Tendencias en publicaciones sobre Inteligencia Artificial explicativa

Lazcano, A., Hidalgo, P., Beleña, L., Universidad Francisco de Vitoria (UFV)

Introducción

Los rápidos avances en el aprendizaje profundo han impulsado importantes progresos en diversos dominios, incluyendo la visión por computador, el procesamiento del lenguaje natural y el reconocimiento de voz. Sin embargo, la naturaleza compleja y opaca de los modelos de aprendizaje profundo ha generado preocupaciones sobre su transparencia e interpretabilidad. La Inteligencia Artificial Explicable (xAI, por sus siglas en inglés) ha emergido como un campo de estudio crucial, orientado a comprender, visualizar e interpretar el funcionamiento interno de los modelos de aprendizaje profundo para fomentar la confianza y la responsabilidad (Samek et al., 2017).

Esta complejidad ha llevado a la inteligencia artificial (IA) explicable a incrementar notablemente su presencia en la literatura científica en el contexto del aprendizaje profundo moderno. Sin un mecanismo explicativo, los resultados de las redes neuronales profundas actuales (DNN) no se pueden explicar ni por la propia red neuronal ni por el desarrollador del modelo. Los diferentes modelos de redes neuronales creados para distintos tipos de problemas, como Perceptrón Multicapa (MLP), Long Short Term Memory (LSTM) o Transformers en la actualidad son consideradas cajas negras, no siendo conocido el proceso de inferencia ni siendo interpretable por humanos (Guidotti et al. 2018).

Por lo general, la explicabilidad de un modelo de aprendizaje automático se considera inversamente proporcional a su precisión predictiva: cuanta mayor precisión, menor capacidad de explicar el modelo (Xu et al. 2019). Un gráfico representativo de esta cuestión es el facilitado por el programa DARPA sobre IA explicable. En él (Figura 1) se muestra cómo los árboles de decisión tienen un gran grado de explicabilidad, pero una peor precisión en la predicción frente a las técnicas de aprendizaje automático que se muestran, mientras que los modelos de aprendizaje profundo tienen una menor capacidad para explicar los resultados.

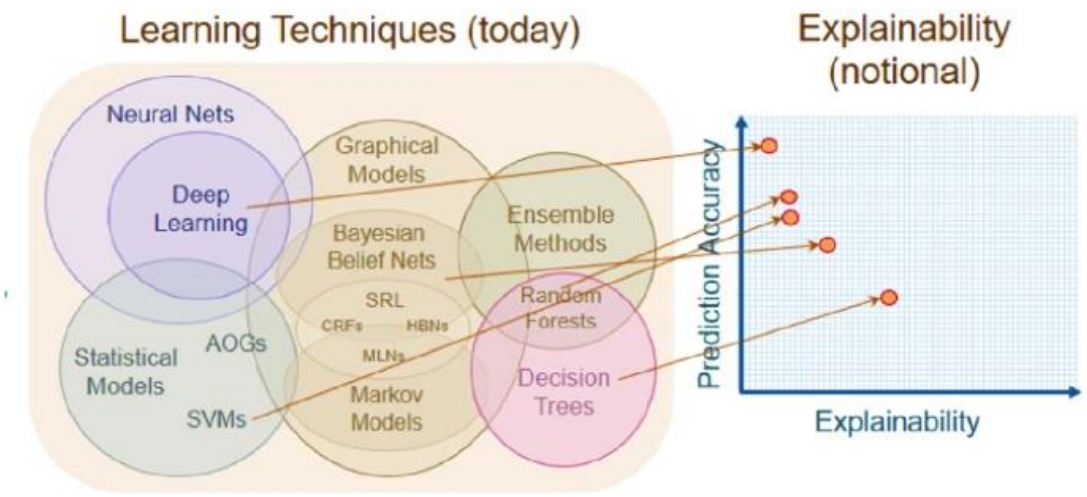

Figura 1. Gráfico DARPA de la explicabilidad de los modelos

Este aumento en la atención a la IA explicable, tanto en la investigación como para el público general llevó a DARPA a financiar el programa “Explainable IA (XAI)” con el objetivo de mejorar la explicabilidad de los modelos de la IA. Mientras que en 2017 el gobierno chino publicó el “Plan de Desarrollo para la nueva generación de IA”, fomentando el desarrollo de modelos IA con una alta explicabilidad. Por parte de Europa, en 2018 se promulgaba el Reglamento General de Protección de Datos (GDPR), reflejando el derecho a una explicación para los ciudadanos por las decisiones tomadas por los algoritmos.

En mayo de 2018, se promulgó el "Reglamento General de Protección de Datos" (GDPR), en el que la Unión Europea otorga a sus ciudadanos el "derecho a una explicación" si se ven afectados por decisiones tomadas mediante algoritmos. La IA explicable será cada vez más importante para todos los grupos de interés, incluidos los usuarios, las personas afectadas y los desarrolladores de sistemas de IA (Goodman and Flaxman, 2017).

En esta investigación se analiza el creciente interés en la Inteligencia Artificial Explicable, no solo en el ámbito científico, mediante el estudio de las publicaciones y las citas relacionadas si no también el interés general mediante la búsqueda de términos en Google Trends.

Metodología

En este informe se recopilan y analizan los artículos científicos y las búsquedas en Google relativas a la explicabilidad de los modelos. Los artículos recopilados se recogen hasta el 18 de enero de 2025 mediante la metodología empleada en https://github.com/alonjacovi/XAI-Scholar, suponiendo una actualización de la investigación e incorporando los resultados obtenidos en Google Trends. El principal objetivo de este análisis es recopilar un amplio set de artículos sobre esta temática, posibilitando el análisis empírico y la identificación de tendencias en esta temática. Los resultados muestran amplias tendencias interdisciplinares.

Como refleja el autor, la investigación en XAI representa varias particularidades, complicando su análisis en comparación con otros campos.

1. Su carácter multidisciplinar da lugar a comunidades relevantes en diversas disciplinas que rara vez interactúan o comparten espacios de colaboración.
2. Los términos empleados para identificar investigaciones como parte de XAI no son exclusivos de este ámbito (por ejemplo, "xai" y "Xai Xai" tienen diversos significados en

publicaciones académicas), y el uso de esta terminología es mucho más reciente que el desarrollo histórico del XAI.

3. Las definiciones más comunes de lo que constituye un artículo sobre XAI suelen incluir trabajos que no se etiquetan a sí mismos como tales, siempre que su objetivo sea explorar formas de explicar las tecnologías de IA.

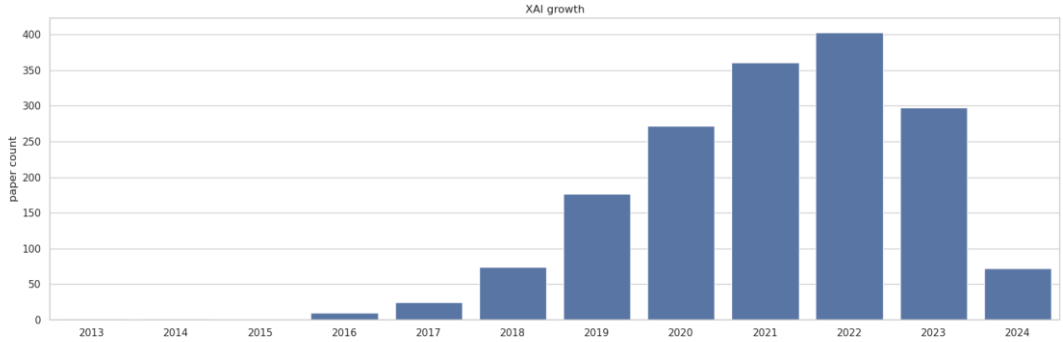

Figura 2. Gráfico del número de artículos científicos publicados en cada año sobre Inteligencia Artificial explicable.

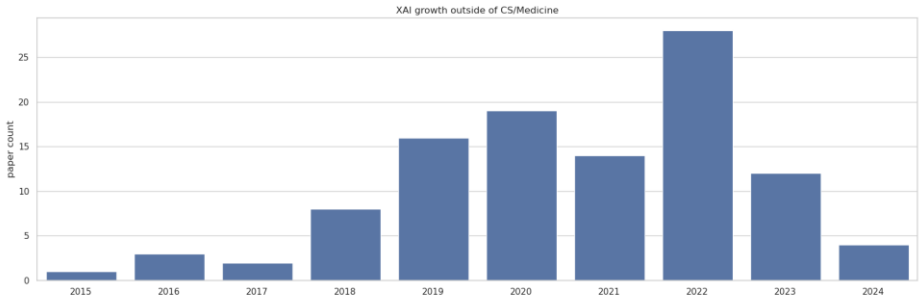

Figura 3. Gráfico del número de artículos científicos publicados en cada año sobre Inteligencia Artificial explicable fuera del ámbito de la informática.

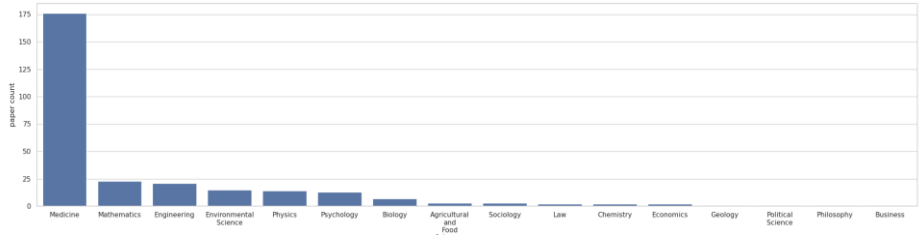

Figura 4. Gráfico del número de artículos científicos publicados en diferentes ámbitos sobre Inteligencia Artificial explicable entre los años 2013 y 2024.

Resultados. Algunos de los resultados obtenidos pueden resumirse en los siguientes puntos:

1. Como puede verse en la figura 2, la publicación de artículos relacionado con la Inteligencia Artificial explicable ha sufrido un crecimiento hasta el año 2022, produciéndose después una desaceleración en los años posteriores.

2. La figura 3 muestra que el mayor pico de crecimiento en la investigación de XAI fuera del ámbito de la informática ocurrieron en los años 2016, 2018 y 2021.

3. Existen diferencias entre los campos de XAI en cuanto a la frecuencia con la que influyen en investigaciones fuera del ámbito de XAI. Las áreas con mayor proporción de influencia externa son XAI-Biología, XAI-Ingeniería y XAI-Derecho, mientras que las de menor impacto en estudios externos son XAI-Psicología, XAI-Negocios y XAI-Filosofía, cuyos aportes tienden a impactar más frecuentemente dentro de la literatura de XAI.

4. El comportamiento de citación también varía significativamente entre disciplinas. Por ejemplo, los artículos de Filosofía más citados por XAI-Filosofía son diferentes de aquellos citados por XAI-Ingeniería Informática (CS), y así sucesivamente. De manera predecible, los trabajos fuera de un campo tienden a enfocarse en un número reducido de artículos dentro de ese campo. Sin embargo, los trabajos que logran "romper" las fronteras tradicionales de su disciplina no siempre son los más citados en su propia área.

5. Las relaciones de citación entre la informática y otros campos del XAI muestran dinámicas variadas. Por ejemplo, el área de XAI en informática cita con más frecuencia a la psicología dentro de XAI que al contrario, mientras que ocurre lo opuesto en el caso de XAI en medicina. Estas direcciones de influencia revelan cuáles disciplinas suelen proporcionar conocimiento e informar a otras dentro del cuerpo actual de literatura.

6. Además, esta colección puede servir como un motor para descubrir artículos relevantes al observar cuáles trabajos de XAI son más influyentes fuera de su propio campo o incluso fuera del ámbito de XAI, o qué artículos no relacionados con XAI en un área específica son más útiles para otro campo.

Datos

Recopilación de la información:

Búsqueda Inicial por Palabras Clave

- **Resultados:** 1723 artículos.
- Se definió un conjunto de palabras clave mediante un proceso iterativo para garantizar alta precisión.
- Las palabras clave debían coincidir al menos dos veces entre el título y el resumen.
- **Palabras clave principales:** *"explainability," "interpretability," "feature importance," "global explanation," "local interpretation," etc.*

Incorporación de Fuentes Curadas

- **Resultados adicionales:** +766 publicaciones (total: 2489).
- Se integraron artículos de bases y colecciones confiables existentes, utilizando coincidencia difusa a través de APIs para validar títulos.

Expansión de Árbol de Citas con Revisión Manual

- **Resultados adicionales:** +648 publicaciones (total: 3137).

- Se analizaron los artículos más citados y se seleccionaron manualmente aquellos relevantes a XAI.

Expansión de Árbol de Citas con Filtros Automatizados

- **Resultados adicionales:** +709 publicaciones (total: 3846).
- Se automatizó el filtrado de citas y referencias, aplicando los criterios de palabras clave de la primera etapa.

Revisión Final y Depuración

- **Resultados eliminados:** -25 publicaciones incorrectas (total final: 3821).
- Se realizó una validación manual para garantizar la calidad de la colección, eliminando entradas mal clasificadas.

Información de los datos

La base de datos contiene un total de 3821 documentos. Para cada uno de estos documentos, la información obtenida a través de la API de SemanticScholar incluye:

- Identificador único y enlace a la página de SemanticScholar
- Nombre completo del artículo
- Resumen de la investigación
- Autores que contribuyeron al trabajo
- Cantidad de citas recibidas por el artículo
- Cantidad de referencias citadas en el artículo
- Año de publicación del artículo
- Revista o conferencia en la que fue publicado
- Categoría académica o disciplina a la que pertenece el artículo
- Resumen breve generado automáticamente por SemanticScholar (tldr)
- Listado completo de las referencias utilizadas
- Número de citas que el artículo ha generado
- Vector de representación utilizado por SemanticScholar

Resultados

El creciente interés en la explicabilidad de los modelos de inteligencia artificial ha dado lugar a un notable aumento en la literatura relacionada, lo que refleja una creciente conciencia sobre la necesidad de comprender cómo los sistemas de aprendizaje automático toman decisiones. Este campo, conocido como inteligencia artificial explicable (XAI), ha experimentado un desarrollo acelerado en los últimos años, impulsado tanto por avances tecnológicos como por un enfoque más riguroso en garantizar que los modelos sean transparentes, justos y comprensibles para los usuarios.

El análisis de la evolución de la literatura muestra que, a medida que las aplicaciones de la inteligencia artificial se expanden a sectores críticos como la salud, las finanzas, la seguridad y la automatización industrial, también ha aumentado el enfoque en la interpretabilidad de los modelos. La necesidad de explicaciones claras y comprensibles sobre cómo un modelo llega a

sus conclusiones es cada vez más urgente, especialmente en aplicaciones que pueden tener un impacto directo sobre la vida humana, como en la medicina o en la toma de decisiones legales.

El volumen de publicaciones en el campo de XAI ha experimentado un crecimiento exponencial, con una cantidad significativamente mayor de artículos científicos que abordan la explicabilidad de modelos de aprendizaje automático. Este incremento no solo se refleja en la cantidad de trabajos publicados, sino también en la cantidad de citas recibidas, lo que evidencia un mayor reconocimiento de la relevancia de la explicabilidad en la comunidad académica. La tendencia es clara: las investigaciones sobre la transparencia y la comprensión de los modelos no solo se están multiplicando en número, sino que también están ganando atención y visibilidad en diversas áreas de estudio. A medida que se amplían las aplicaciones de la inteligencia artificial, también lo hace el enfoque multidisciplinario de la investigación, que ahora incluye áreas como la psicología, la biología, la ingeniería, el derecho y la filosofía, entre otras.

Este auge en las publicaciones ha dado lugar a un ciclo de retroalimentación positiva, donde los artículos más influyentes en el área han generado un número creciente de citas, lo que, a su vez, refuerza la importancia de este campo de estudio. Investigaciones previas sobre explicabilidad, justicia y transparencia han formado la base para nuevas líneas de investigación, consolidando a XAI como una prioridad en la agenda científica y académica. Además, este aumento en el número de citas refleja no solo el interés académico, sino también la integración de los avances en XAI en las aplicaciones industriales y empresariales, que buscan implementar soluciones más transparentes y responsables en el uso de la inteligencia artificial.

Referencias

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

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8* (pp. 563-574). Springer International Publishing.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, *38*(3), 50-57.