PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

DESARROLLO Y EVALUACIÓN DE UN MODELO DE DISEÑO DE VISUALIZACIONES PARA INTELIGENCIA ARTIFICIAL EXPLICABLE

HERNÁN FELIPE VALDIVIESO LÓPEZ

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería

Profesor Supervisor:
DENIS PARRA

Santiago de Chile, Enero 2022

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

DESARROLLO Y EVALUACIÓN DE UN MODELO DE DISEÑO DE VISUALIZACIONES PARA INTELIGENCIA ARTIFICIAL EXPLICABLE

HERNÁN FELIPE VALDIVIESO LÓPEZ

Miembros del Comité:
DENIS PARRA
JORGE MUÑOZ
MARCELO MENDOZA
DIEGO LOPEZ-GARCÍA

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería

Santiago de Chile, Enero 2022

Para todos quienes estuvieron a mi lado

AGRADECIMIENTOS

Quiero partir agradeciendo a mi familia, quienes me apoyaron durante todos estos años en mi carrera universitaria y se preocuparon de darme todas las facilidades posibles para que pueda enfocarme en los estudios y en el magíster.

Agradezco a mi profesor supervisor Denis Parra, por toda la paciencia que me tuvo, por motivarme a salir de mi zona de confort, por guiarme en todo este proceso, por todo las enseñanzas que me otorgó y por todas las oportunidades que me permitió vivir durante el desarrollo del magíster. También quiero agradecer al profesor Jorge Muñoz, quien me motivó y supervisó mis primeras investigaciones en la Escuela de Ingeniería. Tales investigaciones fueron el inicio de mi interés en la academia y mi deseo por intentar un magíster en computación.

Por otro lado, esta tesis no solo es un instrumento para finalizar mis estudios de postgrado, también es un hito del fin de mi viaje como alumno en la Escuela de Ingeniería. Todos estos años están llenos de hermosos recuerdos y personas, quienes les debo todo lo que soy ahora. Un enorme agradecimiento a:

Felipe Q., Kenichi T., Julian F., Matías S. y Vicente A., quienes me invitaban a jugar y sufrir con *League of Legend*, pero que aún así fue una gran fuente de diversión y distracción para todo lo que era universidad. Y una mención honrosa a Cristóbal O. quien es más listo que nosotros y no jugaba *League of Legend*.

Todos los grupos de amigos que fui formando durante el transcurso de los años (CalcetínClub, Zíppèèrš, Rúters Domingueros, La fldsmdfr, entre otros), que me apoyaron durante mi carrera universitaria y me entregaron hermosos recuerdos que nunca olvidaré.

Daniela C. y Francisca I., quienes estuvieron presente cada semana por más de un año en nuestras juntas *online* a ver anime. Estas juntas fueron una gran fuente de relajación y recreación durante el fin de mis estudios y como entretención para sobrellevar la pandemia. Además, agradezco que fueron quienes me apoyaron como ayudantes cuando empecé a ejercer como profesor del Diplomado de *Big Data* en la Universidad Católica.

Agradezco a la fundación Epistemonikos, por todo el apoyo dado durante el estudio de usuario realizado en este trabajo, y al Departamento de Ciencia de la Computación (DCC) por siempre tener un ambiente muy acogedor, tener a un excelente personal que se preocupa por sus alumnos, y tener a los mejores académicos que he conocido hasta el día de hoy.

Finalmente, agradezco a Lisa Thiel, por el diseño de los iconos para OpenMoji, los cuales fueron utilizados en este trabajo.

Este trabajo ha sido parcialmente financiado por Fondecyt Regular 1191791.

Hernán.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ABSTRACT

In recent years, we have witnessed the rapid adoption of AI to automate and solve different tasks. However, the use of AI-based systems is often lacking in explainability, which means letting users understand the rationale behind AI models' predictions. This problem has driven the development of explainable AI (XAI), and a diversity of methods have been proposed to construct explanations. Several methods resort to generating visualizations that support the explanation; thus, XAI does not only involve AI knowledge like the model to be used or algorithms to be explained, it also requires knowledge to design and implement appropriate visualizations. Although there are workflows for designing interactive machine learning (IML) or XAI applications, they focus on the stages of the machine learning (ML) model building process and do not provide guidelines or strategies to design or analyze the visualizations intended for XAI applications. Therefore, in this thesis we propose starting from the XAI task space and connect it to Munzner's widely adopted nested model for visualization design in order to bridge this gap. In this way, we propose the VD4XAI (Visualization Design for XAI) framework to guide the analysis and design process of XAI visualizations for local explanations. Our goal is to bridge the gap between AI/ML experts, designers, domain experts and final users for building effective XAI visualizations. This also will foster the development and application of visual XAI approaches.

Keywords: Information Visualization, Explainable AI.

RESUMEN

En los últimos años, hemos sido testigos de la rápida adopción de la inteligencia artificial (IA) para automatizar y resolver diferentes tareas. Sin embargo, el uso de sistemas basados en IA a menudo carece de explicabilidad, lo que significa permitir que los usuarios comprendan el fundamento detrás de las predicciones de los sistemas de IA. Este problema ha impulsado el desarrollo de la IA explicable (XAI) y se han propuesto diversos métodos para construir explicaciones. Varios métodos recurren a generar visualizaciones que apoyen la explicación; por lo tanto, XAI no solo implica conocimientos de IA como el sistema que se utilizará o los algoritmos que se explicarán, sino que también se necesitarán conocimientos para diseñar e implementar visualizaciones adecuadas. Aunque existen flujos de trabajo para diseñar aplicaciones de aprendizaje automático interactivo (IML) o XAI, estos se centran en las etapas del proceso de creación de modelos de aprendizaje automático (ML) y no proporcionan pautas o estrategias para diseñar o analizar las visualizaciones destinadas a aplicaciones de XAI. Por lo tanto, este trabajo propone comenzar desde el espacio de tareas de XAI y conectarlo al modelo anidado de Munzner ampliamente adoptado para diseñar visualización que permitan cerrar esta brecha. De esta forma, en esta tesis se propone el *framework* VD4XAI (*Visualization Design for XAI*) para guiar el proceso de análisis y diseño de visualizaciones de XAI para explicaciones locales. El objetivo es cerrar la brecha entre los expertos en IA/ML, diseñadores, expertos en dominios y usuarios finales para crear visualizaciones efectivas de XAI. Esto también fomentará el desarrollo y la aplicación de enfoques visuales de XAI.

Palabras Claves: Visualización de Información, IA explicable.

CAPÍTULO 1. INTRODUCCIÓN

Durante la última década, el uso de la Inteligencia Artificial (IA) se ha incrementado en diferentes aspectos de la vida diaria (Pouyanfar et al., 2018), tales como medicina (Miller y Brown, 2018; Reyes et al., 2020), finanzas (Vui, Soon, On, Alfred, y Anthony, 2013), sistemas recomendadores (Messina, Dominguez, Parra, Trattner, y Soto, 2019), navegación automática (Tian, Pei, Jana, y Ray, 2018), traducción (Zhang y Zong, 2015), e incluso se utilizan para automatizar una gama cada vez mayor de tareas en dominios críticos, como la atención médica, el ejército y la ley (Gunning, 2016). El principal objetivo de estos sistemas ha sido lograr una alta precisión en sus predicciones. Sin embargo, en la búsqueda de un mejor rendimiento, muchos de estos sistemas se han convertido en "cajas negras", es decir, no es posible entender, bajo términos compresibles para un ser humano, el funcionamiento del sistema o bajo qué razonamiento el sistema entregó dicha predicción. En consecuencia, la investigación sobre cómo explicar estos sistemas a los humanos se ha vuelto de extrema importancia. Incluso, empresas importantes como Apple, Google y Microsoft han incluido la explicabilidad en sus directrices de interacción humano-IA (Wright et al., 2020).

Bajo este contexto, Gunning (2016) acuñó el termino XAI, por su nombre en inglés *eXplainable Artificial Intelligence*, para referirse a las soluciones que permitan explicar los sistemas de IA. Además, dado el interés de empresas e investigadores en el desarrollo de XAI, en los últimos años se han impulsado diferentes iniciativas para fomentar la investigación y desarrollo en esta área. Un caso conocido es la Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA), una agencia del Departamento de Defensa del gobierno de Estados Unidos, quien planteó un programa en 2018 para financiar y apoyar la investigación de XAI [1]. Por otro lado, el 2021, la Unión Europea publicó una

[1] Recuperado el 27 de octubre de 2021 de `https://www.darpa.mil/program/explainable-artificial-intelligence`.

propuesta para sentar las bases para crear el primer marco legal sobre IA[2]. Por este motivo, bibliotecas de *software* han implementado métodos para explicar estos sistemas de "caja negra". LIME (Ribeiro, Singh, y Guestrin, 2016) y SHAP (Lundberg y Lee, 2017) han sido dos de las bibliotecas principales en Python para enfrentar esta tarea[3]. Uno de los aspectos más atractivos de estas bibliotecas es su representación visual, que puede ayudar al usuario a comprender el sistemas o predicciones particulares, como la visualización de salida de LIME que se muestra en la Figura 1.1 para un texto clasificado automáticamente por un modelo de IA.

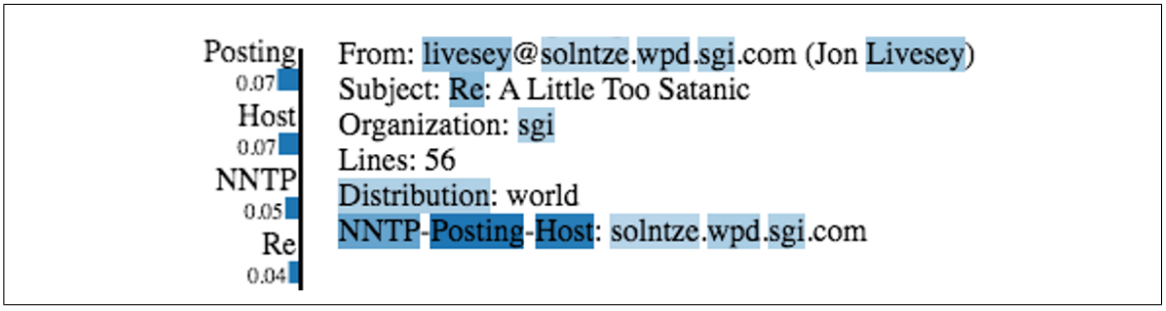

Figura 1.1. Salida de LIME para la tarea de clasificación de texto. En esta se puede observar que las palabras más importantes para la respuesta entregada por el modelo de IA fueron *Posting* y *Host*. Esto significa que la presencia de esas 2 palabras fue lo que más repercutió en la clasificación realizada por el modelo.

Sin embargo, ¿son estas visualizaciones el diseño más efectivo para XAI? ¿Cómo se pueden rediseñar estas visualizaciones para que sean más efectivas, posiblemente mejorando las marcas y los canales utilizados? En esta misma línea, el trabajo de Parra et al. (2019) muestra que existen diferentes alternativas para diseñar una visualización de XAI, pero ¿cómo llegaron a identificar todas esas alternativas? Y ¿cuál es la más efectiva para generar una explicación al usuario final? En los últimos años, investigadores han identificado una brecha entre los dominios de diseño y la IA (Abdul, Vermeulen, Wang, Lim, y Kankanhalli, 2018), y pocos *frameworks* de XAI han intentado conectar estas dos áreas

[2]Recuperado el 27 de octubre de 2021 de `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`.
[3]LIME: `https://github.com/marcotcr/lime`,
SHAP: `https://github.com/slundberg/shap`

(Spinner, Schlegel, Schafer, y El-Assady, 2019; Dudley y Kristensson, 2018). Además, estos *frameworks* se enfocan en el proceso de IA, es decir, en el proceso para confeccionar un sistema de IA que resuelva una tarea, y utilizan las visualizaciones de XAI como una herramienta para apoyar este proceso, pero no ayudan a analizar y potencialmente mejorar las visualizaciones de XAI existentes de una manera estructurada y formal. Dado lo expuesto anteriormente, en esta tesis se propone `VD4XAI` (*Visualization Design for XAI*), una extensión significativa del modelo anidado de Munzner Munzner (2014) para analizar y diseñar visualizaciones de XAI. El objetivo con el *framework* `VD4XAI` es cerrar la brecha entre diseñadores, expertos en XAI, expertos en dominios y usuarios finales, proporcionando orientación sobre el diseño y la validación al crear o analizar una visualización de XAI.

Esta tesis se divide en XX capítulos, incluyendo este. En el Capítulo 2 se presenta el marco teórico de este trabajo. En el Capítulo 3 se analizan investigaciones relevantes sobre el área de XAI, visualización y *frameworks* relacionados a estas dos áreas. El Capítulo 4 detalla las hipótesis de esta investigación y se presenta `VD4XAI`, un *framework* para analizar y diseñar visualizaciones de XAI. En el Capítulo 5 se describe la metodología utiliza en el desarrollo de este trabajo, el conjunto de datos utilizado y el estudio de usuario realizado al final de esta investigación. El capítulo 6 presenta la validación empírica de las hipótesis y los resultados de esta investigación. Finalmente el Capítulo 7 expone un resumen, el trabajo futuro de esta tesis y se concluye este trabajo.

CAPÍTULO 2. MARCO TEÓRICO

En esta sección se introducen los principales conceptos y definiciones necesarias para entender este trabajo.

2.1. Definición de conceptos

- **Aprendizaje automático o *Machine Learning* (ML)**: corresponde a una rama de la Inteligencia Artificial (IA) que realizan estudios de algoritmos informáticos que pueden mejorar automáticamente a través de la experiencia y mediante el uso de datos (Mitchell, 1997).
- ***Interactive Machine Learning* (IML)**: corresponde a una área dentro de la IA que intenta hacer que las técnicas de ML sean más accesibles al enmarcar cuidadosamente el proceso de entrenamiento como una tarea de interacción Humano-Computador (HCI) (Dudley y Kristensson, 2018).
- **Aprendizaje profundo o *Deep Learning* (DL)**: corresponde a una clase de algoritmos de ML que utilizan una secuencia de múltiples capas de procesamiento no lineal para extraer y transformar características de los datos. Cada capa utiliza la salida de la capa anterior como entrada. Estos algoritmos buscan aprender múltiples niveles de representación de los datos, los cuales corresponden a diferentes niveles de abstracción de ellos (Deng y Yu, 2014). Por ejemplo, una capa podría aprender a detectar bordes de una imagen, otra detecta el color rojo y otra capa detecta caras en la imagen.
- **método de IA**: se define como un técnica, dentro del campo de la IA, creada para resolver alguna tarea de forma automática de dicho campo. Ejemplos de métodos de IA son: regresión, árbol de decisión, red neuronal, SVM, entre otros.
- **Sistema de IA**: corresponde a un sistema informático que puede contener uno o más métodos de IA.

- **Interpretable**: hasta la fecha no existe una definición común de sistema interpretable en el área de Inteligencia Artificial (IA) y *Machine Learning* (ML). Sin embargo, este concepto generalmente se asocia con dos términos: transparencia y explicabilidad (Lipton, 2018).
- **Transparencia**: esta propiedad se atribuye a un sistema donde el usuario es capaz de comprender el mecanismo interno por el cual funciona dicho sistema. En otras palabras, el sistema deja de ser percibido como una caja negra. A modo de ejemplo, si se dispone una regresión, como la ecuación 2.1, que predice el precio de una casa en función de su terreno y años de antigüedad.

$$precio = 6.500.000 + 90.000 \times metros\_cuadrados - 50.000 \times años \tag{2.1}$$

  El usuario es capaz de entender que el precio base de la casa es de $6.500.000, el precio de la casa aumenta en $90.000 pesos por cada metro cuadrado de terreno, y por cada año de antigüedad, el precio baja $50.000 pesos. En consecuencia, se puede atribuir la propiedad de transparencia a la regresión lineal.
- **Explicabilidad**: esta propiedad se atribuye a un sistema que puede ser explicado, pero a diferencia de la transparencia, esta explicación no necesariamente revela como funciona internamente el sistema, puede ser una explicación post-hoc de su comportamiento.

  Imagine a un periodista que busca noticias sobre temas económicos. Un algoritmo de clasificación, como una regresión logística, identifica que un documento contiene temas económicos con un 85 % de probabilidad, pero ¿por qué dijo esa respuesta? Con un método explicable como LIME (Ribeiro et al., 2016), se puede encontrar una solución potencial para esta pregunta. En particular, este método entrena un nuevo modelo que permite explicar la respuesta, por ejemplo, una regresión para identificar características, en este caso palabras, que explican la clasificación mediante una ponderación de importancia de cada palabra. La Figura 2.1

presenta una visualización que ocupa las ponderaciones entregadas por la regresión para explicar la clasificación. Se puede observar que el modelo entregó dicha respuesta porque identificó algunas palabras clave, por ejemplo, fondo, propuesta y mercado. Por lo tanto, a pesar de no entender el funcionamiento interno del modelo, este se transformó en uno explicable dado al uso de LIME para identificar las palabras con mayor importancia en la respuesta del modelo.

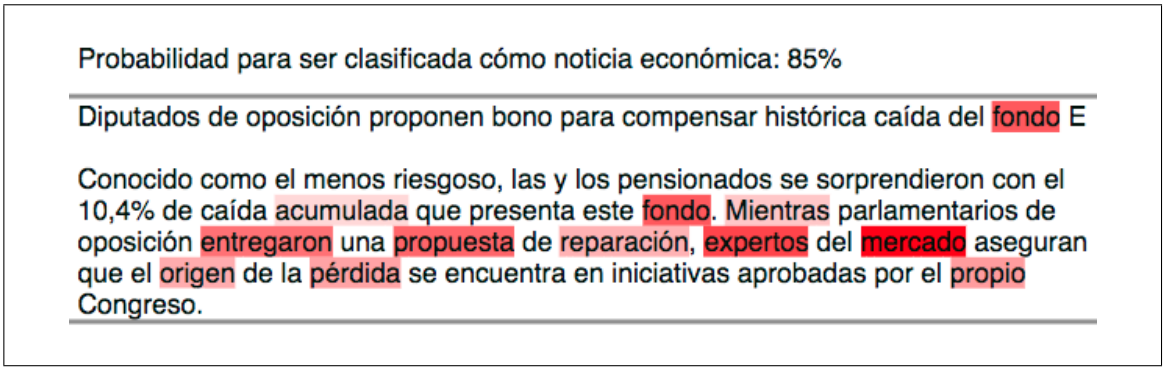

Figura 2.1. Visualización de los pesos de la regresión entrenada bajo el método LIME para explicar la clasificación de una documento como noticia con temas económicos.

2.2. Inteligencia Artificial Explicable (XAI)

Gunning (2016) introduce el término XAI como parte del programa DARPA, dirigido a construir sistemas de IA capaces de explicar su razón de ser, proporcionar detalles de sus fortalezas y debilidades, y transmitir la forma en que se comportarán en el futuro. Bajo este concepto, Gunning propuso diferentes preguntas que un usuario debería ser capaz de responder con la explicación del sistema de IA. Estas preguntas son:

- ¿Por qué hiciste eso? (*Why did you do that?*)
- ¿Por qué no otra cosa? (*Why not something else?*)
- ¿Cuándo lo logras? (*When do you succeed?*)
- ¿Cuándo fallas? (*When do you fail?*)

- ¿Cuándo puedo confiar en ti? (*When can I trust you?*)
- ¿Cómo corrijo un error? (*How do I correct an error?*)

2.3. Explicación local y global en Inteligencia Artificial (IA)

En el área de IA, es posible distinguir dos tipos de explicación: globales y locales (Guidotti et al., 2018). Por un lado, las explicaciones globales permiten comprender el razonamiento general del sistema o modelo de IA que conduce a los diferentes resultados posibles. Dentro de este tipo de explicación se puede encontrar el trabajo de Strobelt, Gehrmann, Pfister, y Rush (2018a), LSTMVis, una herramienta que ofrece una variedad de visualizaciones para entender un modelo de *deep learning* llamada *Long short-term memory (LSTM)* (Hochreiter y Schmidhuber, 1997). Por otro lado, las explicaciones locales permiten comprender las razones de una respuesta específica (clasificación, recomendación, entre otros) que entregó el sistema o modelo de IA, por ejemplo al clasificar objetos en una imagen que se dio de entrada a una red neuronal. LIME, SHAP y Grad-CAM++ son ejemplos de algoritmos que ofrecen este tipo de explicación (Ribeiro et al., 2016; Lundberg y Lee, 2017; Chattopadhay, Sarkar, Howlader, y Balasubramanian, 2018).

2.3.1. Estrategias para explicación local

En la actualidad, existen diversas investigaciones que abordan la explicación global de sistemas de IA (Strobelt, Gehrmann, Pfister, y Rush, 2018b; Strobelt et al., 2019; Snyder, Lin, Karimzadeh, Goldwasser, y Ebert, 2019), pero la información en este tipo de explicación puede ser demasiado específica y heterogénea para ser abordada, en primera instancia, en este trabajo. En consecuencia, esta tesis se basa en el supuesto de que el *framework* `VD4XAI` será utilizado, inicialmente, para diseñar y analizar visualizaciones de XAI acotadas a explicaciones locales.

A continuación se presentan tres estrategias de explicación local expuestas en el trabajo de D. Wang, Yang, Abdul, y Lim (2019), las cuales serán utilizadas dentro de `VD4XAI`:

1. **Importancia de características**: esta estrategia se usa frecuentemente (Bhatt et al., 2020) e indica cuáles datos de entrada son las más importantes para el sistema o modelo de IA, ya que tienen un impacto, positivo o negativo, en el resultado. El mecanismo de atención y SHAP son ejemplos de métodos de explicación basados en esta estrategia (Larochelle y Hinton, 2010; Lundberg y Lee, 2017).
2. **Analogías**: esta estrategia resalta instancias similares o muy diferentes a los datos de entrenamiento para que sirvan de apoyo o contraejemplo en la explicación. El trabajo de Xie, Chen, Kao, Gao, y Chen (2020b), CheXplain, corresponde a una herramienta visual donde se hace uso de esta estrategia, junto a la importancia de características, para explicar la respuesta de su sistema.
3. **Contrafactuales**: esta estrategia busca un punto cercano a los datos de entrada (por ejemplo, algún valor umbral) para los cuales la respuesta del sistema o modelo es alterada (Bhatt et al., 2020; Byrne, 2019). De esta forma, la explicación se centra en ofrecer casos hipotéticos de otras respuestas del sistema en función de una mínima variación de los datos entregados por el usuario. Para esta estrategia, el trabajo de Van Looveren y Klaise (2019) presenta un algoritmo que permite identificar los puntos cercanos.

Otra estrategia para generar explicación consiste en el uso de reglas. En este tipo de explicación, se genera una lista ordenada de reglas del tipo “sí - entonces” (Chen y Rudin, 2018; F. Wang y Rudin, 2015). Por ejemplo: (1) si la edad es mayor a 18, entonces el sistema responde “si”. (2) En otro caso, si la edad es mayor a 5, el sistema responde “no”. (3) Finalmente, si la edad es menos a 5, el sistema responde “si”.

2.3.2. Métodos de explicación basado en importancia de características

Hasta el día de hoy, es posible encontrar una diversidad de investigaciones que utilizan métodos basados en este tipo de explicación para generar una visualización de XAI, tales como los trabajos de Xie et al. (2020b); Cádiz (2021); Morichetta, Casas, y Mellia (2019). No obstante, cada método es diferente, cada uno presenta su propio rango de valores para codificar esta importancia, y la elección del método a utilizar condiciona el espacio de diseño disponible para confeccionar una visualización de XAI. Por lo cual, es necesario poseer una noción inicial de algunos de estos algoritmos. A continuación se presentan tres métodos populares de XAI que se basan en este tipo de explicación:

1. ***Local Interpretable Model-agnostic Explanation (LIME)***: el trabajo de Ribeiro et al. (2016) presenta un método post-hoc y agnóstico para explicar la respuesta de un modelo de clasificación o regresión. Es considerado agnóstico porque es una técnica que puede ser aplicado a cualquier método de IA que cumpla con ser de clasificación o regresión. El funcionamiento de LIME se centra en entrenar un nuevo modelo tradicionalmente explicable g, como una regresión o un árbol de decisión, que se comporte de forma similar al modelo original f para una región cercana al dato que se desea explicar x. Para lograr esto, LIME genera un conjunto de datos z a partir de un muestreo (*sampling*) de perturbaciones del dato que se espera explicar x, es decir, aleatoriamente se generan nuevos datos similares al que se desea explicar. Luego, para cada dato generado z, se obtiene la etiqueta entregada por el modelo original f. Con el conjunto de datos z y sus etiquetas, se entrena el nuevo modelo g para que entregue una respuesta lo más similares posibles a las que entregó el modelo original f. Finalmente, se utiliza el modelo tradicionalmente explicable g para explicar la respuesta x del modelo de clasificación o regresión original f. La Figura 2.2 presenta una descripción gráfica de la intuición de LIME.

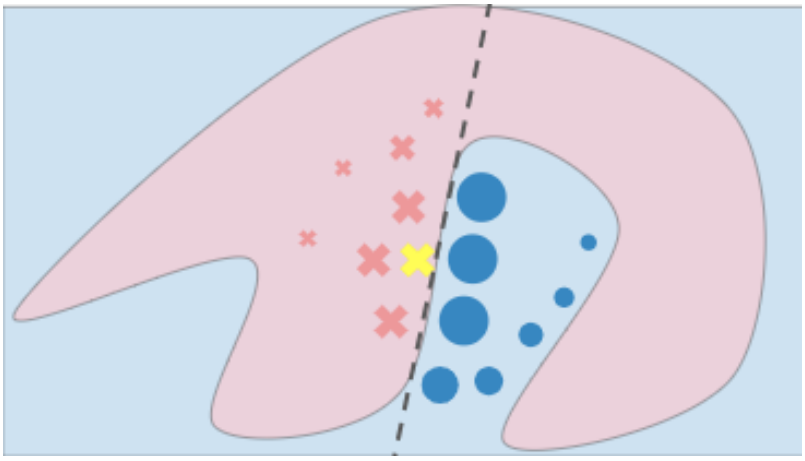

Figura 2.2. Ejemplo de juguete para presentar la intuición de LIME. El modelo original f (caja negra para LIME) está representado como la región roja en un fondo azul, el cual no se puede aproximar bien con un modelo lineal. La equis amarilla es el ejemplo a explicar x. LIME muestrea instancias z (puntos azules y equis rojas), obtiene predicciones usando f y las pesa en función de la proximidad a la instancia que se desea explica x (el peso se representa en la imagen por su tamaño). La línea discontinua es la explicación aprendida g que es localmente fiel al comportamiento del modelo original f.

LIME se puede catalogar como un método de explicación local basada en importancia de características porque uno de los modelos tradicionalmente explicable que se puede utilizar es la regresión. En tal caso, se utilizan los pesos de la regresión como estrategia de explicación. Estos pesos pertenecen en los números reales y representan el impacto que tiene una característica en la etiqueta entregada por el modelo original. Por ejemplo, si se busca explicar la clasificación de un texto, cada peso será el impacto dado por la presencia de cada palabra en el texto. Si este peso presenta un valor positivo, implica que la presencia de dicha palabra provocó que el modelo original entregara la clasificación o regresión que se busca explicar. La Figura 2.3 presenta un ejemplo donde se utiliza este método para explicar un modelo de clasificación de noticias. En particular, se pinta con tonalidades de rojo las palabras que contribuyeron positivamente a que el clasificador indicara que la noticia era de la categoría económica.

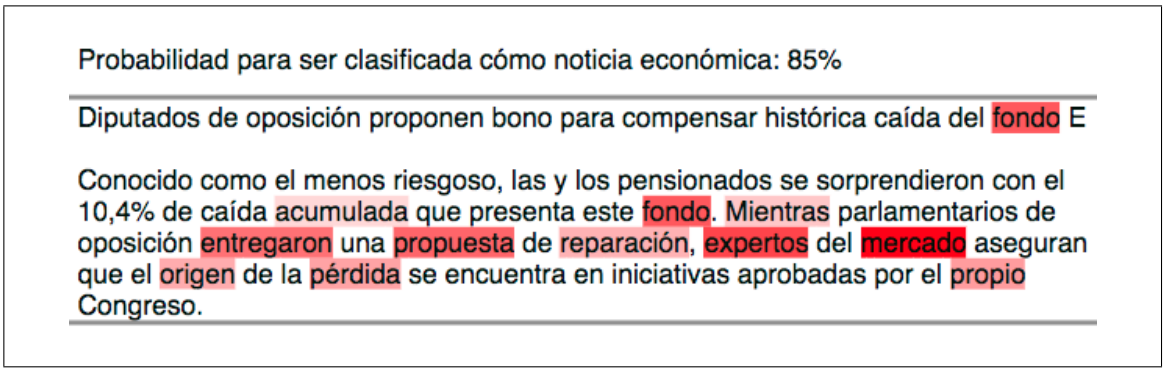

Figura 2.3. Ejemplo de visualización utilizando el método LIME para explicar la clasificación de una documento como noticia con temas económicos. Aquí se destacan palabras claves para el modelo como mercado, fondo y expertos.

2. ***SHapley Additive exPlanations (SHAP)***: el trabajo de Lundberg y Lee (2017) consiste en un método agnóstico basado en la teoría de juegos para explicar una respuesta de un modelo de IA. Para lograr esta explicación, se realizan diferentes perturbaciones a un dato de entrada, eliminando ciertos fragmentos del dato, por ejemplo, se eliminan píxeles de la imagen, algunas palabras del texto o algunas columnas de un dato tabular. Con estas perturbaciones, se determina la variación en la respuesta del modelo. Esa variación es identificada como la contribución marginal de la información eliminada. Con dicha contribución marginal, SHAP atribuye un peso a cada fragmento del dato. Este peso pertenece al conjunto de los números reales y permite determinar el impacto que tiene la presencia de cada fragmento en la respuesta del modelo. La figura 2.4 muestra un ejemplo visual de la explicación entregada por SHAP.
3. **Mecanismo de atención**: el trabajo de Larochelle y Hinton (2010) introduce una técnica considerada como una de las más significativas para las redes neuronales profundas en los últimos años. La idea detrás de este mecanismo se inspira en el sistema visual del ser humano, ya que ellos se enfocan selectivamente en partes de la imagen en lugar de la imagen completa (Mnih, Heess, Graves, y Kavukcuoglu, 2014). En consecuencia, este mecanismo permite que el modelo de DL se concentre en un subconjunto de los datos de entradas cuando es entrenado en una tarea.

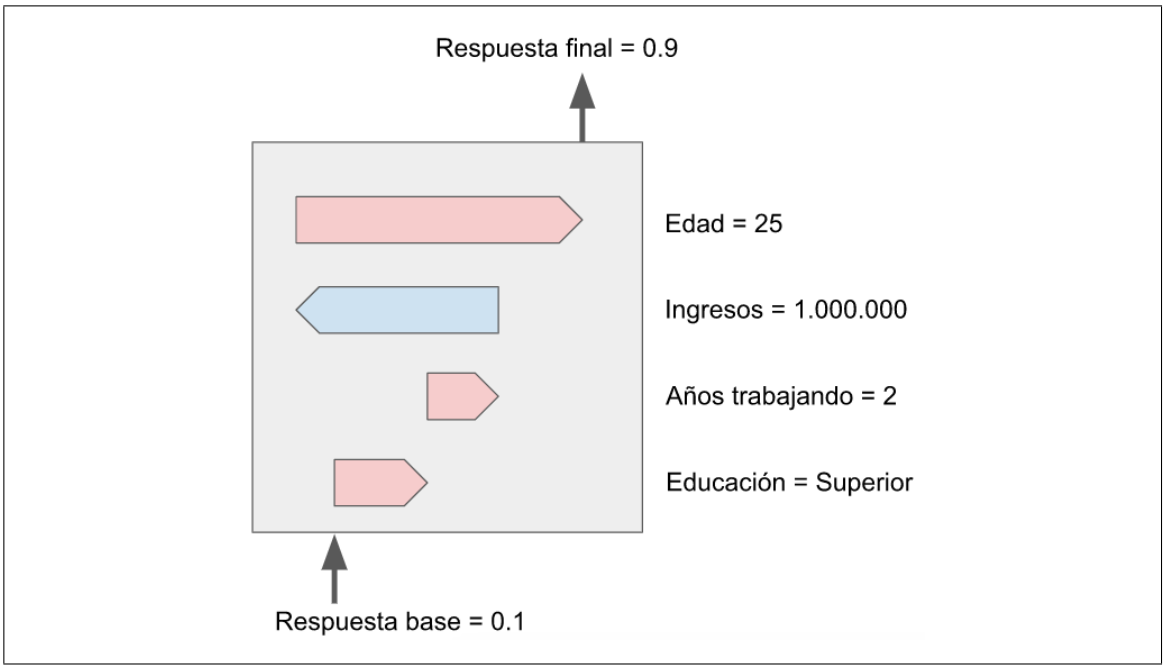

Figura 2.4. Ejemplo de visualización utilizando el método SHAP para explicar la respuesta de un modelo que predice la probabilidad de tener crédito. El modelo responde, por defecto, que un usuario tiene un 10 % de probabilidad de tener un crédito, pero dado que el usuario trabaja desde hace dos años, tiene una educación superior y tiene 25 años, el modelo de IA cambió su respuesta a un 90 % de tener un crédito. Además, se observa que la característica de tener 25 años es la que más impacto tiene en la respuesta del modelo.

Este método retorna un peso de atención para cada dato de entrada, este peso presenta un valor entre cero y uno, donde cero implica una nula atención en dicho dato y uno significa que el dato recibió toda la atención del modelo. Adicionalmente, los modelos de DL que utilizan este mecanismo pueden estar construidos en capas, lo que implica que cada capa puede utilizar este mecanismo de atención. En base a lo anterior, para cada dato de entrada es posible disponer de múltiples pesos de atención. La Figura 2.5 muestra un ejemplo visual de la explicación entregada por el mecanismo de atención. En el apéndice A se presenta el detalle matemático de una versión de este método propuesto en el trabajo de Vaswani et al. (2017).

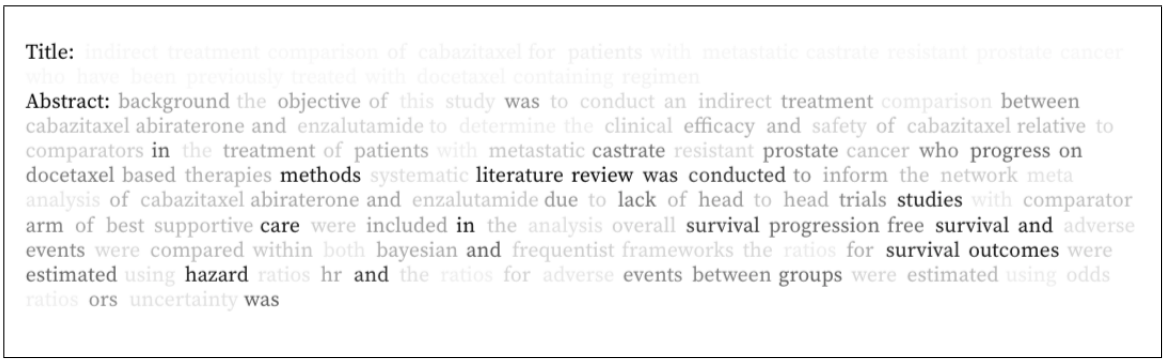

Figura 2.5. Ejemplo de visualización utilizando el mecanismo de atención para explicar la clasifica de un documento. El clasificación dada por el modelo se debe a la presencia de ciertas palabras como *literature*, *review*, *conducted*, *methods* y *survival*.

Finalmente, estas técnicas de explicación local han sido introducidas en los últimos años. Esto implica que las visualizaciones que hacen uso de dichas técnicas no están totalmente depuradas. Por ejemplo, el trabajo de Parra et al. (2019) da evidencia que el mecanismo de atención, aplicado en textos, todavía presenta diferentes alternativas para la confección de visualizaciones.

2.4. Visualización de información

La visualización se puede definir como la comunicación de información utilizando representaciones gráficas (Ward, Grinstein, y Keim, 2010). Mientras que la visualización de información es definida por Tamara Munzner, autora del libro *Visualization Analysis and Design* (2014), como sistemas de visualización creados por una computadora que brindan una representación visual de los *datasets* (conjunto de datos) y que están diseñados para ayudar a las personas a realizar tareas más eficazmente. A pesar de que las visualizaciones son aplicadas para diferentes propósitos (visualizar fluidos, ventas en una empresa, mapas, entre otros), al momento de construir una visualización, se comparten las mismas decisiones de diseño. Con el fin que estas decisiones sean efectivas, es necesario recurrir a diferentes áreas de conocimientos, como la ciencia de la computación, interfaces de

usuario, psicología, percepción y estadísticas (Grinstein y Wierse, 2002). Diversos *frameworks* y metodologías para diseñar una visualización son creados a partir de considerar lo expuesto de estas áreas de conocimiento.

Un *framework* reconocido para analizar y diseñar visualizaciones corresponde al trabajo de Munzner (2014). Este trabajo se concentra en responder tres preguntas fundamentales: (1) ¿qué datos se van a visualizar? (*what?*), (2) ¿por qué el usuario necesita usar la visualización? (*why?*) y (3) ¿cómo será diseñada la visualización? (*how?*). Adicionalmente, este mismo trabajo presenta un modelo anidado, compuesto por 4 niveles, para guiar el proceso de diseño: situación de dominio, abstracción de datos y tareas, codificación visual e interacciones y algoritmo. En la Figura 2.6 se pueden observar los 4 niveles, la dependencia que tiene uno con el otro, y en qué niveles se abordan las tres preguntas fundamentales.

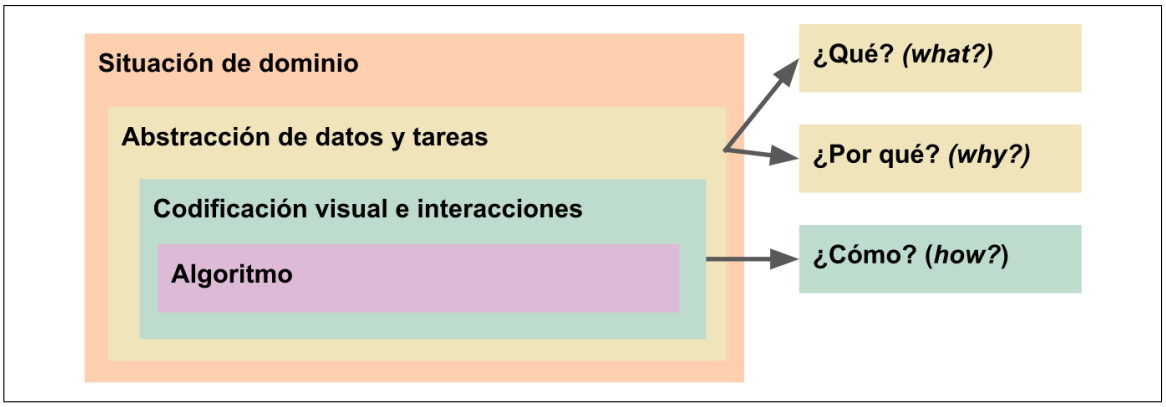

Figura 2.6. Figura basada en el libro de Tamara Munzner (2014), donde se muestran la dependencia de los cuatro niveles del modelo anidado y las tres preguntas fundamentales del *framework*.

1. **Situación de dominio**: corresponde al primer nivel del modelo y en este se define el grupo de usuarios objetivos, las preguntas que quieren solucionar y los datos que se dispone. Con tal información, se espera entender en su completitud el problema que el usuario quiere solucionar mediante el uso de visualización. Además,

en este nivel se estudian las diversas herramientas o visualizaciones existentes para enfrentar este problema y cuál es el uso que le dan los usuarios. Es posible que dichas herramientas ya solucionen el problema o no las estén usando por algún motivo específico. Es relevante conseguir toda esa información para diseñar una visualización que no presente las mismas dificultades que las herramientas anteriormente diseñadas.

2. **Abstracción de datos y tareas**: luego de conocer en profundidad las necesidades del usuario, comienza el segundo nivel de este modelo. En este nivel se responden dos de las tres preguntas fundamentales del *framework*: ¿qué? y ¿por qué?. Para lograr este objetivo, se mapean los datos y las necesidades del usuario a un vocabulario específico del *framework* que permite abstraerse del caso particular. Por ejemplo, si una empresa desea ver cómo cambian las ventas de un cierto producto en el tiempo, o si es un investigador que desea ver cómo cambian las precipitaciones en el tiempo, ambas situaciones se mapean a un vocabulario en común: identificar la tendencia de un atributo cuantitativo en el tiempo. De este modo, las decisiones de diseño aplicadas a un problema pueden llegar a ser aplicados a otros si es que comparten la misma abstracción de datos y/o tareas.
   - **Abstracción de datos**: esta abstracción consiste en identificar el tipo de *dataset* (tabular, geométrico, red, entro otros), el tipo de dato (ítem, atributo, enlace, posición, entre otros), la disponibilidad de los datos (estático o dinámico) y para el caso de los atributos, una descripción más detallada de ellos (categórico, ordenado o cuantitativo).
   - **Abstracción de tareas**: esta abstracción se divide en dos elementos: el propósito de la visualización y el aspecto del dato de interés para el usuario objetivo. En relación al propósito de la visualización, el trabajo de Munzner propone tres niveles: tipo de análisis (presentar, descubrir, anotar, grabar, disfrutar), tipo de búsqueda (localizar, navegar, explorar, buscar) y el tipo de consulta (identificar, comparar, resumir). En el caso de aspectos del dato de interés, este puede

ser un aspecto que involucra toda la información (característica, patrón, datos atípicos), solo información de un atributo (extremos, distribución) o información de más de un atributo (correlación, similaridad, dependencia). En caso de ser un dato geométrico o de red, el usuario puede estar interesado en ver la forma, topología o caminos en dichos datos.

3. **Codificación visual e interacciones**: en este nivel se responde la última pregunta fundamental, ¿cómo?. En otras palabras, se decide la forma específica de crear y manipular la visualización para mostrar los datos a partir de las tareas abstractas identificadas en el nivel anterior. Existen dos principales elecciones a realizar en este nivel: la codificación visual y las interacciones.
   - **Codificación visual**: corresponde al mapeo de los datos a alguna representación visual de ellos, es decir, se determina cómo se verán los datos en la visualización. En este punto, se puede elegir entre mostrar los datos con una cierta disposición (separados, ordenados, alineados, geo-referenciados, entre otros), o con algún canal visual (color, largo, volumen, forma, entre otros).
   - **Interacción**: corresponde a determinar cómo el usuario controla lo que verá. En este punto se puede decir entre permitir una manipulación de los datos (navegar, seleccionar o cambiar), utilizar diferentes vistas (yuxtaposición, superposición o partición) y/o aplicar una reducción de la información (filtros, agregación o uso de ventanas embebidas).
4. **Algoritmo**: en este último nivel se trabaja con la confección de la visualización a partir de la codificación visual, y los mecanismos de interacción elegidos en el nivel anterior. En este nivel es necesario estudiar formas que aseguren un manejo eficiente de los datos, la codificación visual y los mecanismos de interacción. Una visualización bien diseñada, pero poco eficiente puede ocasionar una mala experiencia al usuario, provocando un rechazo a la visualización.

Un factor importante de este *framework* es que identifica posibles amenazas o problemas que pueden surgir durante el diseño de la visualización, y entrega diferentes validaciones para mitigar estas amenazas. Algunas de estas validaciones se pueden realizar de forma inmediata antes de seguir con los siguientes niveles, mientras que otras validaciones requieren de haber diseñado y programado la visualización para verificar si la amenaza fue mitigada. En la figura 2.7 se puede observar las amenazas identificadas en cada nivel y las validaciones propuestas.

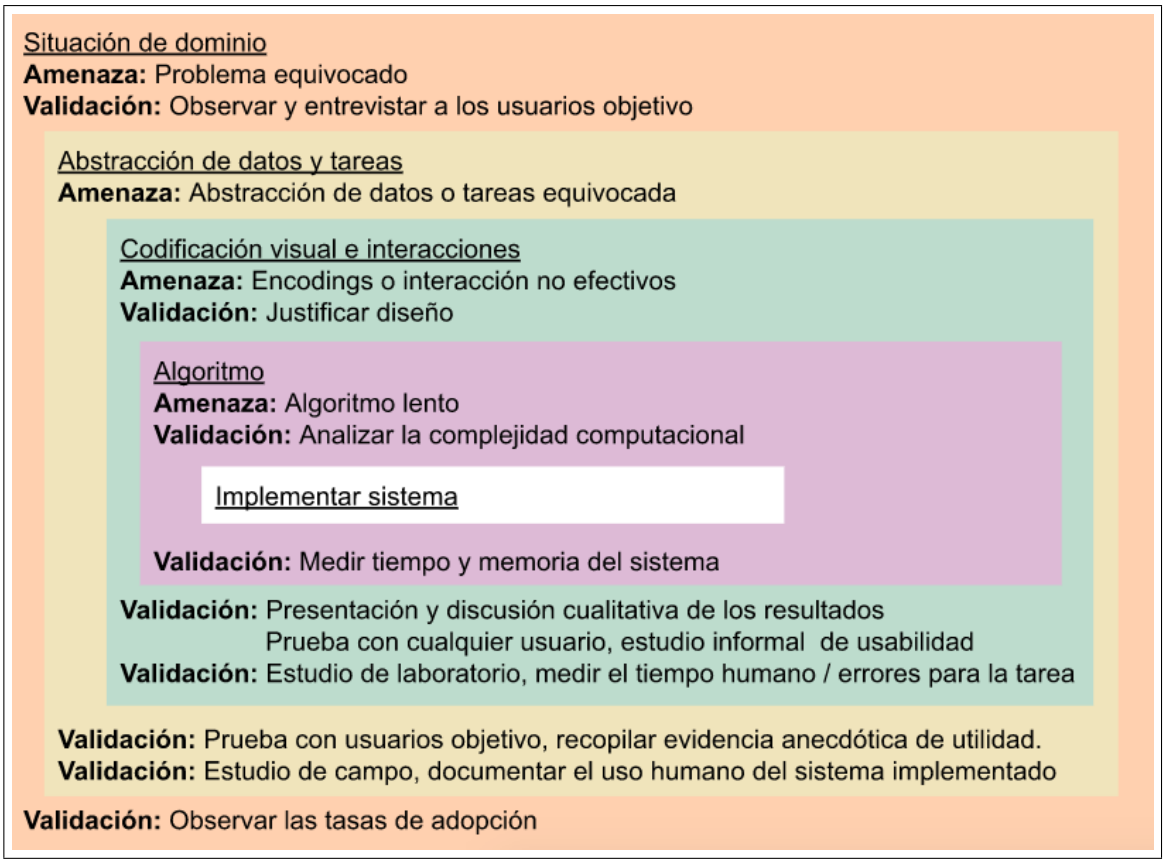

Figura 2.7. Figura basada en el libro de Tamara Munzner (2014) donde se presentan las amenazas y validaciones del *framework*.

CAPÍTULO 3. TRABAJO RELACIONADO

3.1. Inteligencia Artificial Explicable (XAI)

En el capítulo 2 se explicó en qué consiste concepto de XAI y cómo se originó el 2016. Desde entonces, varios métodos de XAI han sido creados, así como trabajos que los recopilan y resumen (Guidotti et al., 2018; Adadi y Berrada, 2018; Mohseni, Zarei, y Ragan, 2018). Incluso, investigaciones muestran que empresas importantes como Apple, Microsoft y Google han incluido la explicabilidad en sus guías sobre interacción Humano-IA (Wright et al., 2020). Del mismo modo, Shneiderman (2020) ha recomendado añadir la explicabilidad en los sistemas de IA para aumentar la fiabilidad, seguridad y confiabilidad de la IA centrada en el ser humano (HCAI). Estos hechos dan evidencia de la gran relevancia que tienen los sistemas explicables en la investigación actual sobre IA.

3.2. *Frameworks* de visualización, IML y XAI

Algunos *framewoks* abordan el proceso de visualización, IML y XAI, pero ninguno de ellos los integra todos al mismo tiempo. Por un lado, en visualización de información, (i) Upson et al. (1989) provee uno de los *frameworks* iniciales con cuatro componentes principales para diseñar una visualización: fuente de datos, mapeo y filtro, representación y salida. Luego, (ii) Munzner (2014) introdujo un modelo anidado para analizar y diseñar visualizaciones. Por otro lado, (iii) Hohman, Kahng, Pienta, y Chau (2019) resumieron a fondo el rol de la analítica visual en la investigación de *deep learning*. Esta encuesta analizó cómo se están utilizando las visualizaciones en el área de *deep learning* y propuso un *framework* para organizar la investigación sobre este tema. A pesar que estos *frameworks* guían el proceso para analizar y/o diseñar visualizaciones, ninguno de estos fue construido para aplicar sus propuestas a un dominio tan específico como XAI. Por ejemplo, ninguno considera el tipo de explicación como un componente al momento de diseñar o analizar las visualizaciones de XAI.

Desde el lado de IML, (i) el estudio de Jiang, Liu, y Chen (2019) sintetizó diversos trabajos donde los usuarios pueden controlar de forma iterativa los sistemas de IA mediante el uso de visualizaciones interactivas. (ii) Dudley y Kristensson (2018) estudiaron diferentes investigaciones sobre el diseño de interfaz para IML. Con tal estudio, ellos propusieron un flujo generalizado de IML que se extiende del modelo de Fails y Olsen (2003). (iii) También se encuentra el trabajo de Zhang, Wang, Molino, Li, y Ebert (2019), que propusieron Manifold, un *framework* agnóstico para interpretar y diagnosticar modelos de ML. Este *framework* utiliza las visualizaciones para ayudar en todo el proceso de IML, pero no provee detalles del proceso para validar las visualizaciones utilizadas. (iv) Otro *framework* propuesto recientemente para conectar el área de XAI y las visualizaciones es explAIner (Spinner et al., 2019), el cual integra las etapas de IML y un gran número de métodos de XAI con sus respectivas visualizaciones. No obstante, este trabajo no incluye ninguna etapa para evaluar el diseño, modificar o crear nuevas visualizaciones para los métodos de XAI. Por ejemplo, tal como muestra la Figura 3.1, el sistema explAIner permite utilizar el método LIME (Ribeiro et al., 2016) para generar explicaciones. En el caso de clasificación de imágenes, utiliza el color rojo y azul para indicar las regiones de la imagen más importantes. No obstante, se podrían utilizar otros colores como una paleta óptima para daltónicos, pero explAIner no permite eso. Otro caso es la clasificación de texto, se utiliza la saturación del color de fondo para presentar las palabras más importantes. Sin embargo, existen otros canales para codificar esta importancia, tales como el tamaño de la letra o su luminosidad, y ninguno de estos es sugerido o es posible de utilizar como alternativa en explAIner.

Por otro lado, se han realizado múltiples esfuerzos para proporcionar orientación al crear sistemas de XAI:

1. El trabajo de D. Wang et al. (2019) utiliza teorías del razonamiento humano y enfatiza cómo los elementos de XAI apoyan esos procesos y ayudan a mitigar

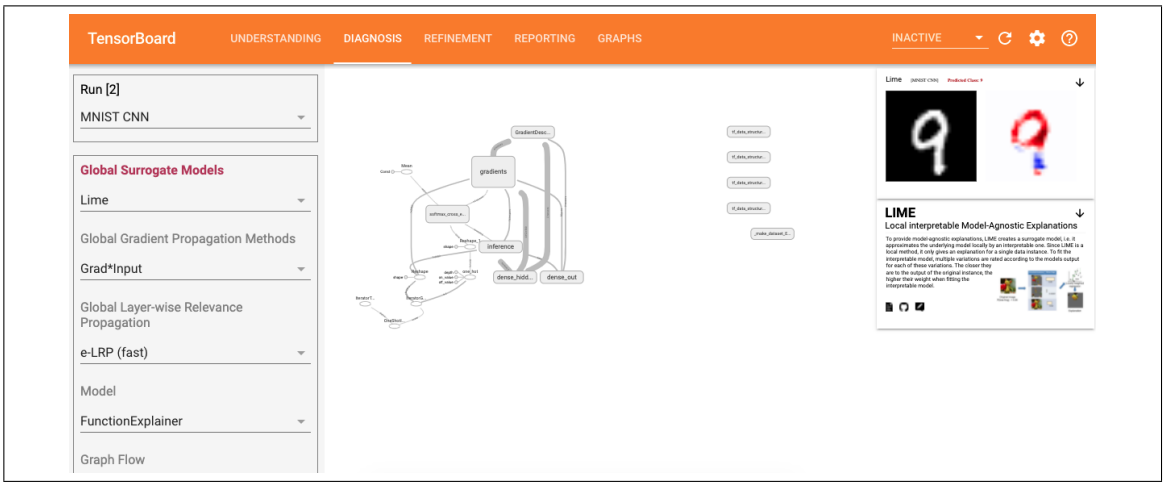

Figura 3.1. Ejemplo de una de las interfaces de eXplAIner para analizar una modelo de *deep learning*. Foto generada a partir del código fuente provisto por los autores. Recuperada el 23 de noviembre del 2021 de `https://github.com/dbvis-ukon/explainer`.

errores. Su trabajo ofrece un *framework* para vincular las técnicas de explicación de IA con diversos patrones cognitivos.

2. Ras, van Gerven, y Haselager (2018) identifican diferentes tipos de usuarios de estos sistemas y con tal trabajo proponen una taxonomía para clasificar y analizar los diferentes métodos de explicación.
3. Mohseni et al. (2018) presenta un *framework* para unificar los esfuerzos de diferentes disciplinas involucradas con los sistemas de XAI. Este *framework* anidado permite diseñar y evaluar sistemas de XAI, y se compone de tres capas: sistema, interfaz y algoritmo. Adicionalmente, describe tres formas de explicación: visual (Simonyan, Vedaldi, y Zisserman, 2013; Zeiler y Fergus, 2014), verbal (Lakkaraju, Bach, y Leskovec, 2016; Berkovsky, Taib, y Conway, 2017; Herlocker, Konstan, y Riedl, 2000; Rosenthal, Selvaraj, y Veloso, 2016; Myers, Weitzman, Ko, y Chau, 2006) y analítica (Hohman et al., 2019; Strobelt et al., 2018a; Goodall et al., 2019), en donde la primera y última forma utilizan las visualizaciones para generar explicaciones entendibles al humano.
4. Ribera y Lapedriza (2019) proponen un *framework* centrado en el usuario para crear explicaciones. En este trabajo se identifican tres usuarios finales diferentes y

cuatro áreas de decisión (objetivo de la explicación, contenido, tipo de explicación y evaluación).

5. Barda, Horvat, y Hochheiser (2020) proponen un *framework* que añade dos áreas de decisión (dónde y cuándo la explicación será utilizada) al trabajo propuesto por Ribera y Lapedriza (2019). Ellos explican que los enfoques anteriores se han centrado solo en el usuario que recibirá la explicación, pero no en el contexto donde el usuario las consumirá.

Todos los *frameworks* presentados anteriormente poseen un punto en común, y es que utilizan las visualizaciones como una opción, entre muchas, para crear explicaciones. No obstante, **ninguna de estas propuestas presentan un proceso formal para diseñar dichas visualizaciones que apoyen las explicaciones**, es decir, no conectan la presentación de los métodos de XAI con el diseño de su representación visual.

A raíz de este problema, el trabajo de Liao, Pribić, Han, Miller, y Sow (2021) propone un proceso basado en preguntas que fomentan la interacción entre los diseñadores, el experto en IA y el usuario final. Con estas preguntas se guía el proceso para confeccionar una interfaz que permita al usuario final entender diferentes aspectos de los datos y del sistema de IA, es decir, proponen un proceso basado en preguntas para confeccionar una experiencia de usuario (UX) en XAI. Si bien esta interfaz utiliza visualizaciones para explicar diferentes aspectos de los datos o el sistema de IA, los mismos autores declaran que su propuesta no ofrece una guía sobre cómo se debe diseñar o evaluar esta experiencia de usuario. En consecuencia, todavía no existe una conexión bien definida entre el conocimiento de los expertos en IA y los diseñadores visuales (Abdul et al., 2018). Los investigadores y los profesionales necesitan un proceso de diseño correctamente definido y validado en diferentes áreas de XAI. Con esto, podrán crear y analizar una visualización de información en el dominio de XAI de manera eficaz ya que, como afirma Munzner (2014), “el espacio de diseño visual es enorme y la mayoría de los diseños son ineficaces.”

Por lo tanto, hasta donde se sabe, todavía no existe un enfoque, *framework* o flujo de trabajo formal que conecte los métodos actuales de XAI con el proceso de diseño y análisis de visualizaciones de XAI.

3.3. Sistemas visuales de XAI

El número de visualizaciones aplicadas a problemas enmarcados en el contexto XAI ha aumentado en los últimos tres o cuatro años. Algunos trabajos han centrado sus aportes en IML y otros en la explicación del sistema (local o global). Estos trabajos utilizan la visualización principalmente de dos formas: para respaldar la creación de modelos mostrando el rendimiento del sistema, sus fortalezas y debilidades, y comparaciones con modelos anteriores. (Pezzotti et al., 2018; M. Liu, Shi, Cao, Zhu, y Liu, 2018; Dingen et al., 2019; Q. Wang et al., 2019; Sun et al., 2020; Puhringer, Hinterreiter, y Streit, 2020); y para proporcionar explicaciones locales y/o globales del sistema que entregan información para mejorar dicho sistema (Huang et al., 2021; Strobelt et al., 2019; Ming, Xu, Cheng, Qu, y Ren, 2020; Wexler et al., 2019; Ma, Xie, Li, y Maciejewski, 2020; Sahoo y Berger, 2020; Das et al., 2020; Marai et al., 2019; X. Zhao, Wu, Lee, y Cui, 2019; Ming, Qu, y Bertini, 2019; Kwon et al., 2019; Z. J. Wang et al., 2021; Snyder et al., 2019). Un ejemplo del primer grupo es DeepEyes (Pezzotti et al., 2018), un sistema visual que analiza al sistema de IML durante la fase de entrenamiento. Las visualizaciones utilizadas ayudan a identificar problemas con el sistema, como capas o filtros que no se desempeñan como uno espera. Además, estas visualizaciones también ayudan a entender qué información es capturada por el sistema. En relación al segundo grupo, un ejemplo es el sistema SMART 2.0 (Snyder et al., 2019), el cual aplica las visualizaciones en diferentes etapas del proceso de IML. En particular, este trabajo utiliza la visualización para mostrar la relevancia de cada *tweet* para cada tópico. Con esta información, el usuario es capaz de corregir la clasificación realizada por el sistema. Otro ejemplo de este grupo es Seq2Seq-Vis (Strobelt et al., 2019), un sistema para explicar visualmente la respuesta de un modelo neuronal

Sequence-to-Sequence (Sutskever, Vinyals, y Le, 2014). La Figura 3.2 presenta la interfaz diseñada por los autores de Seq2Seq-Vis para explicar la traducción de una frase en alemán a inglés.

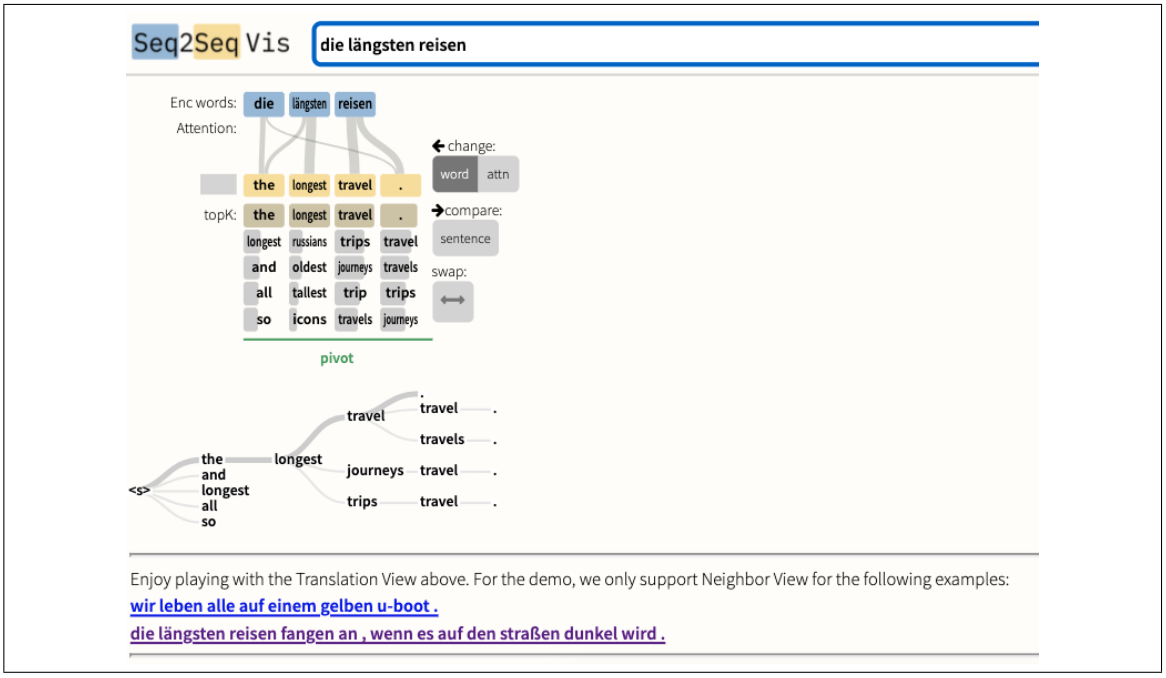

Figura 3.2. Ejemplo de la interfaz de Seq2Seq-Vis para explicar la traducción de una frase en alemán a inglés. Foto extraída de la demo provista por los autores. Recuperada el 23 de noviembre del 2021 de `http://ganter.ailab.res.ibm.com/s2s/client/index.html?in=die%20l%C3%A4ngsten%20reisen`.

Dentro de las visualizaciones diseñadas específicamente para explicar los resultados de un modelo, estas se pueden dividir en dos clases. La primera clase corresponde a las visualizaciones que ayudan a interpretar sistemas no supervisados, es decir, sistemas que no reciben una etiqueta asociada a cada dato durante el proceso de entrenamiento, sino que realizan su tarea solo utilizando los datos de entrada (Lin et al., 2018; Cavallo y Demiralp, 2019; J. Wang, Gou, Shen, y Yang, 2019; Chatzimparmpas, Martins, y Kerren, 2020; Berger, 2020; Wentzel et al., 2020; El-Assady, Sperrle, Deussen, Keim, y Collins, 2019; Cavallo y Demiralp, 2018). Por ejemplo, el trabajo de El-Assady et al. (2019) presenta visualizaciones que revelan cómo funcionan modelos de tópicos, y Cavallo y Demiralp

(2018) proponen un sistema que le permite al usuario cambiar la entrada y salida de una reducción de dimensionalidad y examinar el efecto de estos cambios. La segunda clase corresponde a visualizaciones para sistemas supervisados, es decir, sistemas que reciben el dato y la etiqueta a la que deben llegar durante el proceso de entrenamiento. Esta clase se puede dividir en visualizaciones para explicaciones locales y globales (Liang et al., 2020; Xie, Chen, Kao, Gao, y Chen, 2020a; Strobelt et al., 2018b; Kahng, Andrews, Kalro, y Chau, 2018; Kahng, Thorat, Chau, Viegas, y Wattenberg, 2019; S. Liu et al., 2019; Hohman, Park, Robinson, y Chau, 2020; J. Zhao, Dai, Xu, y Ren, 2020; Ming et al., 2017). Muchas de estas visualizaciones no ofrecen una validación de su diseño, la cual podría involucrar al experto en IA y a los diseñadores, y podría evaluar si el gráfico elegido es óptimo para la tarea. El siguiente ejemplo ilustra este punto: LSTMVis (Strobelt et al., 2018b) presenta varias visualizaciones para comprender la pregunta de alto nivel: "*What information does an RNN capture in its hidden feature-states?*". Cómo se observa en la Figura 3.3, la paleta de colores de este sistema no es apta para algunos tipos de daltónicos y esto afecta la accesibilidad de las visualizaciones. Un diseñador puede seleccionar una paleta segura para daltónicos o advertir sobre las limitaciones de accesibilidad que pueden afectar algunas tareas visuales.

Finalmente, existen diversos sistemas visuales de XAI que incluyen las decisiones de diseño de la visualización. (Kwon et al., 2019; Huang et al., 2021; Strobelt et al., 2019; Ming et al., 2020; Wexler et al., 2019; Ma et al., 2020; Sahoo y Berger, 2020; Das et al., 2020; Marai et al., 2019; X. Zhao et al., 2019; Ming et al., 2019; Z. J. Wang et al., 2021; Ming et al., 2017). Estos trabajos incluyen una justificación del diseño de su sistema para validar la visualización, pero cada sistema lo justifica con una estructura diferente. Las justificaciones se presentan generalmente en dos partes:

1. Los autores declaran lo que el usuario quiere lograr. Por ejemplo, los autores de Seq2Seq-Vis (Strobelt et al., 2019) presentan metas y tareas que ayudan a lograr

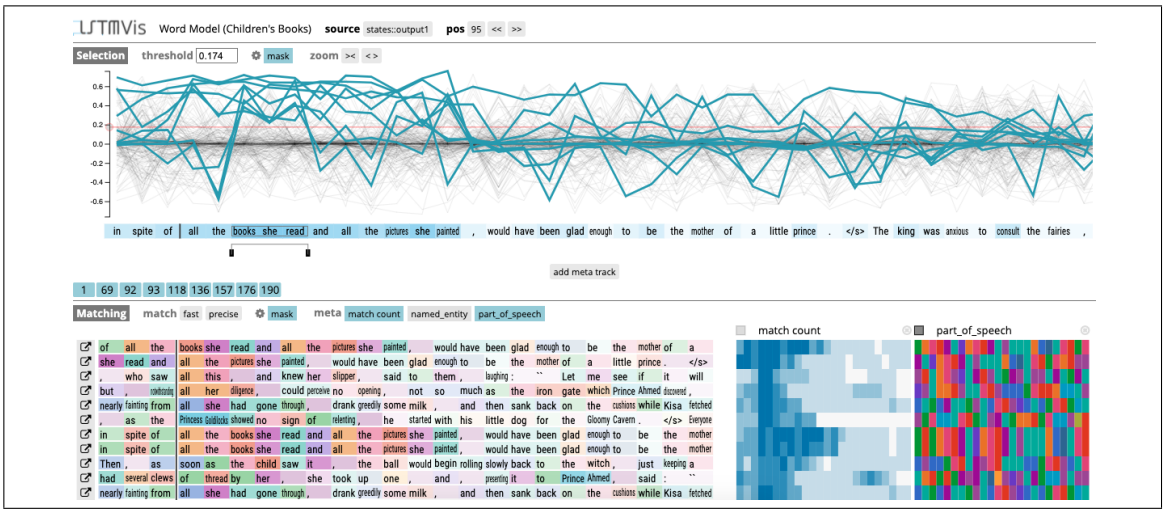

Figura 3.3. Ejemplo de la interfaz de LSTMVis para analizar una LSTM. Foto generada a partir del código fuente provisto por los autores. Recuperada el 23 de noviembre del 2021 de `https://github.com/HendrikStrobelt/LSTMVis`.

esas metas. No obstante, las tareas son presentadas como requerimientos de *software* y no como tareas visuales. Otro ejemplo similar es RuleMatrix, (Ming et al., 2019). Para diseñar esta interfaz (Figura 3.4), los autores presentan preguntas que el usuario busca responder con el apoyo de la visualización y presentan requerimientos a cumplir, pero estos requerimientos están escritos bajo los mismos términos que el *framework* de Munzner.

2. Los autores presentan el diseño de sus sistemas visuales. Describen cómo cada componente puede cumplir con los objetivos de los usuarios previamente definidos. Para algunos componentes de la visualización, los autores proporcionan ejemplos de posibles gráficos con su respectiva argumentación para descartarlos.

Si bien estos trabajos presentan partes del proceso que siguieron, no brindan pautas para reproducir su proceso de diseño en otro problema o contexto. Además, como se explicó anteriormente, utilizan diferentes formas para presentar el proceso que siguieron. Dado lo anterior, uno de los objetivos de este trabajo es proporcionar un *framework* que sintetice y unifique la forma de realizar y presentar las decisiones de diseño de estos trabajos. De esta forma, es posible refinar y organizar estas etapas de diseño.

Figura 3.4. Ejemplo de la interfaz de RuleMatrix para explicar un clasificador de plantas. Foto extraída de la demo provista por los autores. Recuperada el 23 de noviembre del 2021 de `http://iml-test2.herokuapp.com/#rule-surrogate-iris-svm`.

CAPÍTULO 4. OBJETIVOS Y SOLUCIÓN PROPUESTA

En el área de IA, ha surgido una gran relevancia de proveer sistemas explicables (Shneiderman, 2020; Wright et al., 2020). De forma paralela, el uso de visualizaciones se ha vuelto una herramienta con gran potencial para ayudar a presentar las explicaciones (Reyes et al., 2020). Sin embargo, investigaciones demuestran que todavía falta una conexión entre el conocimiento de los expertos en IA/XAI y los diseñadores de visualizaciones (Abdul et al., 2018). Por este motivo, el objetivo general de este trabajo consiste en crear un *framework* inspirado en uno ya existente para que ayude a cerrar la brecha entre los diseñadores y los expertos en IA. En particular, este nuevo *framework* (`VD4XAI`) permitirá analizar y diseñar de forma sistemática y justificada visualizaciones en el dominio de XAI. Como primera versión del *framework*, este estará acotado a explicaciones locales y para tareas de clasificación y recomendación. Para lograr este objetivo, el trabajo de Tamara Munzner (2015) es utilizado como base para su extensión. En consecuencia, este trabajo plantea dos hipótesis:

1. Incluir un nuevo espacio de tareas de XAI al espacio de diseño del *framework* de Tamara Munzner permitirá un análisis íntegro de las visualizaciones de XAI para explicaciones locales, considerando las dimensiones visuales y de XAI.
2. Con una extensión de los pasos del modelo anidado de Tamara Munzner, será posible diseñar de forma sistemática y justificada visualizaciones de XAI para explicaciones locales.

Cómo se mencionó anteriormente, en este trabajo se propone `VD4XAI`, una extensión del *framework* de Munzner para poder ser aplicado en el dominio de XAI. Este trabajo se centra en el análisis y diseño de visualizaciones de XAI y no en el proceso de IA o ML, que ya ha sido abordado por modelos anteriores (Fails y Olsen, 2003; Dudley y Kristensson, 2018). Por lo tanto, `VD4XAI` presenta las siguientes suposiciones: (i) el sistema de IA que se explicará ya ha sido entrenado y (ii) que el propósito de la visualización de XAI es

explicar el resultado de una clasificación o recomendación para un sistema de IA a partir de una instancia de datos en particular (explicación local). Para lograr analizar y diseñar estas explicaciones visuales, VD4XAI consiste en extender cuatro puntos del trabajo original: (i) precisar los roles presentes en el proceso de diseño, (ii) agregar un nuevo espacio con vocabulario específico de XAI, (iii) agregar nuevos niveles en el modelo anidado y (iv) definir la interacción de roles en las diferentes niveles de VD4XAI.

4.1. Roles de usuarios

La primera extensión consiste en precisar los roles de los usuarios para las diferentes actividades consideradas en el *framework*. Originalmente, el modelo anidado presenta dos roles: usuario final y diseñador. En esta extensión se incluyen dos roles: **experto en IA** y **experto en dominio de aplicaciones**. Es importante tener en consideración que, en algunos casos, una misma persona puede asumir varios roles. Por ejemplo, el experto en el dominio también puede ser el usuario final de una aplicación de XAI. En aras de la generalización, este trabajo los considerará como sujetos diferentes para VD4XAI.

Por un lado, el objetivo de una visualización de XAI es, en la mayoría de los casos, explicar el sistema de IA (explicación global) o la respuesta dada por un sistema de IA (explicación local). Dado que los métodos de IA y XAI pueden tener varios niveles de complejidad, es imperante incluir el rol de experto en IA en la definición de tareas, codificación visual a usar, interactuar con el sistema de IA y validaciones posteriores. En particular, el experto en IA puede extraer los datos del método de XAI, validar el uso correcto del sistema de IA y verificar que el método de XAI se pueda aplicar al sistema de IA. Por ejemplo, si un sistema corresponde a una regresión lineal, el experto en IA sabe que no es posible utilizar el mecanismo de atención (Larochelle y Hinton, 2010) como método de XAI. O si el sistema era un *transformer* (Vaswani et al., 2017) y se utiliza un mecanismo de atención para explicar la clasificación de un documento, el experto en IA sabe de qué parte del sistema debe extraer los valores de atención.

Por otro lado, el sistema de IA, en muchos casos, resolverá una tarea en un dominio específico como biología, finanzas, derecho, entre otros, y el conocimiento de estos dominios no es necesariamente manejado por un diseñador, el experto en IA o el usuario final. Esto implica que se vuelve necesario incluir a expertos en el dominio para tomar decisiones en función de las visualizaciones propuestas por el diseñador de visualización junto al experto en IA. Por ejemplo, explicar la respuesta de un sistema de IA que recibe una imagen de rayos X e identifica una enfermedad en el paciente, requiere un cierto nivel de conocimientos médicos y el usuario final, es decir, el paciente, no necesariamente posee dichos conocimientos. Por esta razón, un usuario experto en el dominio, como un médico, puede justificar y validar decisiones de diseño que permitirán comunicar eficazmente la información al usuario final.

4.2. Espacio de XAI

En la segunda extensión, se define un nuevo espacio con vocabulario específico al dominio de XAI. Este espacio se denominó espacio de XAI y permite analizar de forma sistemática los elementos relacionados al aspecto XAI de una visualización. Para lograr este análisis, se identificaron cinco dimensiones que se deben considerar: (i) el tipo de datos que utiliza el sistema de IA y el método de XAI, (ii) la tarea de XAI a resolver, (iii) la estrategia de explicación utilizada, (iv) el método de XAI y (v) el sistema de IA que puede utilizar la visualización. Se propuso separar el espacio de XAI de las tres preguntas fundamentales del modelo anidado de Munzner porque el diseñador usará la salida del Espacio de XAI para iniciar el proceso de diseño de la visualización de XAI. Por ejemplo, la abstracción de datos se aplicará a los datos obtenidos del método de XAI. Además, el rol principal para definir cada espacio es diferente. En el espacio de XAI está el experto en IA y en el modelo anidado de Munzner está el diseñador. A continuación se detalla cada dimensión del espacio de XAI:

1. **Tipo de dato**: se determina el tipo de dato utilizado por el sistema de IA y el método de XAI. Por ejemplo, texto, imágenes, datos tabulares, grafos, entre otros.
2. **Tarea de XAI**: se determina una o más tareas que se esperan resolver con las explicaciones. Estas tareas se basan en las preguntas propuestas en el trabajo de Gunning (2016). Las posibles tareas identificadas en esta dimensión son: (i) entender por qué el sistema de IA ofrece cierta respuesta, (ii) entender por qué el sistema IA no responde otra cosa, (iii) comprender cuándo tiene éxito el sistema de IA y (iv) comprender cuándo falla el sistema de IA.
3. **Tipo de explicación**: se determinan una o más estrategias utilizadas por la visualización para generar las explicaciones que resuelven las tareas de XAI. Las posibles estrategias a elegir son (i) importancia de la característica, (ii) analogías (ejemplos basados en similitudes o ejemplos de adversarios), (iii) contrafactual o (iv) reglas.
4. **Método de XAI**: se determinan los métodos computacionales que aplican las estrategias mencionadas en la dimensión anterior para generar las explicaciones. La salida de este método se utilizará en el modelo anidado para diseñar la visualización de XAI. No obstante, si el sistema de IA es uno tradicionalmente explicable, como un árbol de decisión, y el experto en IA utiliza dicho sistema para generar la explicación, se puede omitir esta dimensión porque no se requiere usar otro método computacional, como LIME (Ribeiro et al., 2016), SHAP (Lundberg y Lee, 2017) o el mecanismo de atención (Larochelle y Hinton, 2010), para generar la explicación. En otras palabras, un modelo tradicionalmente explicable se podría considerar, en esta dimensión, como el método de XAI.
5. **Sistema de IA**: se determinan los posibles sistemas de IA que se pueden utilizar junto con la visualización. En este punto, ya pueden ser modelos explicables como una regresión o un árbol de decisión, o pueden ser modelos más complejos como una red neuronal o un *random forest* (Breiman, 2001).

4.3. Extender modelo anidado

La tercera extensión de VD4XAI consiste en agregar tres niveles al modelo anidado. Como se muestra en la Figura 4.1, esta extensión se posiciona entre el primer nivel (situación del dominio) y el segundo nivel (abstracción de datos y tareas). Esta extensión guía al experto en IA en el proceso de definir el espacio de XAI de la visualización. Para guiar este proceso, se proponen dos niveles en los que el experto en IA toma decisiones, seguido de un nivel de validación. En estos niveles, se identifican posibles amenazas que pueden aparecer, como el uso de un método de XAI que no aborda la tarea de XAI, y se proponen alternativas para mitigar estas amenazas, por ejemplo, utilizar la literatura para justificar la elección del método de XAI. En el nivel de validación, el experto en IA y el diseñador deben analizar el resultado entregado por el método de XAI y justificar la necesidad de utilizar una visualización para abordar las tareas de XAI. Finalmente, se modifican dos validaciones propuestas en el *framework* original para adaptarlas a este nuevo dominio. En la Figura 4.2 se presenta un resumen de las amenazas y validaciones para cada fase de VD4XAI.

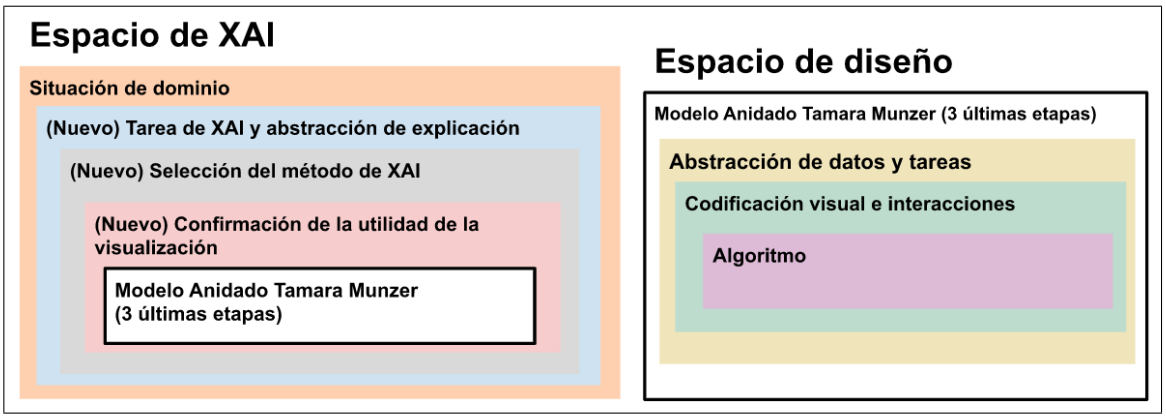

Figura 4.1. Niveles de VD4XAI.

Tarea de XAI y abstracción de explicación

Este nivel recibe el problema que el usuario busca resolver, los datos disponibles y el sistema de IA utilizado. Con esta información, el experto en IA puede identificar una o más tareas XAI y el tipo de explicación para resolver esas tareas. Una posible amenaza que surge a este nivel es que el tipo de explicación no es apropiado para el problema del usuario. Por ejemplo, si se utiliza una explicación del tipo contrafactual para explicar el resultado, pero el usuario no comprende este tipo de explicación para el problema particular que tiene, será necesario iterar el tipo de explicación que se está utilizando. Para mitigar esta amenaza, se proponen tres validaciones: una inmediata y dos posteriores a la implementación de la visualización de XAI.

1. **Validación inmediata**: consiste en justificar el tipo de explicación con la literatura. Por ejemplo, buscar casos donde se presenta la misma tarea de XAI, y exista evidencia de que el tipo de explicación seleccionada fue efectiva para el usuario final.
2. **Validaciones después de implementar la visualización de XAI**: la primera validación corresponde a realizar pruebas con usuarios finales y expertos en el dominio para incluir evidencia anecdótica de la utilidad de la explicación visual. La segunda validación consiste en un estudio de campo para documentar el uso de la visualización de XAI. Originalmente, estas dos validaciones estaban incluidas en el nivel de “Abstracción de datos y tareas” y solo validaban la visualización. No obstante, se trasladaron dichas validaciones a este nuevo nivel para agregar el aspecto de explicabilidad en los estudio. De este forma, solo se realiza un estudio de campo o pruebas con los usuarios finales y expertos en el dominio de aplicación para mitigar ambos problemas: (i) abstracción de datos o tarea equivocada, y (ii) una mala elección de explicación.

Selección del método de XAI

Este nivel recibe el tipo de explicación y la tarea de XAI. Con esta información, el experto en IA define el método de XAI que permita generar la explicación que cumpla con las tareas de XAI. A diferencia del nivel "algoritmo" propuesto en el *framework* original, este nivel consiste en determinar el método computacional que genera la explicación y usar los datos que entrega dicho algoritmo para los siguientes niveles. Mientras que en el nivel de "algoritmo" se determinan la forma para codificar los datos y sus interacciones en la visualización.

Además, en este nivel, el experto en IA podría definir un método de XAI incorrecto para la tarea de XAI y dicha amenaza debe mitigarse en este nivel. Por ejemplo, si las tareas de XAI son (i) entender por qué el sistema de IA ofrece cierta respuesta y (ii) entender por qué el sistema IA no responde otra cosa, un posible tipo de explicación es la importancia de la característica y visualizar dicha relevancia. La amenaza puede surgir si el experto en IA decide utilizar el mecanismo de atención (Larochelle y Hinton, 2010). Este método solo retorna la relevancia de las características para la salida proporcionada al sistema de IA, lo que conlleva a no poder abordar la segunda tarea de XAI. En consecuencia y de manera análoga al nivel anterior, se propone incluir una validación inmediata para mitigar esta amenaza. Esta validación consiste en utilizar evidencia en la literatura que respalde el uso del método de XAI elegido para las tareas de XAI definidas en el nivel anterior.

Confirmación de la utilidad de la visualización

Este último nivel corresponde a uno de validación. En este nivel, se utilizan los datos provistos por el método de XAI para justificar la necesidad de utilizar una visualización. Para lograr este objetivo, se propone que el experto en IA y el diseñador analicen la disponibilidad de los datos proporcionados por el método de XAI y evalúen si esos datos permiten continuar con los siguientes niveles de `VD4XAI`. Por ejemplo, si la estrategia de explicación para un clasificador de texto son analogías, es posible proporcionar textos similares para explicar la predicción del sistema de IA sin necesidad de visualización. En ese caso, es imposible seguir aplicando el *framework* para crear una visualización. En consecuencia, será necesario cambiar uno o más elementos de los definidos previamente o no utilizar una visualización.

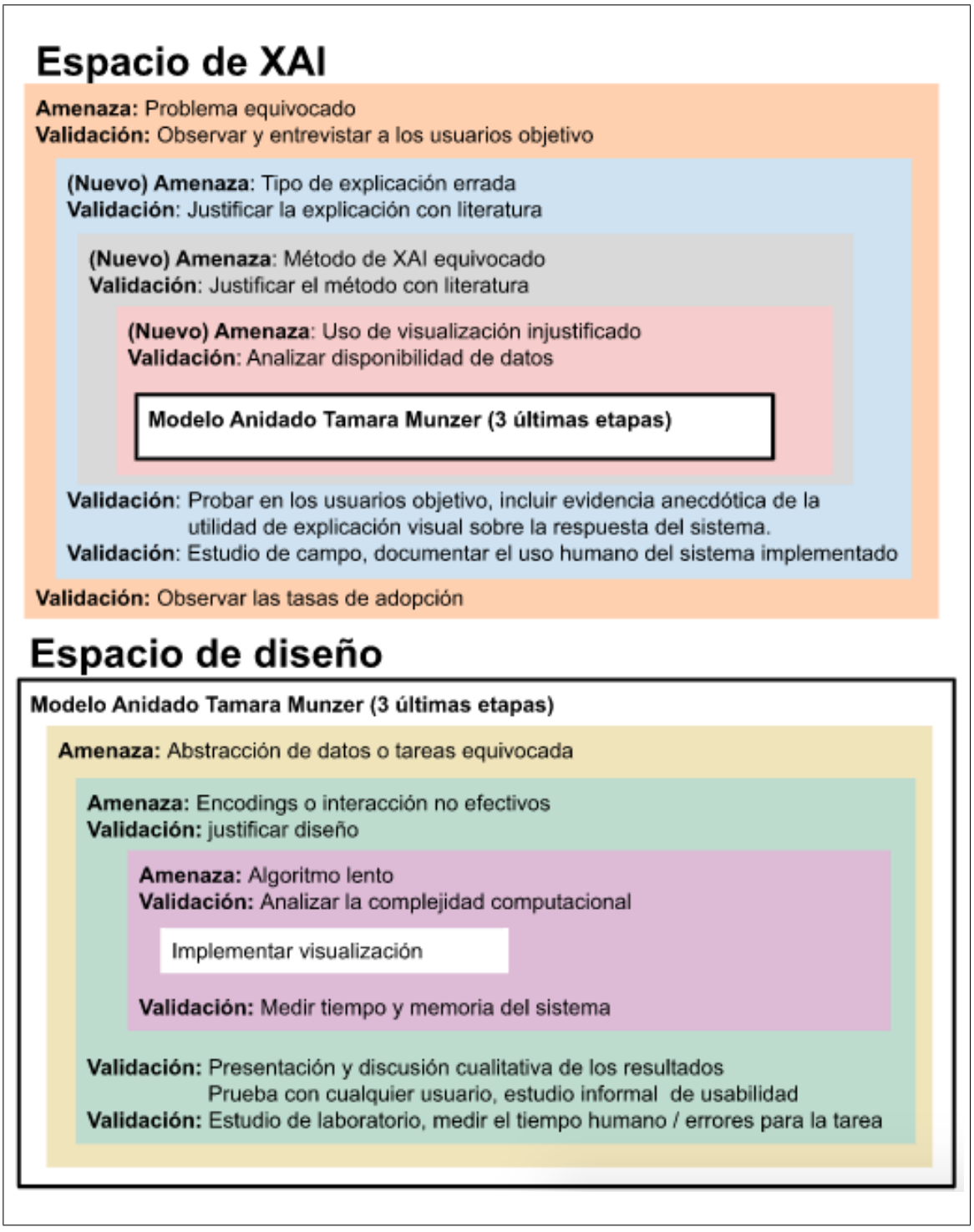

Figura 4.2. Amenazas y validaciones en VD4XAI.

4.4. Interacción de los roles

La interacción entre los nuevos roles (experto en IA y experto en el dominio de aplicación), el diseñador y el usuario final es clave para elegir las codificaciones visuales e interacciones más efectivas. Por lo tanto, la última extensión de este *framework* consiste en presentar la interacción entre los diferentes roles en cada nivel de VD4XAI. La figura 4.3 resume la participación de cada rol dentro de los niveles de este *framework*.

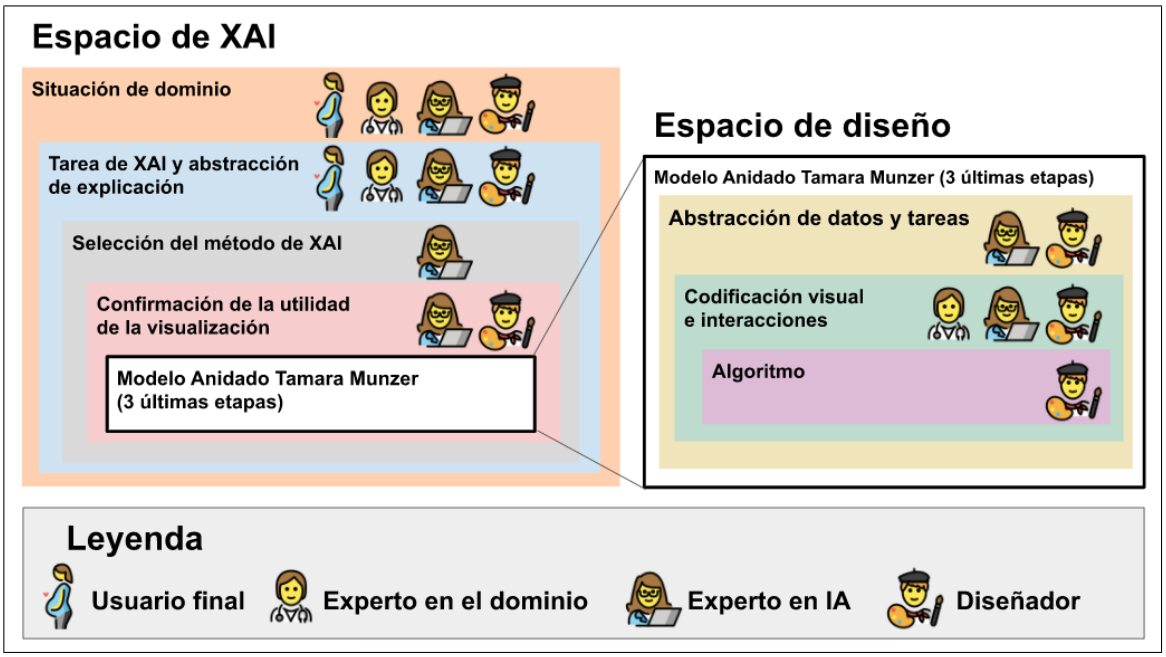

Figura 4.3. Participación de nuevos y originales roles en VD4XAI.

1. El primer nivel (Situación de Dominio en la Figura 4.3) describe la situación del dominio específico. Este nivel abarca el grupo de usuarios objetivo, su dominio de interés, sus preguntas y sus datos. En el modelo anidado original, el usuario final y el diseñador ya estaban involucrados. No obstante, el problema y los datos están dentro de un contexto de una aplicación de XAI en un dominio específico, lo que implica que es necesario involucrar los dos nuevos roles. De esta manera, el usuario de IA puede comprender el problema de XAI que se abordará y el

experto en el dominio puede proporcionar más información sobre el dominio de la aplicación y sus datos.

2. En el segundo nivel (Tarea de XAI y abstracción de explicación en la Figura 4.3), el experto en IA determina las tareas de XAI y el tipo de explicación. Luego, tras haber construido la visualización de XAI, se solicita que los usuarios finales y los expertos en el dominio de aplicación realicen un estudio de campo y/o estudio de usuario para validar la efectividad de la explicación visual. En consecuencia, se requiere la participación del experto de IA para determinar las tareas de XAI, del diseñador para confeccionar el estudio de usuario y de los usuarios finales y expertos en el dominio de aplicación para participar de los estudios.
3. El tercer nivel (Selección del método de XAI en la Figura 4.3) consiste en determinar el método de XAI (LIME, SHAP, entre otros) más adecuado para las tareas de XAI. Además, este nivel no considera una validación con los usuarios finales o expertos en el dominio de aplicación. Por este motivo, solo es necesaria la participación del experto en IA.
4. En el cuarto nivel (Confirmación de la utilidad de la visualización en la Figura 4.3) se debe justificar la utilidad de la visualización en función del tipo de datos que entrega el método XAI definido en el tercer nivel. Para lograr este objetivo, solo se requiere de la participación del experto en IA y el diseñador para que en conjunto realicen esta tarea.
5. El quinto nivel (Abstracción de datos y tareas en la Figura 4.3) consiste en abstraer los datos y las tareas visuales a los conceptos proporcionados por el modelo. En el modelo anidado original, solo el diseñador era quien se encargaba de realizar tal tarea. No obstante, ahora los datos provienen de un método de XAI, lo que implica que el diseñador puede tener dificultades para entender la semántica de estos datos. En consecuencia, se requiere la participación del experto en IA y del diseñador en este nivel para guiar el correcto proceso de entender y abstraer los datos y tareas.

6. En el sexto nivel (Codificación visual e interacciones en la Figura 4.3), originalmente el diseñador tomaba decisiones en torno al diseño de la visualización. No obstante, ahora la visualización es de XAI y aplicada a un dominio específico, por lo que es necesario incluir los dos nuevos roles para que participen en la validación de las decisiones de diseño:
   - El experto en IA puede validar que los datos relacionados con el sistema o método de XAI se interpretan correctamente en la visualización. Por ejemplo, si los datos corresponden a la probabilidad que da el sistema de IA para clasificar un documento, el experto en IA puede validar que la codificación visual nos permite entender que este valor es una probabilidad.
   - El experto en el dominio de aplicación puede validar que el diseño se centra en los datos relevantes del problema y es comprensible en el dominio específico. Por ejemplo, si se está explicando la clasificación de la imagen de rayos X y se desea enfatizar un órgano, el experto en el dominio puede proporcionar orientación sobre la forma más eficaz de resaltar un órgano en la radiografía y asegurar que el usuario final pueda comprenderlo.
7. El último nivel (Algoritmo la Figura 4.3) consiste en construir la visualización, por lo que no es necesario involucrar a los nuevos roles para realizar esta tarea, es decir, solo es necesario el diseñador o la persona a cargo de programar/construir la visualización en este nivel.

CAPÍTULO 5. METODOLOGÍA

En este capítulo se describe la metodología utilizada para el desarrollo de este trabajo.

5.1. Validación de propuesta

Para validar VD4XAI y las hipótesis planteadas en el capítulo 4, se realizaron dos etapas de validación, una por hipótesis respectivamente.

- La primera etapa consistió en utilizar VD4XAI para ***analizar*** diferentes visualizaciones de XAI ya existentes. Estas visualizaciones debían estar restringidas a ser explicaciones locales y para tareas de clasificación y/o recomendación. Tras una inspección de diferentes visualizaciones existentes, se seleccionaron los siguientes dos casos: una visualización confeccionada con el método LIME para clasificación de texto, y otro caso donde se explican visualmente las recomendaciones de un sistema utilizando el mecanismo de atención.
- Para validar la segunda hipótesis, se utilizó VD4XAI para un caso de uso. Este caso consistió en la **confección** de un sistema de clasificación de texto y la elaboración de una visualización para explicar la respuesta del sistema. En particular, se recopilaron diferentes títulos y *abstract* de documentos médicos que debían ser clasificados entre 5 categorías distintas. Posteriormente, se utilizó VD4XAI para diseñar justificadamente una visualización que permita explicar la respuesta del sistema. Este caso de uso incluyó un estudio de usuario como etapa final para validar la visualización.

5.2. Conjunto de datos para la segunda validación

Para el estudio de usuario, incluido en el caso de uso, se utilizó una base de datos provista por la fundación Epistemonikos. Esta base de datos consiste en un sistema colaborativo multilingüe donde los médicos clasifican artículos bajo la práctica de Medicina basada en evidencia. Para esta ocasión, la base de datos utilizada cuenta con aproximadamente 46.000 títulos y *abstract* de documentos médicos. Estos textos fueron clasificados, según una amplia búsqueda realizada por Epistemonikos, entre 4 categorías posibles: revisión sistemática (*Systematic review*), síntesis amplia (*Broad synthesis*), ensayos randomizados (*Randomised trials*) o ensayos no randomizados (*Non-randomised trials*). Posteriormente, todos los documentos fueron clasificados por un sistema de IA (XLNET) en 5 categorías posibles, las 4 utilizadas por Epistemonikos y una quinta categoría denominada *"excluded"*. Esta última categoría implica que el documento debe ser excluido de la base de datos. En la figura 5.1 se observa la distribución de los textos en función de las categorías indicadas por Epistemonikos y las calculadas con el sistema de IA (XLNET). Se puede apreciar que el sistema de IA está cargado a clasificar los documentos como *Systematic Review*.

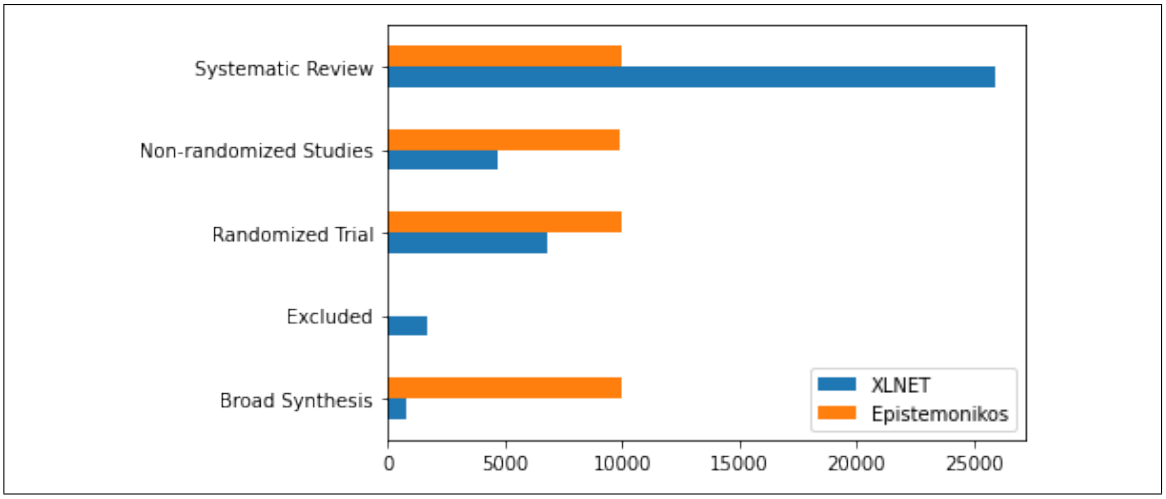

Figura 5.1. Distribución de documentos médicos según clasificación realizada por Epistemonikos y por el sistema de IA (XLNET).

5.3. Diseño de Estudio de Usuario para segunda validación

Para finalizar la validación de la primera hipótesis, se diseñó un estudio de usuario usando el diseño *Within-Subjects Design*, es decir, cada usuario testea cada condición, en este caso, cada visualización posible. Se utilizó la metodología *Latin square* para no tener un sesgo correspondiente al orden de las visualizaciones a estudiar. Este estudio de usuario tiene la función de mostrar diferentes visualizaciones para explicar la respuesta de un sistema de IA, registrar las acciones y respuestas dadas por el usuario, y medir el tiempo requerido para llevar a cabo esta tarea. Luego, para cada visualización, se mide su carga cognitiva en base al test NASA-TLX (Hart y Staveland, 1988) y preguntas en torno a la efectividad de la explicación visual. Finalmente, luego de terminar el estudio, se le indica una última encuesta al usuario para tener más información sobre su opinión en relación al sistema de IA y la explicación visual recibida. Las decisiones de diseño tomadas para el estudio de usuario y el flujo que debe seguir el usuario se detallan en el Capítulo 6.

CAPÍTULO 6. USO DE VD4XAI Y RESULTADOS

6.1. Análisis de casos existentes

Para validar la Hipótesis 1, la cual dice que incluir un nuevo espacio de tareas de XAI al espacio de diseño del *framework* de Tamara Munzner permitirá un análisis íntegro de las visualizaciones de XAI para explicaciones locales, considerando las dimensiones visuales y de XAI, se analizaron dos visualizaciones con `VD4XAI`. El primer caso consiste en una visualización muy popular dentro de las herramientas de XAI, correspondiente al paquete LIME de Python. El segundo caso aborda una visualización utilizada para explicar la recomendación de contenidos dentro de un juego multijugador. Para la presentación de estos análisis, se destacó en **negro** conceptos claves utilizados en el *framework* de Munzner (2014) para evidenciar el uso de dicho trabajo como base del análisis para la dimensión visual. Finalmente, para cada análisis se identificó debilidades en la visualización utilizada actualmente y se proponen mejoras basadas en tareas de XAI o tareas visuales.

6.1.1. Clasificación de texto con LIME

Como se explicó en el marco teórico, el método LIME (Ribeiro et al., 2016) utiliza los coeficientes de ponderación de una regresión para describir la importancia de algunas palabras cuando se clasifica un documento. Con estos pesos, el paquete LIME[1] de Python, que implementa este método, ofrece la siguiente visualización:

[1] `https://github.com/marcotcr/lime`

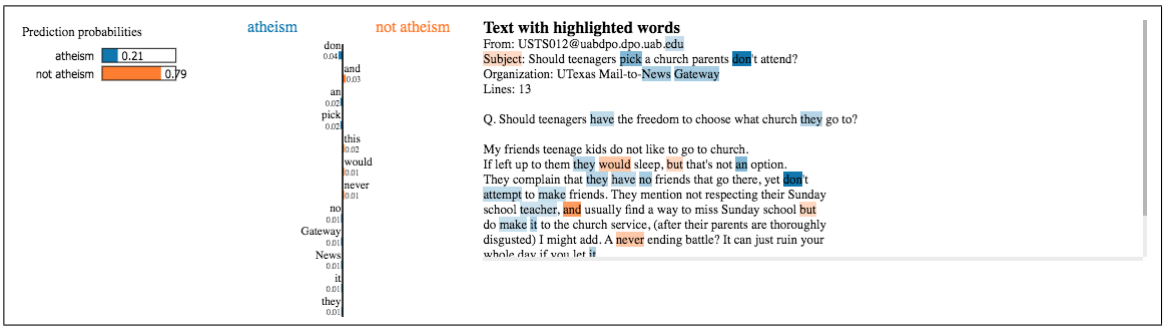

Figura 6.1. Visualización de la biblioteca LIME para explicar una clasificación binaria de un documento. Se pintó en naranjo las palabras que fueron más importantes para que el modelo clasificara el texto como *not atheism* y en azul las palabras más importantes para clasificar el texto como *atheism*. En este caso, el modelo clasificó el texto como *not atheism* y esta clasificación se debió principalmente a la presencia de las palabras *and*, *never* y *Subject*.

Como muestra la Figura 6.1, la visualización provista por el paquete LIME se compone de 3 gráficos: uno de barras con las probabilidades de las clases proporcionadas por el sistema de IA, un gráfico de barras con los pesos de las palabras seleccionadas y un gráfico de texto en donde el color de fondo codifica los pesos. A continuación se presenta el análisis de esta visualización utilizando `VD4XAI`:

1. Espacio de XAI

1. **Tipo de dato**: Texto.
2. **Tarea de XAI**: (a) Entender por qué el sistema de IA ofrece cierta respuesta & (b) Entender por qué el sistema IA no responde otra cosa.
3. **Estrategias de explicación**: Importancia de características.
4. **Método de XAI**: LIME.
5. **Sistema de IA**: Cualquier sistema de clasificación que reciba un texto y devuelva la probabilidad de las diferentes clases.

2. Abstracción de datos (*What?*)

1. La probabilidad de salida de cada clase, que corresponde a un atributo **cuantitativo** con valores continuos entre 0 y 1.
2. Palabras con sus correspondientes pesos. Pueden ser valores positivos o negativos y corresponden a un atributo **cuantitativo** con orden **divergente**. Estos pesos se interpretan como la importancia de la palabra para la clasificación del documento como clase 1 (peso positivo) o clase 0 (peso negativos).

3. Abstracción de tareas (*Why?*)

1. **Comparar** las probabilidades.
2. **Comparar** la importancia de las palabras para cada respuesta.
3. **Resumir** la importancia de cada palabra.
4. **Identificar** valores atípicos o grupos de palabras según su importancia.

4. Codificación visual (*How?*)

La **yuxtaposición** es utilizada, mediante el uso de tres gráficos, para abordar las cuatro tareas indicadas en las sección anterior.

1. Para el gráfico la izquierda, el **largo de la barra** es utilizada para **codificar** las probabilidades de las clases. El **color** es usado como canal categórico para **mapear** cada posible clase en el sistema: 0 o 1.
2. En el gráfico del centro, la **longitud de la barra codifica** la importancia de la palabra para la salida del sistema. Las barras se encuentran **ordenadas** de forma decreciente en función de su valor absoluto. Además, están **separadas** según el signo del peso: si son positivas, las barras apuntan a la derecha; si son pesos negativos, las barras apuntan a la izquierda. Finalmente, se utiliza el **color** para **mapear** el signo del peso.

3. Para el gráfico de texto, se utiliza la **saturación del color** de fondo, en una escala bidireccional, para **codificar** la importancia por palabra. Además, cada palabra está **organizada** en forma de párrafo para mantener la estructura del texto original.

5. Rediseño de la visualización

La visualización provista por el paquete LIME (Figura 6.1) permite comparar y resumir la importancia de cada característica. Debido a que la visualización utiliza la yuxtaposición para construir un gráfico por cada tarea, quienes buscan realizar ambas tareas simultáneamente, presentar un aumento en la división de su atención (Lobo, Pietriga, y Appert, 2015). En otras palabras, los usuarios que buscan algunas palabras en el texto y necesitan comparar la importancia de dichas palabras en el gráfico de barra, verán su carga cognitiva aumentada por requerir intercalar su atención entre un gráfico y el otro. Para buscar la minimización de la división de atención y lograr efectuar ambas tareas simultáneamente, se rediseñó la visualización.

Usando `VD4XAI`, si se desea realizar ambas tareas visuales sin reducir la carga cognitiva, se está alterando el nivel de "Abstracción de tareas". Por lo tanto, los siguientes niveles deberán ser iterados nuevamente. En este caso, para cumplir con este nuevo requerimiento, se exploró el espacio de diseño propuesto en la literatura, tales como las nubes de palabras o atención neuronal en clasificación de texto (Felix, Franconeri, y Bertini, 2017; Parra et al., 2019). La Figura 6.2 muestra dos codificaciones que podrían ser utilizadas en el rediseño de la visualización. En la Figura 6.2.A, se añade una marca de barra detrás de cada palabra y se utiliza el largo de la barra para codificar importancia de dicha palabra. Mientras que en la Figura 6.2.B, se añade una marca de círculo al lado de cada palabra y se utiliza el tamaño del círculo para codificar la importancia de la palabra. Luego, para distinguir entre pesos negativos y positivos, se puede pintar cada marca en dos colores de acuerdo con el signo del peso y además, se pueden elegir colores a prueba de daltónicos, como el naranja y el azul. Con estos cambios, el gráfico de barra se puede omitir y utilizar

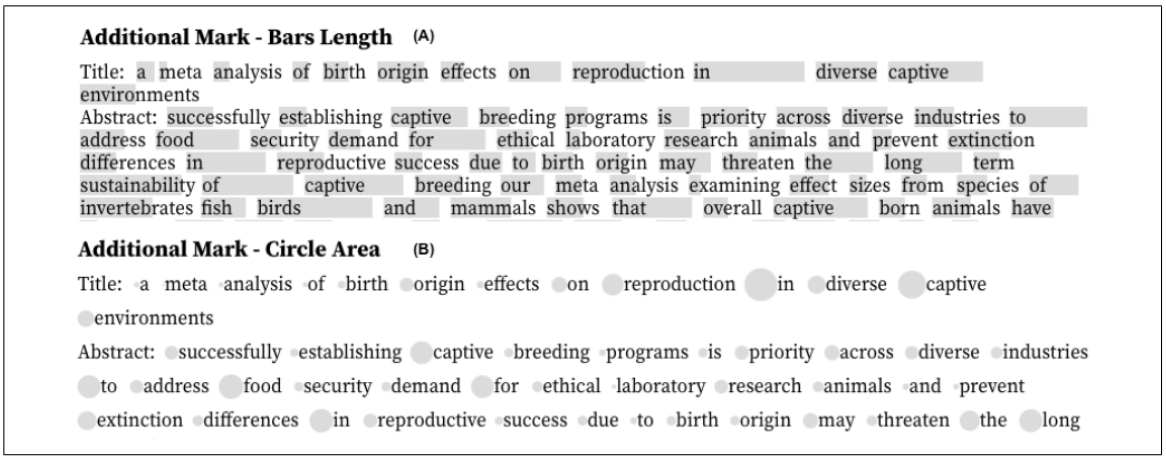

Figura 6.2. Alternativas para codificar la importancia de características para LIME usando (A) barras detrás de cada palabra o (B) círculos al lado de cada palabra (Parra et al., 2019).

solo una visualización de texto para facilitar la comparación de los pesos. De este modo, se logran ejecutar las dos tareas de forma simultánea sin aumentar la división de atención y la carga cognitiva.

Otro caso a analizar es el gráfico de barras del centro que viene con datos ordenados en base al peso absoluto. Es decir, la interfaz ofrece un orden específico para comparar los valores en dicha visualización. Esto produce dos dificultados: (i) el usuario no es libre de elegir la posición de cada característica a comparar en el gráfico de barra, (ii) si el algoritmo llega a utilizar más característica, el gráfico del medio utilizará mucho más espacio en relación a los demás gráficos. En resumen, el diseño no es óptimo cuando la visualización utiliza demasiadas palabras para explicar la clasificación. Para esta situación, si se desea que el usuario final pueda organizar estas características, por ejemplo, qué y cuántas palabras quiere comparar, aplicando `VD4XAI`, será necesario iterar nuevamente el nivel de "Codificación visual e interacciones". Por un lado, una forma de solucionar este requerimiento es utilizar una visualización interactiva donde el usuario puede seleccionar las palabras en el texto y solo desplegar dichas palabras en el gráfico de barra. Por otro lado, el diseñador también podría explorar la literatura para encontrar otros diseños que permitan desplegar estas características, como en *TileBars* (Hearst, 1995), *Ink blots*

(Abbasi y Chen, 2007) o *TextArc* (Paley, 2002). Sin embargo, será necesario validar si estas visualizaciones permiten realizar correctamente las tareas visuales y las tareas de XAI definidas previamente.

Finalmente, utilizando el *framework* propuesto, se desea reducir las dos tareas de XAI a una, en particular, rediseñar la visualización para que cumpla solo con la tarea: *Entender por qué el sistema de IA ofrece cierta respuesta*, además se desea mantener la visualización lo más parecido a la original. Para esto, es necesario revisar nuevamente los siguientes niveles: "Selección del método de XAI", "Confirmar la utilidad de la visualización", "Abstracción de datos y tareas", y "Codificación visual e interacciones". En esta situación, el método LIME todavía es capaz de solucionar la tarea de XAI indicada. Por lo tanto, los dos primeros niveles mencionados anteriormente ("Selección del método de XAI" y "Confirmar la utilidad de la visualización") no requieren ser iterados nuevamente, solo queda procurar que las tareas visuales y la codificación posteriormente elegida estén enfocadas a explicar la respuesta del sistema, esto es, explicar la clase con mayor probabilidad. Dado lo anterior, si se busca abordar solo una tarea de XAI y mantener la visualización con una mínima cantidad de modificaciones, los cambios sugeridos son: (i) el gráfico de probabilidad se puede reemplazar a un texto que indique la probabilidad de la clase más probable, (ii) el gráfico de barras solo debe tener las barras asociadas a una clase, y (iii) el cuadro de texto debe utilizar una escala de colores secuencial y no una divergente. En resumen, los cambios sugeridos implican modificar la visualización para conservar únicamente la información y codificación asociada a la clase más probable del sistema, eliminando toda codificación relacionada a la clase menos probable.

6.1.2. Recomendación de ítems

El siguiente caso trata sobre explicación visual de la respuesta entregada por un sistema de recomendación. El objetivo del sistema es recomendar ítems en el juego en línea llamado *League of Legends*. El sistema utiliza una arquitectura llamada *transformer* y

que fue modificada para recomendar elementos en un contexto de equipos (Villa, Araujo, Cattan, y Parra, 2020). Los datos de entrada corresponden a la información de 10 personajes: 5 por equipo, rojo y azul, en donde el equipo azul es aquel que estará sujeto a las recomendaciones entregadas por el sistema de IA. Dada esa información, el sistema recomienda hasta seis ítems para cada personaje del equipo. Con estos ítems, el personaje mejora sus probabilidades de ganar el juego. A continuación se presenta la visualización propuesta por los autores para explicar la recomendación dada a los usuarios utilizando el mecanismo de atención:

Figura 6.3. Visualización para explicar la recomendación de ítems con un *Transformer for Team-aware Item Recommendation architecture* (Villa et al., 2020). En la imagen se puede observar que para las recomendaciones del primer personaje (primera fila), el sistema puso más atención a los personajes de la columna 1 y 5, los cuales corresponden a Garen y Sona respectivamente. En otras palabras, la presencia de esos 2 personajes fueron los que más impulsaron al sistema de recomendar esos 6 ítems. De forma análoga, para la recomendación del último personaje (última fila), el sistema puso la mayor atención en el personaje de la segunda columna, Syndra, para generar la recomendación de los ítems.

Como se muestra en La Figura 6.3, esta se compone a la izquierda de una grilla que codifica los pesos de atención utilizados por el sistema recomendador, y a la derecha se presentan el conjunto de ítems recomendados para cada personaje del equipo azul. A continuación se presenta el análisis de la visualización utilizando `VD4XAI`:

1. Espacio de XAI

1. **Tipo de dato**: Lista de datos tabulares.
2. **Tarea de XAI**: Entender por qué el sistema de IA ofrece cierta respuesta.
3. **Estrategias de explicación**: Importancia de características.
4. **Método de XAI**: Mecanismo de atención.
5. **Sistema de IA**: *Transformer for Team-aware Item Recommendation architecture (TTIR)*.

2. Abstracción de datos (*What?*)

Campo escalar donde cada celda contiene la atención prestada por el sistema de IA a los datos de entrada para generar la recomendación. La atención es un atributo **cuantitativo** con un valor de orden **secuencial** entre 0 y 1. Este valor representa la importancia del valor de entrada cuando el sistema entrega la recomendación.

3. Abstracción de tareas (*Why?*)

1. **Resumir** la atención puesta por el sistema de IA.
2. **Identificar** valores atípicos o *clusters* en los valores de atención.

4. Codificación visual (*How?*)

Uso de una **alineación de matriz 2D**. La celda de la matriz **codifica** el valor de atención con un paleta de **colores secuencial**. Cada columna es un dato de entrada, mientras que cada fila es la recomendación que se hace a cada jugador del equipo azul.

5. Rediseño de la visualización

Esta visualización (Figura 6.3) permite una exploración de los valores de atención utilizados por el sistema para la recomendación. Sin embargo, existen otras tareas que el

usuario final podría desear, tales como: (i) identificar valores extremos y (ii) comparar con alta precisión los valores de atención en las diferentes celdas. Cuando se ocupa el color para codificar los valores de atención, el usuario es incapaz de ejecutar las dos tareas mencionadas anteriormente de forma eficiente. Por lo tanto, se va a rediseñar la visualización para permitir la ejecución de estas dos tareas eficazmente.

El rediseño consiste principalmente en alterar las tareas definidas en el nivel de "Abstracción de tareas". Por este motivo, `VD4XAI` indica que se debe iterar las siguientes niveles, es decir, el nivel de "Codificaciones visuales e interacciones". Una propuesta de diseño es cambiar el color por un canal más eficiente para representar datos numéricos. Como tal, una sugerencia podría ser un gráfico de ejes paralelos como el mostrado en la Figura 6.4. En esta visualización, el canal de posición vertical se utiliza para codificar los valores de atención colocados por cada recomendación en una entrada. La posición horizontal separa los datos de entrada y el canal de color asigna cada recomendación. Estas codificaciones permiten al usuario final comparar con mayor precisión e identificar extremos en cada fila y columna, es decir, cumplir con las 2 tareas visuales definidas en este proceso de rediseño.

Adicionalmente, el contexto donde se utiliza esta visualización es mientras el juego está cargando, es decir, existe un tiempo limitado para analizar la visualización antes que el juego empiece. Además, usualmente el usuario no desea ver las recomendaciones de los demás jugadores, solo la de su personaje y una explicación para su recomendación. En este escenario, tareas como ver únicamente la información para el personaje que el jugador está utilizando serán más difíciles de lograr. La visualización actual requiere buscar la fila correspondiente a un personaje específico y luego explorar las celdas en esa fila para encontrar a los personajes que el sistema le puso más atención al momento de generar la recomendación. Por lo tanto, se requiere modificar las tareas visuales para mejorar el tipo de búsqueda que realiza el usuario.

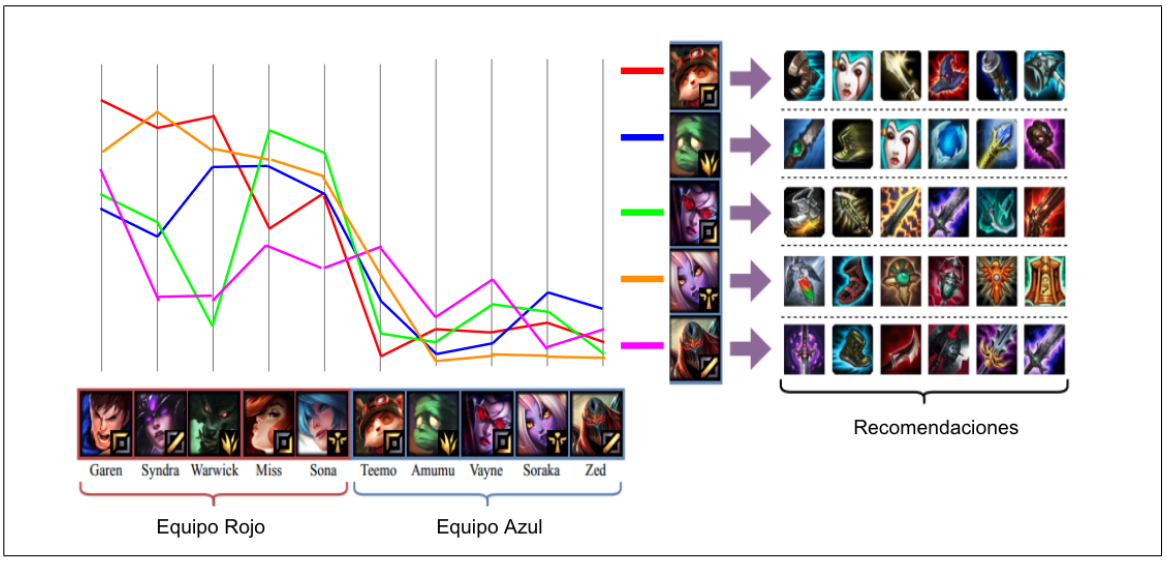

Figura 6.4. Alternativa para comparar los valores de atención y explicar la recomendación de todos los personajes con un *Transformer for Team-aware Item Recommendation architecture*.

Al igual al caso anterior, se está alterando las tareas definidas en el nivel de "Abstracción de tareas", por lo tanto, se debe iterar el nivel de "Codificaciones visuales". La Figura 6.5 presenta una solución para satisfacer la tarea de encontrar el personaje con mayor atención, y solo se muestre información sobre un personaje en específico. En este gráfico interactivo, el usuario selecciona el personaje y la visualización utiliza un gráfico de barra para codificar los valores de atención. Además, las barras se ordenan de mayor a menor. Con esta codificación, el usuario final solo debe seleccionar el personaje que desea y verá únicamente dicha información. Luego, este usuario debe leer el gráfico de derecha a izquierda para identificar los personajes que el sistema le puso mayor atención. Adicionalmente, el usuario final podrá comparar de forma efectiva la magnitud de ese valor con los otros personajes del juego, tarea que no era posible con la visualización propuesta por los autores.

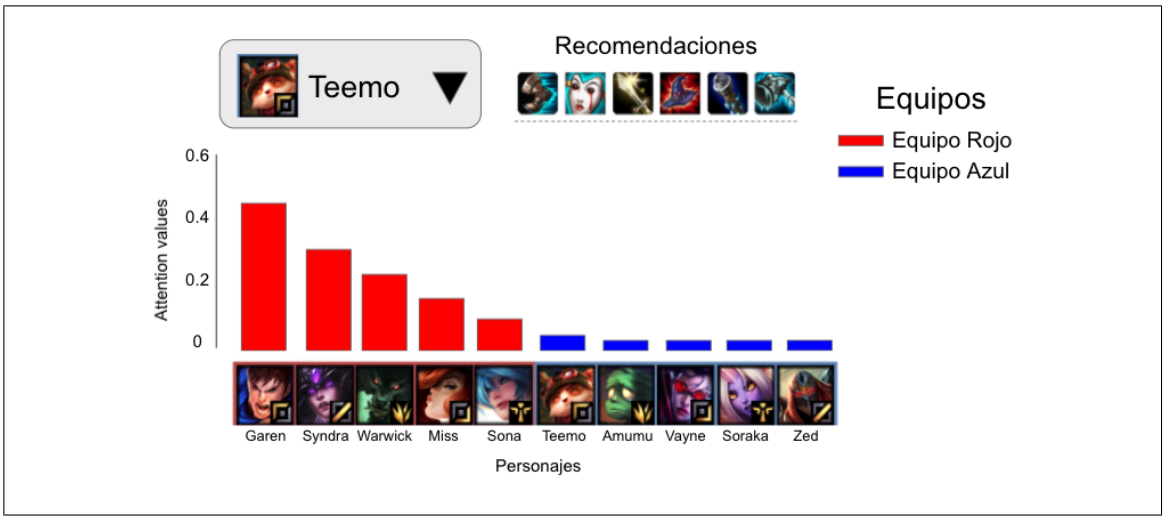

Figura 6.5. Alternativa para codificar los valores de atención y explicar solo una recomendación con un *Transformer for Team-aware Item Recommendation architecture*.

6.2. Diseño de visualización

Para validar la Hipótesis 2, la cual dice que con una extensión de los pasos del modelo anidado de Tamara Munzner, será posible diseñar de forma sistemática y justificada visualizaciones de XAI para explicaciones locales, se utilizó `VD4XAI` con el propósito de **guiar el proceso de diseño y validación de una visualización de XAI**. Esta visualización será diseñada para explicar los resultados de un sistema de IA confeccionado para clasificar documentos médicos.

6.2.1. Uso de VD4XAI

A continuación se detalla el proceso que se llevó a cabo en cada nivel de `VD4XAI` con el fin de mitigar las amenazas durante el diseño de la visualización.

1. Situación de dominio

Los usuarios clasifican día a día una gran cantidad de documentos médicos y determinan cuales deben ser excluidos de la base de datos de Epistemonikos. Cómo se explicó

en el capítulo 5. Esta base de datos consiste en un sistema usado por médicos para clasificar artículos bajo la práctica de Medicina basada en evidencia. Dada esta extenuante tarea, se confeccionó un sistema de IA que clasifica los documentos. Este sistema consiste en una XLNET, un modelo en *deep learning* propuesto por Yang et al. (2020). Los datos utilizados corresponden a textos con el título y *abstract* de documentos médicos. Con esta información, el sistema de IA clasifica el documento y entrega dos datos: la probabilidad de cada categoría posible y los pesos de atención utilizados en cada capa de la red neuronal.

Por un lado, el interés principal es que el usuario comprenda la respuesta que entregó el sistema y entienda el motivo detrás de dicha predicción. Con tal información, se espera que el usuario sea capaz de clasificar el documento de forma más eficiente y menos extenuante. Por otro lado, los expertos en este dominio (médicos) indican que una heurística para clasificar documentos se basa en identificar palabras claves del documento. Por lo tanto, una explicación similar a dicha heurística sería más intuitiva de entender para ellos. En consecuencia, para una explicación basada en palabras claves, el interés principal consiste en lograr identificar las palabras más importantes para el sistema al momento de clasificar un documento.

2. Tarea de XAI y abstracción de explicación

En base a las necesidades del usuario, se puede extrapolar que la tarea de XAI es entender por qué el sistema de IA ofrece cierta respuesta. Además, la heurística indicada por los médicos permite identificar la explicación basada en importancia de características (palabras) como una estrategia adecuada para el usuario final. Para validar que el tipo de explicación no esté incorrecto, se buscó literatura al respecto. En particular, el trabajo de Jeyakumar, Noor, Cheng, Garcia, y Srivastava (2020) muestra que gran parte de las personas prefieren una explicación basada en importancia de características cuando el sistema

de IA utiliza texto como dato de entrada. Por lo tanto, existe un respaldo en la literatura para utilizar esta estrategia de explicación para este caso de uso.

Finalmente, para asegurar una mitigación total de la amenaza de este nivel y de los niveles interiores del *framework*, se diseñó un estudio de usuario, el cual va a permitir, potencialmente, validar la utilidad la explicación y de la visualización desarrollada en los siguientes niveles. El detalle de este estudio se encuentra en la Sección 6.2.2.

3. Selección del método de XAI

Posterior a definir que la explicación sería basada en importancia de características, es necesario definir el método específico de XAI que permita efectuar esta explicación. Para este caso decidimos utilizar el mecanismo de atención porque diferentes trabajos, como los de Bahdanau, Cho, y Bengio (2014), Cádiz (2021) y Parra et al. (2019), han presentado evidencia que el mecanismo de atención ha sido de utilidad para entender por qué el sistema de IA ofrece cierta respuesta, es decir, existen diversas fuentes en la literatura que respaldan la decisión de utilizar el mecanismo de atención como método de XAI.

4. Confirmación de la utilidad de la visualización

Luego de identificar el método de XAI a utilizar, se debe validar que los datos provistos por dicho método justifiquen construir una visualización. En este caso, la misma literatura presentada en la etapa anterior da evidencia que estos datos permiten construir una visualización de XAI. En particular, el mecanismo de atención entrega un numero escalar por cada característica (palabra), lo cual permite diseñar una visualización que codifique este número escalar con algún canal (tamaño, saturación, luminosidad, entre otros). Por lo tanto, es posible continuar con los siguientes niveles de `VD4XAI` para diseñar una o más visualizaciones de XAI efectivas para este caso de uso.

5. Abstracción de datos y tareas

Dado el método de XAI utilizado, se desprende que los datos a utilizar corresponden a las palabras con su respectivo peso de atención. Este peso corresponde a un atributo cuantitativo con un valor de orden secuencial entre 0 y 1. Este valor representa la importancia de cada palabra cuando el sistema clasifica el documento. Por otro lado, en base a las necesidades del usuario, se espera que la visualización permita, mediante una exploración, descubrir e identificar las palabras más importantes para el sistema de IA.

6. Codificación visual e interacciones

Para la elección de las codificaciones visuales, se utilizó el trabajo de Parra et al. (2019) donde se presentaron diferentes alternativas para codificar los pesos de atención cuando el tipo de dato era un texto. De dicho espacio de diseño posible, se seleccionan tres alternativas que cumplen con la tarea visual: identificar las palabras más importantes del sistema. En la figura 6.6 se muestra un ejemplo de cada visualización.

- **Saturación de color de fondo**: esta es la forma tradicional de presentar la atención, como en el trabajo de Yang et al. (2016). Utilizar el color de la letra para codificar la atención puede esconder las palabras con menor importancia, por lo que esta codificación usa el color de fondo para resaltar las palabras con mayor atención sin opacar aquellas con menor atención.
- **Luminosidad de la palabra**: esta forma cambia la luminosidad de las palabras para dejar las palabras más importantes con poca luminosidad (negro puro) y aquellas menos importantes, llevarlas a un gris claro. Esta forma permite identificar las top N palabras utilizadas para clasificar un documento gracias al efecto de opacar las con menor atención.
- **Largo de barra**: esta forma incluye una barra detrás de cada palabra para codificar el peso de atención. Mientras más larga sea la barra, el peso presenta un mayor valor. Esta marca permite una comparación eficiente entre las barras gracias al uso de canales efectivos para representar valores cuantitativos (Munzner, 2014). De

esta forma, la visualización cumple con la tarea de identificar palabras mas importantes y adicionalmente permite realizar comparaciones precisas en la magnitud de los pesos.

7. Algoritmo

Finalmente, las visualizaciones fueron implementadas con D3.js versión 6.7.0, una biblioteca en Javascript para la visualización de datos. El procedimiento consistió en primero determinar el tamaño, en píxeles, de cada letra. Con esa información se calcula el ancho de cada palabra. La ecuación 6.1 presenta matemáticamente el cálculo del tamaño de una palabra, en donde w corresponde a una palabra, e es cada carácter de dicha palabra y $S(x)$ es una función que retorna el tamaño, en píxeles, de una carácter.

$$Tamaño(w) = \sum_{e \in w} S(e) \tag{6.1}$$

Luego, se posiciona una palabra al lado de otra para mantener la estructura de párrafo, y cuando una línea de palabras iba a llegar al máximo de ancho permitido en la visualización, se posicionan las palabras en la siguiente línea. Además, cuando se identifica una palabra clave como “OBJETIVO”, “CONCLUSION” o la palabra final del título, se incluye un nuevo salto de línea para dejar de forma más estructurada y presentable el texto. Para el caso de la visualización que incluye barras, el ancho de cada palabra se calcula como el máximo entre el ancho original de la palabra y el tamaño de la barra. Posteriormente, se implementa una interacción en donde pasar el cursor sobre una palabra, es decir, realizar el evento de *hover* sobre una palabra, destaca dicha palabra en cada parte del texto. Esta interacción se incluye con el fin de facilitar al usuario la tarea opcional de buscar una misma palabra en todo el documento. Todo el procedimiento descrito anteriormente implicó que el tiempo para confeccionar la visualización era aproximadamente de un segundo.

6.2.2. Diseño de Estudio de Usuario

Para finalizar el caso de uso, se diseñó un estudio de usuario con dos objetivos, (a) identificar si una visualizaciones, de las tres confeccionadas, presentaba mayor utilidad para el usuario objetivo, y (b) evaluar la efectividad de la explicación en la tarea de clasificación. En este contexto, la explicación será efectiva si el usuario es capaz de identificar apropiadamente las palabras más importantes para el sistemas, y además, le permita entender el razonamiento utilizado por el sistema de IA para clasificar el documento médico. En relación al objetivo (a), la Figura 6.6 muestra las tres visualizaciones posibles (largo de barra, intensidad del color de fondo y luminosidad de las palabras) y el caso de control (texto plano). En relación al objetivo (b), el estudio disponía de los pesos de atención extraído de la última capa de la red neuronal utilizada por el sistema. De esta forma, se estudió si dicha capa ofrecía una explicación efectiva para el usuario final. En la Figura 6.7 se muestra una visualización de documento junto a los pesos de atención ofrecidos por la última capa de la red neutonal. Se puede observar que el sistema le puso atención las últimas palabras del documento (*sarcome*, *envolving* y *thyroid*)

Para este estudio, se construyó una extensión para *Google Chrome* con el lenguaje de programación Javascript. La Figura 6.8 muestra cómo se ve la extensión una vez instalada en *Google Chrome*. Esta extensión realizaba dos principales acciones: (i) registrar las acciones y respuestas del usuario, y (ii) modificar el sitio web de Epistemonikos, donde los usuarios clasificaban los documentos, para incluir la explicación del sistema y preguntas adicionales que debe responder el usuario. Adicionalmente, la extensión ofrece visualizar el progreso del usuario en cualquier minuto del estudio. En la Figura 6.9 se puede ver la interfaz original utilizada por los usuarios, mientras que la Figura 6.10 muestra la interfaz modificada por la extensión.

Las acciones registradas por la extensión son:

- *Click* en las palabras de la visualización.

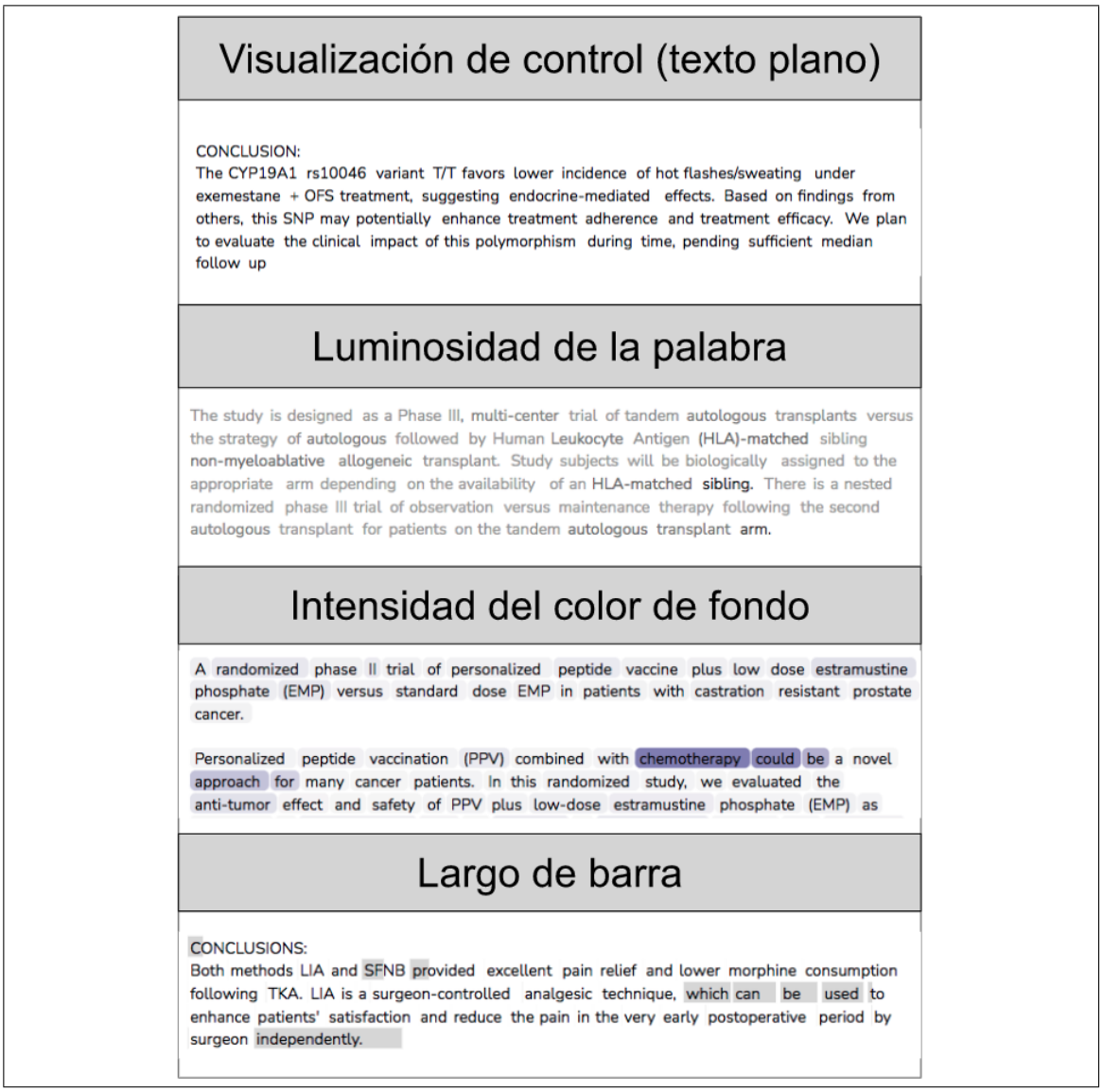

Figura 6.6. Ejemplo de las visualizaciones a utilizar en el estudio de usuario.

- *Hover* en las palabras de la visualización, es decir, mantener el cursos sobre la palabra. Para esta acción, se utilizó el trabajo de Feng, Deng, Peck, y Harrison (2016) para definir que el usuario revisó una palabra solo si se mantuvo el cursor por 0.5 segundos en dicha palabra.

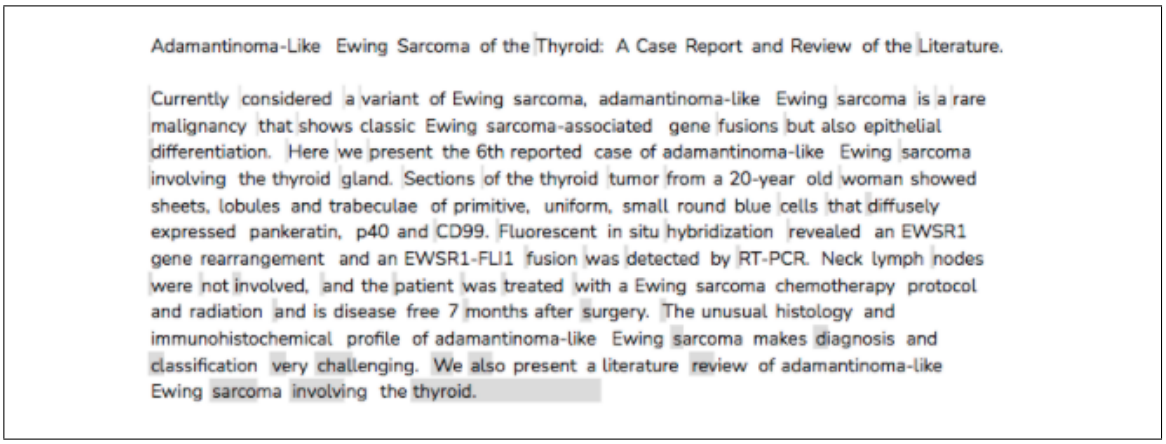

Figura 6.7. Visualización de los pesos de atención utilizando la última capa de la red neuronal.

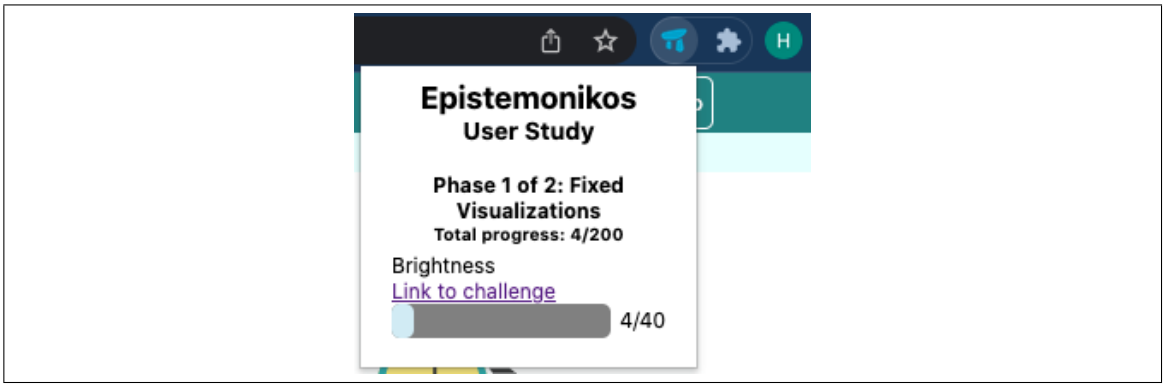

Figura 6.8. Foto de la extensión construida para el estudio de usuario. Cuando se presiona sobre ella, se visualiza el progreso del usuario y se incluye un hipervínculo a la página donde el usuario puede continuar con el estudio.

- *Scroll* en la página, es decir, desplazarse hacia arriba o hacia abajo en la página. Esta acción solo se registraban cada 5 segundos para evitar saturar la base de datos con demasiados registros.
- Si el usuario hacía *click* en algún *links* externos que ofrecía el sitio web.
- La respuesta dada por el usuario, si cancela su respuesta o si la cambia.
- Tiempo tomado para realizar cada acción.
- Opinión del usuario sobre la utilidad de la respuesta del sistema y de las palabras destacadas en la visualización.

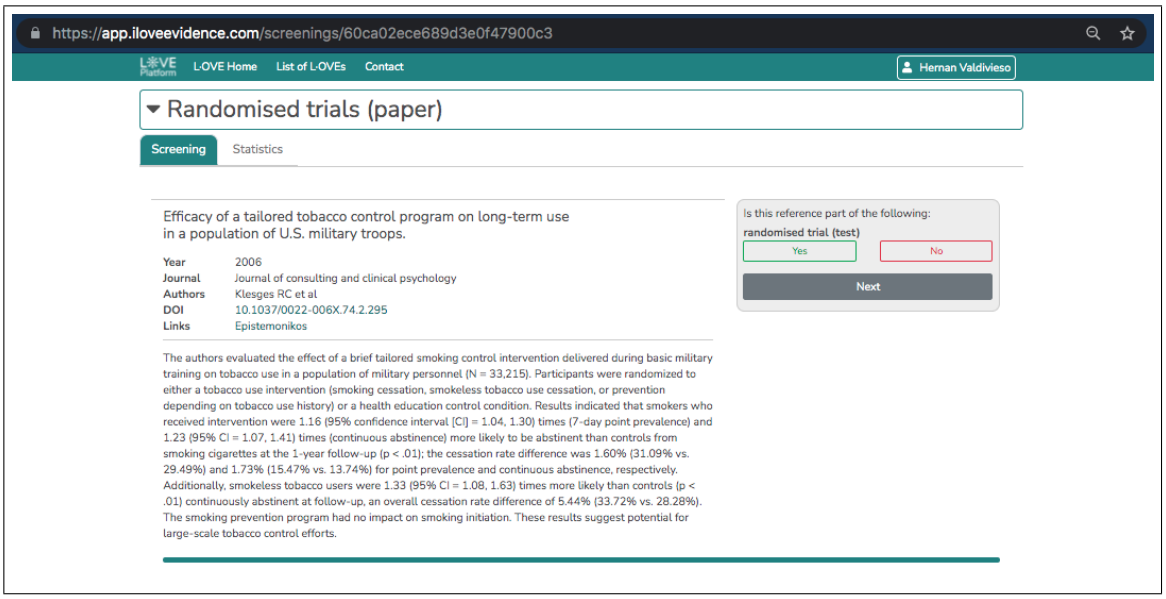

Figura 6.9. Foto de la interfaz original para clasificar documentos en Epistemonikos.

Flujo del Estudio de Usuario

Antes de comenzar el estudio, los usuarios deben leer y aceptar un consentimiento informado (apéndice B). Luego, deben responder un encuesta inicial con preguntas relacionadas a sus conocimientos previos sobre la clasificación de documentos y el uso de visualizaciones de texto (apéndice C). Posteriormente, empieza el estudio que se compone de dos fases. En la Figura 6.11 se puede observar un diagrama que expone el flujo que seguía el usuario en el estudio.

1. En la primera fase, el usuario está condicionado a utilizar una visualización específica para clasificar los documentos. Como se observa en la Figura 6.11, esta fase se compone de cuatro etapas. Cada una corresponde a una visualización distinta y al caso de control. Por otro lado, cada etapa se compone de cuatro *challenges* donde cada uno de estos corresponde a un conjunto de documentos agrupados bajo una misma categoría según la clasificación realizada por Epistemonikos. El

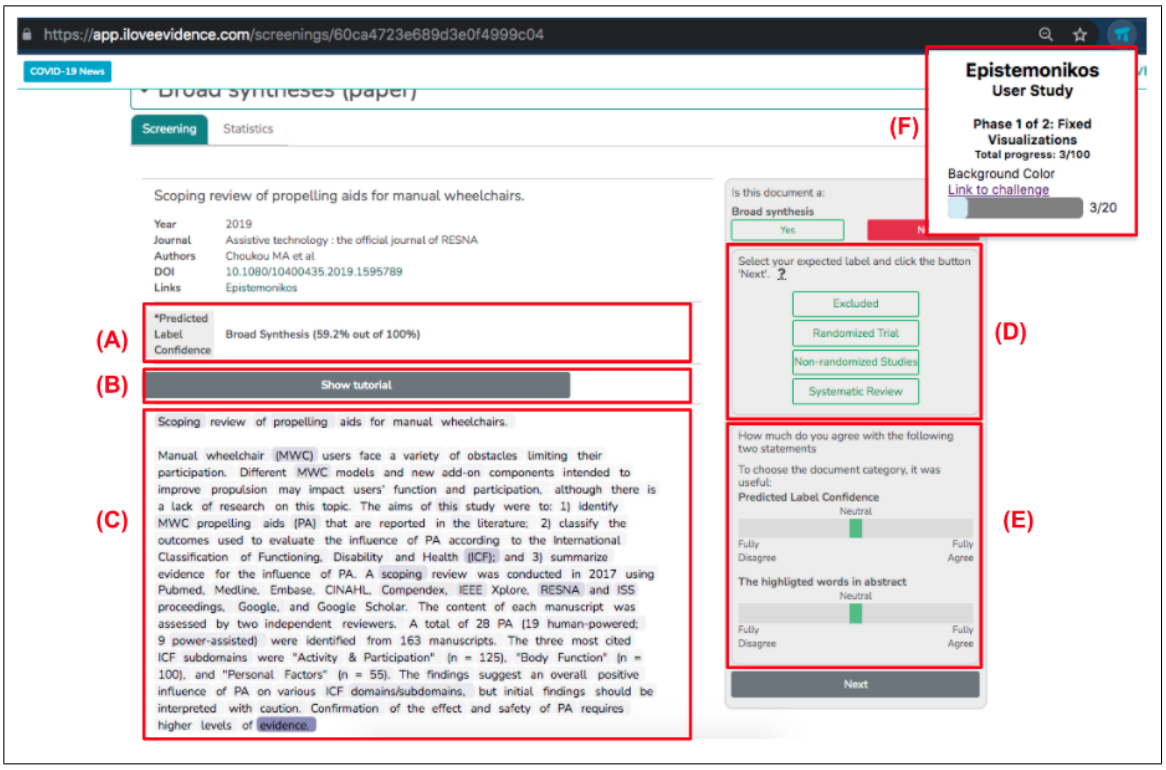

Figura 6.10. Interfaz modificada por la extensión para clasificar documentos en Epistemonikos. Incluye la respuesta del sistema y su probabilidad (A), un botón para ver un tutorial sobre la interfaz modificada (B), la visualización del documento (C), la opción de indicar la etiqueta esperada cuando se oprime el botón "No" (D), preguntas sobre la opinión del usuario respecto a la utilidad de la respuesta del modelo y la visualización (E), y el progreso del usuario en el estudio (F).

orden de las visualizaciones y el orden de los *challenges* están definido por la metodología *Latin square*. Finalmente, tras acabar cada etapa, el usuario responde una encuesta donde se mide su carga cognitiva mediante el test NASA-TLX (Hart y Staveland, 1988). Además, se le consulta sobre su experiencia con la visualización utilizada (apéndice D).

2. En la segunda fase, el usuario es libre de utilizar una de las tres visualizaciones o apagar la explicación visual. Esta fase se compone de una única etapa con los cuatro *challenges*. Luego de finalizar dicha etapa, el usuario responde una encuesta final con su experiencia utilizando el sistema y las visualizaciones (apéndice E).

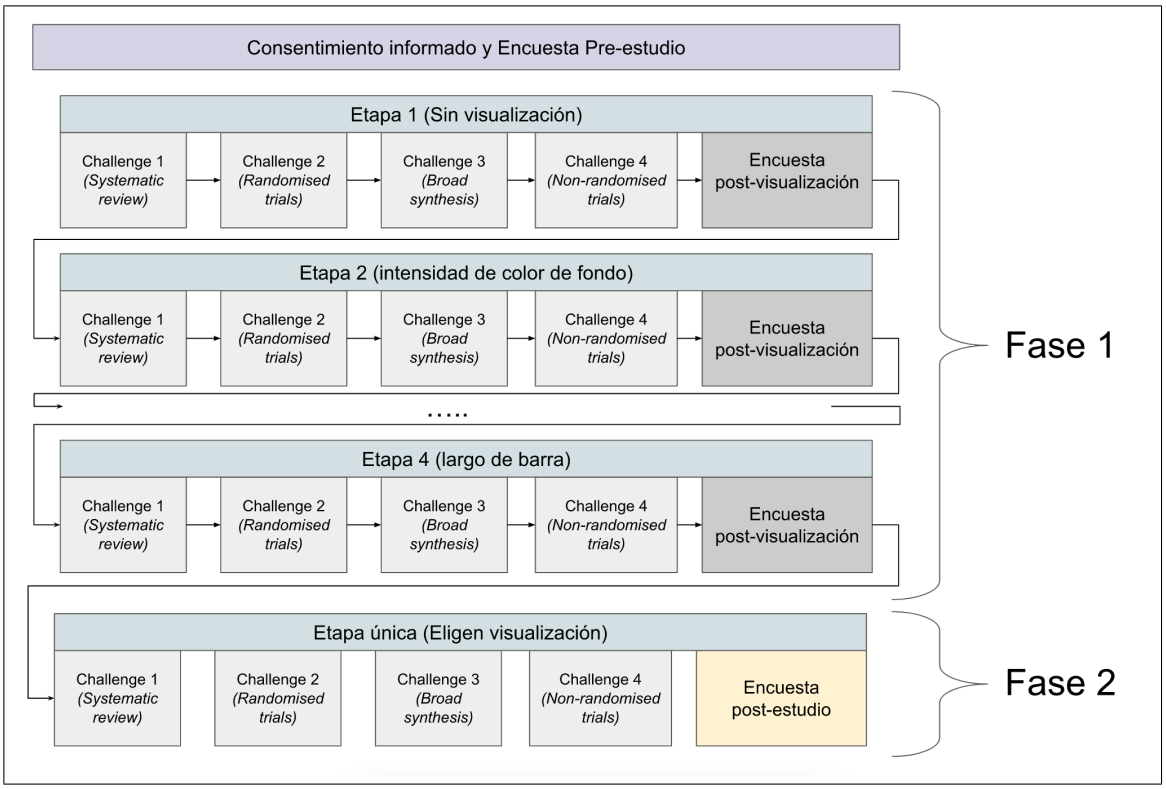

Figura 6.11. Presentación de un posible flujo realizado en el estudio de usuario. Tanto las condiciones de las etapas 1 a 4, como el orden de los *challenges* son aleatorizadas. Un usuario podría responder primero con una visualización de largo de barra y el *challenge* de *Randomised trials*, mientras que el de la figura parte con la interfaz sin visualización y en el *challenge* de *Systematic review*.

6.2.3. Escenarios posibles del Estudio de Usuario

Al momento de escribir esta sección, el estudio de usuario todavía se está llevando a cabo. Incluso, todavía se están buscando usuarios que puedan participar en el estudio para alcanzar una cantidad adecuada de respuestas. Todo este proceso ha extendido la investigación más de lo planificado. Si se espera al término del estudio para incluir un análisis de los resultados obtenidos de este, el presente trabajo se extendería más allá de los tiempos originalmente estipulados. Por lo tanto, debido a los límites de tiempo, se optó por dejar fuera los resultados del estudio de usuario y hacer un análisis exhaustivo de los posibles escenarios. Además, la principal contribución de esta tesis no es la conclusión que

se pueda obtener a partir del análisis de los resultados, sino que proponer un *framework* que permita diseñar visualizaciones de XAI y entregar evidencia que el presente trabajo es capaz de guiar correctamente este proceso de diseño. Realizar un análisis de los resultados nos permitirá enfrentar, con `VD4XAI`, sólo un posible escenario de los varios existentes, por lo cual aquí se describen diferentes escenarios posibles para lograr dos objetivos: (a) preparar el análisis empírico de los resultados y (b) mostrar evidencia que este *framework* está diseñado para guiar los diferentes escenarios. A continuación se detallarán qué acciones se pueden tomar, utilizando `VD4XAI`, para diversos escenarios que puedan surgir a partir de los resultados del estudio de usuario. De esta forma, este trabajo proporciona evidencia que este *framework* está diseñado para guiar correctamente el proceso para confeccionar una visualización de XAI bajo diferentes escenarios. En particular, los posibles escenarios que fueron analizados son: (a) el método de explicación y una o más visualizaciones son efectivas, (b) ninguna visualización es de utilidad para el usuario, (c) el método de explicación no es efectivo, (d) la respuesta del sistema no es efectiva o (e) dos o tres de los escenarios (b), (c), (d) ocurren simultáneamente.

(a) El método de explicación y una o más visualizaciones son efectivas

Este escenario se considera el *happy path* (Peter, 2010) del estudio de usuario, es decir, el método de explicación utilizado, última capa del mecanismo de atención, y alguna de las visualizaciones fueron efectivas para las tareas visuales y de XAI del usuario final. Tal como se mencionó anteriormente, una visualización de XAI efectiva es aquella que permite, al usuario final, identificar apropiadamente las palabras más importantes para el sistemas, y además, permite entender el razonamiento utilizado por el sistema de IA para clasificar el documento médico. Asimismo, con la explicación entregada del sistema, el usuario también será capaz de anticipar la respuesta que dará el sistema cuando tenga que clasificar un nuevo documento médico. Bajo este escenario, no se requiere iterar ninguno de los niveles anteriores y se puede pasar a la validación final de `VD4XAI`: observar la tasa de adopción de este sistema de XAI en la plataforma de Epistemonikos.

De forma alternativa, se podría evaluar esta explicación visual frente a otros criterios diferentes a la efectividad antes de continuar a la última validación. Tal como indica el trabajo de Bertini, Correll, y Franconeri (2020), la efectividad no debe ser el único criterio a utilizar cuando se diseña una visualización. Por este motivo, como en este escenario se asume que la visualización ya es efectiva, se podrían considerar otros criterios adicionales, como el tiempo utilizado para clasificar cada documento o la carga cognitiva del usuario cuando utiliza la visualización. Un caso hipotético podría ser que el nivel de carga cognitiva aumente considerablemente con esta explicación visual. Por lo cual, a pesar de que el usuario sea capaz de entender el razonamiento del sistema y de anticipar su respuesta en función a la explicación, la cantidad de documentos clasificados por usuario podría a ser menor a la esperada debido al desgaste mental de este. Bajo este caso hipotético, se debería poner en discusión si la reducción de documentos clasificados es una desventaja crítica que implique re-evaluar la explicación visual diseñada.

(b) Ninguna visualización es de utilidad para el usuario

Este escenario se produce cuando ninguno de los tres diseños propuestos (saturación de color de fondo, luminosidad de la palabra o largo de barra) fue de utilidad para el usuario. Esta situación puede ocurrir por diversos motivos: la elección de colores no permitió destacar correctamente las palabras más importante, los canales elegidos no fueron los indicados, entre otros. Dado lo anterior, la etapa de "Codificación visual e interacciones" debe ser iterada nuevamente. Para iterar correctamente, es necesario conversar con el usuario y registrar su experiencia con las visualizaciones. En caso que el usuario no sea capaz de dar una retroalimentación sobre las visualizaciones o se requiera de mayor información sobre los diseños confeccionados, se puede recurrir a métricas calculadas a partir de las respuestas dadas por los usuarios durante el estudio, por ejemplo, con cuál visualización el usuario respondió más veces de forma correcta o con cual visualización le consumió menos tiempo entregar una respuesta correcta. En función a la información obtenida del

usuario y de estas métricas, se toman las decisiones para modificar las codificaciones elegidas o definir nuevas codificaciones. Por ejemplo, mantener los canales pero utilizar otros colores, utilizar otra visualización de las ofrecidas en el espacio de diseño de Parra et al. (2019), o explorar otros diseños como *TileBars* (Hearst, 1995), *Ink blots* (Abbasi y Chen, 2007) o *TextArc* (Paley, 2002). Finalmente, se deberá realizar nuevamente un estudio de usuario para validar los nuevos diseños.

(c) El método de explicación no es efectivo

Este escenario consiste en que el método de explicación utilizado no fue efectivo para generar la explicación. Un motivo para que ocurra este escenario es porque las palabras destacadas por la última capa de atención no coincidía con las palabras claves esperadas por el usuario final. Para esta situación, `VD4XAI` indica que se debe iterar nuevamente desde el nivel que falló, es decir, la selección del método de XAI. En este caso se pueden evaluar múltiples opciones para solucionar el problema, como utilizar otra capa de atención del sistema, utilizar todas las capas del sistema o recurrir a otro método de explicación alineado con el tipo de explicación elegido previamente. Es importante destacar que si la iteración de este nivel provoca un cambio en lo datos entregados por el método de XAI, por ejemplo, si ahora se utiliza LIME que entrega datos positivos y negativos, será necesario iterar nuevamente los siguientes niveles como la "Abstracción de datos y tareas", y "Codificación visual e interacciones", para definir tareas visuales y codificaciones acorde a los nuevos datos generados.

Adicionalmente, en este caso de uso se utilizó únicamente los pesos de atención provistos por la última capa del sistema, situación que se realiza normalmente cuando se ocupa este mecanismo para generar explicación. No obstante, existen trabajos como el de Alammar (2021), `Ecco`, donde se ofrece una interfaz que permite visualizar cada capa de un sistema como el utilizada en el estudio de usuario. Por lo tanto, si bien es normal utilizar la última capa del sistema, se podría experimentar con este método de XAI y validar

posteriormente con el estudio de usuario si era un método efectivo para la tarea de XAI, y en caso de ser un escenario donde el método no fuera efectivo, volver a iterar dicho nivel con VD4XAI.

(d) La respuesta del sistema no es efectiva

Cómo se explicó en el capítulo 5, el sistema de IA permite predecir la categoría de un documento médico entre cinco categorías posibles: revisión sistemática (*Systematic review*), síntesis amplia (*Broad synthesis*), ensayos randomizados (*Randomised trials*), ensayos no randomizados (*Non-randomised trials*) o excluido (*excluded*). Este cuarto escenario se produce cuando la predicción entregada por el sistema de IA no es efectiva para el usuario final. Por ejemplo, que el sistema prediga todos los documentos bajo una misma categoría, como *Systematic Review*, cuando estos deberían ser de categorías diferentes. Provocando una desconfianza en el sistema y en su explicación. Esta situación puede ocurrir por diversos motivos: los datos de entrenamiento estaban sesgados, el sistema no generalizó correctamente, los datos de entrenamiento provienen de una distribución diferente a los datos que está siendo testeado el sistema, entre otros.

Ante esta situación, antes de seguir utilizando el *framework* para resolver la tarea de XAI definida inicialmente, primero será necesario entender el origen de la falla. Por lo tanto, utilizando VD4XAI, se debe iterar desde el nivel de "Tarea de XAI y abstracción de explicación" para diseñar una visualización acorde a la tarea de XAI: "Comprender cuándo falla el sistema de IA". Luego de identificar y solucionar la o las fallas, se debe utilizar nuevamente VD4XAI para resolver la tarea de XAI inicial. Como ya se identificó el tipo de explicación y la tarea de XAI ya está definida, la iteración se continua desde el nivel "Selección del método de XAI" y evaluar si el método de XAI elegido inicialmente es aplicable después de solucionar las fallas o se debe definir otro.

(e) Dos o tres de los escenario (b), (c), (d) ocurren simultáneamente

Esta situación puede ocurrir cuando los resultados muestran que más de un elemento de la explicación visual no fue efectivo. Por ejemplo, tanto el diseño de la visualización como las palabras destacadas no fueron de utilidad para el usuario final. En esta situación, se debe volver a iterar desde el nivel más externo que falló. Utilizando el ejemplo anterior, si el diseño y el método de explicación falló, como el nivel de “Selección del método de XAI” está antes que “Codificación visual e interacciones”, se debe volver a iterar desde el nivel de “Selección del método de XAI” y luego, en el nivel de “Codificación visual e interacciones” se debe tener en consideración que las visualizaciones actuales no fueron de utilidad, para no volver a reutilizarlas sin aplicar una iteración adicional a su diseño.

CAPÍTULO 7. TRABAJO FUTURO Y CONCLUSIONES

En base al análisis de las visualizaciones de XAI, fue posible identificar rápidamente el propósito con el cual fueron creadas, sus limitaciones y falencias. Por ejemplo, la visualización del sistema recomendador no fue planificada para comparar precisamente los valores de atención, puesto que no dispone de canales efectivos para dicha tarea. Si se utiliza esta visualización y la tarea visual involucra comparar los pesos de atención, se deberá descartar o rediseñar esta visualización, tal como se propuso en el Capítulo 6. En conclusión, este análisis contribuye con una forma ordenada e íntegra de describir una visualización de XAI, lo que permitirá, por ejemplo, mejorar la tarea de seleccionar una visualización de XAI efectiva para el problema de un usuario o facilitar un proceso posterior de rediseñar la visualización en función de sus tareas visuales o de XAI. Por lo tanto, el análisis de las dos visualizaciones de XAI da evidencia que se cumple la Hipótesis 1, la cual plantea que incluir un nuevo espacio de tareas de XAI al espacio de diseño del *framework* de Tamara Munzner permitirá un análisis íntegro de las visualizaciones de XAI para explicaciones locales, considerando las dimensiones visuales y de XAI.

En relación a la Hipótesis 2 presentada en el Capítulo 4, se planteó que extender los pasos del modelo anidado de Tamara Munzner nos permitirá diseñar de forma sistemática y justificada visualizaciones de XAI para explicaciones locales. Esta hipótesis fue corroborada con el caso de uso en donde se confeccionó una visualización para explicar la clasificación de documentos médicos. Luego, se validó justificadamente su proceso de diseño con los diferentes niveles de `VD4XAI`, lo que incluyo diseñar y ejecutar un estudio de usuario para validar la efectividad de la explicación visual. Este estudio consideró registrar los tiempos y diversas acciones del usuario, tales como navegar por la página, pasar el cursor por alguna palabra, acceder al documento completo, cambiar su respuesta, entre otros. Con esta información se espera estudiar la efectividad visual de las visualizaciones diseñadas. A pesar de no incluir los resultados finales de dicho estudio, se entregó evidencia que `VD4XAI` está creado para guiar el proceso de diseño frente a diversos escenario

que pueden ocurrir después de realizar el estudio de usuario. Para todos los escenario presentados en donde algún factor de la explicación visual no fue efectivo, se explicó cómo utilizar VD4XAI para iterar el proceso de diseño de forma justificada y así lograr diseñar una visualización de XAI efectiva para el caso de uso.

En vista de los validaciones presentadas en los capítulos anteriores, este trabajo presenta evidencia de su contribución para ayudar a cerrar la brecha entre los expertos en inteligencia artificial, los diseñadores de visualización, los expertos en el dominio y los usuarios finales. Sin embargo, todavía no cubre todos los aspectos en XAI o visualización. En particular, en este trabajo no se abordan visualizaciones con explicaciones globales. En la actualidad existen diversas investigaciones que se centran en la explicación global, pero la información utilizada en este tipo de explicación puede ser demasiado específica y heterogénea para ser abordada, en primera instancia, en este trabajo. Además, se realizó un supuesto inicial sobre la existencia de estrategias específicas de explicación (por importancia de característica, por analogía, por contrafactuales o por reglas), pero pueden existir más alternativas identificadas en otras investigaciones, y esto puede aumentar el espacio de XAI definido en este tesis. Adicionalmente, no hay ninguna tarea de XAI que aborde la pregunta de XAI propuesta por (Gunning, 2016): "¿Cómo corrijo un error?". Por último, VD4XAI se centró en visualizaciones de XAI para tareas de clasificación y recomendación. En otras palabras, este *framework* no aborda otras tareas de aprendizaje automático, como el aprendizaje por refuerzo, u otras formas más específicas, como el aprendizaje activo, el co-entrenamiento, entre otros. No obstante, de todas formas es posible aplicar este trabajo en otros sistemas como en uno de *clustering*. Por ejemplo, el sistema propuesto por Morichetta et al. (2019) explica un modelo de *clustering* utilizando LIME (Ribeiro et al., 2016). Con el uso de VD4XAI, se podría guiar el proceso de diseño de este sistema y analizar la visualización confeccionada, pero como esta tesis no consideró inicialmente este tipo de tarea, no es posible asegurar que el *framework* abordará completamente este caso.

En función de lo investigado en esta tesis y las limitaciones identificadas previamente, se identificaron tres direcciones futuras para mejorar esta propuesta:

- **Definir posibles guías de diseño**: se necesita validar si es posible construir guías generales que ayuden a elegir visualizaciones para sistemas y tareas populares. Además, se debe evaluar evaluar si este proceso puede inspirarse en pautas existentes como los trabajos presentados por Lu, Garcia, Hansen, Gleicher, y Maciejewski (2017) o Endert et al. (2017).
- **Expandir las aplicaciones de IA y XAI abordadas por el *framework***: dentro de las limitaciones se indicó que `VD4XAI` no aborda explicaciones globales, la preguntas de XAI "¿Cómo corrijo un error?" o tareas de aprendizaje automático tales como co-entrenamiento, aprendizaje por refuerzo, aprendizaje no supervisado, entre otros. Como trabajo futuro, se puede investigar más sobre estas otras aplicaciones para considerarlas dentro de `VD4XAI`. De este modo, se pueden modificar el espacio de XAI y/o incorporar algunos niveles en el modelo anidado de este trabajo para abordar totalmente y con mayor claridad más aplicaciones de IA y XAI. Además, sería interesante estudiar cómo proporcionar información de forma visual para fomentar al usuario a dar retroalimentación que, eventualmente, pueda mejorar la precisión de los sistemas, y así permitir que este *framework* aborde la pregunta de XAI "¿Cómo corrijo un error?".
- **Integración con otras propuestas de investigación**: durante el término de esta investigación se publicó el trabajo de Liao et al. (2021). Como se explicó en el Capítulo 3, este trabajo propuso un modelo basado en preguntas para la confección de una experiencia de usuarios (UX) en XAI. Aunque su propuesta es diferente a la presentada en este trabajo, sería interesante comparar ambas investigaciones con un estudio de usuario, analizando la retroalimentación entregada por estos para evaluar posibles cambios e integraciones de ambos trabajos. De esta forma se

podrían reducir las limitaciones de cada investigación con algunos de los aspectos propuestos en la investigación del otro. Por ejemplo, algunas de las preguntas propuestas en el trabajo de Liao et al. (2021), como *What are the potential limitations/biases of the data?*, se podrían incluir en la abstracción de datos de esta investigación para incluir un aspecto ético durante el diseño de las visualizaciones de XAI.

Para concluir, en este trabajo se identificó una brecha entre el proceso del diseño de visualización y la investigación de XAI. Luego de analizar diferentes *frameworks* de diseño para solucionar este problema, se identificó y reconoció diversas dificultades de las aplicaciones de XAI para diseñar sus visualizaciones, mientras que sistemas visuales de XAI no poseen un estándar para presentar su proceso de diseño o no brindan pautas para reproducir su proceso de diseño en otro problema o contexto. Por este motivo, en esta tesis se propuso un nuevo *framework* (`VD4XAI`) como una significativa extensión del modelo anidado de Munzner, con el fin de aplicarlo en el contexto de visualizaciones de XAI para explicaciones locales. De este modo, este trabajo contribuye en ayudar a cerrar la brecha entre los expertos en IA, los diseñadores de visualización, los expertos en el dominio y los usuarios finales. Además, se aplicó este *framework* en dos etapas de validación. La primera consistió en analizar visualizaciones existentes de XAI, mientras que la segunda etapa consistió en diseñar una visualización de XAI para un caso de uso, lo cual incluyó el diseño, ejecución de un estudio de usuario y análisis de los posibles escenarios que pueden ocurrir tras ejecutar el estudio de usuario.

Por un lado, el análisis de las dos visualizaciones existentes da evidencia que `VD4XAI` es capaz de analizar de forma sistemática el aspecto visual y de XAI de la visualización. Mediante el uso de l espacio de XAI, se puede describir la dimensión XAI de la visualización y con el uso del *framework* de Munzner (¿Qué? ¿Por qué? y ¿Cómo?) se aborda el aspecto visual. Por lo tanto, este trabajo genera un aporte desde el lado de generar análisis íntegros de las visualizaciones de XAI. Por otro lado, el caso de uso, el diseño del

estudio de usuario y la identificación de los escenarios posibles del estudio proporcionan evidencia que VD4XAI permite diseñar y validar de forma sistemática y justificada visualizaciones de XAI para explicaciones locales. De esta forma, esta tesis ofrece una pauta para diseñar visualizaciones, presentar el proceso de diseño para su reproducción y cómo deben interactuar los diferentes roles presentes en el proceso de diseño. Como trabajo futuro, se espera integrar VD4XAI con otras investigaciones para formalizar un *framework* que considere nuevas dimensiones, como el aspecto ético del modelo, o que entregue guías de cómo diseñar la visualización de XAI. Además, se espera expandir este *framework* para permitir que VD4XAI contemple más aplicaciones de IA y XAI como aprendizaje no supervisado (*clustering*) y explicaciones globales.

REFERENCIAS

Abbasi, A., y Chen, H. (2007). Categorization and analysis of text in computer mediated communication archives using visualization. En *Proc.of the 7th acm/ieee-cs joint conf. on digital libraries* (pp. 11–18).

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., y Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. En *Proc.of the 2018 chi conf. on human factors in computing systems* (p. 582).

Adadi, A., y Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138-52160.

Alammar, J. (2021, agosto). Ecco: An open source library for the explainability of transformer language models. En *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). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.30

Bahdanau, D., Cho, K., y Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barda, A. J., Horvat, C. M., y Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making*, *20*(1), 1–16.

Berger, M. (2020, octubre). Visually analyzing contextualized embeddings. En *2020 ieee visualization conference (vis)* (p. 276-280). IEEE Computer Society. doi: 10.1109/VIS47514.2020.00062

Berkovsky, S., Taib, R., y Conway, D. (2017). How to recommend? user trust factors in movie recommender systems. En *Proc.of the 22nd international conf. on intelligent user interfaces* (pp. 287–300).

Bertini, E., Correll, M., y Franconeri, S. (2020). Why shouldn't all charts be scatter plots? beyond precision-driven visualizations. En *2020 ieee visualization conference (vis)* (p. 206-210). doi: 10.1109/VIS47514.2020.00048

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... Eckersley, P. (2020). Explainable machine learning in deployment. En *Proc. of the 2020 conf. on fairness, accountability, and transparency* (p. 648–657). ACM.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. En *Ijcai* (pp. 6276–6282).

Cavallo, M., y Demiralp, Ç. (2018, abril). A visual interaction framework for dimensionality reduction based data exploration. En *Proc.of the 2018 CHI conf. on human factors in computing systems.* ACM.

Cavallo, M., y Demiralp, C. (2019, enero). Clustrophile 2: Guided visual clustering analysis. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 267–276.

Chattopadhay, A., Sarkar, A., Howlader, P., y Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. En *2018 ieee winter conf. on applications of computer vision (wacv)* (p. 839-847).

Chatzimparmpas, A., Martins, R. M., y Kerren, A. (2020, agosto). t-viSNE: Interactive assessment and interpretation of t-SNE projections. *IEEE Trans. on Vis. and Comp. Graph.*, *26*(8), 2696–2714.

Chen, C., y Rudin, C. (2018). An optimization approach to learning falling rule lists. En *International conf. on artificial intelligence and statistics* (pp. 604–612).

Cádiz, A. (2021). *Deep neural network models with explainable components for urban space perception* (Tesis de Master, Pontificia Universidad Católica de Chile. Escuela de Ingeniería). Descargado de `https://repositorio.uc.cl/handle/11534/52735`

Das, N., Park, H., Wang, Z. J., Hohman, F., Firstman, R., Rogers, E., y Chau, D. H. P. (2020, octubre). Bluff: Interactively deciphering adversarial attacks on deep neural networks. En *2020 IEEE visualization conf. (VIS)*. Institute of Electrical and Electronics Engineers (IEEE).

Deng, L., y Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, *7*(3–4), 197–387.

Dingen, D., van't Veer, M., Houthuizen, P., Mestrom, E. H. J., Korsten, E. H., Bouwman, A. R., y van Wijk, J. (2019, enero). RegressionExplorer: Interactive exploration of logistic regression models with subgroup analysis. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 246–255.

Dudley, J. J., y Kristensson, P. O. (2018, junio). A review of user interface design for interactive machine learning. *ACM Trans. Interact. Intell. Syst.*, *8*(2).

El-Assady, M., Sperrle, F., Deussen, O., Keim, D., y Collins, C. (2019, enero). Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 374–384.

Endert, A., Ribarsky, W., Turkay, C., Wong, B. W., Nabney, I., Blanco, I. D., y Rossi, F. (2017, marzo). The state of the art in integrating machine learning into visual analytics. *Computer Graphics Forum*, *36*(8), 458–486.

Fails, J. A., y Olsen, D. R. (2003). Interactive machine learning. En *Proc. of the 8th intl. conf. on intelligent user interfaces* (p. 39–45). New York, NY, USA: ACM.

Felix, C., Franconeri, S., y Bertini, E. (2017). Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(1), 657–666.

Feng, M., Deng, C., Peck, E. M., y Harrison, L. (2016). Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE transactions on visualization and computer graphics*, *23*(1), 351–360.

Goodall, J. R., Ragan, E. D., Steed, C. A., Reed, J. W., Richardson, G. D., Huffer, K. M., ... Laska, J. A. (2019, enero). Situ: Identifying and explaining suspicious behavior in networks. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 204–214.

Grinstein, U. M. F. G. G., y Wierse, A. (2002). *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., y Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Gunning, D. (2016). *Explainable artificial intelligence (xai)* (Inf. Téc.). Defense Advanced Research Projects Agency (DARPA).

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

Hearst, M. A. (1995). Tilebars: Visualization of term distribution information in full text information access. En *Proc.of the sigchi conf. on human factors in computing systems*

(pp. 59–66).

Herlocker, J. L., Konstan, J. A., y Riedl, J. (2000). Explaining collaborative filtering recommendations. En *Proc.of the 2000 acm conf. on computer supported cooperative work* (pp. 241–250).

Hochreiter, S., y Schmidhuber, J. (1997, noviembre). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Hohman, F., Kahng, M., Pienta, R., y Chau, D. H. (2019, agosto). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(8), 2674–2693.

Hohman, F., Park, H., Robinson, C., y Chau, D. H. P. (2020, enero). Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Trans. on Vis. and Comp. Graph.*, *26*(1), 1096–1106.

Huang, X., Jamonnak, S., Zhao, Y., Wang, B., Hoai, M., Yager, K., y Xu, W. (2021, febrero). Interactive visual study of multiple attributes learning model of x-ray scattering images. *IEEE Trans. on Vis. and Comp. Graph.*, *27*(2), 1312–1321.

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., y Srivastava, M. (2020). How can i explain this to you? an empirical study of deep neural network explanation methods. En H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, y H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 4211–4222). Curran Associates, Inc. Descargado de `https://proceedings.neurips.cc/paper/2020/file/2c29d89cc56cdb191c60db2f0bae796b-Paper.pdf`

Jiang, L., Liu, S., y Chen, C. (2019, abril). Recent research advances on interactive machine learning. *J. Vis.*, *22*(2), 401–417.

Kahng, M., Andrews, P. Y., Kalro, A., y Chau, D. H. (2018, enero). ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(1), 88–97.

Kahng, M., Thorat, N., Chau, D. H. P., Viegas, F. B., y Wattenberg, M. (2019, enero). GAN lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 310–320.

Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., ... Choo, J. (2019, enero). RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 299–309.

Lakkaraju, H., Bach, S. H., y Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. En *Proc.of the 22nd acm sigkdd international conf. on knowledge discovery and data mining* (pp. 1675–1684).

Larochelle, H., y Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order boltzmann machine. En J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, y A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 1243–1251). Curran Associates, Inc.

Liang, Y., Fan, H. W., Fang, Z., Miao, L., Li, W., Zhang, X., ... Chen, X. '. (2020, abril). OralCam: Enabling self-examination and awareness of oral health using a smartphone camera. En *Proc.of the 2020 CHI conf. on human factors in computing systems.* ACM.

Liao, Q. V., Pribić, M., Han, J., Miller, S., y Sow, D. (2021). *Question-driven design process for explainable ai user experiences.*

Lin, H., Gao, S., Gotz, D., Du, F., He, J., y Cao, N. (2018, julio). RCLens: Interactive rare category exploration and identification. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(7),

2223–2237.

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.

Liu, M., Shi, J., Cao, K., Zhu, J., y Liu, S. (2018, enero). Analyzing the training processes of deep generative models. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(1), 77–87.

Liu, S., Li, Z., Li, T., Srikumar, V., Pascucci, V., y Bremer, P.-T. (2019, enero). NLIZE: A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 651–660.

Lobo, M.-J., Pietriga, E., y Appert, C. (2015). An evaluation of interactive map comparison techniques. En *Proc.of the 33rd annual acm conf. on human factors in computing systems* (p. 3573–3582). New York, NY, USA: Association for Computing Machinery.

Lu, Y., Garcia, R., Hansen, B., Gleicher, M., y Maciejewski, R. (2017, junio). The state-of-the-art in predictive visual analytics. *Comput. Graph. Forum*, *36*(3), 539–562.

Lundberg, S. M., y Lee, S.-I. (2017). A unified approach to interpreting model predictions. En *Proc. of the 31st intl. conf. on neural information processing systems* (p. 4768–4777). Red Hook, NY, USA: Curran Associates Inc.

Ma, Y., Xie, T., Li, J., y Maciejewski, R. (2020, enero). Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE Trans. on Vis. and Comp. Graph.*, *26*(1), 1075–1085.

Marai, G. E., Ma, C., Burks, A. T., Pellolio, F., Canahuate, G., Vock, D. M., ... Fuller, C. D. (2019, abril). Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(4), 1732–1745.

Messina, P., Dominguez, V., Parra, D., Trattner, C., y Soto, A. (2019, abril). Content-based artwork recommendation: Integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction*, *29*(2), 251–290.

Miller, D. D., y Brown, E. W. (2018). Artificial intelligence in medical practice: the question to the answer? *The American journal of medicine*, *131*(2), 129–133.

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., y Qu, H. (2017). Understanding hidden memories of recurrent neural networks. En *2017 ieee conf. on visual analytics science and technology (vast)* (pp. 13–24).

Ming, Y., Qu, H., y Bertini, E. (2019, enero). RuleMatrix: Visualizing and understanding classifiers with rules. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 342–352.

Ming, Y., Xu, P., Cheng, F., Qu, H., y Ren, L. (2020, enero). ProtoSteer: Steering deep sequence model with prototypes. *IEEE Trans. on Vis. and Comp. Graph.*, *26*(1), 238–248.

Mitchell, T. (1997). *Machine learning*. McGraw hill Burr Ridge.

Mnih, V., Heess, N., Graves, A., y Kavukcuoglu, K. (2014). Recurrent models of visual attention. En *Proceedings of the 27th international conference on neural information processing systems - volume 2* (p. 2204–2212). Cambridge, MA, USA: MIT Press.

Mohseni, S., Zarei, N., y Ragan, E. D. (2018). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv preprint arXiv:1811.11839*.

Morichetta, A., Casas, P., y Mellia, M. (2019). Explain-it: towards explainable ai for unsupervised network traffic analysis. En *Proc.of the 3rd acm conext workshop on big data, machine learning and artificial intelligence for data communication networks* (pp. 22–28).

Munzner, T. (2014). *Visualization analysis and design*. CRC Press. Descargado de https://books.google.cl/books?id=dznSBQAAQBAJ

Myers, B. A., Weitzman, D. A., Ko, A. J., y Chau, D. H. (2006). Answering why and why not questions in user interfaces. En *Proc.of the sigchi conf. on human factors in computing systems* (pp. 397–406).

Paley, W. B. (2002). Textarc: Showing word frequency and distribution in text. En *Poster presented at ieee symposium on information visualization* (Vol. 2002).

Parra, D., Valdivieso, H., Carvallo, A., Rada, G., Verbert, K., y Schreck, T. (2019). Analyzing the design space for visualizing neural attention in text classification. En *Proc. ieee vis workshop on vis x ai: 2nd workshop on visualization for ai explainability (visxai).*

Peter, B. (2010, enero). Bpmn: A meta model for the happy path. *Maastricht : METEOR, Maastricht Research School of Economics of Technology and Organization, Research Memoranda*.

Pezzotti, N., Hollt, T., Gemert, J. V., Lelieveldt, B. P., Eisemann, E., y Vilanova, A. (2018, enero). DeepEyes: Progressive visual analytics for designing deep neural networks. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(1), 98–108.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... Iyengar, S. S. (2018, septiembre). A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, *51*(5).

Puhringer, M., Hinterreiter, A., y Streit, M. (2020, octubre). InstanceFlow: Visualizing the evolution of classifier confusion at the instance level. En *2020 IEEE visualization conf. (VIS).* Institute of Electrical and Electronics Engineers (IEEE).

Ras, G., van Gerven, M., y Haselager, P. (2018). Explanation methods in deep learning:

Users, values, concerns and challenges. En H. J. Escalante et al. (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Cham: Springer International Publishing.

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., Tengg-Kobligk, H. v., ... Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, *2*(3), e190043.

Ribeiro, M. T., Singh, S., y Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. En *Proc. of the 22nd acm sigkdd intl. conf. on knowledge discovery and data mining* (p. 1135–1144). ACM.

Ribera, M., y Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable ai. En *Iui workshops*.

Rosenthal, S., Selvaraj, S. P., y Veloso, M. M. (2016). Verbalization: Narration of autonomous robot experience. En *Ijcai* (Vol. 16, pp. 862–868).

Sahoo, S., y Berger, M. (2020, octubre). Visually analyzing and steering zero shot learning. En *2020 IEEE visualization conf. (VIS)*. Institute of Electrical and Electronics Engineers (IEEE).

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Trans. on Interactive Intelligent Systems (TiiS)*, *10*(4), 1–31.

Simonyan, K., Vedaldi, A., y Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Snyder, L. S., Lin, Y.-S., Karimzadeh, M., Goldwasser, D., y Ebert, D. S. (2019). Interactive learning for identifying relevant tweets to support real-time situational awareness. *IEEE Trans. on Vis. and Comp. Graph.*, 1–1.

Spinner, T., Schlegel, U., Schafer, H., y El-Assady, M. (2019). explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Trans. on Vis. and Comp. Graph.*, 1–1.

Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., y Rush, A. M. (2019, enero). Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 353–363.

Strobelt, H., Gehrmann, S., Pfister, H., y Rush, A. M. (2018a, enero). LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(1), 667–676.

Strobelt, H., Gehrmann, S., Pfister, H., y Rush, A. M. (2018b, enero). LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Trans. on Vis. and Comp. Graph.*, *24*(1), 667–676.

Sun, D., Feng, Z., Chen, Y., Wang, Y., Zeng, J., Yuan, M., ... Qu, H. (2020, abril). DFSeer: A visual analytics approach to facilitate model selection for demand forecasting. En *Proc.of the 2020 CHI conf. on human factors in computing systems.* ACM.

Sutskever, I., Vinyals, O., y Le, Q. V. (2014). Sequence to sequence learning with neural networks. En *Proc. nips.* Montreal, CA. Descargado de `http://arxiv.org/abs/1409.3215`

Tian, Y., Pei, K., Jana, S., y Ray, B. (2018). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. En *Proc.of the 40th international conf. on software engineering* (pp. 303–314).

Upson, C., Faulhaber, T. A., Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., ... van Dam, A. (1989). The application visualization system: a computational environment for scientific visualization. *IEEE Computer Graphics and Applications*, *9*(4), 30-42.

Van Looveren, A., y Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need.*

Villa, A., Araujo, V., Cattan, F., y Parra, D. (2020). Interpretable contextual team-aware item recommendation: Application in multiplayer online battle arena games. En *Fourteenth acm conf. on recommender systems* (pp. 503–508).

Vui, C., Soon, G., On, C., Alfred, R., y Anthony, P. (2013, noviembre). A review of stock market prediction with artificial neural network (ann). En (p. 477-482).

Wang, D., Yang, Q., Abdul, A., y Lim, B. Y. (2019, mayo). Designing theory-driven user-centric explainable AI. En *Proc.of the 2019 CHI conf. on human factors in computing systems.* ACM.

Wang, F., y Rudin, C. (2015). Falling rule lists. En *Artificial intelligence and statistics* (pp. 1013–1022).

Wang, J., Gou, L., Shen, H.-W., y Yang, H. (2019, enero). DQNViz: A visual analytics approach to understand deep q-networks. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 288–298.

Wang, Q., Ming, Y., Jin, Z., Shen, Q., Liu, D., Smith, M. J., ... Qu, H. (2019, mayo). ATMSeer. En *Proc.of the 2019 CHI conf. on human factors in computing systems.* ACM.

Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., ... Chau, D. H. P.

(2021, febrero). CNN explainer: Learning convolutional neural networks with interactive visualization. *IEEE Trans. on Vis. and Comp. Graph.*, *27*(2), 1396–1406.

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

Wentzel, A., Canahuate, G., van Dijk, L. V., Mohamed, A. S., Fuller, C. D., y Marai, G. E. (2020, octubre). Explainable spatial clustering: Leveraging spatial data in radiation oncology. En *2020 IEEE visualization conf. (VIS).* Institute of Electrical and Electronics Engineers (IEEE).

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., y Wilson, J. (2019). The what-if tool: Interactive probing of machine learning models. *IEEE Trans. on Vis. and Comp. Graph.*, 1–1.

Wright, A. P., Wang, Z. J., Park, H., Guo, G., Sperrle, F., El-Assady, M., … Chau, D. H. (2020). *A comparative analysis of industry human-ai interaction guidelines.*

Xie, Y., Chen, M., Kao, D., Gao, G., y Chen, X. '. (2020a, abril). CheXplain. En *Proc.of the 2020 CHI conf. on human factors in computing systems.* ACM.

Xie, Y., Chen, M., Kao, D., Gao, G., y Chen, X. A. (2020b). Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. En *Proc. of the 2020 chi conf. on human factors in computing systems* (p. 1–13). ACM.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., y Le, Q. V. (2020). *Xlnet: Generalized autoregressive pretraining for language understanding.*

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., y Hovy, E. (2016, junio). Hierarchical attention networks for document classification. En *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human*

language technologies (pp. 1480–1489). San Diego, California: Association for Computational Linguistics. doi: 10.18653/v1/N16-1174

Zeiler, M. D., y Fergus, R. (2014). Visualizing and understanding convolutional networks. En *European conf. on computer vision* (pp. 818–833).

Zhang, J., Wang, Y., Molino, P., Li, L., y Ebert, D. S. (2019, enero). Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 364–373.

Zhang, J., y Zong, C. (2015, septiembre). Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, *30*(05), 16-25.

Zhao, J., Dai, Z., Xu, P., y Ren, L. (2020, octubre). ProtoViewer: Visual interpretation and diagnostics of deep neural networks with factorized prototypes. En *2020 IEEE visualization conf. (VIS).* Institute of Electrical and Electronics Engineers (IEEE).

Zhao, X., Wu, Y., Lee, D. L., y Cui, W. (2019, enero). iForest: Interpreting random forests via visual analytics. *IEEE Trans. on Vis. and Comp. Graph.*, *25*(1), 407–416.

APÉNDICE

A. MECANISMO DE ATENCIÓN

A continuación se presenta el funcionamiento del mecanismo de atención y del mecanismo de auto-atención propuesto en el trabajo de Larochelle y Hinton (2010). Primero se explicará el mecanismo tradicional (atención) y luego se detallará la diferencia que tiene este mecanismo con la auto-atención.

En el mecanismo de atención, se dispone de un vector $y_i \in \mathbb{R}^d$ donde d es la dimensión/largo del vector. El objetivo de este mecanismo es actualizar el vector y_i a partir de n otros de vectores, $X = \{x_1, x_2, ..., x_n\}$. Para lograr esta actualización, se dispone de 3 matrices (W^Q, W^K, W^V) de tamaño $d \times d'$ y 3 vectores (b^Q, b^K, b^V) de tamaño d'. Tanto las matrices cómo los vectores viven en los reales ($\mathbb{R}$) y sus valores son aprendidos por el sistema que ocupe este mecanismo. Finalmente, se dispone de 3 funciones ($q(y), k(x), v(x)$) definidas del siguiente modo:

$$\begin{aligned} q(y) &= yW^Q + b^Q \\ k(x) &= xW^K + b^K \\ v(x) &= xW^V + b^V \end{aligned} \tag{A.1}$$

Este mecanismo determina cuales son los vectores en X con mayor importancia para actualizar y_i. Para lograr esto, se calcula un coeficiente a_i, j para cada vector $x_j \in X$ que representa la importancia de x_j para el vector y_i. Este coeficiente se calcula como se indica en la ecuación (A.2).

$$a_{i,j} = \underset{x_j \in X}{\text{softmax}} \left(\frac{q(y_i)k(x_j)^T}{\sqrt{d'}} \right) \tag{A.2}$$

Tras calcular los coeficientes, cuyo valores están entre 0 y 1, estos son utilizados para considerar con mayor relevancia los $V(x_j)$ en la ecuación (A.3). Luego, se utiliza la matriz

W^O de tamaño $d' \times d$ y el vector b^O de dimensión d para calcular el nuevo valor de y_i (y_actualizado$_i$). Los valores de la matriz y del vector son aprendidos por el sistema.

$$\text{y\_actualizado}_i = \left(\sum_{j=i}^{n} a_i, jv(x_j) \right) W^O + b^O \tag{A.3}$$

Lo descrito anteriormente corresponde al mecanismo de atención y los coeficientes de atención $a_{i,j}$ corresponden a los pesos entre 0 y 1 que son utilizados posteriormente para visualizar la importancia que tenía cada dato de entrada cuando el sistema entrega una respuesta.

Finalmente, para utilizar el mecanismo de auto-atención, solo es necesario incluir y_i dentro de los vectores X, es decir, $\{y_i, x_1, x_2, ..., x_n\} \in X$. De este modo, se añade un nuevo coeficiente de atención que indica la importancia del dato original cuando se intentó actualizar.

B. ESTUDIO DE USUARIO: CONSENTIMIENTO INFORMADO

Esta documento fue presentado a los usuarios en español y en ingles.

B.1. Versión español

Consentimiento informado

El propósito de esta información es ayudarle a tomar la decisión de participar en una investigación científica. Solo el consentimiento informado está disponible también en español. Todas las encuestas y aplicación web a utilizar estarán en ingles.

Tome el tiempo que estime necesario para decidir participar, lea con detención este documento y pregunte lo que sea necesario al personal del estudio.

Objetivos de investigación

Este estudio forma parte del FONDECYT "Harnessing Information Visualization and Interactivity to Develop Interfaces for Explainable Artificial Intelligence", cuyo próposito es investigar qué aspectos de visualización de información e interactividad pueden ser útiles para el desarrollo de sistemas de inteligencia artificial con características de explicabilidad de sus predicciones, considerando como área de aplicación principal la medicina. El objetivo principal de nuestro estudio es estudiar si incluir cierto tipo de visualizaciones mejoran el desempeño y disminuyen la carga cognitiva de expertos en salud para la práctica de medicina basada en la evidencia.

Para realizar la evaluación reclutaremos entre 10 y 15 expertos en salud que trabajen en la selección y filtrado de documentos en la Fundación Epistemonikos o que sean alumnos del curso de Medicina basada en la evidencia UC, dictado por el profesor Gabriel Rada, co-IR de este proyecto.

Procedimientos de la investigación

Si usted acepta formar parte de este estudio le solicitaremos:

1. Utilizar dos aplicaciones web. Una alojada en los servidores de la Pontificia Universidad Católica de Chile para responder las encuestas, y otra alojada en los servidores de Epistemonikos para realizar la tarea de filtrado y clasificación de documentos en distintas categorías con la ayuda de un algoritmo automático con y sin visualizaciones que serán alternadas siguiendo la metodología latin square (`https://paasp.net/within-subject-study-designs-latin-square/`).
2. Para realizar lo anterior se le entregará una extensión de Google Chrome. Tendrá un plazo de 2 semanas para realizarla pero puede hacerlo en varias sesiones. Puede entrar a cualquier hora o pausar su trabajo actual para luego continuar. No le tomará más de 1 hora en total.
3. La información obtenida será solo para el propósito de esta investigación. Además la única información almacenada serán *clicks* y tiempo de realización de la tarea, así como las respuestas a los cuestionarios. Generaremos un identificador para cada usuario pero no lo enlazaremos con datos personales como RUT, edad, etc.
4. Publicaremos los resultados de forma agregada, considerando las diferentes visualizaciones del estudio. No se publicarán datos desagregados ni personales.

Beneficios

Usted al formar parte de este estudio participará por una giftcard en Amazon por un monto de CLP$100.000.- o su equivalente en dólares de EEUU.

Además se verá beneficiado de poder probar tecnologías de punta en el área de inteligencia artificial y visualización de datos para clasificación de textos médicos.

Alternativas

Dado que esta investigación no implica ni diagnóstico ni tratamiento de alguna condición médica o exposición a daños, su alternativa es la de no participar en esta investigación científica.

Riesgos

Su participación puede provocar malestar si tiene una mala postura frente al computador o cansancio visual por la atención requerida frente a la pantalla del computador.

Costos asociados

Formar parte de este estudio no tendrá ningún costo para usted.

Cobertura de daños

No se contempla cobertura de daños.

Compensaciones

Este estudio no cuenta con ningún tipo de compensación económica por participar de él.

Confidencialidad de la información

Los datos personales serán encriptados y anonimizados. Además si los resultados son publicados en una conferencia o journal serán mostrados de manera agregada, sin mostrar nombres ni nada que pueda identificarlo. Se solicitará e-mail del usuario para notificar al ganador de la gift card, pero una vez realizado el sorteo, se eliminarán los e-mails de los participantes de la base de datos.

Voluntariedad

Su participación en este estudio es completamente voluntaria. Usted tiene el derecho a no aceptar participar o a retirar su consentimiento en cualquier momento. Su decision no tendrá impacto en eventual evaluaciones académicas y/o profesionales.

Si retira su consentimiento su información será eliminada inmediatamente, para no ser utilizada con posterioridad.

Preguntas

Si tiene alguna pregunta sobre este estudio puede contactar a Andrés Carvallo o a Denis Parra investigadores responsables del estudio al teléfono 6xxxxxxx o a los e-mails axxxxxxxx@xx.xx o dxxxxxxx@xx.xx.

Si tiene preguntas sobre sus derechos como participante de una investigación, usted puede llamar a la Dra Claudia Uribe, Presidente del Comité Ético Científico de Ciencias de la Salud, Pontificia Universidad Católica de Chile, al teléfono 2xxxxxxxx, o enviar un correo electrónico a etxxxxxxxxxxxxxxxxxx@xx.xx

Declaración de consentimiento

- Se me ha explicado el propósito de esta investigación, los procedimientos, los riesgos, los beneficios y los derechos que me asisten y me puedo retirar de ella en cualquier momento.
- Firmo este documento sin ser forzado/obligado a hacerlo.
- No estoy renunciando a ningún derecho que me asista.
- Se me ha informado que tengo el derecho a reevaluar mi participación en esta investigación según mi parecer y en cualquier momento que lo desee.

Considerando lo anterior ¿Está Ud. dispuesto a participar en nuestro estudio de usuario?

Si es así, por favor haga clic en el botón respectivo:

- Acepto participar
- No acepto participar

B.2. Versión ingles

Informed consent

The purpose of this information is to help you in deciding to participate in scientific research. Only informed consent is also available in Spanish. All surveys, and web application to be used will be in English.

Take your time to decide to participate, read this document carefully, and ask the study's staff as necessary.

Research objectives

This study is part of the FONDECYT "Harnessing Information Visualization and Interactivity to Develop Interfaces for Explainable Artificial Intelligence", whose purpose is to investigate what aspects of information visualization and interactivity can be helpful for the development of artificial intelligence systems with descriptive characteristics of their predictions, considering medicine as the main application area. The main objective of our research is to study whether including certain types of visualizations improves the performance and decreases the cognitive load of health experts for the practice of evidence-based medicine.

To carry out the evaluation, we recruit between 10 and 15 health experts who work in selecting and filtering documents at the Epistemonikos Foundation and/or students of the UC Evidence-Based Medicine course, taught by Professor Gabriel Rada, co-LR of this project.

Research procedure

If you agree to be part of this study, we will ask you to:

1. Use two web applications. One hosted on the servers of the Pontificia Universidad Católica de Chile to answer the surveys, and another hosted at the Epistemonikos servers to perform the task of filtering and classifying documents in different categories with the help of an automatic algorithm with and without visualizations that They will be alternated following the Latin square methodology (`https://paasp.net/within-subject-study-designs-latin-square/`).
2. To do the above, you need to install a Google Chrome extension. You have up to 2 weeks time to do it, but you can split it into several sessions. You can enter at any time or pause your current work and then continue. It will not take more than 1 hour in total.
3. The information obtained will be only used for this study. In addition, the only information stored will be clicks and time to complete the task, as well as the answers to the questionnaires. We will generate an identifier for each user, but we will not link it with personal data such as ID, age, among others.
4. We will publish the results in an aggregated way, considering the different visualizations of the study. No disaggregated or personal data will be published.

Profits

By being part of this study, you will participate for a gift card by Amazon for an amount of CLP 100,000.- or it's equivalent in US dollars.

You will also benefit from testing state-of-the-art technologies in artificial intelligence and data visualization for the classification of medical texts.

Alternatives

Since this research does not imply the diagnosis or treatment of any medical condition or exposure to harm, your alternative is not to participate in this scientific research.

Risks

Your participation can cause discomfort if you have a bad posture in front of the computer or visual fatigue due to the attention required in front of the computer screen.

Associated costs

Taking part in this study will be at no cost to you.

Damage coverage

Damage coverage is not contemplated.

Compensation

This study does not have any kind of financial compensation for participating in it.

Confidentiality of information

Personal data will be encrypted and anonymized. Also, if the results are published in a conference or journal, they will be shown in an aggregated way, without showing names or anything that can identify you. The user's e-mail will be requested to notify the winner of the gift card, but once the raffle has been carried out, the participants' e-mails will be eliminated from the database.

Willfulness

Your participation in this study is completely voluntary. You have the right not to agree to participate or to withdraw your consent at any time. Your decision will have no impact on eventual academic and/or professional evaluations.

If you withdraw your consent, your information will be deleted immediately, not to be used later.

Questions

If you have any questions about this study, you can contact Andrés Carvallo or Denis Parra, the researchers responsible for the study, by calling 6xxxxxxx or by e-mails axxxxxxxx@xx.xx or dxxxxxxx@xx.xx.

If you have questions about your rights as a research participant, you can call Dr. Claudia Uribe, President of the Scientific Ethics Committee of Health Sciences, Pontificia Universidad Católica de Chile, at 2xxxxxxxx, or send an e-mail to etxxxxxxxxxxxxxxxxxx@xx.xx

Declaration of consent

- The purpose of this research, procedures, risks, benefits, and rights have been clearly explained.

- I sign this document without being forced/compelled to do so.
- I know I can cancel this study whenever I want.
- I am not waiving any rights to assist me.
- I have been informed that I have the right to reevaluate my participation in this research whenever I want and at my discretion.

Considering the above, are you willing to participate in our user study?

If so, please click the respective button:

- I agree to participate
- I do not agree to participate

C. ESTUDIO DE USUARIO: ENCUESTA PRE-ESTUDIO

Esta encuesta fue realiza en ingles.

Please read each statement and indicate to what extent you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

- I know what an automatic classifier is and how it works.
- I usually work with text data visualizations.
- I quickly understand text data visualizations.
- I can read research articles in my working field in English without a problem.
- I have extensive knowledge of COVID-19 evidence treatments.
- I know what a randomized controlled trial is.
- I know what a systematic review is.
- I know that studies based on randomized controlled trials are more robust than other types of studies.
- I know the reasons why documents are excluded from Epistemonikos databases.

Additional questions

1. How many research documents have you reviewed for classification before this study?
   - None
   - around 20
   - around 50
   - around 70
   - More than one hundred studies
2. How long, on average, does it take you to classify one document into the different categories (systematic review, randomized controlled trial, etc.)?

- Less than 1 minute
- 1 minute
- 3 minutes
- 5 minutes
- More than 5 minutes

3. Are you a Medical student or a graduated physician ?
   - Medical Student
   - Graduated Physician
   - Other clinician
   - Others
4. If you have finished your studies, please indicate your specialization.

D. ESTUDIO DE USUARIO: ENCUESTA POST-VISUALIZACIÓN

Esta encuesta fue realiza en ingles.

Please read each question and answer these on a scale of 1 (low) to 100 (high).

- How mentally demanding was the task?
- How physically demanding was the task?
- How hurried or rushed was the pace of the task?
- How successful were you in accomplishing what you were asked to do?
- How hard did you have to work to accomplish your level of perfomance?
- How insecure, discouraged, irritated, stressed, and annoyed were you?

Please read each statement and indicate to what extent you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

1. Compared to an interfaces without visualization, this visualization improved my performance for reviewing documents.
2. This visualization helped me understand why documents were automatically classified as a particular category.
3. This visualization colors and look & feel were appropriate to understand.

E. ESTUDIO DE USUARIO: ENCUESTA POST-ESTUDIO

Esta encuesta fue realiza en ingles.

Please read each statement and indicate to what extent you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

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

Additional questions

1. Do you have any comments about the activity? (optional)