ISSN: 2951-6641 (papel) 2951-8156 (online)
https://doi.org/10.54790/rccs.51

ARTÍCULO/ARTICLE

Big data en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático

Big Data in Social Sciences. An Introduction to the Automation of Textual Data Analysis Using Natural Language Processing and Machine Learning

Alba Taboada Villamarín
Universidad Autónoma de Madrid, España
alba.taboada@uam.es

Recibido/Received: 11/1/2023
Aceptado/Accepted: 25/7/2023

RESUMEN

Las innovaciones en el campo de la ingeniería computacional y la inteligencia artificial brindan nuevas oportunidades metodológicas para la investigación científica, permitiendo el estudio de fenómenos sociales emergentes que nacen y habitan en los espacios virtuales. El propósito de este trabajo es familiarizar al científico social con los procesos ampliamente establecidos en el análisis masivo de texto mediante técnicas de aprendizaje automático que dan lugar a lo que hoy conocemos como procesamiento de lenguaje natural (PLN). En primer lugar, se lleva a cabo un breve recorrido por la historia del PLN y su relación con el análisis de texto en las ciencias sociales. Luego, en cada sección del texto, se valoran los pasos a seguir cuando se aplica PLN a investigaciones de carácter social, proporcionando información sobre programas informáticos, herramientas, fuentes de datos y enlaces útiles, con el propósito de ofrecer una guía introductoria y simplificada que sirva como acercamiento inicial a esta disciplina. Por último, se examinan y evalúan los principales desafíos que las ciencias sociales enfrentan al implementar técnicas de PLN.

PALABRAS CLAVE: datos masivos; procesamiento de lenguaje natural; ciencias sociales; aprendizaje automático, minería de texto.

CÓMO CITAR: Taboada Villamarín, A. (2024). *Big data* en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático. *Revista Centra de Ciencias Sociales*, 3(1), 51-75. https://doi.org/10.54790/rccs.51

English version can be read on https://doi.org/10.54790/rccs.51

ABSTRACT

Innovations in the field of computer engineering and artificial intelligence provide new methodological opportunities for scientific research, enabling the study of emerging social phenomena that are born and inhabit virtual spaces. The purpose of this paper is to familiarise the social scientist with the widely established processes in massive text analysis using machine learning techniques that give rise to what we know today as natural language processing (NLP). First, a brief overview of the history of NLP and its relation to text analysis in the social sciences is given. Then, in each section of the text, the steps to follow when applying NLP to social research are assessed, providing information on software, tools, data sources and useful links, with the aim of offering an introductory and simplified guide to serve as an initial approach to this discipline. Finally, the main challenges that the social sciences face when implementing NLP techniques are examined and assessed.

KEYWORDS: big data; natural language processing; social sciences; machine learning, text mining.

1. Introducción: Una aplicación del *big data* a las ciencias sociales. El procesamiento de lenguaje natural (PLN)

El procesamiento de lenguaje natural (PLN) hace referencia al campo de estudio de las ciencias computacionales que, en convergencia con la lingüística, permite a determinados sistemas informáticos procesar y «entender» el lenguaje humano (Bird, Klein y Loper, 2009). El lenguaje, en forma de texto escrito, constituye una fuente primordial de documentación humana de gran importancia en los contextos de investigación social. El análisis de texto ha experimentado una trayectoria amplia, involucrando numerosas técnicas de investigación y herramientas metodológicas que han permitido afinar el uso de esta información, tanto como fuente de datos primaria como secundaria, especialmente en el ámbito de los enfoques cualitativos.

No obstante, la importancia del texto como unidad de análisis se extiende a diversas disciplinas de conocimiento. Las ciencias computacionales han demostrado un creciente interés en automatizar y desarrollar máquinas capaces de acercar el lenguaje humano al «lenguaje máquina». Los esfuerzos por extraer información sustantiva de corpus textuales, proceso conocido como minería de texto (Justicia de la Torre *et al.*, 2018), esbozan trayectorias paralelas en ambas disciplinas, con puntos de convergencia que resultan determinantes para los avances metodológicos en la investigación social.

La literatura especializada sugiere que la emergencia de la Guerra Fría generó la necesidad de desarrollar máquinas capaces de realizar traducciones automáticas de texto en diversos idiomas, destacando la traducción del ruso al inglés. Como respuesta a esta demanda, surgieron los primeros sistemas simbólicos de análisis textual mediante máquinas (Johri *et al.*, 2021). Paralelamente, aunque el análisis de texto ya tenía una trayectoria consolidada en las disciplinas de antropología y sociología, hacia el final de la Segunda Guerra Mundial, los estudios que examinaban la correspondencia entre migrantes y soldados adquirieron significatividad metodológica en los

trabajos pioneros de la Escuela de Chicago (Abbott, 1997). En aquel entonces, la sistematización del análisis de datos textuales, tanto de fuentes primarias como secundarias, implicaba un esfuerzo considerable en términos de etiquetado, organización y gestión de los textos. Esta labor manual dio lugar al desarrollo de diversas ramas metodológicas que hasta hoy día marcan las distintas líneas analíticas presentes en la investigación cualitativa.

En sus etapas iniciales, el procesamiento de lenguaje natural fundamentó su trabajo en las teorías chomskynianas de estructuras sintácticas, las cuales recibieron contundentes críticas por parte de otros lingüistas (véanse Radick, 2016; Hockett, 2020). Estos críticos argumentaron que el lenguaje humano involucra complejidades que exceden las reglas de asociación y los modelos comparativos, fundamentales en la lógica computacional en sus primeras etapas. Esta discusión, aún hoy en activo, se inclinó hacia los avances en computación cuando se incorporaron los primeros cálculos estadísticos en el procesamiento del lenguaje humano mediante máquina (Bitter *et al*., 2010). Estos desarrollos permitieron abordar las características peculiares y variables del lenguaje humano, superando las limitaciones de los enfoques basados únicamente en reglas sintácticas.

Este cambio de paradigma alcanzó su punto álgido en la década de los años noventa, en estrecha relación con el crecimiento de las telecomunicaciones y la amplia difusión de los ordenadores personales, configurando lo que actualmente conocemos como la Sociedad de la Información (Castells, 1997). La «etapa estadística» permitió la complejización de los análisis de texto, dando lugar al lanzamiento de los primeros programas informáticos especializados en el análisis de datos cualitativos (CADQAS), tales como Atlas.Ti (1993) o Nvivo (1999). Por primera vez fue factible llevar a cabo tareas de etiquetado de texto de manera semiautomatizada. El recuento de palabras y la capacidad de realizar cálculos con sus frecuencias estimularon enfoques como el análisis de contenido y técnicas aproximadas al análisis mixto o a la triangulación.

Desde las primeras técnicas rudimentarias hasta la época actual, la sociedad de la información ha experimentado cambios significativos que configuran un nuevo escenario para el PLN y, por ende, un posible avance en los métodos y técnicas de análisis textual dentro de la investigación social. Tres ejes principales establecen los retos sobrevenidos por los desarrollos en las áreas de las TICs, propiciando la creación de nuevos modelos sociales, científicos y técnicos.

En primer lugar, la intensa competencia en el mercado internacional, debido a la inclusión de economías orientales y del sur global, ha resultado en una reducción sin precedentes en los costos de los materiales en la industria tecnológica. Esto ha posibilitado la expansión e interconexión de las estructuras de telecomunicaciones, que se han convertido en la columna vertebral de la sociedad virtual. En segundo lugar, los avances en áreas como la computación, las matemáticas aplicadas, la estadística y la robótica han impregnado de inteligencia a estas interconexiones, dando lugar a lo que conocemos como inteligencia artificial. Esta inteligencia artificial se aleja del enfoque clásico de la informática basado en la acción y reacción, y adopta modelos interactivos que pueden generar respuestas múltiples ante una amplia variedad de entradas. Por último, el elemento central

que alimenta y se ve influenciado por estas dos infraestructuras es el denominado *big data*.

En el ámbito de la discusión científica a nivel internacional, se ha observado un creciente volumen de investigaciones que abordan este objeto de estudio. Desde la perspectiva del científico social, los *big data* se refieren a la totalidad de los rastros digitales generados por las interacciones entre seres humanos, entre humanos y máquinas, y entre máquinas en el espacio virtual. Investigaciones previas (Gualda *et al*., 2023) han resaltado que el análisis de texto se ha convertido en uno de los enfoques metodológicos más populares al combinar tecnologías *big data* con las ciencias sociales. Esta elección se debe principalmente a que un considerable porcentaje de estos rastros digitales se almacena en formato de texto.

En la extensa red de internet se registran continuamente miles de interacciones provenientes de plataformas como redes sociales, blogs personales, páginas web, mensajería instantánea y foros virtuales. Esta información constituye un reflejo de nuevas narrativas, discursos, representaciones sociales, interacciones y relaciones que se extienden tanto en el entorno en línea como fuera de él, propiciando fenómenos característicos de nuestra contemporaneidad como la propagación de noticias falsas, discursos de odio, tendencias virales, polarización de la información, desconfianza en los sistemas democráticos y científicos, relaciones virtuales y redes de influencia, entre otros.

Si bien la exploración de los nuevos formatos de sociabilidad y sus estructuras reviste una importancia esencial para la investigación social, con frecuencia los científicos enfrentan desafíos al tratar de acceder a estas nuevas realidades. Los problemas que suelen plantear estos tipos de datos incluyen la gestión de grandes volúmenes de información, la velocidad vertiginosa con la que se generan, el formato no estructurado en el que se almacenan y las cuestiones relacionadas con la extracción y propiedad de los mismos (Gillingham y Graham, 2017; Gualda y Rebollo, 2020). Además, la falta de equipos interdisciplinares y el desconocimiento de las herramientas disponibles condicionan considerablemente este tipo de investigaciones.

En el contexto del análisis de texto, el giro estadístico ha sido remplazado por el enfoque de las redes neuronales y el aprendizaje automático, que, en su doble lectura, atiende a una mayor complejidad de análisis pero a una simplificación en su aplicación. Es por ello que este trabajo se esfuerza por reducir las carencias técnicas que en la actualidad se hacen patente en las ciencias sociales, animando a explorar recursos que nos acercan a los problemas sociales emergentes y que tienden puentes con otras disciplinas y objetos de estudio.

En las siguientes páginas se detallan, de forma introductoria y didáctica, los pasos necesarios para aplicar el procesamiento de lenguaje natural (PLN) en una investigación. Se proporciona información práctica sobre los procedimientos y recursos para llevar a cabo estos análisis, incluyendo las fuentes de datos de texto disponibles, las técnicas de extracción de información, la limpieza y el tratamiento de los datos, así como los principales tipos de análisis que se pueden realizar. Por último, se examinarán los principales desafíos que las ciencias sociales enfrentan al implementar técnicas de PLN.

2. Programas para trabajar con lenguaje natural o minería de textos digitales

Cuando se busca llevar a cabo análisis de datos masivos o procedentes de fuentes digitales, es común recurrir a software y entornos de programación que tradicionalmente no formaban parte de la capacitación de los científicos sociales. No obstante, los avances en computación y análisis de datos han simplificado la complejidad de la programación, haciéndola más accesible a todo tipo de usuarios.

Las nuevas herramientas de programación marcan una diferencia significativa al revolucionar la capacidad para analizar datos. Por un lado, presentan un formato de código abierto que permite la descarga gratuita de los entornos y extensiones, además de contar con una comunidad virtual en constante intercambio de información y recursos. Por otro lado, ofrecen una mayor velocidad en la computación y aplicación de cálculos estadísticos, brindando mayor autonomía y control en el refinamiento de los algoritmos. También son capaces de manejar volúmenes de datos más grandes y conectarse fácilmente a diversas fuentes y recursos digitales. Además, estos enfoques incorporan técnicas estadísticas innovadoras de naturaleza predictiva, las cuales no se encontraban previamente disponibles en los programas estadísticos tradicionales.

En la actualidad se distinguen dos enfoques predominantes para el procesamiento de lenguaje natural, según la preferencia por el uso de lenguajes de programación o por la utilización de programas con interfaces de usuario que no exigen la implementación de código. Este último enfoque se presenta como una alternativa más accesible para aquellos investigadores que carecen de conocimientos en informática, pero que desean aplicar análisis de esta índole.

Los dos lenguajes de programación más reconocidos y ampliamente adoptados en el ámbito del análisis de datos son R (*The R Project for Statistical Computing*) y Python. Ambos se caracterizan por ser lenguajes de programación de alto nivel que presentan una sintaxis más accesible y cercana al lenguaje humano que al lenguaje de máquina. Para el lenguaje de programación R es común utilizar el entorno de desarrollo integrado *RStudio*, el cual dispone de diversos cursos y manuales gratuitos dirigidos a principiantes. Los investigadores sociales suelen favorecer el lenguaje de programación R debido a su potencial en el análisis estadístico y la visualización de datos. Existe una amplia disponibilidad de código compartido en la web que puede ser aplicado fácilmente a distintos tipos de datos. Por otro lado, en el caso de Python, se recomienda el uso del entorno *Jupyter Notebook*, que proporciona una forma interactiva de ejecutar el código. Python ofrece una amplia gama de posibilidades tanto en el ámbito estadístico como en el visual. En términos de comparación en el análisis de datos, tanto R como Python presentan capacidades similares, si bien Python es el recurso más ampliamente utilizado por los científicos de datos.

Para aquellos que optan por no utilizar lenguajes de programación existe una amplia variedad de programas que permiten ejecutar los mismos algoritmos sin necesidad de escribir código. En este artículo se proponen dos programas completamente gratuitos y cada vez más populares entre la comunidad académica: *Orange Data Mining*

y *Rapidminer*. Estos programas solo requieren ser descargados en un entorno local, y en sus respectivas páginas web ofrecen tutoriales en vídeo cortos para realizar una amplia gama de análisis.

3. Fuentes de datos textuales

Como se ha indicado, la aplicación de técnicas y métodos de procesamiento de lenguaje natural (PLN) se centra en el análisis de unidades de texto o palabras escritas en lenguaje humano. La variedad y origen de estos textos es amplia, ya que pueden provenir de diversas fuentes donde se emplea el habla o la escritura. No obstante, para poder utilizar técnicas de minería de texto, es necesario que los datos se encuentren en formato digital. Bajo esta premisa, se pueden clasificar tres tipos de fuentes o recursos originales de datos de texto que involucran el lenguaje humano, teniendo en cuenta su naturaleza y contexto de creación. Cada uno de estos tipos implica diferentes procedimientos y consideraciones: 1. Texto analógico, 2. Texto transcrito audiovisual y 3. Texto digital.

3.1. Procedentes de archivos analógicos

En primer lugar, contamos con los recursos «analógicos», es decir, todos aquellos textos clásicos e históricos que se escribieron a mano o a máquina, están impresos y aún no han sido digitalizados. Algunos de los ejemplos son archivos históricos sobre ciudades, correspondencias entre miembros de la realeza o funcionarios, manuscritos, certificados, etc. En las ciencias sociales son clasificados como fuentes de datos secundarios y con frecuencia utilizados en disciplinas como historia, filología o antropología, aunque también son altamente valiosos en cualquier línea de investigación donde se pretende estudiar un momento histórico concreto o recuperar información previa al uso del ordenador.

Este tipo de recursos tienen un alto potencial cuando se aplican métodos y técnicas de PLN, sin embargo, para su explotación, es requisito indispensable que previamente hayan sido digitalizados. Gran parte de los proyectos pioneros que han sentado las bases de las humanidades digitales se han centrado, precisamente, en la digitalización de fuentes históricas (Piotrowski, 2012). Gracias a ello, en la actualidad contamos con múltiples bases de datos abiertas que pueden ser utilizadas por investigadores de todo el mundo de forma gratuita y con herramientas altamente efectivas para la digitalización de textos en el caso de no encontrarlos. Este tipo de recursos, además, pueden resultar de gran valor para el aprendizaje automatizado de la máquina y la comprensión de tipos de lenguajes específicos en épocas concretas.

«Culturomics» (Michel *et al.*, 2011) fue una de las primeras experiencias que alcanzó gran popularidad gracias al uso de los 15 millones de libros digitalizados por Google —actualmente disponibles en Google Books— a partir de los cuales generaron una base de datos de más de 500.000 millones de palabras procedentes de libros escritos entre los años 1500 y 2008 con recientes actualizaciones hasta 2019. El propósito subyacente consistió en generar una suerte de «memoria colectiva digital» en la que poder explotar el uso de palabras y recursos lingüísticos a lo largo de la historia.

De manera similar a este ejemplo, se pueden encontrar numerosas fuentes de textos históricos, tales como bases de datos que albergan periódicos redactados hace varias décadas, que brindan a los investigadores la posibilidad de realizar indagaciones sobre eventos específicos en épocas pasadas. No obstante, antes de poder utilizar este texto en línea, se requiere un proceso previo de tratamiento para almacenarlo en formatos de archivo compatibles con los lenguajes de programación. Una opción recomendada es la de almacenar los archivos de texto sin formato, comúnmente conocido con la extensión «.txt».

3.2. Procedentes de archivos audiovisuales

En segundo lugar, encontramos aquellos textos transcritos que provienen de medios audiovisuales. Es posible querer trabajar con texto adquirido mediante notas de voz o audios, canciones, guiones televisivos, programas de radio, conversaciones producidas en entornos *offline*, etc. Es el caso de las transcripciones que clásicamente se han usado en la investigación cualitativa. Para estos ejemplos, siempre que la transcripción finalmente llegue a ser plasmada de forma digital, las herramientas de procesamiento de PLN serán igualmente válidas.

En este sentido, resultan destacables los recursos que facilitan la conversión automática de archivos audiovisuales a texto digital. Múltiples herramientas ofrecidas de forma gratuita por plataformas como Zoom o Google Docs pueden resultar útiles en este proceso. Una vez obtenido el archivo de texto, se hace necesario, al igual que en el caso de los recursos analógicos, convertirlo a un formato compatible con el programa a utilizar. Los formatos más comúnmente empleados son los archivos de texto sin formato con extensión *.txt*, los archivos «JavaScript Object Notation» conocidos por la extensión *.json*, y los archivos separados por comas «Comma Separated Values» reconocidos por su extensión *.csv*. En ocasiones, estos archivos podrán requerir una estructuración previa por parte del investigador.

3.3. Procedente de archivos digitales

Por último, se dispone de los textos digitales presentes en la web 2.0. Estos recursos se caracterizan por haber sido originalmente redactados en línea y estar accesibles total o parcialmente en el espacio virtual. En el ámbito del procesamiento de lenguaje natural, estos recursos son ampliamente utilizados debido a su diversidad, cantidad y facilidad de acceso. Estas características, en combinación con los atributos clásicos de los grandes conjuntos de datos, convierten a los recursos digitales en los más adecuados para la automatización de procesos y el análisis textual mediante computación. En este contexto se pueden encontrar fuentes de datos tanto secundarias como primarias para investigaciones, si bien la característica principal de los textos en línea radica en la abundancia de datos no solicitados (Ruelens, 2022).

Los recursos digitales que albergan la información sujeta a la explotación mediante técnicas de procesamiento de lenguaje natural (PLN) se distribuyen principalmente en los siguientes ámbitos: 1. Redes sociales, 2. Blogs de opinión y foros virtuales,

3. Páginas web, 4. Periódicos en línea, 5. Bases de datos científicas y enciclopedias en línea, 6. Herramientas asociadas a motores de búsqueda y 7. Aplicaciones de mensajería instantánea. En la mayoría de estos casos, los datos no estarán específicamente orientados a una investigación concreta, siendo responsabilidad del investigador recuperar la información y conferirle pertinencia en función de sus objetivos. No obstante, es factible que algunos de estos recursos actúen como fuentes de datos primarios, pues es posible fomentar la generación de texto en entornos digitales particulares. Entre los métodos más reconocidos se encuentran las entrevistas estructuradas realizadas en aplicaciones de mensajería instantánea como Gmail u Outlook. Asimismo, se han observado casos donde un investigador plantea una pregunta en un foro virtual y posteriormente analiza las respuestas escritas por los usuarios (véanse ejemplos en Dahlin, 2021; Holtz *et al.*, 2012; Murthy, 2008).

4. Mecanismo de selección y extracción de la información

Una vez identificada la fuente o repositorio que alberga los datos relevantes para la investigación, el siguiente paso implica la descarga y almacenamiento de dichos datos. El procedimiento y el medio de extracción dependen en gran medida de dos elementos fundamentales: en primer lugar, el tipo de acceso otorgado a estos datos, y en segundo lugar, la plataforma o programa utilizado para llevar a cabo el análisis y la explotación del texto.

En relación al tipo de acceso, la información a extraer puede encontrarse en dominios de carácter: 1) privado, lo cual implica la necesidad de solicitar previamente el acceso a dichos dominios, adquirirlos mediante pago o acceder únicamente si se es propietario o se pertenece a una comunidad específica; 2) semiprivado, en el caso de que los propietarios de los datos ofrezcan interfaces de programación de aplicaciones (APIs), permitiendo un acceso parcial a los mismos; o 3) abiertos, donde los datos son completamente públicos y están disponibles para su descarga y utilización por parte de cualquier persona interesada.

4.1. Descarga mediante API

La Interfaz de Programación de Aplicaciones (API) es un método de acceso a software que permite a usuarios externos extraer información específica (Qiu, 2017). En esencia, una API sirve como una clave de comunicación proporcionada por los propietarios del programa, permitiendo un acceso directo y la obtención de cierto tipo de información. Siempre existen limitaciones de acceso determinadas por los permisos otorgados por el propietario, aunque, en la mayoría de los casos, la API ofrece los permisos necesarios para extraer la información requerida en la investigación.

Una ventaja destacada de trabajar con APIs es la conexión directa y establecida entre el programa al que se le solicita la información y el programa solicitante, utilizando códigos específicos que permanecen constantes. Esto facilita la consulta y extracción de información. Sin embargo, uno de los inconvenientes de las APIs es la necesidad de obtener claves de acceso y autenticación proporcionadas únicamente por los propie-

tarios, lo que puede dar lugar a solicitudes no autorizadas. Por último, las APIs pueden ser gratuitas o de pago. En el Anexo 1 se proporcionan los enlaces a las APIs gratuitas de las principales redes sociales.

4.2. *Web Scraping*

La práctica conocida como *web scraping* se refiere al proceso de extracción de información de páginas web mediante el uso de software o código de programación (Vilkova, 2020). La información que se obtiene consiste en una réplica exacta de lo que está escrito en el dominio al que se accede, y generalmente se realiza de manera automatizada utilizando robots que simulan el comportamiento humano en la página web. Esta técnica puede aplicarse a diversos tipos de sitios web, redes sociales e incluso resultados de motores de búsqueda.

Una ventaja significativa del *web scraping* en comparación con las APIs es que no requiere una solicitud previa de acceso, lo que lo hace posible en prácticamente cualquier página web. Sin embargo, esta técnica también presenta desafíos adicionales y requiere de un código más complejo. Aunque existen aplicaciones y sitios web que ofrecen la funcionalidad del *web scraping* sin necesidad de escribir código, estos suelen ser programas de pago o con períodos gratuitos limitados. En el área de investigación social, puede ser interesante emplear *web scraping* para acceder a plataformas de opinión, extraer información de blogs o foros cuando no facilitan una API.

4.3. Descarga directa de archivos

La opción más práctica y conveniente consiste en utilizar fuentes de datos que permitan la descarga directa. En tales casos, el propietario del dominio proporciona una interfaz de usuario amigable que permite seleccionar filtros de búsqueda y descargar manualmente los datos en diversos formatos. No obstante, es frecuente encontrarse con limitaciones en las descargas debido a la capacidad limitada de los servidores. Cuando se requiera realizar descargas masivas sin que ello represente una carga significativa para el investigador, se recomienda buscar APIs que posibiliten el establecimiento de una conexión directa con los datos, lo cual facilita la generación de consultas automatizadas.

Es importante considerar también el tipo de formato en el que se ofrecen los datos. En el caso de los archivos almacenados como hojas de cálculo tipo Excel, se puede manejar la información de manera sencilla siempre que no supere el millón de registros (Microsoft, 2022). No obstante, cuando se plantee trabajar con un número considerable de datos, es recomendable almacenar la información en formatos como «Comma-Separated Values» (.csv), es decir, datos separados por comas o en «JavaScript Object Notation» (.json), que contienen datos en un formato legible por máquina basado en pares de clave y valor. Cuando la opción de descarga solo esté disponible en formato de documento digital tipo PDF (Portable Document Format), se recomienda convertir el archivo a texto plano.

5. Limpieza y preparación de textos para el análisis

Una vez que los datos han sido almacenados de manera adecuada, el siguiente paso consiste en llevar a cabo la limpieza y preparación del texto, con el fin de facilitar el desarrollo de análisis posteriores. En el ámbito de la ciencia de datos, esta fase adquiere una importancia fundamental y demanda la mayor parte del tiempo invertido cuando se trabaja con conjuntos de datos de gran envergadura. Una buena limpieza de los datos determinará el desempeño satisfactorio de los algoritmos aplicados y, por tanto, la calidad y veracidad de los resultados (Bird *et al*., 2009). Para el caso concreto de PLN, las estrategias a llevar a cabo dependen principalmente del grado de complejidad del análisis que se desee aplicar. Sin embargo, existen dos acciones ineludibles en el tratamiento de los datos textuales independientemente de las herramientas y objetivos que tengamos con ellos.

5.1. Eliminación de palabras vacías

El primer paso implica la exclusión de «palabras vacías» o «*stopwords*». El lenguaje humano, en general, se caracteriza por la presencia de conectores y palabras auxiliares que otorgan coherencia y continuidad a la comunicación. Sin embargo, las ideas principales y el significado de las oraciones se encuentran en sustantivos, adjetivos, verbos y, en algunos casos, adverbios, que contienen la sustancia de lo que se pretende expresar o comunicar. En el contexto de la minería de texto, se busca simplificar el contenido y mantenerlo homogéneo. Para ello se procederá a convertir todo el texto a minúsculas y se eliminarán enlaces externos, como vínculos, emoticonos o caracteres similares, en caso de existir.

En segundo lugar, se eliminarán las palabras vacías como conjunciones (y, ni, sino, igual que, porque, etc.) o preposiciones (a, ante, bajo, cabe, con, contra, de, etc.). Usualmente, los paquetes estadísticos para el procesamiento de lenguaje natural incorporan esta función, de manera que permite la eliminación de palabras vacías de forma automática. Aún así, se aconseja realizar algún tipo de prueba que indique qué otras palabras no incluidas en los diccionarios genéricos representan poco valor en el corpus.

Después de llevar a cabo el proceso inicial de depuración, se puede proceder a realizar un análisis de frecuencia de las palabras utilizadas con el propósito de identificar conceptos irrelevantes. No obstante, es necesario minimizar la eliminación de palabras para preservar la integridad de la información original. La exclusión de estas palabras durante la etapa posterior de depuración debe ser claramente documentada en la sección de metodología, proporcionando una justificación adecuada para las decisiones tomadas por parte del investigador. Un ejemplo ilustrativo de cómo se presenta un texto tras la limpieza automática se muestra en la figura 1.

Figura 1
Ejemplo de limpieza de texto automatizada con la librería NLTK en Python

Texto original:
"Existing big datasets of biological data brings a big challenge for the traditional computational algorithms. To have a better understanding of complex biological networks and existing relationships among the components, network models have been using for a long time." (Alinejad-Rokny, 2016).

Texto tras limpieza automática:
"existing big datasets biological data brings big challenge traditional computational algorithms better understanding complex biological networks existing relationships among components network models using long time".

Fuente: elaboración propia.

5.2. Tokenización y creación de diccionario

En segundo lugar, será necesario la tokenización del texto. La tokenización es el proceso por el cual el texto se fracciona en palabras únicas o frases (Saleem *et al.*, 2021). Se trata de reducir el corpus textual a la unidad mínima que será tratada como un dato único. Siguiendo con el ejemplo anterior, sería posible fragmentar el texto en palabras únicas conocidas como *uni-grams* de forma que la primera frase que reciba la máquina se verá como se muestra en la figura 2.

Figura 2
Ejemplo de frase tokenizada de forma automática mediante uni-grams

Tokenización de frase mediante uni-grams:
["existing", "big", "datasets", "biological", "data", "brings", "big", "challenge", "traditional", "computational", "algorithms"].

Fuente: elaboración propia a partir de la librería NLTK en Python.

También puede interesar tratar la unidad mínima como palabras combinadas, ofreciendo un mayor contexto de las palabras analizadas. La fragmentación en pares de palabras se conoce como *bi-grams*. La figura 3 muestra un ejemplo de este tipo de tokenización.

Figura 3

Ejemplo de frase tokenizada de forma automática mediante bi-grams

Tokenización de frase mediante bi-grams:

[“existing big”, “big datasets”, “ dataset biological”, “biological data”, “data brings”, “brings big”, “big challenge”, “ challenge traditional”, “traditional computational”, “computational algorithms”].

Fuente: elaboración propia a partir de la librería NLTK en Python.

Las formas tradicionales de tokenización son *uni-grams* y *bi-grams* tal y como se ha ejemplificado, sin embargo, será posible tokenizar por tantos conjuntos de palabras como se desee, es decir, por «n-grams» (Saleem *et al*., 2021). La decisión estará guiada por la naturaleza del corpus con el que trabajemos. Se recomienda, además, llevar a cabo los análisis trabajando con diferentes grados de tokenización para comparar y dar consistencia a los resultados. La tokenización será un proceso llevado a cabo de forma automatizada. Por lo general, el software o código permitirá asignar el número de *grams* por el cual se desea tokenizar el texto.

Por último, estos «tokens» se transformarán a lenguaje numérico, esto es, se convertirán en vectores con el objetivo de que la máquina pueda contabilizar y realizar cálculos con las palabras. El proceso de vectorialización puede realizarse con múltiples librerías y paquetes estadísticos, debiendo encontrar la transformación que más se adapte a nuestras necesidades. Es importante en este paso tener en cuenta la cantidad de texto, la capacidad de computación de nuestro disco duro o la velocidad. De esta vectorialización resultará un diccionario que de forma simplificada asignará a cada token un número único y un segundo número que señalará las veces que se repite este token en la frase. Siguiendo con el ejemplo, a la palabra «existing» se le asignaría el número 1, a la palabra «big» el número 2 y así sucesivamente.

Figura 4

Ejemplo de creación de diccionario

Frase 1. “existing”, “big”, “datasets”, “biological”, “data”, “brings”, “big”, “challenge”, “traditional”, computational”, “algorithms”

De modo que el diccionario se vería de la siguiente forma:

Entrada 1. (1,1), (2,2), (3,1), (4,1), (5,1), (6,1), (7,1), (8,1), (9,1) (10,1)

*Véase que en el caso de “big” 2, se le añade el recuento de dos ya que aparece dos veces en la frase.

Fuente: elaboración propia a partir de la NLTK.

6. Aplicación de algoritmos mediante técnicas de aprendizaje automático para el análisis de texto

El aprendizaje automático se establece en una rama de la inteligencia artificial donde confluyen esencialmente la ingeniería computacional y la estadística matemática (James *et al.*, 2013). La principal diferencia entre la programación clásica deriva en que hasta ahora las acciones de una máquina venían determinadas por un programador que establecía una regla concreta donde cada entrada tenía asociada una salida específica. El aprendizaje automatizado, por el contrario, permite que la máquina aprenda por sí misma la variabilidad de entradas y salidas, pudiendo establecer nuevos patrones de salida sin que necesariamente haya una programación previa (Müller y Guido, 2016). Este tipo de ingeniería es lo que ha posibilitado que el procesamiento de lenguaje natural a través de ordenadores pueda incrementar la complejidad de los análisis textuales, refinando la variedad del léxico y sus múltiples relaciones que permiten acercarse cada vez más al uso humano que hacemos del lenguaje.

Como ya se ha comentado, el análisis de datos de texto cuenta con una amplia tradición en las disciplinas de ciencias sociales. Por lo general, el uso de este tipo de datos se establece en enfoques cualitativos o mixtos, que a su vez han constituido distintos enfoques metodológicos. El objetivo de la investigación, así como los medios y la naturaleza de los datos, marcan las diferentes vías de análisis. No obstante, la mayoría de autores con experiencia en este campo confluyen en la idea de que el análisis de datos cualitativos y, por ende, gran parte de los análisis de texto, necesariamente implican dos aspectos: «El manejo de los datos y la interpretación» (Gibbs, 2012, p. 23).

El manejo de los datos atiende a una necesaria clasificación, etiquetación, ordenación y reorganización de la información que establecerá las bases metodológicas de la posterior interpretación. En la actualidad, esta primera fase «administrativa» se realiza con la ayuda de los programas privados para el análisis cualitativo asistido por computadora (CAQDAS) como Atlas.Ti o NVivo, junto con el criterio teórico de los investigadores. Aunque estos programas han experimentado grandes desarrollos en los últimos años, muchas de las limitaciones que encontramos en ellos pueden ser solventadas bajo técnicas de PLN y herramientas de aprendizaje automático.

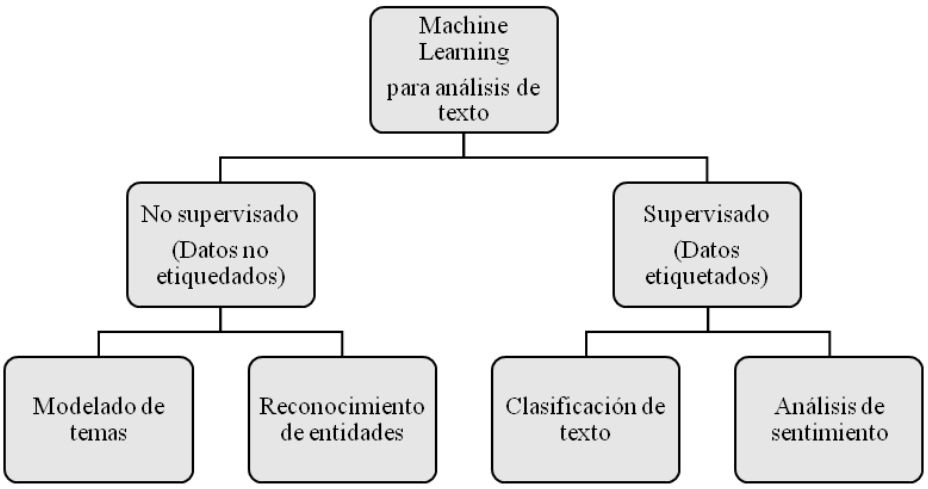
Las técnicas de aprendizaje automático ofrecen diversas posibilidades para el manejo de grandes corpus textuales o datos no estructurados, lo cual también abre nuevas oportunidades para el aprovechamiento de dichos datos en la realización de análisis estadísticos, como correlaciones y series temporales. Estas técnicas no solo contribuyen a los análisis clásicos de contenido o categorización, que se centran en enfoques cuantitativos, sino que también facilitan el desarrollo de análisis discursivos al descubrir patrones y conceptos clave en conjuntos de datos voluminosos y complejos. En última instancia, la elección de una herramienta específica dependerá de la naturaleza de la investigación y sus objetivos, así como de la oportunidad de triangular y contrastar métodos y resultados.

El aprendizaje automático, a su vez, se ramifica en dos tipos de aprendizaje según las características de los datos (*inputs*) que el investigador utilice y la intención del resultado (*outputs*) que se desee obtener. En el caso de los datos que carecen de información asociada

para su categorización o clasificación, se recurre al empleo del aprendizaje no supervisado. Por otro lado, cuando se dispone de una porción de los datos etiquetados y se desea predecir la etiqueta de observaciones futuras, se recurre al aprendizaje supervisado. A continuación, se describen las características distintivas de cada enfoque de aprendizaje, así como los principales tipos de análisis que pueden ser llevados a cabo (véase la figura 5).

Figura 5

Principales tipos de análisis de texto con aprendizaje automático



Fuente: elaboración propia.

6.1. Aprendizaje no supervisado

El aprendizaje no supervisado se aplica cuando se trabaja con datos no «etiquetados», es decir, aquellos que no están clasificados o presentan respuestas previas asociadas a cada dato u observación. El objetivo de este enfoque consiste en agrupar las observaciones en función de su similitud o reconocimiento de entidades utilizando cálculos de distancia estadística. De esta manera, la máquina nos proporciona patrones que no son perceptibles a simple vista (Müller y Guido, 2016). Esta metodología resulta especialmente útil en el análisis de datos de texto, ya que permite un primer acercamiento a corpus no clasificados o de los cuales se carece de información previa sobre su contenido.

Entre los múltiples ejemplos que encontramos en la literatura actual destacan los trabajos que utilizan estos enfoques para descubrir discursos y narrativas presentes en las redes sociales, así como para llevar a cabo revisiones masivas de la literatura. Sirva de ejemplo el trabajo realizado por Lindstedt (2019) en el cual se aplicó el aprendizaje no supervisado para identificar los principales temas investigados en la literatura relacionada con los movimientos sociales durante el período comprendido entre 2005 y 2017.

O la investigación llevada a cabo por los investigadores Pavlova y Berkers (2020), cuyo objetivo fue analizar los discursos sobre salud mental presentes en la red social Twitter.

6.1.1. Modelado de temas

El modelado de temas o *topic modeling* es una técnica de análisis de texto basada en el aprendizaje automático no supervisado, ampliamente utilizada en la minería de texto (Nikolenko *et al*., 2017). Su propósito radica en descubrir los temas estadísticamente significativos presentes en los textos analizados para, como en los ejemplos anteriormente citados, poder obtener de forma genérica la información sustancial de un conjunto de textos.

Esta técnica puede ser implementada utilizando diversos algoritmos, siendo el Latent Direct Allocation (LDA) el más ampliamente utilizado en la literatura científica. El LDA agrupa las palabras más frecuentes en base a su similitud. No obstante, investigaciones recientes han comparado el rendimiento de diferentes algoritmos, revelando las limitaciones del LDA. En los últimos años se ha propuesto el algoritmo BERTopic como una alternativa prometedora, aunque su adopción en trabajos de ciencias sociales aún no está generalizada (Egger y Yu, 2022).

Independientemente del algoritmo empleado, con la aplicación del modelado de temas se obtiene, por un lado, la clasificación de los textos que se agrupan por semejanza y se diferencian de los otros por lejanía, y por otro, una serie de palabras claves de cada grupo, las cuales, según el criterio del investigador, deberán recibir asignaciones descriptivas acordes al conjunto de palabras en cuestión.

Figura 6

Ejemplo de salida del algoritmo LDA aplicado a un corpus textual

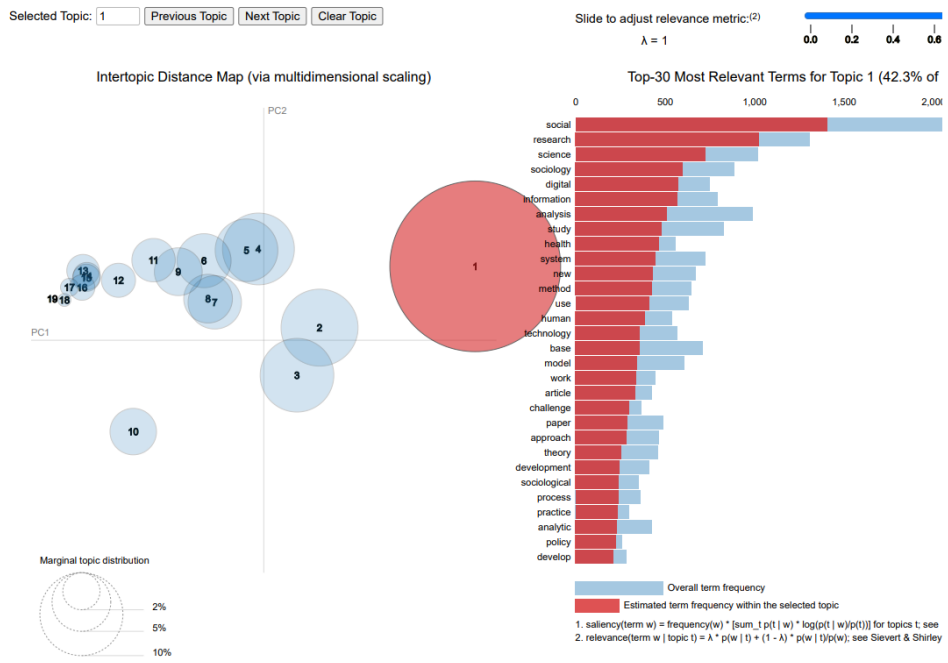
```
[(0,
  '0.028*"collaborative" + 0.028*"recommendation" + 0.024*"system" + '
  '0.023*"filtering" + 0.022*"user" + 0.017*"base" + 0.017*"algorithm" + '
  '0.016*"datum" + 0.012*"recommender" + 0.011*"method"'),
 (1,
  '0.019*"system" + 0.015*"recommendation" + 0.014*"model" + 0.013*"base" + '
  '0.011*"use" + 0.010*"collaborative" + 0.010*"user" + 0.009*"image" + '
  '0.009*"algorithm" + 0.009*"content"'),
 (2,
  '0.024*"datum" + 0.012*"review" + 0.012*"research" + 0.011*"literature" + '
  '0.011*"quality" + 0.011*"model" + 0.010*"analysis" + 0.010*"management" + '
  '0.010*"use" + 0.008*"study"'),
 (3,
  '0.026*"analysis" + 0.026*"research" + 0.015*"social" + 0.010*"datum" + '
  '0.010*"study" + 0.009*"network" + 0.009*"technology" + 0.009*"use" + '
  '0.007*"information" + 0.007*"bibliometric"'),
 (4,
  '0.014*"datum" + 0.013*"system" + 0.012*"study" + 0.011*"analytic" + '
  '0.009*"learn" + 0.008*"research" + 0.008*"information" + 0.008*"management" '
  '+ 0.007*"collaboration" + 0.007*"decision"')]
```

Fuente: elaboración propia en Python Jupyter Notebook.

En la figura 6 se presenta un ejemplo concreto que ilustra el resultado obtenido tras la aplicación del análisis de Latent Dirichlet Allocation (LDA) a un corpus textual compuesto por diversas investigaciones en los campos de negocios y tecnología. Este caso específico fue implementado utilizando el lenguaje de programación Python dentro del entorno Jupyter Notebook, mediante el uso de la biblioteca Gensim. No obstante, como se sintetiza en la tabla 1, es factible llevar a cabo este procedimiento en varios entornos e, incluso, utilizando los softwares mencionados en la sección 2 sin requerir conocimientos de programación.

Figura 7

Ejemplo de visualización gráfica del algoritmo LDA



Fuente: elaboración propia a partir de la librería pyLDAvis en Python.

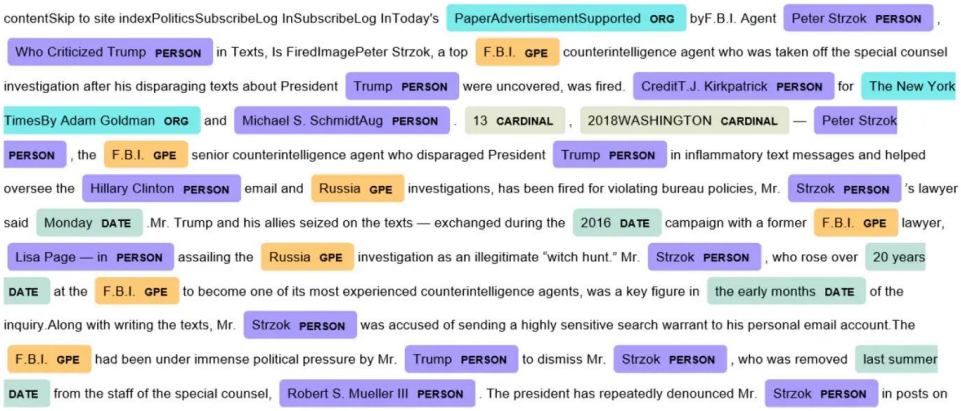
En la figura 7 se presenta una propuesta de visualización de este algoritmo mediante el uso de la biblioteca pyLDAvis. Esta herramienta permite obtener una mejor comprensión de la distribución de los clústeres o temas, así como de las palabras más relevantes asociadas a cada uno de ellos. El siguiente paso consistirá en extraer la información y, a través de la supervisión de una muestra de los textos seleccionados en cada grupo, realizar inferencias temáticas y asignar un nombre a cada clúster.

6.1.2. Reconocimiento de caracteres o entidades

Otra de las técnicas más usadas bajo el aprendizaje no supervisado es el reconocimiento de entidades o *Named Entity Recognition* (NER). Para esta ocasión, dado un corpus textual, la máquina es capaz de detectar entidades predefinidas. El algoritmo puede detectar en qué parte del texto se habla de personas, lugares, empresas, números, etc. (Calzolari, 2020). Esto puede ser interesante para etiquetar textos y organizarlos de forma rápida y eficaz. El reconocimiento de estas entidades viene dado por la existencia de diccionarios previos que permiten al algoritmo detectar en el corpus nombres que ya han sido etiquetados previamente en otros corpus textuales. Es por ello que aunque nuestra forma de proceder con esta técnica sea la de un aprendizaje no supervisado, la realidad es que se trata de un método mixto entre supervisado y no supervisado.

Figura 8

Ejemplo de salida aplicando NER



Fuente: Li (2018).

En la figura 8 se presenta un ejemplo concreto de la salida visualizada en pantalla al aplicar la biblioteca Spacy para la visualización en el lenguaje de programación Python dentro del entorno Jupyter Notebook. Estas bibliotecas también ofrecen la capacidad de almacenar las clasificaciones en listas o bases de datos, lo que permite su posterior explotación y procesamiento.

En ambos casos, modelado de temas y reconocimiento de entidades, el resultado será la etiquetación y clasificación de los textos que de forma original se encontraban sin estructurar. Como se ha comentado, estos análisis pueden ser aplicados tanto en entornos que requieren programación como en softwares libres de codificación. La ventaja de este tipo de técnicas es la de establecer mapas de temáticas y conceptos de forma rápida y genérica. No obstante, el papel del investigador tendrá un gran peso a la hora de interpretar las temáticas y generar inferencia de resultados.

Es importante resaltar que en el caso del modelado de temas (*topic modeling*), el número de clústeres en los que se clasifican los textos es una decisión que corresponde al investigador. Aunque existen métricas de evaluación disponibles para comparar estadísticamente los resultados, siendo la métrica de coherencia la más popular (Stevens *et al*., 2012), la literatura actual defiende que estas métricas deben ser consideradas de manera orientativa. En última instancia, será el criterio teórico del investigador el que determine el sentido y significado de los resultados obtenidos.

6.2. Aprendizaje supervisado

El aprendizaje supervisado se refiere al enfoque utilizado en el procesamiento de datos etiquetados o clasificados, con el propósito de realizar predicciones sobre la clasificación de observaciones futuras bajo los mismos criterios y características (Shahbaz *et al*., 2022). Mediante este tipo de algoritmo se instruye a la máquina utilizando un conjunto de datos de entrenamiento, en el cual se proporcionan las características de entrada y las correspondientes etiquetas de salida. El objetivo principal es permitir que la máquina sea capaz de predecir la etiqueta correspondiente a una nueva observación. El aprendizaje supervisado puede llegar a alcanzar niveles de complejidad considerables, y las posibilidades actuales en este campo son amplias y diversas.

Este tipo de enfoque se muestra especialmente útil en la identificación de categorías específicas, la predicción de resultados y la asignación de etiquetas a nuevos textos basándose en patrones previamente establecidos. A diferencia del aprendizaje no supervisado, que se centra en análisis globales y descriptivos, el aprendizaje automático supervisado permite aplicaciones más precisas y orientadas a objetivos específicos. En la literatura científica se han empleado técnicas de aprendizaje automático supervisado en diversos contextos. Por ejemplo, Naseeba *et al*. (2023) utilizaron este enfoque para la clasificación de artículos de periódicos, mientras que Khanday *et al*. (2022) se enfocaron en la detección de discursos de odio en redes sociales. Por su parte, Mbona y Eloff (2023) aplicaron técnicas de aprendizaje automático supervisado para la identificación de *bots*, y Shevtsov *et al*. (2023) exploraron su utilidad en el análisis electoral.

6.2.1. Clasificación de texto

La clasificación de textos puede ser de gran utilidad en estudios que involucran grandes volúmenes de datos. En este escenario, es necesario realizar una clasificación previa de un conjunto de textos con las salidas esperadas, para luego aplicar este conocimiento al resto del corpus y permitir que el algoritmo realice la clasificación de forma automática. En este sentido, es recomendable partir del conocimiento previo sobre posibles categorías o tipologías teóricas, generando etiquetas que puedan respaldar o refutar hipótesis. Por tanto, la tarea de etiquetado previo llevada a cabo por el investigador debe estar sólidamente fundamentada y seguir un código de codificación exhaustivo, lo que permitirá a los algoritmos realizar predicciones con mayor precisión. Esta etiquetación deberá ir precedida del diseño de una guía de codificación clara y detallada que capacite a los codificadores para seguir un criterio homogéneo y sólido en la clasificación de la unidad de textos que se desea analizar.

La clasificación de textos puede ser realizada mediante diversos algoritmos. Al igual que en el caso del modelado de temas, encontramos en la literatura reciente múltiples trabajos que evalúan el desempeño de diferentes algoritmos de clasificación (Dogra *et al.*, 2022). En el área de las ciencias sociales, uno de los algoritmos más usados en la investigación científica es el *Support Vector Machines* (SVMs), que permite la clasificación de grupos textuales con alta dimensionalidad (Joachims, 1998).

Es importante tener en cuenta que, cuando se usan algoritmos de clasificación, los datos empleados para el entrenamiento deberán tener las mismas características que los datos a los que se les desean aplicar posteriormente el algoritmo. Por ejemplo, si el algoritmo ha sido entrenado para obtener dos clases de respuesta, como «extremista» o «moderado», la salida que proporcionará este algoritmo siempre corresponderá a estas dos etiquetas. Por lo tanto, si se sospecha que nuevos conjuntos de textos contienen categorías adicionales no previstas en la muestra etiquetada, será necesario adquirir una nueva muestra en la que se etiqueten las categorías adicionales y volver a entrenar el algoritmo.

6.2.2. Análisis de sentimiento

El análisis de sentimiento es un tipo de análisis con gran recorrido en el área de negocios y otras ciencias sociales. Clásicamente se ha llevado a cabo a través de encuestas de satisfacción del cliente. Sin embargo, el aumento de plataformas de comercio digital ha propiciado el desarrollo de nuevos métodos que, a su vez, han permitido la expansión de estos análisis a diversas áreas de investigación como el análisis electoral, los estudios sobre migración y discursos de odio o estudios de comunicación, entre otras. El análisis de sentimiento a través de aprendizaje automático nos permite identificar emociones en grandes corpus textuales. Este funciona con la misma lógica que la clasificación de texto, no obstante, al igual que sucedía con la técnica NER, existen múltiples diccionarios ya entrenados que, sin necesidad de etiquetar previamente una muestra de la base de datos, el algoritmo puede reconocer el sentimiento en nuestro corpus textual.

En ambos casos obtendremos un corpus de texto etiquetado bajo criterios teóricos previamente establecidos, a diferencia de lo que ocurría en el análisis no supervisado. El objetivo cuando generamos un algoritmo de clasificación es obtener el mayor porcentaje de precisión *(accuracy)* en el modelo de predicción que hemos generado. La evaluación de un modelo de aprendizaje supervisado sigue un procedimiento establecido.

En primer lugar, se extrae del corpus textual una muestra que será etiquetada. Esta muestra se divide en un conjunto de entrenamiento (*training*) y en un conjunto de prueba *(test)*, que generalmente aplica la división 70%-30% respectivamente. El modelo se entrenará con el 70%, al que se le muestra la observación o dato y posteriormente la respuesta de salida. Una vez el modelo queda entrenado, se evalúa su capacidad de clasificación utilizando el otro 30% del conjunto de prueba. Basándonos en el porcentaje de acierto obtenido en esta evaluación, se toma la decisión sobre si el modelo es lo suficientemente preciso para clasificar el resto de los datos no vistos anteriormente.

Para lograr una adecuada generalización del modelo, es necesario contar con un corpus textual de gran volumen. Además, es importante explicitar en la metodología de la investigación el nivel de precisión con el que se ha clasificado el modelo.

Tabla 1

Principales librerías para cada tipo de análisis y lenguaje de programación

Tipo de Análisis	Lenguaje de programación Python	Lenguaje de programación R
Modelado de temas	-Gensim -Sklearn	-topicmodels -lda
Reconocimiento de entidades	-SpaCy -NLTK	-spacyr -openPLN
Clasificación de texto	-NLTK -tensorflow (Keras)	-tm -Caret
Análisis de sentimiento	-scikit-learn -Keras	-tidytext -quanteda

Fuente: elaboración propia.

7. Consideraciones finales

A lo largo de este trabajo se ha desarrollado una guía sintetizada de los pasos esenciales para aplicar el procesamiento de lenguaje natural (PLN) en investigaciones de ciencias sociales dentro del marco del aprendizaje automático. En primer lugar, se proporcionó una visión general de la historia y origen del procesamiento de lenguaje natural, seguido de los aspectos prácticos de la aplicación de técnicas de PLN. Se detallaron algunos de los lenguajes de programación y software disponibles para llevar a cabo los análisis, así como las diversas fuentes de texto y los métodos para extraerlos y almacenarlos. A continuación, se dedicó una sección al proceso de limpieza y tratamiento del texto, seguido de una descripción de los diversos tipos de técnicas de análisis que se pueden aplicar para procesar el lenguaje natural.

Si bien en la literatura científica internacional se observa un creciente interés por la integración transdisciplinar entre las áreas de ciencias sociales y las técnicas de computación, resulta evidente la necesidad de continuar desarrollando investigaciones y trabajos pedagógicos que fomenten y motiven a los científicos sociales a adoptar y emplear activamente este tipo de técnicas en sus estudios. La implementación de estas metodologías puede brindar beneficios significativos, como el análisis más profundo y riguroso de los datos, la identificación de patrones y tendencias ocultas, y una comprensión más completa y enriquecedora de los fenómenos sociales.

La presente guía se concibe como un recurso introductorio para investigadores interesados en adentrarse en el ámbito del aprendizaje automático y el PLN. Su propósito fundamental radica en brindar una orientación clara y efectiva para la selección de enfoques, programas y algoritmos ampliamente validados en un contexto caracterizado por la multiplicidad de opciones disponibles, ya que este amplio abanico de alternativas puede resultar abrumador para aquellos que se encuentran en las etapas iniciales de su formación en este campo de estudio.

Es relevante señalar que, si bien los trabajos más recientes que emplean procesamiento de lenguaje natural (PLN) en objetos de investigación de ciencias sociales respaldan la eficacia de las técnicas presentadas, la discusión actual identifica desafíos significativos que los investigadores sociales deberán afrontar en el futuro cercano.

En primer lugar, uno de los principales desafíos reconocidos por los investigadores sociales es la existencia de sesgos tanto en los modelos utilizados como en su proceso de entrenamiento. Es esencial tener en cuenta que las máquinas tienden a reproducir los prejuicios inherentes a los seres humanos. Por ende, resulta fundamental evitar dichos sesgos desde la etapa de extracción de datos hasta el propio entrenamiento de los algoritmos (Zwilling, 2023).

Otro desafío reside en la actual incapacidad de los modelos para «comprender» las particularidades culturales y las expresiones propias de las jergas presentes en los datos. Estas particularidades y expresiones suelen tener un valor significativo para los investigadores sociales, y su correcta interpretación resulta crucial para un análisis preciso y relevante (Sambeek, 2021).

Por último, cabe resaltar la notable carencia de conjuntos de datos debidamente anotados que sean aptos para el entrenamiento de modelos supervisados en el contexto de las ciencias sociales. En muchas ocasiones, se hace evidente la falta de datos etiquetados para áreas específicas o temas de interés para campos como la sociología o las ciencias políticas. Este hecho conlleva la necesidad de recurrir a técnicas de transferencia de aprendizaje y la formulación de estrategias específicas destinadas a abordar esta limitación, con el objetivo primordial de alcanzar resultados que sean tanto fiables como representativos.

Estos desafíos subrayan la necesidad de continuar investigando y desarrollando el PLN aplicado a las ciencias sociales, a fin de superar las limitaciones actuales y garantizar un análisis riguroso y sólido de los textos en este campo. Además, resulta fundamental destacar la importancia de hacer que estos análisis sean replicables y reproducibles, señalando en la sección de métodos, las librerías, herramientas y entornos utilizados, así como las verificaciones y modificaciones realizadas, desde la limpieza del corpus textual hasta la evaluación y visualización de los resultados.

8. Anexo

Anexo 1. Enlaces a las APIs de las redes sociales más populares

Plataforma	Api
Facebook	https://developers.facebook.com/docs/graph-api
Instagram	https://developers.facebook.com/docs/instagram
Twitter	https://developer.twitter.com/en/docs
Youtube	https://developers.google.com/youtube/v3/docs/
Reddit	https://www.reddit.com/dev/api/

Fuente: elaboración propia.

9. Bibliografía

Abbott, A. (1997). Of Time and Space: The Contemporary Relevance of the Chicago School. *Social Forces, 75*(4), 1149. doi: 10.2307/2580667.

Ajmal, S., Khan, S., Hossain, M., Lomonaco, V., Cannons, K., Xu, Z. y Cuzzolin, F. (2022). International Workshop on Continual Semi-Supervised Learning: Introduction, Benchmarks and Baselines. *Continual Semi-Supervised Learning*, Vol. 13418 (pp. 1-14). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-17587-9_1

Alinejad-Rokny, H. (2016). Proposing on Optimized Homolographic Motif Mining Strategy Based on Parallel Computing for Complex Biological Networks. *Journal of Medical Imaging and Health Informatics*, *6*(2), 416-424. https://doi.org/10.1166/jmihi.2016.1707

Bird, S., Klein, E. y Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

Bitter, C., Elizondo, D. A. y Yang, Y. (2010). Natural language processing: A prolog perspective. *Artificial Intelligence Review*, *33*(1-2), 151-173. https://doi.org/10.1007/s10462-009-9151-4

Calzolari, N. (2020). *LREC 2020 Marseille Twelfth International Conference on Language Resources and Evaluation$dMay 11-16, 2020, Palais Du Pharo, Marseille, France: Conference Proceedings*. Paris: The European Language Resources Association (ELRA).

Castells, M. (2018). *La era de la información: economía, sociedad y cultura. Vol. 3, Fin de milenio*. 4ª ed., 2ª reimpr. Madrid: Alianza Editorial.

Dahlin, E. (2021). Email Interviews: A Guide to Research Design and Implementation. *International Journal of Qualitative Methods,* 20:160940692110254. doi: 10.1177/16094069211025453.

Dhiraj, M. (2008). Digital Ethnography: An Examination of the Use of New Technologies for Social Research. *Sociology, 42*(5), 837-855. doi: 10.1177/0038038508094565.

Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J. y Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. En S. K. Sah Tyagi (Ed.), *Computational Intelligence and Neuroscience* (pp. 1-26). doi: 10.1155/2022/1883698.

Egger, R. y Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7:886498. doi: 10.3389/fsoc.2022.886498.

Gibbs, G. (2012). *El análisis de datos cualitativos en investigación cualitativa*. Madrid: Ediciones Morata.

Gillingham, P. y Graham, T. (2017). Big Data in Social Welfare: The Development of a Critical Perspective on Social Work's Latest «Electronic Turn». *Australian Social Work*, *70*(2), 135-147. https://doi.org/10.1080/0312407X.2015.1134606

Gualda, E., Taboada Villamarín, A. y Rebollo Díaz, C. (2023). Big data y ciencias sociales: Una mirada comparativa a las publicaciones de antropología, sociología y trabajo social. *Gazeta de Antropología*, *39* (1).

Gualda, E. y Rebollo, C. (2020). Big data y Twitter para el estudio de procesos migratorios: Métodos, técnicas de investigación y software. *Empiria. Revista de metodología de ciencias sociales*, *46*, 147. https://doi.org/10.5944/empiria.46.2020.26970

Hockett, C. F. (2020). The state of the art. *The State of the Art*. De Gruyter.

Holtz, P., Kronberger, N. y Wagner, W. (2012). Analyzing Internet Forums: A Practical Guide. *Journal of Media Psychology*, *24*(2), 55-66. https://doi.org/10.1027/1864-1105/a000062

James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013). *An Introduction to Statistical Learning* (vol. 103). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S. y Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. En A. Abraham, O. Castillo y D. Virmani (Eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks* (vol. 167, pp. 365-375). Springer Singapore. https://doi.org/10.1007/978-981-15-9712-1_31

Justicia de la Torre, C., Sánchez, D., Blanco, I. y Martín-Bautista, M. J. (2018). Text Mining: Techniques, Applications, and Challenges. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *26*(04), 553-582. https://doi.org/10.1142/S0218488518500265

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R. y Malik, S. H. (2022). Detecting Twitter Hate Speech in COVID-19 Era Using Machine Learning and Ensemble Learning Techniques. *International Journal of Information Management Data Insights*, *2*(2), 100120. doi: 10.1016/j.jjimei.2022.100120.

Li, S. (2018). *Named Entity Recognition and Classification with Scikit-Learn.* https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

Lindstedt, Nathan C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005-2017. *Social Currents,* *6*(4), 307-318. doi: 10.1177/2329496519846505.

Maud, R. y Blanchard, A. (2022). The Framing of Health Technologies on Social Media by Major Actors: Prominent Health Issues and COVID-Related Public Concerns. *International Journal of Information Management Data Insights,* *2*(1), 100068. doi: 10.1016/j.jjimei.2022.100068.

Mbona, I. y Eloff, J. H. P. (2023). Classifying Social Media Bots as Malicious or Benign Using Semi-Supervised Machine Learning. *Journal of Cybersecurity*, *9*(1), tyac015. doi: 10.1093/cybsec/tyac015.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. y Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176-182. https://doi.org/10.1126/science.1199644

Microsoft (2022). *Especificaciones y límites de Excel*. https://support.microsoft.com/es-es/office/especificaciones-y-l%C3%ADmites-de-excel-1672b34d-7043-467e-8e27-269d656771c3

Morimoto, J. y Ponton, F. (2021). Virtual reality in biology: Could we become virtual naturalists? *Evolution: Education and Outreach*, *14*(1), 7. https://doi.org/10.1186/s12052-021-00147-x

Müller, A. C. y Guido, S. (2016). *Introduction to aprendizaje automático with Python: A guide for data scientists*. O'Reilly Media, Inc.

Naseeba, B., Challa, N. P., Doppalapudi, A., Chirag, S. y Nair, N. S. (2023). Machine Learning Models for News Article Classification. *5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1009-1016). Tirunelveli, India: IEEE. https://doi.org/10.1109/ICSSIT55814.2023.10061095

Nikolenko, S. I., Koltcov, S. y Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88-102. https://doi.org/10.1177/0165551515617393

Pavlova, A., y Berkers, P. (2020). Mental Health Discourse and Social Media: Which Mechanisms of Cultural Power Drive Discourse on Twitter. *Social Science & Medicine,* 263, 113250. doi: 10.1016/j.socscimed.2020.113250.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Cham: Springer. https://doi.org/10.1007/978-3-031-02146-6

Radick, G. (2016). The unmaking of a modern synthesis: Noam Chomsky, Charles Hockett, and the politics of behaviorism, 1955-1965. *Isis*, *107*(1), 49-73. https://doi.org/10.1086/686177

Ruelens, A. (2022). Analyzing user-generated content using natural language processing: A case study of public satisfaction with healthcare systems. *Journal of Computational Social Science*, *5*(1), 731-749. https://doi.org/10.1007/s42001-021-00148-2

Saleem, Z., Alhudhaif, A., Qureshi, K. N. y Jeon, G. (2021). Context-aware text classification system to improve the quality of text: A detailed investigation and techniques. *Concurrency and Computation: Practice and Experience*. https://doi.org/10.1002/cpe.6489

Sambeek, I. (2021). Natural Language Processing & Social Sciences. Towards Data Science. https://towardsdatascience.com/natural-language-processing-social-sciences-94a35a8a7c78

Shevtsov, A., Oikonomidou, M., Antonakaki, D., Pratikakis, P. y Ioannidis, S. (2023). What Tweets and YouTube Comments Have in Common? Sentiment and Graph

Analysis on Data Related to US Elections 2020. *PLOS ONE, 18*(1), e0270542. doi: 10.1371/journal.pone.0270542.

Thorsten, J. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. En C. Nédellec y C. Rouveirol, *Aprendizaje automático: ECML-98*. Vol. 1398, *Lecture Notes in Computer Science* (pp. 137-142). Berlin, Heidelberg: Springer. https://doi.org/10.1007/BFb0026683

Vilkova, O. (2020). Web Scraping as a Method of Data Extraction in Sociological Studies: On Scientific Applicability. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya, sotsiologiya, politologiya,* (54), 163-175. doi: 10.17223/1998863X/54/16.

Yuanbo, Q. (2017). The Openness of Open Application Programming Interfaces. *Information, Communication & Society,* 20(11), 1720-36. doi: 10.1080/1369118X.2016.1254268.

Zwilling, Moti (2023). Big Data Challenges in Social Sciences: An NLP Analysis. *Journal of Computer Information Systems*, *63*(3), 537-554. doi: 10.1080/08874417.2022.2085211.

Alba Taboada Villamarín

Estudiante de doctorado en el programa de Economía y Empresa por la Universidad Autónoma de Madrid. Ha obtenido una beca como Personal Investigador Predoctoral en Formación (FPI) vinculado al proyecto I+D CONCERN(PID2020-115095RB-I00), además de formar parte del equipo de trabajo del proyecto I+D NON-CONSPIRA-HATE!(PID2021-123983OB-I00). Se graduó en Sociología por la UCM y realizó un Máster en Big Data Science por la Universidad de Navarra. Ha sido becaria en el Centro de Investigaciones Sociológicas (CIS), promoción 2022. Actualmente investiga nuevos enfoques metodológicos a través de Big Data y Machine Learning aplicados a las Ciencias Sociales.