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 of all, I want to thank my wife and parents for their unconditional support throughout this process, which was quite long and full of challenges. Also, I would like to thank the members of IALab, that although we were physically distanced due to the pandemic, we have continued actively collaborating, and Ivania Donoso and Hernán Valdivieso for their support on this thesis.

I would like to thank all the professors and research collaborators who helped me inspire and correct mistakes and bring the maximum benefit from my work and perseverance. Additionally, I want to thank the committee members Marcelo Mendoza, Gabriel Rada, and Hans Lobel, who provided comments that allowed me to improve my work and find new ways to tackle the main problem, and also thank Gabriel for having the willingness to help from the Epistemonikos Foundation and their collaborators.

I would also like to thank my advisor Denis Parra, for trusting in my qualities to carry out this work and for his patience during the whole process, namely iterations thorough the first drafts of the first journal article, this thesis and other related publications.

This work was funded by the Millenium Institute Fundamentals on Data and by CONICYT FONDECYT Regular Project number 1191791.

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.

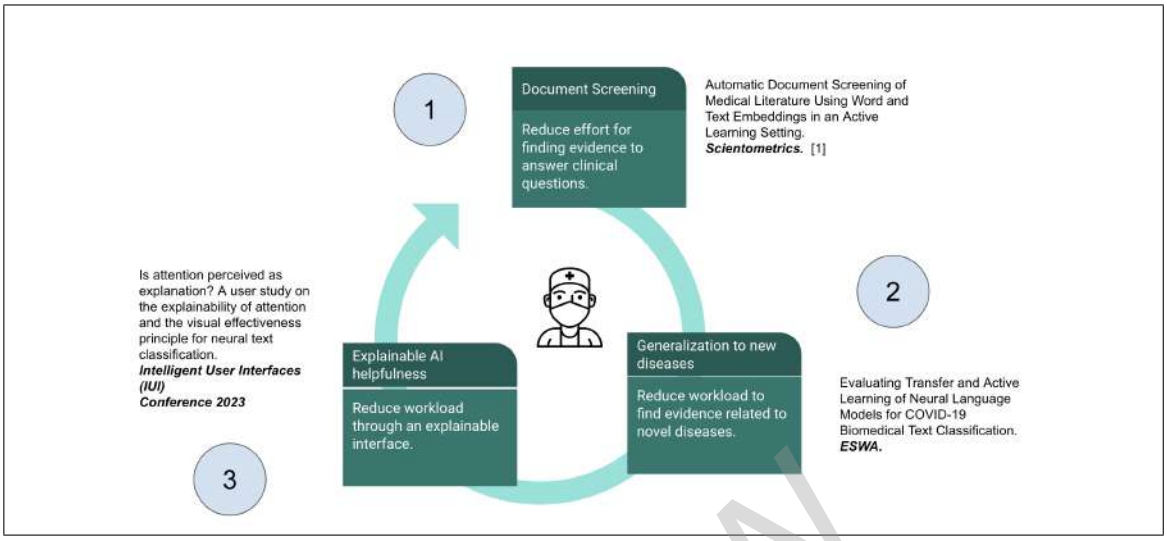

Figure 1.1. Overall diagram: This diagram shows how the thesis and its chapters are organized, with their corresponding publication.

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;