**TABLE OF CONTENTS**

[1 INTRODUCTION 3](#_Toc143010806)

[2 OBJECTIVE AND RELEVANCE 4](#_Toc143010807)

[3 METHODOLOGY 5](#_Toc143010808)

[3. 1 Data sources 5](#_Toc143010809)

[3. 2 Sample 5](#_Toc143010810)

[3. 3 Data extraction 6](#_Toc143010811)

[3.3.1 Google News 6](#_Toc143010812)

[3.3.2 RSelenium 7](#_Toc143010813)

[3. 4 Data cleaning 7](#_Toc143010814)

[4 ANALYSIS 8](#_Toc143010815)

[4.3 Text analysis 8](#_Toc143010816)

[4.4 Data visualization 8](#_Toc143010817)

[5 RESULTS 8](#_Toc143010818)

[6 CONCLUSIONS AND LIMITATIONS 8](#_Toc143010819)

[7 REFERENCES 8](#_Toc143010820)

**Abstract**

# INTRODUCTION

+1200

En la última década en España, al igual que en el resto del mundo occidental, se ha experimentado un creciente interés por el desarrollo de la Inteligencia Artificial. Este interés ha venido precedido por el lanzamiento de diversas herramientas que incorporan esta tecnología. Ejemplo de ello son los asistentes virtuales inteligentes como Siri de Apple, Google Assistant, Alexa de Amazon o Microsoft Cortana. También empresas como Tesla han avanzado mucho en el desarrollo de vehículos autónomos que incorporan la IA para reconocer el entorno e incluso tomar decisiones autónomas. Otros campos en los que esta tecnología ha supuesto una revolución son los sistemas de recomendación en diversas plataformas de entretenimiento (por ejemplo, Netflix o Spotify); en la medicina, donde se han desarrollado algoritmos para el diagnóstico médico; y en el procesamiento del lenguaje natural que ha permitido el desarrollo de chatbots como Chat-GPT o Midjourney.

La irrupción de la IA en la sociedad ha puesto sobre la mesa diversos desafíos éticos y sociales derivados de su uso. Cada vez más gente muestra interés por esta tecnología y se pregunta cosas sobre ella. Los medios de comunicación han plasmado este debate a la vez que han sido líderes de opinión mediante la labor divulgativa de algunos temas que todavía no habían llegado a la sociedad. Además, en los últimos años hemos visto como esta cuestión ha llegado cada vez más lejos, haciendo que también los representantes políticos se vean en la necesidad de sentarse a tomar decisiones sobre la regulación de la IA. En el contexto de la Unión Europea se está debatiendo la propuesta de la Ley de Inteligencia Artificial (cita) que tiene entre otros propósitos el de establecer un sistema para evaluar el riesgo que una tecnología que incorpore la IA puede suponer para la salud y la seguridad de las personas (Newtral).

Este Trabajo de Final de Master se adentra en el debate a través de las noticias sobre Inteligencia Artificial publicadas en dos medios de comunicación digitales españoles. A través de técnicas de escrapeo web se obtendrán los textos de las noticias que posteriormente se analizarán mediante análisis de texto. Este trabajo pretende contribuir de forma empírica al análisis de noticias, es por ello que todo el código usado para realizar el proceso de extracción de datos, limpieza, análisis y visualización quedará adecuadamente recopilado en varios archivos formato rmd que estarán subidos en el siguiente enlace de GitHub. El trabajo se estructura en seis apartados diferentes: primero se explican los objetivos y la relevancia de este, haciendo un pequeño análisis de la literatura existente sobre el tema. Posteriormente se presentará de forma breve la metodología y las fuentes de datos empleadas. En el apartado principal del trabajo se explica cómo se ha realizado la extracción de datos, su limpieza, el análisis de texto y la visualización. Finalmente, se exponen los resultados y se esbozaran las conclusiones.

# OBJECTIVE AND RELEVANCE

El objetivo de este trabajo es realizar una aportación empírica en el ámbito de la extracción de datos de noticias de periódicos digitales y su análisis de texto a través del entorno de programación R. El código creado para realizar la extracción del contenido de las noticias será replicable, por lo que cualquier persona que quiera realizar un análisis similar podrá hacer uso de este código y adaptarlo para extraer la selección de noticias que necesite. Esto es importante porque actualmente estos periódicos no tienen sus noticias en ninguna API ni en otro formato accesible, por lo que la extracción de datos mediante la técnica de web scraping es la única forma posible de analizar el contenido. Por otra parte, el análisis de texto a través de R también facilita mucho poder extraer conclusiones de grandes volúmenes de texto sin tener que recurrir a su lectura pormenorizada, lo cual supondría invertir mucho tiempo. R cuenta con una gran cantidad de paquetes especializados en el procesamiento del lenguaje natural (NLP), que permiten procesar y analizar los textos de manera más rigurosa y eficiente, así como herramientas de visualización de datos textuales que ayudan a comunicar estos resultados de forma comprensible. Este código también estará disponible para quien quiera realizar el análisis de otra base de datos de texto diferente a la que se va a usar en este trabajo.

El análisis de las noticias publicadas en los medios de comunicación es útil especialmente cuando no se dispone de otra fuente de datos sobre la opinión de los ciudadanos en torno a un tema concreto. Como la irrupción de la Inteligencia Artificial es relativamente reciente, no hay apenas encuestas ni material cualitativo (entrevistas, grupos de discusión…) que permita conocer qué piensa la ciudadanía de esta tecnología. En este contexto, analizar la prensa permite obtener una información muy valiosa sobre cuáles son los temas principales de los que se habla, qué tipo de datos se divulgan y cuáles son los temas conflictivos que se están poniendo sobre la mesa. Además, este trabajo va a incluir noticias de periódicos de diferentes ideologías, lo que también permite comparar el posicionamiento de cada uno de ellos respecto de la inteligencia artificial. En definitiva, el análisis del texto de las noticias sobre inteligencia artificial que se va a realizar en este trabajo va a permitir identificar tendencias y temas principales, evaluar el tono y la opinión predominante, detectar sesgos y analizar la cobertura mediática de este tema en el periodo de tiempo analizado.

Existen ejemplos de otros trabajos de investigación que también han aplicado técnicas de análisis de texto para estudiar el contenido de las noticias de medios de comunicación. Añadir papers.

El trabajo de investigación se estructura en cuatro apartados diferentes. En primer lugar se va a explicar la metodología empleada y las fuentes de donde se ha extraído la información. Este apartado es especialmente importante puesto que al tratarse de un trabajo empírico tiene una especial relevancia el trabajo de extracción de datos realizado. Por ello, se dedicará un apartado propio a explicar cómo se ha llevado a cabo el proceso de extracción de datos mediante el lenguaje de programación R y la herramienta RStudio. Posteriormente, se explicará cómo se han limpiado los datos. El cuarto apartado consistirá en un análisis de texto donde también expondrán las visualizaciones realizadas que permitirán profundizar en ese análisis. El trabajo finaliza con un breve apartado con las conclusiones y las limitaciones del mismo.

# METHODOLOGY

+400

## 3. 1 Data sources

En este trabajo de investigación se usarán datos extraídos de dos fuentes primarias, la web de dos periódicos digitales españoles: elDiario.es (<https://www.eldiario.es/>) y El Mundo ([elmundo.es](https://www.elmundo.es/)). Se ha escogido estos dos periódicos porque se trata de dos periódicos generalistas de ámbito nacional, por tanto, son comparables entre sí. Además, tienen una línea editorial ideológicamente diferente, por lo que la comparación puede resultar más interesante. Mientras que elDiario.es se sitúa a la izquierda en el espectro ideológico, El Mundo se sitúa a la derecha.

La extracción de los datos se ha realizado mediante técnicas de “data harvesting” con el programa RStudio. Para ello se ha elaborado un código mediante el cual se han podido extraer tres variables diferentes de cada artículo publicado en estos dos periódicos: el título, el contenido del artículo y la fecha de publicación. En el apartado 3 se explicará con mayor detalle el proceso de extracción de los datos.

## 3. 2 Sample

La muestra empleada en este trabajo está compuesta de todos los artículos publicados en la página web de los periódicos elDiario.es y El Mundo que contienen en el texto del artículo las palabras en español “inteligencia artificial”. Además, se ha filtrado por todos los artículos publicados a partir de julio de 2014, para tener una muestra comparable en ambos periódicos. Aunque el artículo de esas características más antiguo publicado por elDiario.es se remonta al año 2009, en el caso de El Mundo la fecha más antigua es julio de 2014. Es por eso por lo que se ha decidido eliminar todas las filas que corresponden artículos publicados en fechas previas. Por otra parte, el artículo más reciente recogido en la base de datos creada para este trabajo fecha del 18 de mayo de 2023. Ambos periódicos tienen al menos un artículo publicado en esa fecha.

La muestra de artículos no incorpora aquellos que tienen acceso restringido. Ambos periódicos digitales tienen algunos artículos para cuyo acceso es necesario pagar una subscripción. Se trata de los artículos más recientes, ya que una vez pasado un tiempo los ponen a disposición del público general. Es por ello por lo que al tratarse de un número pequeño y limitado de artículos no supone un problema para la investigación.

En total se han analizado 1251 artículos de El Mundo y 1056 de elDiario.es. En la siguiente tabla se muestra el número de artículos analizados por periódico y año de publicación:

Tabla 1: Número de artículos publicados por periódico y año de publicación

|  |  |  |
| --- | --- | --- |
| **Year** | **elDiario.es** | **El Mundo** |
| 2014 | 4 | 1 |
| 2015 | 21 | 10 |
| 2016 | 50 | 111 |
| 2017 | 69 | 229 |
| 2018 | 103 | 174 |
| 2019 | 99 | 175 |
| 2020 | 212 | 157 |
| 2021 | 122 | 137 |
| 2022 | 224 | 188 |
| 2023 | 152 | 69 |

## 3. 3 Data extraction

+1200

Como se ha explicado en apartados anteriores, los datos analizados en este trabajo de investigación se han obtenido mediante técnicas de “web-scraping” realizadas a través de la herramienta RStudio. Se trata de una forma de obtención de datos no tan convencional pero muy útil cuando los datos no están accesibles de forma directa. Es decir, en este caso no había forma de descargar los textos de los artículos desde la página web de los periódicos ni desde ninguna otra web. Por ello, la única alternativa posible sería guardar toda la información de forma manual (copiando y pegando el texto en un dataframe). Sin embargo, dada la gran cantidad de noticias publicadas por los periódicos esto llevaría mucho tiempo. Por otra parte, una de las principales ventajas de crear un código en R reproducible es que se puede usar para realizar extracciones de otro tipo de noticias y publicadas en distintas fechas. Para ello solo habría que realizar pequeñas modificaciones en el código original.

Tal y como se ha explicado anteriormente todo el código empleado en este trabajo está accesible a través de GitHub, en formato rmd. El archivo que contiene la parte de extracción de datos se llama “web\_scraping\_ai.rmd”. En este archivo está todo el código empleado para generar un archivo csv con la extracción del texto de los artículos, el título y la fecha de publicación. El archivo se encuentra estructurado en 14 apartados diferentes en los que se va explicando paso a paso el proceso de extracción de datos. Se trata, por tanto, de un código fácilmente replicable en caso de que alguien necesite realizar una extracción similar en algunos de los dos periódicos empleados en este trabajo. A continuación, se procederá a explicar cómo se ha realizado este proceso y algunos de los problemas que han surgido durante el mismo.

### Google News

En un primer momento la idea era extraer el texto de los artículos a través de Google News. Google News es un buscador de noticias que permite introducir un término o varios y un rango de fechas (por ejemplo: *eldiario.es: "inteligencia artificial" site:eldiario.es after: 2023-04-03 before: 2023-05-03*) y te devuelve una serie de links con los artículos filtrados por el periódico o la web que desees. La ventaja de usar este motor de búsqueda es que el código para ambos periódicos sería muy similar. Además, los criterios de búsqueda de artículos también serían parecidos lo que lo hace más comparable. Aunque en el archivo rmd se muestra todo el código necesario para extraer los artículos de esta forma, finalmente se optó por no emplear esos datos. Esto es debido a que Google News ofrece en sus búsquedas un número limitado de links. Es decir, por cada búsqueda que se hace en este motor de búsquedas aparecen un máximo de 30 páginas con 10 links por página. Esto limita el número final de artículos a 300. Sin embargo, se ha considerado útil mantener el código que permite la extracción de los datos de esta manera, por si alguien necesita utilizarlo.

### 3.3.2 RSelenium: extract the links

Una vez descartada la opción de extraer los datos a través de Google News la única opción posible era emplear los motores de búsqueda propios de cada periódico. Para ello se introducen en los motores de búsqueda las palabras entrecomilladas “inteligencia artificial”[[1]](#footnote-1). En ambos casos aparece una página en la que se muestra el título (con un hipervínculo a la página donde está el artículo completo) y la fecha de publicación de una serie de artículos. Para poder ver más hay que pulsar el botón de “siguientes” en la parte inferior de la página. Lo primero que hay que hacer, por tanto, es crear un dataframe con tres variables: el título, la fecha de publicación y el link al artículo completo. Para poder hacer esto es necesario usar un paquete específico de R llamado “RSelenium” que permite interactuar desde RStudio con la página web. Es decir, permite navegar y pulsar en los enlaces y botones de la web de forma remota.

Antes de empezar a usar RSelenium es conveniente instalar un Docker (<https://www.docker.com/>). Se trata de un contenedor donde se pueden ejecutar aplicaciones y que hace que sea más sencillo y ordenado el proceso de interactuar con un sitio web cuando se emplea RSelenium. Una vez instalado e inicializado RSelenium es necesario crear un código que permita leer el html de cada página, extraer la información necesaria a través de los XPaths y pulsar en el botón “siguiente” para pasar de página y repetir el proceso. Este proceso se hace a través de un loop que al finalizar guarda toda la información en un dataframe creado previamente. Además, previamente hay que utilizar la “herramienta para desarrolladores” en la página web de la búsqueda de cada periódico para poder localizar en el html dónde está la información que se quiere extraer y poder crear el XPath adecuado en R. Esto se hace con la ayuda de la librería “scrapex” que incorpora funciones que permiten la lectura y extracción de datos de páginas web.

Cómo cada periódico tiene un diseño web diferente, el código no es intercambiable, sino que hay que crear un XPath propio para cada variable de cada periódico. Por tanto, también hay que crear dos loops que generen dos dataframes, uno para cada periódico. Además, ambas webs tienen una forma distinta de mostrar las noticias una vez se pulsa en el botón “siguientes”. Mientras que en la web del periódico El Mundo cada vez que se pulsa este botón aparece una página diferente con los nuevos enlaces, en la web de elDiario.es aparecen nuevos enlaces a continuación de los anteriores. De esta forma, en el primer caso hay que extraer la información y después pulsar el botón y volver a extraer. En el segundo caso hay que pulsar el botón “siguientes” tantas veces como se pueda y después extraer toda la información de una sola vez. Para poder interactuar con la web, en el caso de elDiario.es también es necesario aceptar las cookies. Para ello hay que buscar el XPath que identifique el botón de aceptar y pulsarlo de forma remota.

### 3.3.3 Extract the text of the articles

Una vez que se ha creado el dataframe con el link a cada artículo se trata de localizar de nuevo en el html la parte en la que se encuentra el texto que interesa extraer y crear el XPath adecuado. Cuando ya se ha verificado que el XPath localiza la información necesaria, hay que crear de nuevo un loop que itere por cada link en el vector creado con los links y extraiga el texto y lo guarde en un dataframe. Una vez realizado este proceso para los dos periódicos ya se dispone de un dataframe con las tres variables necesarias. Sin embargo, al revisar los datos extraídos surge una complicación: hay una gran cantidad de filas vacías o “NAs” en la columna “texto” del conjunto de datos de El Mundo. Esto es debido a que probablemente por un cambio en la web no en todos los artículos se extrae el texto con el mismo XPath. Es por esto que es necesario crear un nuevo loop que itere por los links que no tienen texto y que lo extraiga con el nuevo XPath.

## 3. 4 Data cleaning

+ 300

Una vez que se dispone de los dos dataframes con el texto de los artículos, la fecha de publicación y el título hay que limpiar los datos para que se puedan realizar visualizaciones y análisis de texto. El código para realizar esta limpieza de datos está disponible en el archivo rmd “data\_cleaning\_ai”. Antes de realizar la limpieza de datos se eliminan las filas duplicadas y se juntan los dataframes: el que contiene el texto y el que contiene la fecha de publicación y el título del artículo. También se añaden las filas que inicialmente contenían valores vacíos en la columna del texto. Posteriormente, se transforman las fechas a un formato adecuado para R.

# ANALYSIS AND RESULTS

## 4.4 Data visualization: trends in the number of articles published between 2015 and 2023

+ 500

Para poder ver el número de artículos publicados a lo largo del tiempo por ambos periódicos se ha realizado un gráfico de líneas. Primero se han filtrado los datos por la columna “texto” para sólo mantener aquellos que contengan las palabras “inteligencia artificial” en su contenido, y que así sea comparable. Además, como se explicaba en el apartado en el que se habla de la muestra, se han limitado los artículos a aquellos publicados entre el 28-07-2014 y el 30-04-2023. Como el periodo de tiempo representado en el gráfico es bastante amplio, también se ha creado una columna que indica el cuatrimestre del año en el que está incluida cada fecha y el año. Esta variable es la que se ha representado en el eje X. Por otra parte, en el eje Y se representa el número absoluto de artículos publicados durante cada semestre. La línea verde representa los artículos publicados por elDiario.es y la línea roja representa los publicados por El Mundo.

En el gráfico también se han incluido las fechas clave de lanzamiento de algunas herramientas de inteligencia artificial. Esto permite ver si el patrón de publicaciones de artículos varía en relación con estos sucesos. Todo el código necesario para realizar esta visualización se encuentra disponible en el archivo rmd “visualizations\_ai.rmd”.

Figure 1: Number of articles published per quarter and year (2014-2023)

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Source: own elaboration

Lo primero que se observa al ver el gráfico es que ambos periódicos han seguido una tendencia de publicación de artículos sobre inteligencia artificial diferente. Si bien es cierto que el número de artículos publicados entre 2015 y 2016 es muy bajo en ambos casos, a partir de 2016 en el caso de El Mundo se incrementa significativamente el número absoluto de publicaciones. Pasa de tener cerca de 25 artículos publicados a comienzos de 2016 a tener más del triple en el primer cuatrimestre de 2017 y de 2018. La tendencia durante esos años de elDiario.es es también al alza, pero de forma mucho más pausada. No es hasta 2020 que la tendencia experimenta un crecimiento muy significativo llegando a alcanzar los casi 90 artículos publicados durante el tercer cuatrimestre de ese año.

## 4.3 Text analysis

+ 700

En este apartado se va a proceder a explicar cómo se ha realizado el análisis de texto en R y a exponer los principales resultados. Se han realizado algunas visualizaciones de datos relativas al análisis de texto que permiten entender mejor cómo son es el contenido de las noticias que se están analizando.

4.3.1 Tokenize the text

Para poder realizar el análisis de texto en R el primer paso ha sido limpiar la columna de texto de algunas filas de noticias. En el caso de las noticias del periódico elDiario.es se ha eliminado el texto referido a una agencia de noticias internacional (“Agencia EFE”), así como la ubicación de la publicación, ya que no aportaba información útil. Posteriormente, se ha separado el texto del artículo en palabras (“tokens”), de tal forma que se ha generado un nuevo dataframe con una fila para cada palabra (“Word” columna). El siguiente paso ha sido eliminar aquellas palabras que no son útiles para el análisis de texto. Esto se hace mediante una librería (“stopwords”) que permite guardar en un vector las palabras innecesarias en español. En esta lista de palabras se incluyen distintas formas verbales (estar, ser, haber, tener y hacer), cerca de un 20% de preposiciones, conjunciones (8%), y artículos (7%).

4.3.2 Word frequencies

Una vez que el texto se ha separado en palabras, es interesante ver la frecuencia de palabras en números absolutos por periódico. Para ello se ha creado un gráfico de barras en el que se filtra por las palabras que se han repetido más de 750 veces en total.

# CONCLUSIONS AND LIMITATIONS

+500

# REFERENCES

* [**https://www.newtral.es/ley-inteligencia-artificial-ia-chatgpt-union-europea-ue/20230427/**](https://www.newtral.es/ley-inteligencia-artificial-ia-chatgpt-union-europea-ue/20230427/)

1. <https://www.eldiario.es/busqueda/%22inteligencia%20artificial%22>

   <https://ariadna.elmundo.es/buscador/archivo.html?q=%22inteligencia+artificial%22&b_avanzada=> [↑](#footnote-ref-1)