. Furthermore, in the case of background color for broad synthesis, it can be observed that the perceived helpfulness of this visual encoding has a neutral score as the second most popular after the lowest score. In the case of the bar length encoding, we observe a more skewed distribution towards the lowest score regardless of the type of document being analyzed. Although we see that the lowest score

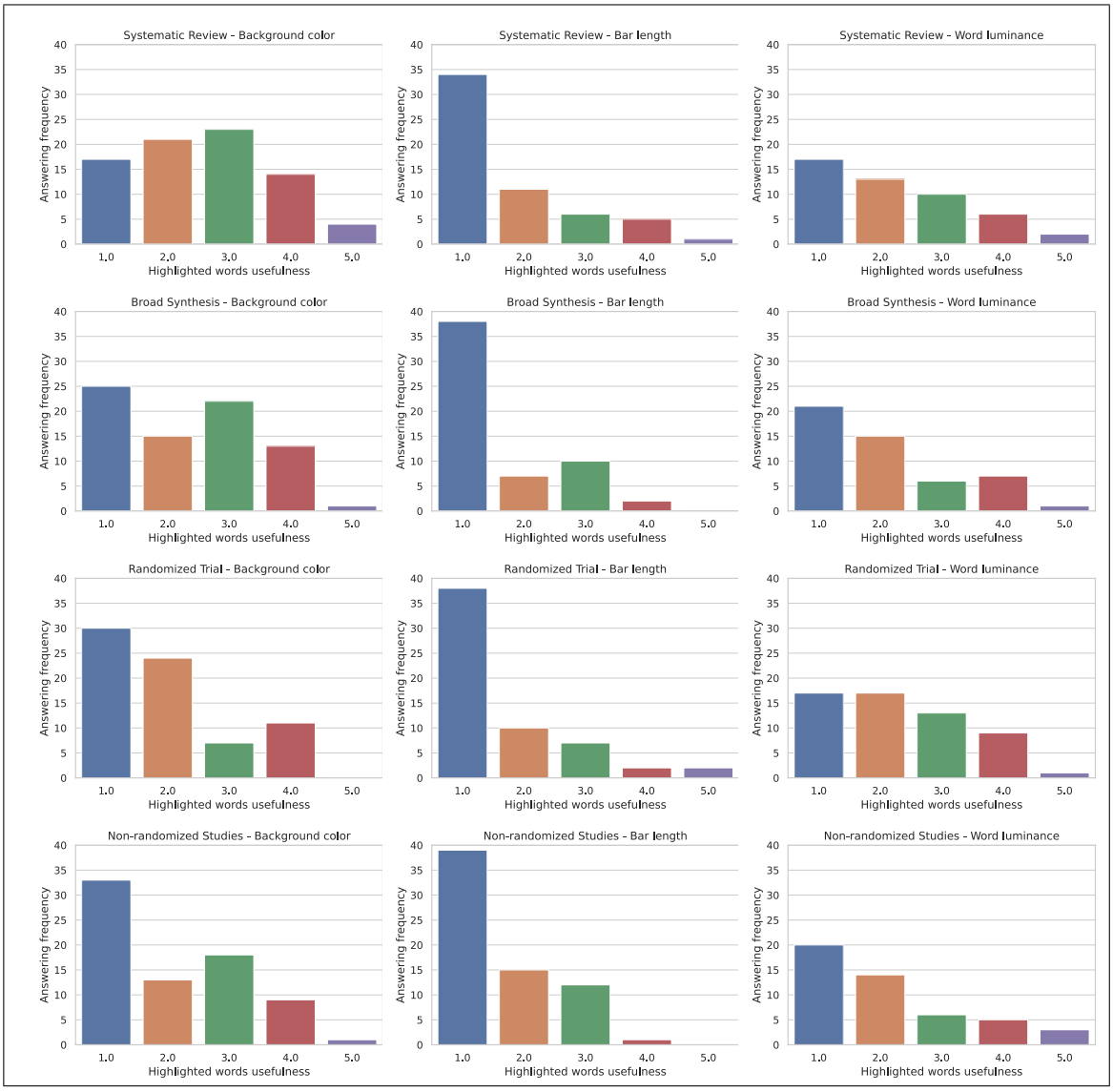


Figure 5.6. Highlighted word visualization helpfulness evaluation for each visual encoding and type of document. The x-axis indicates the user's agreement with the given statement, where 1 completely disagrees, and that of 5 totally agree. The y-axis shows the frequency of answers for each rating provided on highlighted words utility.

prevails in the case of word luminance, the utility valuation scores are more equally distributed than the bar length visual encoding. These results indicate a potential interaction effect which needs further analysis, which will be addressed in an upcoming section.

5.5.2. Perception of helpfulness of model predicted probability

In addition to showing a visualization of highlighted words to which the model placed the most attention, we also show the probability predicted by the model. For this last one, we also want to measure if it is helpful for users as an explanation. To measure this, we ask the user to score (1-5) their perception of the usefulness of the model's predicted probability after reviewing each article. Results are shown in Figure 5.7.

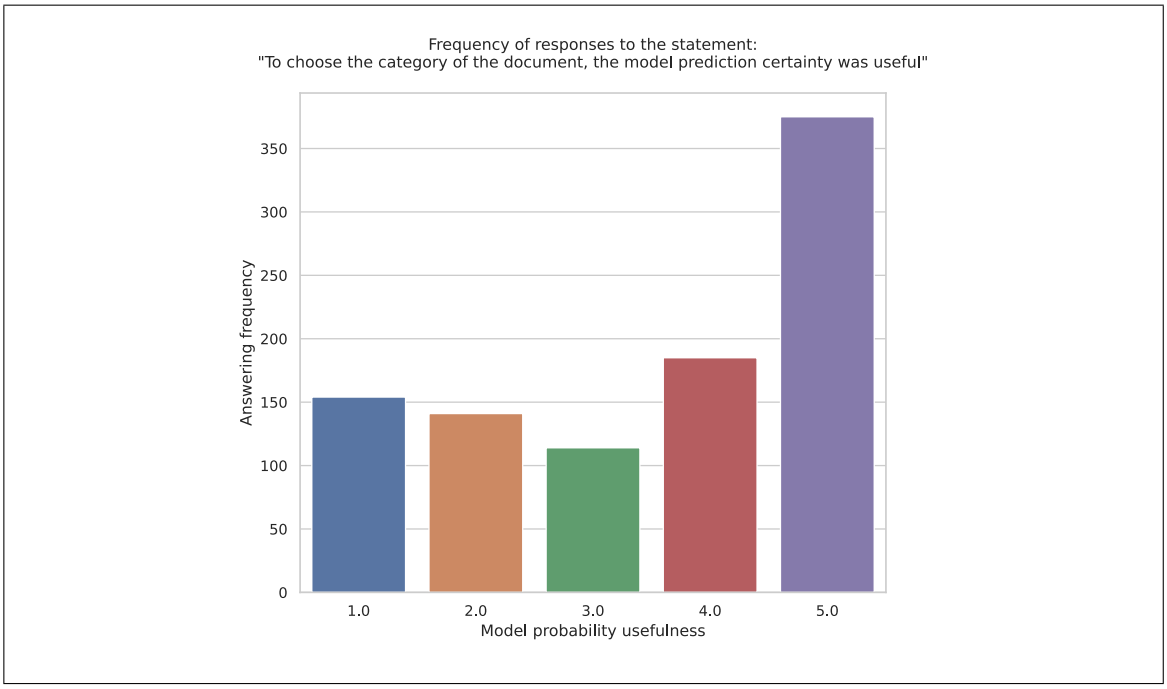


Figure 5.7. Model probability helpfulness evaluation. The x-axis indicates the user's agreement with the given statement, where 1 completely disagrees, and 5 that corresponds to total agreement. The y-axis shows the frequency of answers for each rating provided on model's prediction probability helpfulness.

Figure 5.7 shows that users in most of the cases, corresponding to 572 (58%), found that the model's predicted probability helps them decide to classify a document in a category, since the majority of scores fluctuate between 4 and 5. However in 120 cases (11%),

users were indifferent about the helpfulness of the model's predicted probability to make a decision, with a score of 3. Finally, 302 cases (31%) thought that the model's predicted probability was not valuable or non-useful for making a decision, giving a score of 1 or 2. It can be seen that the distribution of scores on the perception of helpfulness of models predicted probability differs from perception of highlighted words' helpfulness.

Figure 5.8 shows the results of visual explanations as highlighted words for each type of document and visual encoding combination. In general, regardless of the visual encoding and the type of document analyzed, the majority of scores are 4 or 5, indicating a good evaluation on perception of the helpfulness for users of the model's predicted probability information.

5.5.3. Two-Way ANOVA results

A two-way ANOVA was performed to analyze the interaction effect between the type of article and visual encoding on the perceived helpfulnes of model's predicted probability (Figure 5.9). Then we performed another two-way ANOVA to analyze the interaction effect between the type of article and visual encoding on the perceived helpfulnes of highlighted words (Figure 5.10). Results of both two-way ANOVA are shown in Table 5.3 , Table 5.4 (Two-way ANOVA model's predicted probability) and Table 5.5 (Two-way ANOVA highlighted words).

As shown in Figure 5.9 and in Table 5.3, we see that the two-way ANOVA revealed that there is not an interaction effect between visual encoding and type of article on the perception of helpfulness of models predicted probability, since Figure 5.9 and Table 5.3 shows that there is overlap in all cases.

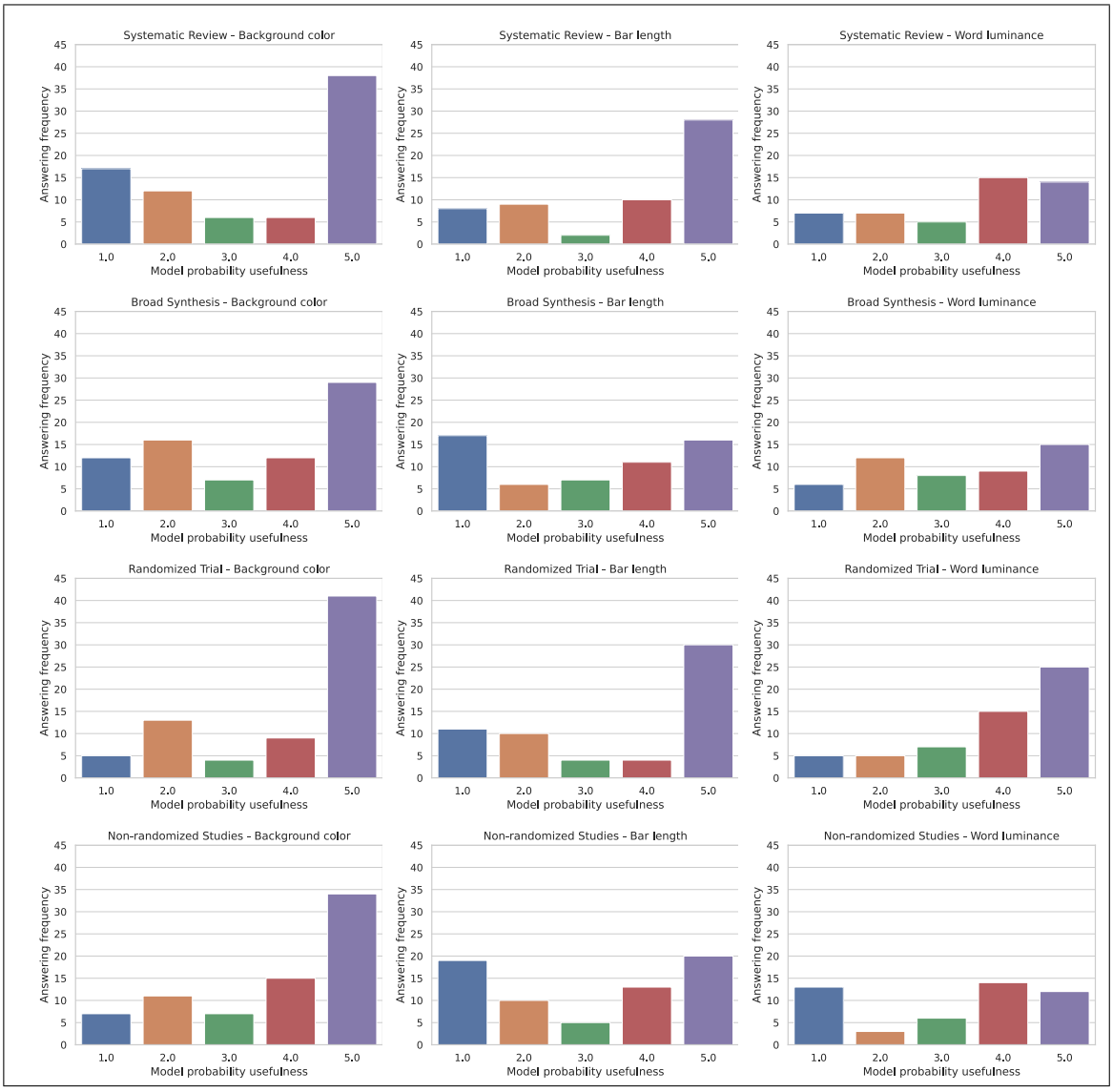


Figure 5.8. Model probability helpfulness evaluation for each visual encoding and type of document. The x-axis indicates the user's agreement with the given statement, where 1 completely disagrees, and that of 5 totally agree. The y-axis shows the frequency of answers for each rating provided on highlighted words utility.

As seen in Table 5.4 the two-way ANOVA revealed that there was not a statistically significant interaction between the effects of type of article and visual encoding (F = 1.36,

		Model's Probability		Highlighted words	
Art Type	Vis	M	SD	M	SD
SR	Background	3.45	1.68	**2.59**	**1.15**
	Bar	3.70	1.55	1.75	1.10
	Luminance	3.32	1.43	2.18	1.13
BS	Background	3.33	1.55	**2.31**	**1.14**
	Bar	3.03	1.64	1.57	.89
	Luminance	3.23	1.41	2.12	1.15
PS-RCT	Background	3.92	1.39	2.00	1.05
	Bar	3.52	1.68	1.62	1.04
	Luminance	3.85	1.30	**2.34**	**1.14**
PS-NRCT	Background	3.77	1.40	**2.09**	**1.14**
	Bar	3.10	1.63	1.63	.82
	Luminance	3.31	1.48	2.03	1.22

Table 5.3. Two-way ANOVA results to analyse interaction effect of type of article and visual encoding on perceived helpfulness of model's predicted probability and on perceived helpfulness of highlighted words. Where M is the mean and SD is standard deviation. Results in **bold** indicate the highest and significant differences for each type of document.

Variable	SS	df	MS	F	p-value
Visual encoding	11.50	2	5.75	2.49	.083
Type of article	33.78	3	11.26	4.87	.002
Interaction	18.90	6	3.15	1.36	.225

Table 5.4. Results of the two-way ANOVA to determine if the type of visual encoding and the type of article effect the perceived helpfulness of model's predicted probability. Where SS = sum of squares, MS = mean squares , df = degress of freedom.

p = .225) on the perception of helpfulness of model's predicted probability. Concerning simple main effects analysis showed that the type of visual encoding did not have a statistically significant effect on the perception of helpfulness of model's predicted probability (p = .083). Furthermore, the type of article did have a statistically significant effect the perception of helpfulness of model's predicted probability (p = .002).

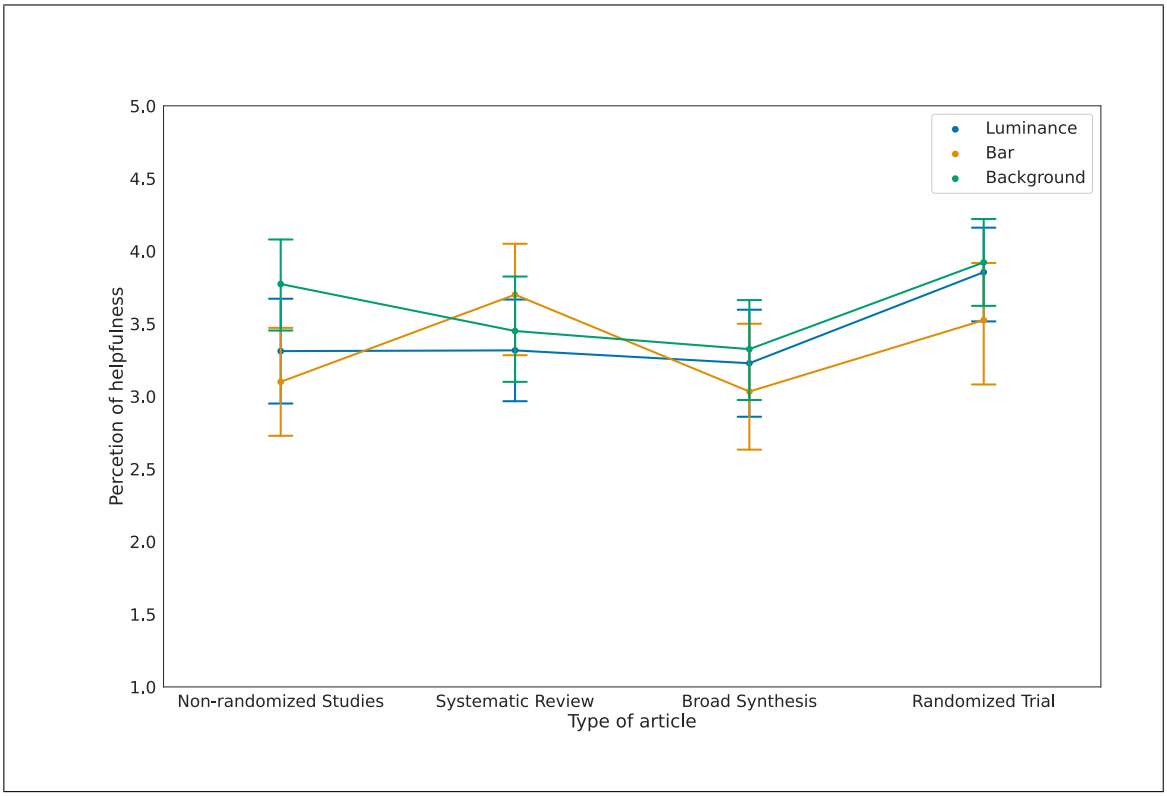


Figure 5.9. Two-way ANOVA to analyse the interaction effect between the type of article and the visual encoding on the perceived helpfulness of model's predicted probability.

The two-way ANOVA to analyze the interaction effect between the type of article and the visual encoding on the perceived helpfulness of highlighted words is shown in Figure 5.10 and Table 5.3. This two-way ANOVA reveals an interaction effect between the type of visual encoding and the type of article on the perceived helpfulness of highlighted words. Concerning SR articles, there are differences only between Background (M=2.58, SD=1.15) and Bar (M=1.75 , SD=1.09). For BS, there are differences between Luminance (M=2.12,SD=1.15) and Bar (M=1.57, SD=0.89) and between Background (M=2.32 , SD=1.14) and Bar (M=1.57, SD=0.89). Regarding RCT, there is a difference between

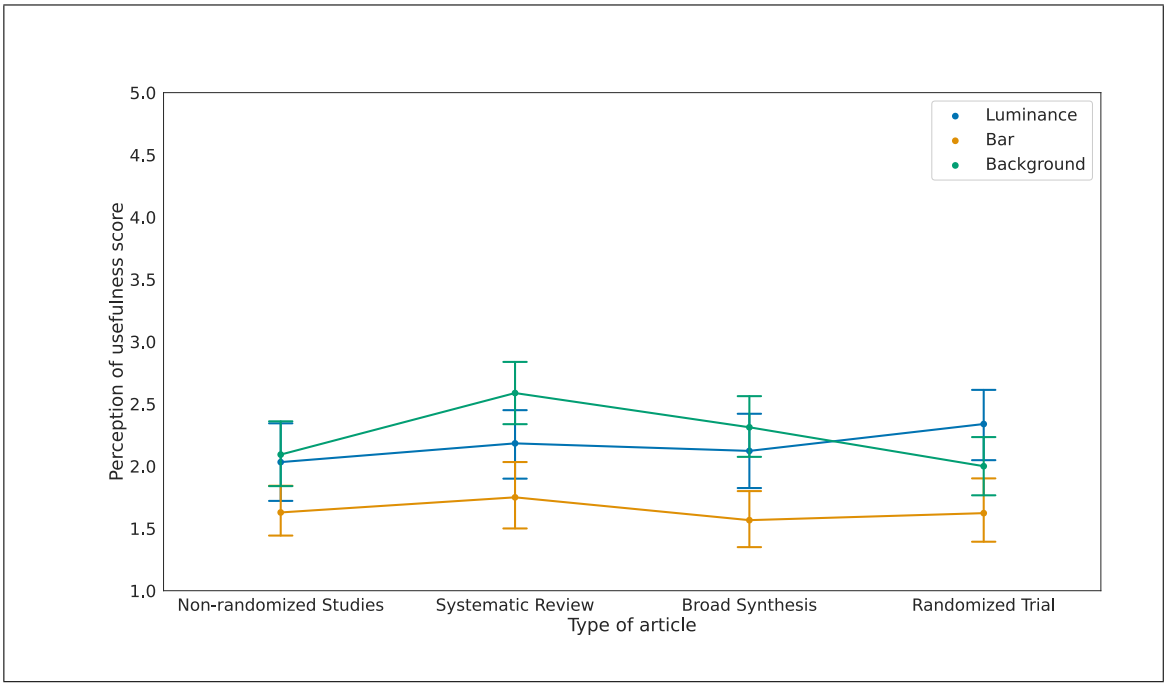


Figure 5.10. Two-way ANOVA to analyse the interaction effect between the type of article and the visual encoding on the perceived helpfulness of highlighted words.

Luminance (M=2.34, SD=1.14) and Bar (M=1.62 , SD=1.04). Finally, regarding non-RCT articles, there are differences between the perceived helpfulness of highlighted words for Luminance (M=2.03 , SD= 1.22) , Bar (M=1.63, SD=.82) and Background (M=2.09, SD=1.14) visual encodings.

Variable	SS	df	MS	F	p-value
Visual encoding	58.15	2	29.07	24.56	0
Type of article	8.54	3	2.84	2.40	.006
Interaction	11.65	6	1.94	1.64	.01

Table 5.5. Results of the two-way ANOVA to determine if the type of visual encoding and the type of article effect the perceived helpfulness of highlighted words. Where SS = sum of squares, MS = mean squares , df = degress of freedom.

As seen in Table 5.5 the two-way ANOVA revealed that there was a statistically significant interaction between the effects of type of article and visual encoding (F = 1.64, p = .01) on the perceived perception of helpfulness of highlighted words. Concerning simple main effects analysis showed that the type of visual encoding did have a statistically significant effect on the perception of helpfulness of highlighted words ($p < .000$). Furthermore, the type of article did have a statistically significant effect the perception of helpfulness of highlighted words (p = .01).

5.5.4. Bootstrap sampling confidence intervals

Despite potential redundancy the results of ANOVA, we conducted an additional analysis of confidence intervals with bootstrap sampling for additional robustness in our conclusions (Dragicevic, 2016).

In Table 5.6 we show the mean score of perceived helpfulness of models predicted probability and highlighted words, along with their corresponding lower (LCB) and upper (UCB) confidence bounds obtained using bootstrap sampling for each visual encoding and type of article.

It can be seen that, in general, the perception of the helpfulness of the model's probability is higher than on highlighted words, among all types of articles and visual encodings,

		Model's Probability			Highlighted words		
Art Type	Vis	LCB	Mean	UCB	LCB	Mean	UCB
SR	Background	3.06	3.46	3.81	**2.32**	**2.58**	**2.82**
	Bar	3.26	3.72	4.07	1.47	1.74	2.02
	Luminance	2.95	3.34	3.67	1.90	2.19	2.48
BS	Background	**3.04**	**3.39**	**3.71**	**2.08**	**2.34**	**2.59**
	Bar	2.61	3.05	3.46	1.35	1.58	1.82
	Luminance	2.84	3.23	3.57	1.82	2.11	2.41
PS-RCT	Background	**3.58**	**3.94**	**4.24**	1.75	1.99	2.24
	Bar	3.08	3.54	3.93	1.41	1.64	1.93
	Luminance	3.49	3.88	4.18	**2.00**	**2.30**	**2.58**
PS-NRCT	Background	**3.43**	**3.78**	**4.07**	**1.82**	**2.08**	**2.34**
	Bar	2.66	3.07	3.43	1.43	1.63	1.84
	Luminance	2.82	3.25	3.60	1.74	2.04	2.37

Table 5.6. Confidence Interval mean (M), lower confidence bound (LCB) and upper confidence bound (UCB) for helpfulness evaluation of models probability and highlighted words using Bootstrap Sampling. Results in **bold** indicate the highest and significant differences for each type of document.

with a maximum mean score (M=3.94) an $UCB = 4.24$ and a $LCB = 3.58$ in the case of PS-RCT using Background visual encoding. However, as seen in the two-way ANOVA there are no significant differences in scores when changing visual encodings and types of articles.

In the case of highlighted words, the highest mean score (M=2.58), with an $UCB = 2.82$ and $LCB = 2.32$ for Systematic Reviews using a Background visual encoding. Regarding the perceived helpfulness of highlighted words, there are differences in scores depending on the type of article and visual encoding. In the case of Randomized Trials, Luminance (M=2.30, CI=(2.00,2.58)) is significantly higher than Bar (M=1.64, CI=(1.41,1.93)). Moreover, for Non-randomized Trials, Background (M=2.08, CI=(1.82,2.34)) is significantly higher than Bar (M=1.63, CI=(1.43,1.84)). For Broad Synthesis there are significant differences between Background (M = 3.39, CI=(3.04, 3.71)) with other visual encodings.

However, in the case of Systematic Reviews, which are articles that summarize evidence, there are no differences in scores when changing the visual encoding.

5.5.4.1. Word attention analysis

In order to gain a deeper understanding of what made people report such low perception of helpfulness of the words highlighted in the abstract, we analyzed the words with higher attention weights when helpfulness was perceived as low (1-2) compared to when it was perceived as high (4-5). Results are shown in Table 5.7.

	helpfulness $>$ 3 , N = 109					helpfulness $<=$ 3 , N = 891				
Doc Type	w1	w2	w3	w4	w5	w1	w2	w3	w4	w5
SR	meta-analyses	meta-analysis	literature	systematic	pubmed	on	all	the	of	a
BS	cadth	methotrexate	gynaecological	literature	search	the	and	of	covid	this
PS-RCT	buccal	pembrolizumab	cochrane	randomized	methylprednisolone	may	was	of	the	is
PS-NRCT	myocardial	resucitative	bmi	panniculitis	thrombo	this	to	all	were	1

Table 5.7. Words with highest attention weights for each type of document when the perceived helpfulness for highlighted words was larger than 3 and when it was smaller than 3.

As Table 5.7 right side indicates, the words more frequently highlighted when scores were low (helpfulness perception less than 3) were stopwords that do not help to discriminate, such as “on, all, the”, etc. On the left side of Table 5.7, we observe very informative words when the helpfulness scores were larger than three. What made the transformer model XLNET pay attention to such uninformative words works such as Wiegreffe & Pinter (2019) and Jain & Wallace (2019) indicate that transformer models can pay attention to words unrelated to the task, revealing the lack of solidity of arguments indicating that human attention would be similar to neural networks attention. Some authors aim at solving this issue, such as the work done by Kobayashi et al. (2020) that studied the words to

which the model gives the highest weight of attention and if they have a resemblance to the functioning of human language, proposing ways to adapt these weights using the norm of the weight vector so that it is as similar as possible, which will be discussed in future work.

5.5.5. Preferred visual encoding

For evaluating the utility of visual encoding alternatives compared to the control group, we evaluated the user behavior on two perspectives: (1) most chosen visual encoding for the second phase of the study, and (2) time required to make a classification decision using each visual encoding.

Concerning results on the most chosen visual encoding for the second phase of document screening are shown on the following figure:

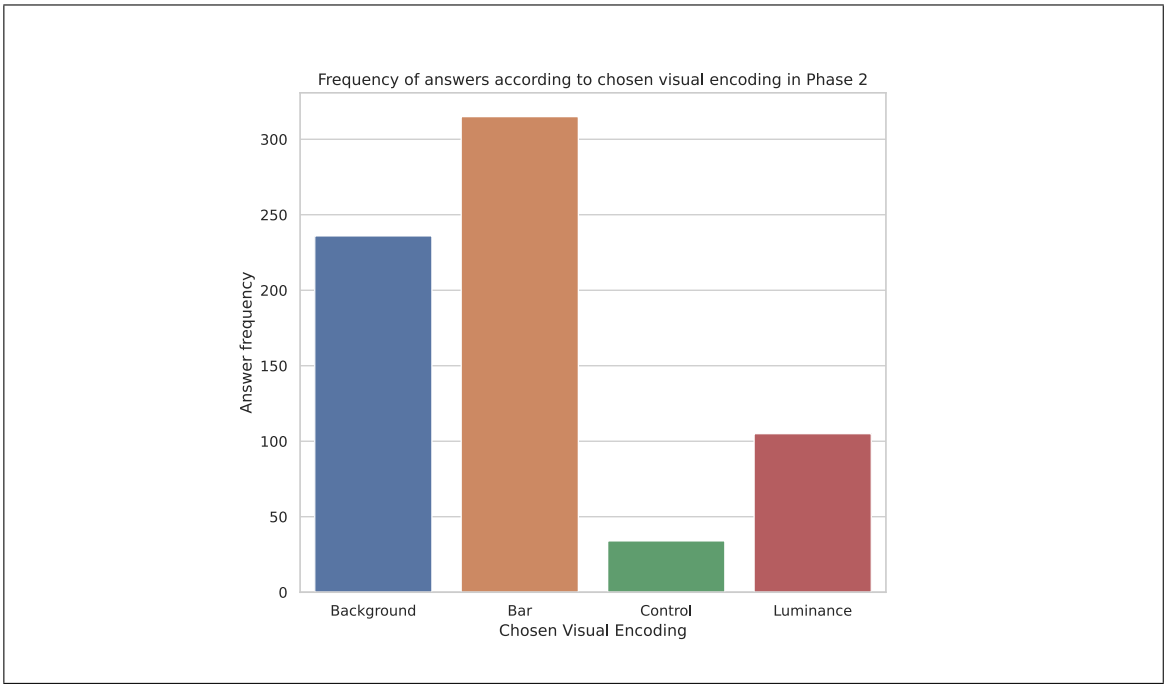


Figure 5.11. Most chosen visual encoding in user study phase 2.

In Figure 5.11, the user generally prefers to use any of the three visual encodings: bar, background, or word luminance, with a total of 95.06% of the documents screened, surpassing the control group (without visualization) that was preferred in only 4.92% of the cases. Among them, the most used visual encoding in the second phase is the bar length chosen in 45.65% of the cases, followed by the background color visual encoding preferred in 34.20% of the cases. The third most selected visualization is luminance (15.21% usage). Finally, the control group without visualization achieved only a 4.92% usage. This result is counterintuitive if we compare them with the first phase results, where users did not have a high perception of helpfulness on visualizations. A possible reason for this behavior is that using some visualization in the text facilitates reading, regardless of how unhelpful the attention was. Further analysis will be presented when analyzing the users' feedback.

5.5.6. Time required for each visual encoding

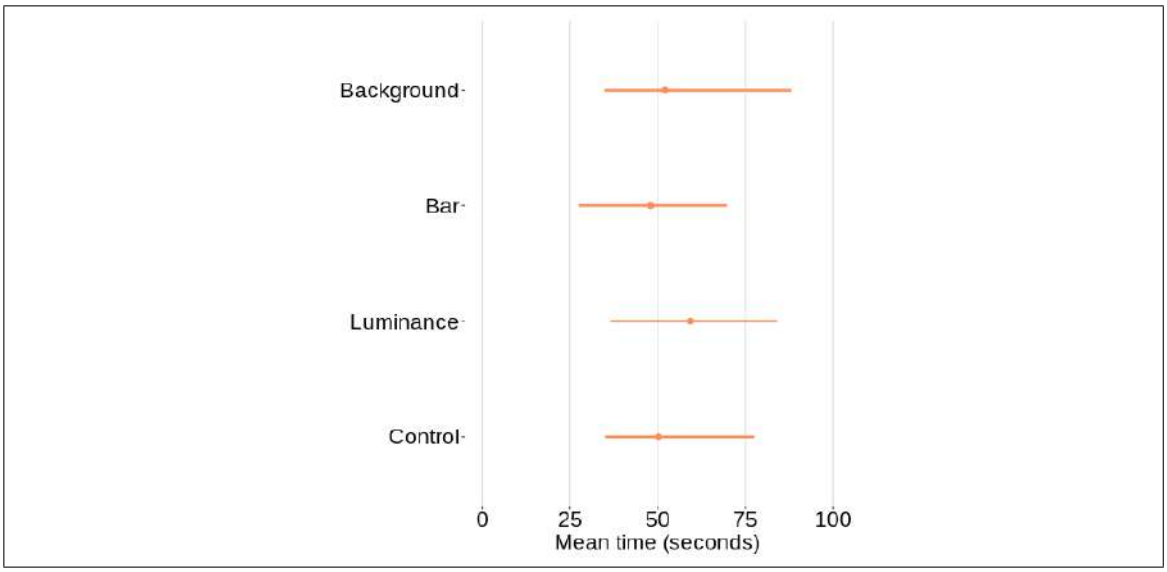


Figure 5.12. Bootstrapped confidence intervals of average time taken for each visual encoding. The X-axis indicates the average taken for each visual encoding. Y-axis shows each visual encoding: background color, bar length, word luminance, and the control group without visualization.

Figure 5.12 shows the mean time taken for each visual encoding with their corresponding upper (UCB) and lower (LCB) bootstrapped confidence bounds. It can be seen that there are no significant differences in the time required for each visual encoding since there is an overlap between all the time confidence intervals.

5.5.7. Cognitive load

In this section, we analyze the responses from users to the NASA TLX survey related to the cognitive load of doing a task. For this, we compare the mean score given to each of the cognitive dimensions obtained from answers on each visual encoding with their

corresponding standard deviation. This section aims to validate a reduction of the cognitive load by including our proposed interface. Results are shown in the following table:

Visual encoding	Mental	Physical	Temporal	Performance	Effort	Frustration
No visualization	46.1 (25.15)	25.3 (11.08)	44.2 (23.54)	**61.6** (16.87)	49.70 (27.32)	**27.8** (16.44)
Background color	**37.2** (26.81)	**24.4** (19.74)	**36.5** (26.12)	55.3 (28.59)	**42.4** (27.73)	30.3 (25.05)
Word luminance	49.1 (30.54)	35.1 (26.13)	49.4 (27.51)	50.6 (25.57)	54.3 (30.71)	43.3 (27.98)
Bar length	48.5 (24.28)	35.4 (22.82)	52.5 (23.58)	59.1 (15.58)	56.5 (24.95)	49.4 (21.66)

Table 5.8. NASA TLX cognitive load mean score and the standard deviation given to each evaluated dimension: Mental demand, Physical demand, Temporal demand, Performance, Effort, and Frustration level. Results in bold mean that it obtained the best score for that given dimension. In the case of Mental, Physical, Temporal, Effort, and Frustration dimensions, the lower the score, the better. For the Performance dimension, a higher score is better.

Results from the NASA TLX shown on Table 5.8 indicates that for mental demand dimension, the background color is the visual encoding with the lowest cognitive load with (M=37.2 and SD=26.81). Regarding the physical effort, the background color has the lowest cognitive load with a (M=24.4 and SD=19.74). Then concerning the temporal effort, the background color have the lowest load, with a (M=36.5 and SD=26.12). Furthermore, in terms of user's perception of performance, the control group has the highest score, which in this case it is better since performance is improved due to the given visualization, with (M=61.6 and SD=16.87). Concerning effort dimension, the background color reaches the lowest score with (M=42.4 and SD=27.73). Finally, the control group is the one that has the lowest levels of frustration with (M=27.8 and SD=16.44).

It can be inferred that having less mental, physical, temporal, and effort load when using background color by facilitating the reading of the article reduces the effort in the four dimensions mentioned above. However, the user's perception of performance and frustration dimensions are better without visualization because it is the interface users have used within their daily work.

5.5.8. Post-study survey

In this section, we detail the results of the post-study survey to know the general feeling of the users regarding the study. The results are shown in the following figure:

N	Question	SD	D	N	A	SA
1	I understood why documents were automatically classified as a particular category.	0	0	4	0	1
2	The suggested classifications seemed accurate, given the content of the document.	0	1	2	2	0
3	I quickly felt familiar with the interface.	0	1	0	1	3
4	I felt the system was easy to use.	0	0	0	2	3
5	I didn't realize how quickly time passed while using the classification interface.	0	2	2	0	1
6	The system made me think that it helped me understand automatic decisions for document classification.	0	1	2	1	1
7	I would use the system again to classify documents	0	0	2	2	1
8	I would recommend the system to a colleague.	0	2	2	0	1
9	I think the system requires other kinds of visualization to understand automated decisions	0	1	2	1	1

Table 5.9. User post-study survey. The results of each question for each multiple selection on strongly disagree (SD), disagree (D), neutral (N), agree (A) or strongly agree (SA) are shown on each column. Empty answers are not considered in this analysis.

Results shown in Table 5.9 from the post-study survey give evidence that firstly, half of the users (50%) who used the platform agree it is helpful to understand the reasons that led the model to make a prediction, corresponding to questions 1 and 6. Followed by a 33.3% neutral, and 16% of the users found the platforms were not valuable. Evaluating the perception of helpfulness of the model's predicted probability, only two users agreed with the predictions, and the other three disagreed or were indifferent. Regarding the ease of use and friendliness of the platform, corresponding to questions 3-5, we obtained that more than 66.6% found the platform to be friendly and easy to use. Then, if we see that users would be willing to use the platform again or recommend it to a colleague, we obtain that

20% would agree, 40% were neutral, and 40% of users would not recommend it. Finally, regarding whether other ways to visualize the explanations improves their performance, 40% agreed, 40% were indifferent, and the rest 20% disagreed.

5.5.9. User's feedback qualitative analysis

This section describes the results obtained by analyzing the feedback received from users who participated in the study. For the qualitative analysis, we recruited all the users once the study was finished and individually interviewed each one of them,

We followed a protocol that consisted first of a presentation of our research project and objectives. We then gave a general description of users' decisions during the study concerning the perception of helpfulness of highlighted words, perception of helpfulness of models predicted probability and their preferred visual encoding on the second phase. Finally, we received feedback and observations from the users about the study.

User opinions of helpfulness of the model's predicted probability, perceived helpfulness of highlighted words, chosen visual encoding in the second phase and user study platform design is summarized as follows:

- **Model predicted probability:** In general, users prefer to have the model's predicted probability. They considered that the level of precision of the model was high, and it helped them confirm their instinct to make a better decision. In addition, the fact of including the predicted probability helped them as a guide to decide when an article was not of a particular category.
- **Perception of helpfulness of highlighted words:** highlighted words were generally not essential for decision making since in most cases, the most relevant

words did not directly correlate with the category predicted by the model. However, when the model highlighted the correct words, it provided robust feedback for the document screening task and helped users boost reading and keyword extraction.

- **Chosen visual encoding:** According to users' feedback, none of them chose the option without visualization in the second phase of the study and chose the other visual encodings, which indicates that users prefer to have highlighted words compared to having nothing to a certain extent. This behavior occurs since the visualizations help users to locate themselves for reading the abstract, even though, in general, the words emphasized with attention are not the most appropriate to classify the reviewed document.
- **Platform Design:** Most users gave superior feedback concerning how intuitive and interactive the platform is for document screening. Furthermore, concerning words highlighted in the abstract facilitated their read, despite highlighted words were not directly correlated to the model's prediction.

In Table 5.10, we show the users' comments on the four points described above: model's predicted probability, visual explanations, chosen visual encoding, and platform design.

Aspect	Comments
• Highlighted words explanations	• *I don't like to have highlighted words. I prefer not having anything.* • *The marked words do not have to do with more important topics.* • *It should have an outstanding level of precision to make a decision, but it was not the case.* • *Although there were words marked without meaning, it was easier for me as a guide for reading the text.* • *The marked words helped me to differentiate words that are not useful to make a decision.*
Model's predicted probability information	• *The probability predicted by the model helped me make a better decision.* • *The probability predicted by the model helped me to have a confirmation of what I was thinking.* • *The probability predicted by the model was helpful as a guide to know when the article was not of a particular category.* • *The probability predicted by the model was pretty accurate. In general, 70 to 80% of the time, it was correct.* • *The probability predicted by the model helped me make decisions, prefer to have it.*
Chosen Visual Encoding	• *In the second phase of the study, I chose bar and background color to be able to compare them.* • *The background color visual encoding was the most intuitive and made it easy for me to read.* • *The luminance display sometimes made it difficult for me to read.* • *The background visualization ends up helping me because sometimes the words I marked made sense to me.* • *The background visualization helped me focus on the text of the article.*
Platform and Study Design	• *In general, the study and the tutorial were easy to follow.* • *The platform was well done. I did not run into any bugs or have to restart the studio because of it.* • *I would recommend this platform, especially the feature of model prediction information, to get a more informed decision.* • *It was easy to follow the flow of the study, especially when given the option to choose a visualization.* • *Although the accuracy of the marked words needs to be improved, it is a good tool for screening medical documents.*

Table 5.10. Users comments on proposed study on four aspects: highlighted words explanations, model predicted probability information, chosen visual encoding, and platform and study design.

It can be seen from Table 5.10 that highlighted words in most of the cases are not helpful in making a decision; however, highlighted words helped users in reading the article. Moreover, as shown by the quantitative analysis from the previous section, the preferred visual encodings were bar and background color since they do not make reading difficult, compared to word-luminance visual encoding where attention values that are too low imply that such words might not be shown. Concerning the model's prediction, probability gave physicians a second opinion to support the final classification of the article. Finally, there are positive comments on the platform and study design since there were no problems with the study instructions and the Google chrome extension.

5.6. Expert evaluation

Given that in all the previous sections, we have evaluated the perceived helpfulness of using visualizations for different types of articles and visual encodings, in this section, we verify with expert labels whether the performance of users improves when using the visual interface for the document screening task.

Visual encoding	Coincidence	Not Coincidence	Accuracy (%)
Background	30	16	65.22%
Luminance	21	9	**70.00%**
Bar	25	13	65.79%
All visual encodings	76	38	**66.67%**
No visualization	19	11	63.33%

Table 5.11. Expert validation. This table shows the coincidence of correctly labeled articles by two users considering the labels of a third expert physician as ground truth. In **bold** we show (1) the visual encoding with higher accuracy and (2) the higher average accuracy comparing the average accuracy of all the visual encodings and not using visualizations.

Table 5.11 shows the expert coincidence of labels on 144 articles reviewed by two physicians. The results give evidence that using Luminance visual encoding achieves the higher accuracy (70%), followed by Bar (65.79%) and Background color (65.22%), where in all of the cases the performance improves compared to the control group without visualization (63.33%). Furthermore, the coincidence of all the visual encodings achieves higher accuracy (66.67%) compared to no visualization (63.33%). These results indicate that users' performance slightly improves when using any of the visualizations.

5.7. Discussion

In this chapter, we evaluated through a user study on an EBM platform for biomedical document screening if including explanations as highlighted relevant words is helpful for users. We also verify if incorporating the model's predicted probability for document classification is beneficial for decision making. Additionally, we compared whether showing the explanations in different ways or visual encodings make a difference in reducing time, cognitive load, and preference. Furthermore, results indicate that there is an interaction effect between the type of article and visual encoding on the perception of helpfulness of highlighted words as an explanation. Finally, we validated through expert labels as ground truth if visualizations improved users performance for the document screening task.

We summarize the main take-aways and some challenging aspects of our research on each of the points mentioned earlier:

Perception of helpfulness of highlighted words: It would have been desirable to have a good perception of the helpfulness of highlighted words as a means of explanation. However, in general, we saw that regardless of the visual encoding, scores lower than three were obtained in most cases, reaching 891 cases (89.63%) compared to only 103 cases (10.36%) where highlighted words have an evaluation score of 4 or 5. These results provide evidence through an empirical evaluation that in most of the cases, attentions are not a suitable means for an explanation (Wiegreffe & Pinter, 2019; Jain & Wallace, 2019; Tutek & Šnajder, 2020; Patro et al., 2020; Bastings & Filippova, 2020).

However, there are many ways to aggregate data to calculate attention weights for each word, which depends on the chosen output layer and (or) attention head. Moreover, other decisions may change the attention values, such as the model's training process, architecture, or chosen meta-parameters for training.

Perception of helpfulness of model's predicted probability: The obtained results indicate that, in general, physicians perceive as helpful the fact of providing information on the predicted probability, giving it a score higher than three in 572 cases (57.5%), compared to only 422 cases (42.4%), which obtained a score lower than three. These results validate that a second opinion made by a model generally allows users more confidence in their decisions (Kompa et al., 2021; Benz & Rodriguez, 2022; Raghu et al., 2019).

Interaction effect on types of articles and visual encodings: The results obtained allowed us to confirm an interaction effect between visual encoding and the type of article on the perceived helpfulness of highlighted words. However, concerning the perceived helpfulness of the model's predicted probability, there is no interaction effect between visual encoding and the type of article on this information.

Expert evaluation: Having labels from an expert user allowed us to validate that using visualizations objectively improves users' performance in the task of document classification. We see that, on average, the coincidence of the labels with the experts using any of the visual encodings is greater than not using any visualization. Most explainable transformers, as highlighted words, have evaluated if attentions are an acceptable mean for explanation (Wiegreffe & Pinter, 2019; Jain & Wallace, 2019). However, neither of them has assessed if attention as visual explanations improves users' performance on a particular task, such as document classification. This is a considerable contribution since our results give evidence that obtained accuracy by using visualizations is 3.5% higher than the reached accuracy when not using any visualization.

Most attended words: The results show that although the XLNET model performed well, and in some cases, it gave more attention to words that did not make sense to physicians in the document classification task. We believe that the model paid attention to these words since attention values depend on the context where words are located in the text

and how the model learned those weights after going through multiple layers and heads in its learning process. According to the work of Jain & Wallace (2019) and Wiegreffe & Pinter (2019), there may be cases in which the attention may not be understandable by humans, but it does not mean that it have worse performance. It is important to study whether fine-tuning a pre-trained language model improves performance on the task but worsens the interpretability of attention. Another research line important to advance is to propose pre-training techniques that improve the model's performance and improve the interpretability of the learned attention.

Time effort: Regarding the time required when using each visual encoding, we obtained that there are no differences between visual encodings in the time required for the task. These results contrast with other works where it has been studied whether including explanations reduces the time required by users in the tasks of analysis of medical images (Holzinger et al., 2019; Folke et al., 2021), and NER from medical health records (Arguello-Casteleiro et al., 2021). However, users' performance on a given task often worsens if explanations are included because they blindly trust the model's criteria (Alufaisan et al., 2020; Linder et al., 2021).

Perception effort: We found that the background color visual encoding was the one that required less cognitive load in terms of mental, physical, temporal, and effort dimensions. We argue this since this type of encoding is the one that less affects reading and gives users a guide for the task of document screening. Furthermore, the control group without visualization reduced the cognitive load concerning frustration and improved performance. This could be caused because, in this case, we are not changing the platform that physicians use in their daily workload.

Users feedback qualitative analysis: After obtaining feedback from the users who were part of this study, we found that the model's predicted probability is perceived

as helpful for the document screening task. Moreover, the study's design and interface achieved a satisfactory evaluation since the platform was intuitive and easy to use. However, users did not highly value the words chosen by the model to make its predictions. In addition, a positive point of the visualizations is that in the second phase of the study, the users chose one of the visual encodings instead of choosing to turn off the visualization. One reason for this behavior is that visualizations allow users to easily extract keywords from the abstract.

5.8. User Study Conclusions

In this chapter, we conducted a user study on an EBM system to validate if explanations are helpful for physicians and which visual encoding is preferred for document screening.

The results allow us to answer the third and fourth questions of this thesis. Concerning the third question: *How does the inclusion of explanations influence the decision and reduce the cognitive effort of health experts?* We answered it since, in a second phase of the study, there is a preference for choosing one of the visual encodings instead of using no visualization, which denotes a preference between having explanations over the platform users interact on their daily work. Moreover, we provide evidence that using an explainable interface allows users to improve their objective performance and reduce their cognitive effort on the task of document screening.

Regarding the fourth question: *How do certain types of visual encoding influence health experts on choosing relevant evidence?*. We found that, in general, model's output attentions are not perceived as useful, since perception scores are in most of the cases lower than three. Furthermore, it can be seen that there is an increase in the objective performance by using a luminance visual encoding. There is also a reduction in cognitive load by using background color.

However, we observed the perception of helpfulness depends on interaction effect between the type of article and the chosen visual encoding. This means that there is no type of visual encoding that works in all cases, the way of visualizing attentions needs to change depending on the type of article being reviewed. For example, frameworks such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017) that propose to always use background color as visual encoding, according to the results obtained, this would not always have the same perception of helpfulness. We then recommend letting the user to choose among different visual encodings and do not keep using only background color.

Concerning information of model's predicted probability in general achieves a high helpfulness score, where in most of the cases scores are higher than three. However, helpfulness scores of this information do not depend on the visual encoding or the type of article being reviewed.

Additionally, we learned through the NASA TLX test that the background color intensity visual encoding is the one that most reduces the cognitive load in terms of mental, physical, temporal, and effort dimensions. Nevertheless, the perception of frustration is reduced and performance is augmented when not using visualizations.

Finally, to verify whether using visualizations improve user performance, we show that accuracy of users using visualization is 5.27% higher than accuracy without visualization.

A limitation of our work is that we used an XLNET model and chose one way to extract the attention values consisting of the last layer and the average attention of the heads since there are many other means to aggregate attention for each word. As future work, we will give the user the possibility of choosing heads to visualize explanations, the work proposed by Alammar (2021) using their proposed framework called Ecco.

6. CONCLUSIONS

In this thesis, we researched different ways to decrease clinicians' workloads for document screening task. We first experimented with different active learning strategies, vector representations of medical text, and machine learning models to select documents to be labeled by health experts and improve the model's performance with the minimum amount of labeled data. We found that the strategy of uncertainty sampling with random forest and a BioBERT representation of medical texts reached the best results among other models, active learning strategies, and representation combinations. At a later stage of this research, two significant events occurred. One was that the Natural Language Processing area evolved considerably, and the other was the COVID-19 pandemic. Given these two new variables, we proposed a state-of-the-art language model to choose relevant evidence related to COVID-19 disease. The appearance of COVID-19 implied in an exponential increase of indexed evidence in online databases (e.g, PubMed and Medline), which given its volume it became unfeasible to manually review by humans and verify relevance of evidence. XLNET was the model that brought the best results among other state-of-the-art language models. Furthermore, this model was brought into production in an evidence-based medicine system, demonstrating that by using a sampling strategy based on its uncertainty to choose certain COVID documents to label, we were able to reduce the workload of physicians by more than 65%.

Finally, we used the model with best performance from the previous chapter to evaluate whether visual explanations of the model's attentions were perceived as helpful for the task of document screening and evaluating if there are differences in the perceived helpfulness by using different visual encodings. We found that words attended by the model are generally not perceived as helpful, however there is an interaction effect between the visual encoding and the type of article being reviewed. Moreover, accuracy of users using

visualization is 5.25% better than accuracy when users do not use any visualization. In addition, an aspect that is generally considered as helpful is the predicted probability of the model.

Regarding the analysis of most attended words, we learned that when the user rates higher the helpfulness of highlighted words, the model pays more attention to domain-specific words for the given article. Nevertheless, as future work we will study whether fine-tuning a pre-trained language model improves performance on the task, but worsens the attention interpretability. One alternative would be to propose pre-training strategies that improve the model's performance and improve the interpretability of learned attentions.

6.1. Future work

As future work within the area of active learning, although in this thesis we evaluate traditional strategies such as uncertainty sampling or random sampling, we will experiment and implement other strategies where a model implicitly learns the sampling mechanism in an adversarial manner (Sinha et al., 2019; Chen et al., 2021; Guo et al., 2021; Mottaghi & Yeung, 2019; Zhang et al., 2020).

Another possible areas of extension of the work on COVID-19 document classification is whether it can be applied to electronic health records to validate if the performance is similar. In addition, a model could be trained for medical record texts in Spanish since, in this work, we use a pre-trained version of BioBERT and XLNET in English language.

Finally, in the area of explainable artificial intelligence, in this thesis, we demonstrate empirically through a user study that the attention learned by a transformer-based model

is not perceived as helpful. However, there is an interaction effect between the visual encoding and the type of article being reviewed.

Nevertheless, there is a trade-off between the model's performance and the model's capability to generate interpretable explanations. In the same spirit, we will explore pre-training strategies that improve both language models' performance and the capability of generating interpretable explanations. A work done by Riquelme et al. (2020) proposed a guided visual attention model to improve both models' performance and interpretability for visual question answering tasks. However, there is room for improvement for language-related tasks such as Text Summarization, Text Classification, Named Entity Recognition, among others.

REFERENCES

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, *41*(4), 1498–1508.

Alammar, J. (2021). Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations* (pp. 249–257).

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2020). Does explainable artificial intelligence improve human decision-making? *arXiv preprint arXiv:2006.11194*.

Arguello-Casteleiro, M., Maroto, N., Wroe, C., Torrado, C. S., Henson, C., Des-Diz, J., ... others (2021). Named entity recognition and relation extraction for covid-19: Explainable active learning with word2vec embeddings and transformer-based bert models. In *International conference on innovative techniques and applications of artificial intelligence* (pp. 158–163).

Bastings, J., & Filippova, K. (2020). The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*.

Bekhuis, T., Tseytlin, E., Mitchell, K. J., & Demner-Fushman, D. (2014). Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PloS one*, *9*(1), e86277.

Benz, N. C., & Rodriguez, M. G. (2022). Counterfactual inference of second opinions. *arXiv preprint arXiv:2203.08653*.

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence

from human reasoning. In *Ijcai* (pp. 6276–6282).

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). Nasa tlx: Software for assessing subjective mental workload. *Behavior research methods*, *41*(1), 113–117.

Carvallo, A., & Parra, D. (2019). Comparing word embeddings for document screening based on active learning. *BIRNDL Workshop, SIGIR 2019*.

Carvallo, A., Parra, D., Lobel, H., & Soto, A. (2020). Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 1–38.

Carvallo, A., Parra, D., Rada, G., Perez, D., Vasquez, J. I., & Vergara, C. (2020). Neural language models for text classification in evidence-based medicine. *arXiv preprint arXiv:2012.00584*.

Chen, F., Zhang, T., & Liu, R. (2021). An active learning method based on variational autoencoder and dbscan clustering. *Computational Intelligence and Neuroscience*, *2021*.

Choi, S., Ryu, B., Yoo, S., & Choi, J. (2012). Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences*, *214*, 76–90.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*, *4*, 129–145.

Dalvi, F., Khan, A. R., Alam, F., Durrani, N., Xu, J., & Sajjad, H. (2021). Discovering latent concepts learned in bert. In *International conference on learning representations.*

Del Fiol, G., Michelson, M., Iorio, A., Cotoi, C., & Haynes, R. B. (2018, Jun 25). A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. *J Med Internet Res*, *20*(6).

Demner-Fushman, D., & Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, *33*(1), 63–103.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donoso-Guzmán, I., & Parra, D. (2018). An interactive relevance feedback interface for evidence-based health care. In *23rd international conference on intelligent user interfaces* (pp. 103–114).

Dragicevic, P. (2016). Fair statistical communication in hci. In *Modern statistical methods for hci* (pp. 291–330). Springer.

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, *26*(11), 1279–1285.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., … others (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS one*, *3*(8), e3081.

Egger, M., Davey-Smith, G., & Altman, D. (2008). *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons.

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, *11*(2), e1001603.

Felix, C., Franconeri, S., & Bertini, E. (2017). Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE transactions on visualization and computer graphics*, *24*(1), 657–666.

Figueroa, R. L., Zeng-Treitler, Q., Ngo, L. H., Goryachev, S., & Wiechmann, E. P. (2012). Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, *19*(5), 809–816.

Folke, T., Yang, S. C.-H., Anderson, S., & Shafto, P. (2021). Explainable ai for medical imaging: explaining pneumothorax diagnoses with bayesian teaching. *arXiv preprint arXiv:2106.04684*.

Francis, S., Van Landeghem, J., & Moens, M.-F. (2019). Transfer learning for named entity recognition in financial and biomedical documents. *Information*, *10*(8), 248.

Gargiulo, F., Silvestri, S., Ciampi, M., & De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, *79*, 125–138.

Geva, M., Caciularu, A., Dar, G., Roit, P., Sadde, S., Shlain, M., ... Goldberg, Y. (2022). Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. *arXiv preprint arXiv:2204.12130*.

Giorgi, J. M., & Bader, G. D. (2018). Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, *34*(23), 4087–4094.

Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., ... Zuccon, G. (2017). Clef 2017 ehealth evaluation lab overview. In *International conference of the cross-language evaluation forum for european languages* (pp. 291–303).

Goodwin, T. R., & Harabagiu, S. M. (2018). Knowledge representations and inference techniques for medical question answering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *9*(2), 14.

Grossman, M. R., Cormack, G. V., & Roegiest, A. (2016). Trec 2016 total recall track overview. In *Trec*.

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, *40*(2), 44–58.

Guo, J., Pang, Z., Bai, M., Xie, P., & Chen, Y. (2021). Dual generative adversarial active learning. *Applied Intelligence*, *51*(8), 5953–5964.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183).

Elsevier.

Hashimoto, K., Kontonatsios, G., Miwa, M., & Ananiadou, S. (2016). Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of biomedical informatics*, *62*, 59–65.

Hassenzahl, M. (2008). User experience (ux) towards an experiential perspective on product quality. In *Proceedings of the 20th conference on l'interaction homme-machine* (pp. 11–15).

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Htut, P. M., Phang, J., Bordia, S., & Bowman, S. R. (2019). Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Stud Health Technol Inform*, *235*, 246–50.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Jin, Q., Dhingra, B., Cohen, W., & Lu, X. (2019). Probing biomedical embeddings from language models. In *Proceedings of the 3rd workshop on evaluating vector space representations for nlp* (pp. 82–89).

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). Clef 2017 technologically assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 1866, pp. 1–29).

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2018). Clef 2018 technologically assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 1866, pp. 1–34).

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2019). Clef 2019 technology assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 2380).

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J.-B. (2009). User experience over time: an initial framework. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 729–738).

Keselman, A., & Smith, C. A. (2012). A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics*, *45*(6), 1151–1163.

Khan, N. M., Abraham, N., Hon, M., & Guan, L. (2019). Machine learning on biomedical images: Interactive learning, transfer learning, class imbalance, and beyond. In *2019 ieee conference on multimedia information processing and retrieval (mipr)* (pp. 85–90).

Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. In *Bmc bioinformatics* (Vol. 12, pp. 1–10).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirk, R. E. (2010). Latin square design. *The Corsini Encyclopedia of Psychology*, 1–2.

Kobayashi, G., Kuribayashi, T., Yokoi, S., & Inui, K. (2020). Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*.

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, *4*(1), 1–6.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995* (pp. 331–339). Elsevier.

Larochelle, H., & Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order boltzmann machine. *Advances in neural information processing systems*, *23*, 1243–1251.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of*

the sigchi conference on human factors in computing systems (pp. 719–728).

Lee, G. E., & Sun, A. (2018). Seed-driven document ranking for systematic reviews in evidence-based medicine. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 455–464).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

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.

Lefebvre, C., Eisinga, A., McDonald, S., & Paul, N. (2008). Enhancing access to reports of randomized trials published world-wide–the contribution of embase records to the cochrane central register of controlled trials (central) in the cochrane library. *Emerging Themes in Epidemiology*, *5*(1), 13.

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Sigir'94* (pp. 3–12).

Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd clinical natural language processing workshop* (pp. 146–157).

Linder, R., Mohseni, S., Yang, F., Pentyala, S. K., Ragan, E. D., & Hu, X. B. (2021). How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, e49.

Lindsey, W. T., & Olin, B. R. (2013). Pubmed searches: Overview and strategies for clinicians. *Nutrition in Clinical Practice*, *28*(2), 165–176.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).

Ma, E. (2019). *Nlp augmentation.* https://github.com/makcedward/nlpaug.

Mayer, T., Cabrio, E., & Villata, S. (2018). Evidence type classification in randomized controlled trials. In *5th argmining@ emnlp 2018.*

Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of health services research & policy*, *10*(1_suppl), 6–20.

Mehmood, T., Gerevini, A. E., Lavelli, A., & Serina, I. (2020). Combining multi-task learning with transfer learning for biomedical named entity recognition. *Procedia Computer Science*, *176*, 848–857.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miwa, M., Thomas, J., O'Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, *51*, 242–253.

Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using lda-based document representations. *Systematic reviews*, *4*(1), 172.

Mottaghi, A., & Yeung, S. (2019). Adversarial representation active learning. *arXiv preprint arXiv:1912.09720*.

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., ... Nweke, H. F. (2019). Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, *116*, 494–520.

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Nadif, M., & Role, F. (2021). Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings in Bioinformatics*, *22*(2), 1592–1603.

Nourani, E., & Reshadat, V. (2020). Association extraction from biomedical literature based on representation and transfer learning. *Journal of theoretical biology*, *488*, 110112.

Nye, B. E., Nenkova, A., Marshall, I. J., & Wallace, B. C. (2020). Trialstreamer: mapping and browsing medical evidence in real-time. *arXiv preprint arXiv:2005.10865*.

Patro, B., Namboodiri, V., et al. (2020). Explanation vs attention: A two-player game to obtain attention for vqa. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 11848–11855).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Peng, Y., Yan, S., & Lu, Z. (2019, August). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th bionlp workshop and shared task* (pp. 58–65). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-5006 doi: 10.18653/v1/W19-5006

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., & Ananiadou, S. (2013). Distributional semantics resources for biomedical text processing. In *Proceedings of lbm 2013* (p. 39-44). Retrieved from http://lbm2013.biopathway.org/lbm2013proceedings.pdf

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., & Kleinberg, J. (2019). Direct uncertainty prediction for medical second opinions. In

International conference on machine learning (pp. 5281–5290).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Riquelme, F., De Goyeneche, A., Zhang, Y., Niebles, J. C., & Soto, A. (2020). Explaining vqa predictions using visual grounding and a knowledge base. *Image and Vision Computing*, *101*, 103968.

Roy, N., & McCallum, A. (2001). *Toward optimal active learning through sampling estimation of error reduction. int. conf. on machine learning.* Morgan Kaufmann.

Sachan, D. S., Xie, P., Sachan, M., & Xing, E. P. (2018). Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In *Machine learning for healthcare conference* (pp. 383–402).

Sackett, D. L. (1997). Evidence-based medicine. In *Seminars in perinatology* (Vol. 21, pp. 3–5).

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

Schein, A. I., & Ungar, L. H. (2007). Active learning for logistic regression: an evaluation. *Machine Learning*, *68*(3), 235–265.

Schmidt, L., Olorisade, B. K., McGuinness, L. A., Thomas, J., & Higgins, J. P. (2021). Data extraction methods for systematic review (semi) automation: A living systematic review. *F1000Research*, *10*, 401.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Settles, B. (2009). Active learning literature survey.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114.

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence

labeling tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 1070–1079).

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 287–294).

Sinha, S., Ebrahimi, S., & Darrell, T. (2019). Variational adversarial active learning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5972–5981).

Strobelt, H., Gehrmann, S., Pfister, H., & Rush, A. M. (2017). Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, *24*(1), 667–676.

Šuster, S., Verspoor, K., Baldwin, T., Lau, J. H., Yepes, A. J., Martinez, D., & Otmakhova, Y. (2021). Impact of detecting clinical trial elements in exploration of covid-19 literature. *arXiv preprint arXiv:2105.12261*.

Tutek, M., & Šnajder, J. (2020). Staying true to your word:(how) can attention become explanation? *arXiv preprint arXiv:2005.09379*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Vermeeren, A. P., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th nordic conference on human-computer interaction: Extending boundaries* (pp. 521–530).

Vig, J. (2019). Bertviz: A tool for visualizing multihead self-attention in the bert model. In *Iclr workshop: Debugging machine learning models.*

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., Schmid, C. H., Bertram, L., ... Trikalinos, T. A. (2012). Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in medicine*, *14*(7), 663.

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). Active learning for biomedical citation screening. In *Proceedings of the 16th acm sigkdd international conference on knowledge discovery and data mining* (pp. 173–182).

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... others (2020). Cord-19: The covid-19 open research dataset. *ArXiv*.

Wang, W., & Siau, K. (2018). Trusting artificial intelligence in healthcare.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., ... Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, *19*(1), 1–13.

Ward, M. O., Grinstein, G., & Keim, D. (2010). *Interactive data visualization: foundations, techniques, and applications*. CRC press.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Wu, Z., Nguyen, T.-S., & Ong, D. C. (2020). Structured self-attention weights encode semantics in sentiment analysis. *arXiv preprint arXiv:2010.04922*.

Yang, Y.-Y., Lee, S.-C., Chung, Y.-A., Wu, T.-E., Chen, S.-A., & Lin, H.-T. (2017). *libact: Pool-based active learning in python.*

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems* (pp. 5753–5763).

Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC medical informatics and decision making*, *19*(3), 31–39.

Zhang, B., Li, L., Yang, S., Wang, S., Zha, Z.-J., & Huang, Q. (2020). State-relabeling

adversarial active learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8756–8765).