UNIVERSIDAD DE GRANADA

TRABAJO FIN DE MÁSTER

MÁSTER EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

# Explicación de la Detección de Mensajes Sexistas en Redes Sociales


**Autora**
Lidia Sánchez Mérida

**Directores**
Alberto Luis Fernández Hilario
Eugenio Martínez Cámara




Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación


—

Granada, Julio de 2023

# Explicación de la Detección de Mensajes Sexistas en Redes Sociales

Lidia Sánchez Mérida (alumna)

**Palabras clave**: sexismo, sexista, mensajes, textos, documentos, reconocimiento, identificación, detección, clasificación, redes_sociales, aprendizaje_automático, aprendizaje_profundo, modelos, clasificadores, arquitecturas, lstm, bert, explicabilidad, interpretabilidad, lime, contrafactuales

## Resumen

Si bien el comportamiento sexista siempre ha estado presente desde tiempos inmemoriables no ha sido hasta hace unas décadas cuando se ha comenzado una lucha para su erradicación. Cuando pensábamos que se habían realizado progresos importantes en esta causa, las nuevas tecnologías emergentes como las redes sociales se han postulado como un nicho en el que desarrollar este tipo de conductas de forma cómoda y anónima. La menospreciación y la irrespetuosidad contra el colectivo femenino en formato escrito conlleva graves consecuencias en el mundo real, tal y como se puede apreciar diariamente en los informativos acerca de los numerosos casos de violencia de género. En este proyecto se experimenta con diferentes técnicas, algoritmos y arquitecturas de diversa naturaleza considerando los ámbitos de Aprendizaje Automático, Procesamiento Natural del Lenguaje y Aprendizaje Profundo. Sin embargo no se podrá confiar en su actuación ni mejorar su rendimiento a menos que podamos interpretar sus decisiones fácilmente. Por ello se han integrado análisis de explicabilidad tanto de desarrollo propio, como mediante la aplicación de metodologías específicas de un nuevo área que trata esta temática conocida como Explicabilidad de la Inteligencia Artificial. De este modo te brinda la capacidad de incluso sugerir la inclusión de procedimientos adicionales con los que paliar los defectos intrínsecos a los modelos a partir de la generación de explicaciones comprensibles por personas.

# Explicability in the Detection of Sexist Messages within Social Media

Lidia Sánchez Mérida (student)

**Keywords**: sexism, sexist, messages, texts, documents, recognition, identification, detection, classification, social_media, machine_learning, deep_learning, models, classifiers, architectures, lstm, bert, explicability, interpretability, lime, counterfactuals

## Abstract

Although sexist behavior has always been present since the beginning of times, it wasn't until a few decades ago that the fight to eradicate it began. Just when we thought that significant progress had been made in this cause, new emerging technologies such as social media have been postulated as a niche in which this type of behavior is still increasing comfortably and anonymously. The disparagement and disrespect against the female collective in written texts has serious consequences in the real world, as can be seen daily in the news about the numerous cases of gender violence. In this project we experiment with different techniques, algorithms and architectures of different nature considering the fields of Machine Learning, Natural Language Processing and Deep Learning. However, it will not be possible to trust their decisions or to improve their performance unless we can easily interpret their responses. For this reason, we have integrated explainable analysis, both self-developed and through the application of specific methodologies of a new area that deals with this subject known as Explainable Artificial Intelligence. In this way it brings the opportunity to suggest the integration of additional procedures to alleviate the intrinsic defects of the models from the generation of human-understandable explanations.

Yo, **Lidia Sánchez Mérida**, alumno de la titulación Máster en Ciencia de Datos e Ingeniería de Computadores de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI -, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Máster en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Lidia Sánchez Mérida

Granada a 9 de Julio de 2023.

D. **Alberto Luis Fernández Hilario (tutor)**, Profesor del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

D. **Eugenio Martínez Cámara (tutor)**, Profesor del Departamento de Informática de la Universidad de Jaén.

**Informan:**

Que el presente trabajo, titulado ***Explicación de la Detección de Mensajes Sexistas en Redes Sociales***, ha sido realizado bajo su supervisión por **Lidia Sánchez Mérida (alumna)**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 9 de Julio de 2023.

**Los directores:**

**Alberto Luis Fernández Hilario, Eugenio Martínez Cámara**

# Agradecimientos

En primer lugar mis mayores agradicimientos van dirigidos a mi madre, mi apoyo incondicional en todas las facetas de mi vida, tanto personales como profesionales.

A continuación me gustaría recordar a mis compañeros de promoción por el ambiente tan familiar y cercano que hemos creado con vínculos que extralimitan el curso académico compartido.

Sin olvidar por supuesto a mis tutores, Alberto y Eugenio, que desde el principio me han guiado en este maravilloso camino aportando todos los recursos a su alcance, además de distintas perspectivas profesionales y académicas. Gracias por todo, siempre recordaré los buenos momentos vividos y la experiencia obtenida por dos magníficos profesionales y personas que sois.

# Índice general

# Índice de figuras

# Índice de tablas

# Capítulo 1

# Introducción

Durante las últimas décadas se han realizado avances muy relevantes a favor de la igualdad de género entre mujeres y hombres. Según estadísticas oficiales del 2022 en las que se incluyen a más de 146 países el **índice de igualdad global se sitúa en el 68.1 %**, aunque sin considerar una gran parte de regiones lationamericanas y africanas que podrían empeorar ampliamente esta cifra. Si bien el continente europeo se encuentra en el segundo puesto del ranking con un valor más favorable del 76.6 %, se calcula que cerrar la **brecha existente entre géneros podría costar más de sesenta años**. Son múltiples las razones que ayudan a agravar cada vez más este problema, siendo una de las más comunes la imposición de roles femeninos en el cuidado del hogar y de la salud, vacantes caracterizadas por unas condiciones laborales y salariales súmamente precarias [1]. Su manifestación ha sido notoria en ciertos estudios americanos que perseguían la identificación de posibles sesgos presentes en la evaluación de opiniones realizadas por un grupo de participantes mixto sobre una variedad de productos y servicios. Como conclusión general se extrae que los **comentarios escritos por hombres son mejor valorados**, mientras que las opiniones originadas por mujeres han sido consideradas mayormente como poco importantes o exageradas. Un hecho remarcable es que el propio colectivo femenino el principal causante de la degradación y denigración hacia su propio género, por lo que no se puede implantar el foco únicamente en la población masculina en la lucha contra este lastre social. Entre las teorías explicativas más destacadas sobre estos fenómenos se encuentran el estereotipado que visualiza al colectivo masculino como más analítico, resolutivo y representativo, mientras que al género femenino se le atribuyen características como la inmadurez, la emocionalidad y una visibilidad inferior [2]. Gracias a la aparición y expansión a nivel mundial de Internet y las nuevas tecnologías emergentes, como las redes sociales, existe un entorno virtual perfecto en el que se continúan desarrollando este tipo de comportamientos en los que se **genera y difunde**

**todo tipo de violencia sexista**.

Por todos los argumentos expuestos hasta el momento, hemos considerado que la aplicación de técnicas de Aprendizaje Automático y Aprendizaje Profundo pueden ser de una inmensa relevancia y utilidad en la construcción de modelos, con diferentes arquitecturas y configuraciones, que actúen como posibles **filtros de contenido enfocados en la detección e identificación de comportamientos sexistas** basados en documentos. Para llevar a cabo este cometido se han tomado un conjunto de entrenamiento y validación que se pueden encontrar libremente en la red y cuyas fuentes de datos residen en redes sociales, unos medios de información muy poderosos y propicios a la aparición de este tipo de conductas tan indeseadas y perjudiciales. Dado el elevado carácter sensible que acompaña a esta temática, consideramos que es **crucial la aplicación de diversas técnicas de explicabilidad** en combinación con el seguimiento, en la medida de lo posible, de las propuestas y los requisitos recopilados en las European Ethic Guidelines [3] y en el Artificial Intelligence Act [4], con las que perseguir dos principales objetivos: mejorar la capacidad de predicción de los clasificadores encontrando las debilidades que les llevan a cometer errores durante el estudio y reconocimiento de los textos para así poder desarrollar procedimientos con los que disminuir su cantidad. Además de la generación de explicaciones que permitan comprender fácilmente el funcionamiento de los sistemas, los pilares en los que fundamentan sus decisiones y procesos de razonamiento, así como los posibles sesgos introducidos para reducir su impacto en el rendimiento de los mismos.

Con el propósito de continuar mi aprendizaje en las áreas mencionadas tanto a nivel académico como profesional, se planifica este proyecto con la ilusión de investigar las técnicas de procesamiento y modelado más propicias para aportar soluciones viables y de calidad en la lucha contra el sexismo escrito. De este modo a continuación se detallan las **hipótesis iniciales** que se han establecido previo al comienzo del desarrollo de este trabajo.

- En la primera teoría se prevee que las arquitecturas pertenecientes al **Aprendizaje Profundo proporcionen un mejor rendimiento** en comparación con los algoritmos más clásicos. Además de haber demostrado su superioridad en otras temáticas, como la salud o la automoción, su capacidad de extraer y analizar características de forma autónoma y empleando funciones complejas no lineales, sin que este proceso dependea totalmente de un humano, demuestra unas competencias inmejorables con las que abordar la identificación y clasificación de prácticamente cualquier tipo de datos. En particular, aquellos modelos que ya se encuentra preentrenados sobre conjuntos de datos masivos parecen ser más ventajosos puesto que brindan la posibili-

dad de partir desde un punto inicial mucho más avanzado que con un modelado propio. Mediante distintos enfoques de ***fine-tuning*** estos sistemas se pueden ajustar a datasets y temáticas específicas minimizando la inversión de recursos necesaria en su planteamiento, desarrollo y construcción desde cero.

- La preparación de documentos también se considera crucial en la construcción de soluciones de calidad para la detección de textos sexistas. Por un lado la integración de técnicas de **tratamiento y limpieza de documentos** pueden contribuir a reducir el ruido intrínseco en los datos, dejando únicamente la terminología que potencialmente contiene información más relevante. Mientras que por otra parte, los mecanismos de generación de muestras sintéticas con los que **aumentar el volumen del conjunto de entrenamiento** suele aportar una mayor cifra de muestras con las que aumentar su representatividad y así otorgar una mayor robustez y capacidad de generalización a los futuros modelos.

- Puesto que los textos aún no pueden ser proporcionados directamente a ninguno de los algoritmos existentes, se apuesta por una codificación mediante **embeddings preentrenados** como la mejor aproximación posible. Además de que estos recursos se encuentran altamente preparados y fundamentados en el entrenamiento de modelos complejos, ajustados y sobre volúmenes inmensos de datos, también conllevan otros beneficios como como el orden de las palabras dentro del documento y la consideración de la similitud entre términos para ser representados con vectores parecidos que reflejen dichas relaciones.

- Como última hipótesis se reflexiona acerca de qué técnicas de explicabilidad pueden facilitar una mejor y mayor interpretación sobre modelos de cajas negras. Las agnósticas locales podrían resultar ser más interesantes gracias a que no contienen restricciones de uso y a que pueden ser efectuadas en base a un subconjunto de ejemplos particular, como por ejemplo los fallos cometidos con los que intentar encontrar las posibles causas de su origen. Mientras que *LIME* se orienta más hacia la generación teórica de explicaciones acerca del comportamiento de un modelos, los *contrafactuales* son un mecanismo práctico que se concentra en la perturbación de documentos con los que estudiar si las modificaciones lideran hacia un cambio de clase. Por lo tanto esta última es más experimental y no se conoce a priori si puede arrojar cierto conocimiento útil acerca del modo de funcionamiento de los sistemas inteligentes, por lo que apostamos por **LIME como la técnica de explicabilidad que más puede ayudar a interpretar los modelos cerrados**.

En particular a continuación se detalla una lista de los **objetivos** establecidos para este proyecto que se apoyan en la motivación explicada anteriormente y en las hipótesis iniciales que se han generado en torno a las herramientas que se consideran pueden ser de mayor utilidad para su cumplimiento.

1. Realizar un análisis exploratorio de un conjunto de datos procedente de redes sociales para comprender las características y patrones del lenguaje sexista en las redes sociales.
2. Evaluar y seleccionar técnicas de Procesamiento de Lenguaje Natural adecuadas para el análisis de los mensajes extraídos de redes sociales.
3. Desarrollar modelos de Aprendizaje Automático para la detección de mensajes sexistas publicados en redes sociales.
4. Evaluar el rendimiento de los modelos anteriores utilizando métricas relevantes de Aprendizaje Automático, como precisión, recall y F1-score.
5. Integrar soluciones de explicabilidad e interpretabilidad sobre los modelos con mejor comportamiento para mejorar la extracción de conocimiento y comprensión del problema.
6. Realizar un análisis detallado de los resultados para identificar fortalezas y debilidades del modelo y sugerir posibles mejoras.

Finalizando este capítulo por último se explica la **estructura de la presente memoria** que puede ser contemplada en tres partes diferentes aunque interconectadas entre sí. Comienza con una introducción hacia la temática del sexismo escrito fundamentándose en los problemas que conlleva su presencia en la sociedad desde hace varias décadas y cómo ha aumentado en los últimos años gracias a la aparición, evolución y anonimato de las redes sociales. A continuación se plantean los objetivos que se pretenden alcanzar con este proyecto y las hipótesis iniciales acerca de las técnicas que se presumen van a ser más ventajosas de aplicar.

A partir del tercer capítulo se detalla el procedimiento por el que se ha llevado a la práctica los distintos hitos establecidos, iniciando la experimentación con un profundo análisis exploratorio de datos e investigaciones varias acerca de la procedencia de los conjuntos de documentos empleados. Prosigue con el modelado realizado empleando en un principio algoritmos clásicos de Aprendizaje Automático y posteriormente elevando el nivel de dificultad y complejidad utilizando arquitecturas y *transformers* propios de Aprendizaje Profundo. En cada una de las soluciones encontradas se adjunta un análisis explicativo de resultados de desarrollo propio a partir de los

conocimientos intrínsecos en la materia y el cruce de información muy útil procedente de la primera etapa del estudio de datos.

En la tercer parte se procede a la aplicación de técnicas de explicabilidad local sobre un subconjunto de errores cometidos por cada mejor modelo encontrado con los propósitos de conocer en qué terminología fundamentan sus decisiones, cuáles pueden ser sus posibles deficiencias que lideran la clasificación hacia categorías erróneas y cómo pueden ser mitigadas sus consecuencias con soluciones adicionales dependiendo de las conclusiones extraídas. Mientras que por un lado se propone una metodología de análisis puro, el segundo enfoque es más práctico en el que se trata de modificar los documentos incorrectamente clasificados para observar si los cambios son suficientemente significativos como para reconducir el comportamiento del modelo. Finalmente se puede un resumen detallado acerca de los resultados y las conclusiones extraídas a lo largo de este proyecto, acabando con varias líneas de investigación futuras para continuar abordando el problema planteado.

# Capítulo 2

# Fundamentos teóricos

En este segundo capítulo se propone como propósito principal una introducción detallada acerca de los pilares teóricos que han posibilitado la ejecución de este proyecto. Comienza con una breve explicación del origen del Aprendizaje Automático, así como su evolución durante las últimas décadas desembocando en una nueva rama más compleja denominada Aprendizaje Profundo. A partir de este momento se pretende informar sobre los distintos progresos realizados en este campo y su especialización en diferentes tareas dependiendo de la naturaleza de los datos objetivo. En nuestro caso particular se trata de Procesamiento del Lenguaje Natural para la que también se integra una descripción de los fines con los que se puede emplear, en combinación con las distintas soluciones localizadas en el estado del arte actual y su relevante aplicación en la resolución de problemas que surgen gracias a la inclusión de nuevas tecnologías. Debido a su creciente opacidad como consecuencia de la investigación y exploración de arquitecturas más enreversadas, por último se destaca el papel vital que desempeña un área recientemente empoderada que persigue el diseño de mecanismos de interpretabilidad capaces de explicar el modo de funcionamiento de los conocidos modelos como cajas negras.

## 2.1. Aprendizaje Automático y Profundo

Una definición popular de Aprendizaje Automático lo describe como una disciplina compuesta por un conjunto de técnicas que son capaces de construir un modelo computacional cuyo objetivo varía dependiendo de la información disponible y del problema que se pretende abordar. En este proyecto se establece el centro de atención en la categoría de Aprendizaje Supervisado, donde se persigue el **modelado de las relaciones existentes entre unos datos proporcionados como entrada y otros que repre-**

**sentan las salidas esperadas**. En particular, puesto que los valores finales pertenecen a intervalos discretos el modelo que se produce como resultado es comunmente denominado clasificador. Por lo tanto la primera etapa siempre se fundamenta en la recolección, almacenamiento, procesamiento y adaptación de un conjunto de muestras representativas del comportamiento que se desea replicar de forma automática. Dependiendo de la naturaleza de los algoritmos algunos de ellos además pueden aportar datos sumamente útiles acerca de cómo se relacionan los diferentes ejemplos, cuán relevantes son las características disponibles con respecto a la variable dependiente, así como métricas de rendimiento generales y específicas que evalúan la bondad del modelo. Al no existir unos estándares globales que guien este procedimiento es una práctica habitual la experimentación de varios de ellos para luego realizar una comparativa con la que ponderar su usabilidad y calidad en función de los criterios de aceptación establecidos dentro del problema que se pretenda abordar [6]. No obstante existen ciertos inconvenientes relacionados con el uso de algoritmos clásicos que han estado cada vez más presentes gracias a la evolución de las tecnologías, la **generación de cantidades ingentes de datos heterogéneos y el planteamiento de nuevos retos más complejos**. Por un lado la mayoría de técnicas pertenecientes al Aprendizaje Automático asumen que las clases a predecir son linealmente separables por un hiperplano, un fenómeno que raramente se produce en los problemas actuales. Asimismo, la **selección de características** es una etapa obligatoria en todos los mecanismos contemplados y una correcta elección requiere una enorme inversión de recursos personales con los que disponer del conocimiento experto necesario para comprender el ámbito del problema, los datos disponibles, si proporcionan o no información útil para su inclusión, los vínculos que los relacionan entre sí y con la variable dependiente. Una elección errónea puede producir un impacto sumamente negativo en el rendimiento del clasificador produciendo una solución de pésima calidad y un gasto temporal y económico desmesurado [7]. Sin embargo los **algoritmos clásicos aún suponen una herramienta popularmente empleada como primera vía de exploración** para tantear la dificultad del problema, la calidad de los datos y de las características y su grado de separabilidad, postulándose como una buena práctica para comenzar el modelado de un conjunto de datos. Durante este procedimiento es necesario tomar en consideración que se debe elegir la solución más sencilla que satisfaga todos los requisitos impuestos, por lo tanto se recomienda iniciar la búsqueda con la aplicación de las técnicas más sencillas conocidas para posteriormente aumentar el nivel de complejidad si los resultados proporcionados no son suficientemente buenos.

Basado en los progresos efectuados dentro del área de la Neurociencia acerca del funcionamiento del cerebro humano, en 1940 se fabricó el primer prototipo de una red neuronal simple en forma de circuito eléctrico única-

mente compuesto por una unidad lógica que simulaba el comportamiento inteligente de una neurona con salida binaria. Como consecuencia producía como resultado un cero en caso de que la suma ponderada de pesos se encontrase por debajo de un umbral establecido, y un uno en caso contrario. Dependiendo de la diferencia existente entre la salida producida y la esperada, se ajustaban los valores de los pesos hasta acertar con la predicción. Esta idea fue extendida en 1949 hacia un modelo más complejo con múltiples neuronas y capas, además de la adición de mecanismos matemáticos que le permitían actualizar los pesos de sus enlaces conforme más casos de uso se le proporcionaban. Así se introdujo una nueva disciplina más avanzada denominada Aprendizaje Profundo que resuelve los inconvenientes descritos anteriormente gracias a su habilidad para representar la distribución subyacente de los datos suministrados en distintos niveles con los que **extraer características y patrones automáticamente** modelando detalladamente su contenido [7]. De acuerdo a los elementos y a los mecanismos que componen a las redes neuronales profundas pueden ser clasificadas dentro de diversas categorías. La más sencilla se denomina *Feedforward Neural Network* cuya tipología permite la transmisión de la información en una única dirección desde las neuronas de una determinada capa a las que conforman la siguiente. Sin embargo presentan una deficiencia muy notoria en la que la red se encuentra limitada a encontrar solamente un óptimo a nivel local, perdiendo el contexto suministrado hasta el momento por los datos analizados previamente, y por tanto decreciendo su precisión y capacidad de generalización futura. Una segunda tipología popularmente conocida a nivel global en los últimos años pretende resolver esta problemática incluyendo relaciones cíclicas en su arquitectura. Las redes pertenecientes a la categoría de *Redes Neuronales Recurrentes* permiten alimentar a las neuronas de una capa con las propias salidas que genera, reteniendo parte de los datos previos para generar una especie de memoria que almacena ciertos estados anteriores con los que fundamentar mejor las futuras predicciones [8]. De este modo mitigan uno de los defectos más influyentes de las Redes Neuronales Recurrentes conocido como el desvanecimiento del gradiente. Se puede apreciar durante la construcción de un modelo cuando deja de considerar las muestras ya estudiadas como consecuencia del suministro de nuevos datos provocando una pérdida importante en su capacidad de predicción al olvidar los patrones extraídos al comienzo de su entrenamiento. Gracias a esta cualidad este tipo de arquitecturas han resultado ser de gran ayuda en el análisis y aprendizaje de conjuntos de datos secuenciales cuyos elementos se encuentran relacionados entre sí, como es la comprensión de contenido escrito [9].

## 2.2. Procesamiento del Lenguaje Natural

Se trata de una disciplina dentro del área de la Inteligencia Artificial que está especializada en un conjunto de técnicas cuyo objetivo consiste en representar y analizar documentos orales o escritos a diversos niveles lingüísticos **simulando la producción y comprensión del lenguaje natural que realizan las personas**. Existen dos principales enfoques que se practican en este proyecto: el procesamiento del lenguaje y la generación de lenguaje. Mientras que el primero juega un papel de lector u oyente disponiendo de metodologías con las que observar los ejemplos acústicos o escritos para transformarlos en valores interpretables por una máquina. El rol de la generación del lenguaje está mayormente reflejado como el de un escritor u orador que necesita una planificación previa con la que guiar su contenido.

Sus orígenes se remontan a la década de los cuarenta en la que comenzaron las investigaciones acerca de máquinas de traducción con las que poder descifrar los códigos de los enemigos durante la Segunda Guerra Mundial utilizando fundamentos teóricos de criptografía y teoría de la información aplicada al lenguaje natural. Una primera aproximación para llevar esta idea a cabo se apoyó en el uso de diccionarios en los que buscar los términos suministrados para ser reemplazados por los propios del lenguaje objetivo considerando el orden natural del mismo. Sin embargo los resultados proporcionados no fueron de buena calidad debido a las múltiples situaciones en las que la **ambigüedad del lenguaje provocaba traducciones incorrectas**. Condicionado además por la pobre tecnología disponible en aquella época y por las dificultades que encontraron al formular una teoría del lenguaje con la que crear algoritmos para construir una máquina de traducción, el Comité de la Academia Nacional de Ciencias recomendó cesar su investigación en 1966. A pesar de esta decisión, los progresos no decayeron en las siguientes décadas en las que se trabajó sobre mejores representaciones del lenguaje natural desde distintos niveles sintácticos, morfológicos y semánticos, haciendo especial énfasis en terminología anómala y estructura gramatical. Así surgieron nuevos sistemas que intentaban contradecir al informe redactado por el comité, como ELIZA un agente conversacional capaz de parafrasear el discurso de un paciente para reflejar el papel de un psicólogo, o como PARRY que en lugar de conceptos independientes empleaba conjuntos de palabras clave y sinónimos [10]. La segunda etapa de la evolución de PLN fue posible gracias tanto al incremento de los recursos computacionales con componentes más veloces y de mayor capacidad de almacenamiento, así como a los importantes avances realizados en la generación y explotación de datos y en el desarrollo de algoritmos de Aprendizaje Automático. Al contrario que en sus orígenes los nuevos métodos que se desarrollaron eran de una naturaleza más empírica, dejando de lado los cálculos estadísticos

y fundamentándose en una **simplificación máxima del funcionamiento de la mente humana**. Mediante operaciones de asociación, reconocimiento de patrones y métodos de generalización pudieron diseñar modelos y métodos de PLN tales como el cálculo de la máxima entropía en arquitecturas SVM, de la máxima información mutua, del mínimo error de clasificación y la construcción de un tipo de redes neuronales conocido como perceptrón. Por último en la tercera fase es en la que se involucra al nuevo paradigma del Aprendizaje Profundo gracias al desmesurado potencial que albergan tanto los algoritmos disponibles como los recursos computacionales existentes en la actualidad. De este modo se amplía considerablemente la diversidad de datos que pueden ser empleados dentro de las técnicas y los procedimientos contenidos en este área. En particular, la causa por la que se introdujo el Aprendizaje Profundo en el ámbito del PLN reside en las teorías de los investigadores que afirmaban esperar una **mejora notable en el rendimiento de los modelos**, que si bien podría requerir un mayor volumen de datos también pudiese disminuir la necesidad del conocimiento experto que es tan laborioso y costoso. Por un lado prácticamente la totalidad de los problemas planteados en PLN se pueden solventar mediante procesos de consulta, traducción, clasificación y predicción estructurada o secuencial. Mientras que asimismo los documentos escritos suelen ser representados mediante cadenas de *tokens* que son procesados y tratados igualitariamente independientemente de su contenido o temática, siendo una frase la unidad de procesamiento utilizada por defecto [11].

### 2.2.1. Técnicas de procesamiento de textos

Tal y como ha sido comentado anteriormente aún no es posible proporcionar los documentos escritos directamente a los algoritmos y modelos para su análisis y clasificación. Por lo tanto la integración de ciertos métodos especializados en el tratamiento de información textual es una fase obligatoria que debe aparecer en todo problema de PLN. Si bien existe una enorme multitud de metodologías diferentes que persiguen distintos propósitos, no se dispone de una combinación generalizada que garantice un buen resultado en todas las temáticas. Así su selección **dependerá de las características de los datos y de los objetivos que se deseen alcanzar**. A continuación se describen algunas de las más popularmente extendidas en idiomas sencillos, como inglés y español, que han sido establecidas como una guía a seguir para comenzar a trabajar con este tipo de información.

- **Eliminación de caracteres no alfabéticos**. En esta primera etapa se pretende suprimir todos aquellos caracteres pertenecientes a signos de puntuación, dígitos y símbolos que no contienen ninguna información útil aunque su procesamiento sí que consume recursos compu-

tacionales y temporales. Asimismo facilita la **tokenización** en la que únicamente se considera el espacio en blanco como separador para dividir un texto en sus diferentes términos individuales.

- **Palabras de parada**. Se trata de un conjunto de palabras que permiten interconectar los distintos elementos de un párrafo, tales como artículos, preposiciones, conjunciones y determinantes. Por lo general tampoco aportan información útil acerca del contexto de un documento, aunque su eliminación sí que altera la estructura del mismo pudiendo impactar negativamente en el rendimiento de los modelos. Como consecuencia este procedimiento no ha formado parte del procesamiento de datos previo a su modelado.

- **Lematización**. Es un proceso que tiene como objetivo principal la simplificación de toda la terminología que pueda ser expresada en plural o en diversas conjungaciones, como los verbos. Así reemplaza cada uno de estos vocablos por su palabra singular o su forma infinitiva favoreciendo posteriormente la codificación y comprensión por algoritmos automáticos.

- **Estemización**. Se trata de una versión más restrictiva que la técnica anterior en la que directamente se reemplaza cada vocablo por su raíz, eliminando los prefijos y/o sufijos que presumiblemente no aportan ninguna información adicional. Sin embargo, una transformación tan notable como esta puede influir negativamente en la capacidad de predicción de los modelos puesto que pueden ocurrir dificultades al intentar codificar los documentos y analizar sus significados. Por lo tanto la estemización no se ha incluido en el tratamiento de datos en este proyecto.

- **Detección y corrección de faltas de ortografía**. Es un fenómeno muy común la aparición de errores ortográficos en los textos procedentes de redes sociales. Por lo tanto una posible solución que se puede emplear para reducir el impacto de estos errores en la transformación de documentos y extracción de sus características reside en intentar detectarlos y reemplazar las palabras erróneamente redactadas por las correctas. Para llevar a cabo esta tarea existen multitud de enfoques que pueden ser muy sencillos, como la búsqueda parcial de cada término en un diccionario de palabras, hasta el uso de modelos inteligentes capaces de comprender el contexto y sugerir el vocablo más semejante al identificado minimizando el impacto en su significado. En capítulos posteriores se ejemplificarán ambas orientaciones mostrando cuán diferentes pueden ser en términos de los resultados que proporcionan.

### 2.2.2. Técnicas de codificación de textos

Como ya es conocido no es posible aún proporcionar los documentos directamente como entrada a ningún algoritmo o arquitectura, deben ser transformados a vectores numéricos que representen su contenido. Dependiendo del procedimiento que se aplique, las matrices matemáticas resultantes reflejarán su significado en mayor o menor medida. A continuación se detallan aquellas metodologías que han sido experimentadas en este proyecto.

- **Bolsa de palabras**. Es una técnica orientada a la extracción de características de un conjunto de textos que convierte a cada elemento en un vector de longitud fija calculando la frecuencia de sus términos. Si bien se trata de un procedimiento muy sencillo y rápido de aplicar, también conlleva algunos inconvenientes relevantes, como el hecho de no respetar el orden de los términos ni considerar el contexto o significado de los documentos.
- **TF-IDF**. Su codificación consiste en calcular la frecuencia absoluta y relativa de cada palabra considerando así tanto la importancia de un término dentro de los documentos en los que aparece, como con respecto a la población completa. Así se determina el grado de relevancia y representatividad de cada concepto con el que se confirma o no su presencia en los vectores numéricos. Este método suele favorecer a los vocablos con un menor número de ocurrencias puesto que les otorga un valor calculado más elevado.
- **Word embeddings**. El cometido de esta técnica se fundamenta en la codificación de documentos como vectores numéricos capturando sus características contextuales dependiendo de las relaciones que presenten entre sí, las definiciones de los términos que contienen, considerando el nivel de similitud sintáctica y semántica. De este modo dos términos con significados parecidos serán transformados a representaciones numéricas semejantes. Para su uso se han ideado dos planteamientos diferentes con el propósito de comparar su rendimiento en función de la calidad de los modelos.
  - **Generación de word embeddings**. Existen multitud de técnicas con las que entrenar modelos para fabricar word embeddings a partir del vocabulario de un conjunto de datos particular. En este proyecto se ha decidido experimentar con dos de las más popularmente utilizadas a nivel general.
    - ***Word2Vec***. La codificación de documentos se realiza a nivel de palabras a partir de la selección de uno de los siguientes algoritmos para llevar a cabo el entrenamiento del modelo generador de word embeddings.

- ◇ *Continuous Bag of Words*. El algoritmo CBWO trata de predecir el siguiente término dado un contexto determinado. Se encuentra recomendado para conjuntos de datos voluminosos ya que únicamente considera un vecindario de palabras y tiende a representar mejor los conceptos más frecuentes.
- ◇ *Skip-Gram*. Esta segunda aproximación intenta predecir los distintos contextos asociados a un único vocablo proporcionado como entrada. Puede ser empleado en volúmenes de datos menores para detectar distintos significados de una misma palabra.

- ○ ***Doc2Vec***. Se trata de una versión muy similar a la técnica anterior aunque añadiendo un nuevo vector numérico que representa los párrafos contenidos en los documentos suministrados. De nuevo, en esta técnica también se debe escoger uno de los dos siguientes enfoques disponibles.
  - ◇ *Distributed Memory Version of Paragraph Vector*. El algoritmo PV-DM emplea algoritmos de Aprendizaje No Supervisado para extraer vectores de características de tamaño fijo a partir de los distintos tipos de elementos, tales como frases, párrafos y textos completos. Así genera un contexto que le ayuda a predecir la secuencia de términos que lo representa. Se trata de un enfoque muy parecido al algoritmo CBWO de la técnica anterior.
  - ◇ *Distributed Bag of Words Version of Paragraph Vector*. El modo de funcionamiento del algoritmo PV-DBOW es prácticamente idéntico al algoritmo Skip-Gram de la metodología previa, en la que la fase de entrenamiento se apoya en la predicción de distintos contextos vinculados a un conjunto de términos.

### 2.2.3. Arquitecturas para textos

En este apartado se pretende ejemplificar los distintos recursos de Aprendizaje Profundo de los que se puede hacer uso para abordar problemas de clasificación de documentos escritos. Las siguientes arquitecturas forman parte de la rama de Aprendizaje Supervisado por su volumen de neuronas y capas, además de por la complejidad de sus interconexiones y mecanismos que las han situado en las primeras posiciones del ranking de soluciones más propicias para trabajar con textos dentro del estado del arte actual.

- **Redes Neuronales Convolutivas**. Este tipo de redes neuronales emplean operaciones de convolución con las que extraer los mapas

de características de los datos proporcionados, mientras que con sus capas de *pooling* minimizan su tamaño reduciendo la dimensionalidad de las entradas. Son utilizadas masivamente en problemas de detección, procesamiento, análisis y clasificación de imágenes, textos y videos.

- **Redes Neuronales Recurrentes**. Su origen reside en la necesidad que surge de contemplar el orden en el que los términos aparecen en una frase o documento para que el significado de los conceptos anteriores puedan contribuir en el análisis de los futuros. Este fenómeno ayuda a identificar dependencias entre términos permitiendo generar un contexto que se mantiene en una memoria limitada almacenando un subconjunto de entradas previamente procesadas. Otra ventaja de esta tipología de red reside en la capacidad de extraer características observando una muestra bidireccionalmente, de izquierda a derecha y viceversa. Como consecuencia la probabilidad de alcanzar una mejor representación aumenta considerablemente, además de propagar esta ventaja hacia el estudio de futuras entradas. Dentro de esta categoría se encuentra un ejemplo particular denominado ***Long Short-Term Memory*** (LSTM) que apareció en 1997 con un funcionamiento que se compone de un conjunto de subredes conectadas denominadas bloques de memoria. Cada uno contiene uno o más celdas de memoria interconectadas y tres puertas que permiten la escritura (*input*), lectura (*output* y eliminación de información (*forget*). Las interacciones entre estos tres elementos le otorgan a las redes LSTM las habilidades de **almacenar y acceder a datos relevantes y previamente analizados** durante varias iteraciones sin ser continuamente reemplazados por nuevos ejemplos.

- **Mecanismos de atención**. En lugar de utilizar arquitecturas con codificadores y decodificadores que únicamente transforman unas entradas en vectores de longitud fija sin importar la relevancia de cada elemento, una alternativa reside en los mecanismos de atención. Estos permiten utilizar capas densas adicionales con las que ayudar a la red a prestar más atención a unos términos u otros dependiendo de los objetivos que se deseen perseguir. Por ejemplo, en el procesamiento de documentos resulta beneficioso modificar el nivel de importancia en cada palabra para que así pueda aprender a identificar qué papel juega dentro de una frase y qué relación tiene con el objetivo del problema que se aborda.

- **Transformers**. Un nuevo tipo de arquitectura conocida por esta nomenclatura se ha posicionado dentro del estado del arte actual como una de las mejores soluciones que se pueden aplicar en tareas específicas de PLN. Su secreto reside en la combinación de redes codificadoras-decodificadoras con mecanismos de atención propios en cada una con

los que aprender a distinguir los elementos más relevantes, siendo sus salidas finalmente combinadas también con mecanismos de atención adicionales especialmente configurados para ello [12]. Dentro de esta categoría se encuentra un modelo particular empleado masivamente desde su origen en 2017: ***Bidirectional Encoder Representations from Transformers*** (BERT). Generalmente su modo de entrenamiento se fundamenta en dos principales etapas: *Masked Lanaguage Modeling* (MLM) cuyo objetivo consiste en predecir los términos que han sido enmascarados aleatoriamente, mientras que la segunda fase conocida como *Next Sentence Prediction* (NSP) se encarga de predecir si dos fragmentos escritos son adyacentes dentro de un determinado contexto. Esta arquitectura dispone de un volumen cuantioso de mecanismos de atención capaces de analizar una entrada suministrada para codificarla como un *embedding*. Previo a esta tarea se debe efectuar un procesamiento que segmenta el documento en *tokens* o palabras individuales insertando tokens especiales como [SEP] que permiten separarlos entre sí. A continuación cada token es codificado como un vector jerárquico de longitud fija que depende del contexto general del documento al que pertenece. En la estructura de datos fabricada incluye conocimiento acerca de los siguientes niveles lingüísticos.

- Información sintáctica. Varios estudios demostraron que los modelos BERT son capaces de efectuar análisis sobre las distintas partes que componen un párrafo o documento, almacenando así el tipo de terminología y el orden de aparición. Si bien su información parece no ser similar a la que pueden extraer las personas con un estudio sintáctico, sí que disponen de la habilidad de inferir este tipo de conocimiento a partir de su representación matricial.
- Información semántica. Asimismo existen suficientes pruebas que confirman la capacidad de los sistemas BERT de almacenar también conocimiento semántico tal como el tipo de entidad de cada vocablo, sus relaciones y roles que desempeñan dentro de un fragmento de texto [13].

## 2.3. Explicabilidad en Inteligencia Artificial

La necesidad de poder interpretar las decisiones que toman los sistemas inteligentes se fundamenta en una multitud de razones que van desde la comprensión de sus fallos con la que mejorar su rendimiento, hasta el nivel de confianza que se les puede otorgar cuando su campo de aplicación es tan crítico como podría ser el ámbito de la salud o la automoción. El primer planteamiento de este problema data en 1991 cuando se efectuó el primer ensayo de un vehículo totalmente autónomo. Su principal compo-

nente era una cámara con la que percibía y analizaba su entorno mediante una red neuronal. En cada prueba el sistema era entrenado durante solo unos minutos proporcionándole los movimientos que realizaba una persona conduciendo para posteriormente ser liberado sin ninguna intervención. La mayor parte de la experimentación fue un éxito hasta que un día casi ocurrió un accidente cuando circulaba por un puente. Sin embargo los investigadores no conocían las causas de esta indeseada conducta puesto que el modelo era considerado como una **caja negra** por lo que su modo de funcionamiento era totalmente opaco. Tras una prolongada temporada de estudios exhaustivos pudieron comprobar que el sistema se había sobreajustado a un determinado tipo de carretera, por lo tanto un contexto de distintas características como un puente asfaltado parece ser que le habría confundido provocando un fallo en la conducción. A partir de este momento se sucedieron una variedad de aproximaciones con las que intentar resolver esta problemática. El primero de ellos consistió en modificar la etapa de entrenamiento de una red neuronal en la que **en lugar de suministrarle la salida esperada, trataban de reconstruirla** a partir del conocimiento experto. Así se podía analizar las representanciones que el modelo realizaba de las características señaladas maximizando la probabilidad de que cada neurona aportase su dato correspondiente. Sin embargo, pronto descubrieron que este enfoque no era el correcto para erradicar el problema de las cajas negras puesto que los modelos eran fácilmente engañables con imágenes similares a las salidas esperadas [14]. Con el origen del Aprendizaje Profundo el inconveniente de la explicabilidad se hizo aún más patente puesto que las nuevas técnicas y arquitecturas que surgieron eran incluso más complicadas de comprender y menos transparentes en las decisiones que proporcionaban.

A raíz del incremento masivo de la automatización de todo tipo de procesos, surgen cuestiones acerca de si el nivel de confianza otorgado a estos sistemas es suficientemente elevado como para permitirles una actuación totalmente autónoma. Conforme más temerarios son los riesgos y más perjudiciales son las consecuencias que pueden desencadenarse tras la construcción de modelos de aprendizaje autónomos, más relevancia adquiere el tópico de la explicabilidad. Así surgió una nueva área dentro de la inteligencia artificial conocida como ***Explanaible Artifitial Intelligence*** con la que se concluyeron dos variantes de generación de explicabilidad orientadas a distintos propósitos. Mientras que la primera de ellas se basa en cuál es el funcionamiento del sistema a partir de la relación que presentan sus entradas y salidas, el segundo enfoque se mantiene en la construcción de modelos interpretables, como Árboles de Decisión, que aproximen el comportamiento al sistema que se desea explicar intentando conocer por qué toma las decisiones que manifiesta con respecto a las características de los datos y su configuración. Detallando aún más ambas propuestas finalmente se estableció una clasifiación de las técnicas disponibles en función de los siguientes

criterios: si su aplicación es generalizada a todos los algoritmos y arquitecturas, y si proporcionan explicaciones a nivel global o local a un conjunto de predicciones seleccionadas.

- **Modelos específicos globales**. Esta primera categoría asegura la interpretabilidad de un modelo al integrar condiciones a su estructura para conseguir una serie de características deseables que permitan su explicabilidad por personas. Entre ellas se encuentra la reducción del número de características que puede utilizar y que las relaciones que presenten con las salidas deseadas sean monotónicas. Adicionalmente se pueden incluir otro tipo de restricciones como semánticas para limitar el nivel de abstracción que se genera a partir de la extracción de información de las entradas.

- **Modelos específicos locales**. Emplea mecanismos de atención con los que mostrar al modelo cuáles son las características más relevantes en las que debe prestar mayor atención para aprender las características las instancias. Sin embargo, al contrario que la metodología anterior, en este caso es un requisito indispensable la selección de un conjunto de muestras sobre las que ejecutar este procedimiento.

- **Modelos agnósticos globales**. Esta técnica es de libre uso por lo que puede ser aplicada en cualquier tipo de algoritmo y arquitectura, ya que únicamente utiliza los datos de entrada y las predicciones que genera un modelo. Así existen dos principales enfoques, donde el primero intenta construir un sistema intrínsecamente interpretable basado en las relaciones que manifiestan la información entrante y saliente para intentar conocer qué factores han desencadenado las predicciones producidas. Mientras que el segundo consiste en utilizar técnicas de diagnósitco con las que descubrir la relevancia de características específicas en la generalización de las variable dependientes. Para su selección se pueden integrar gráficos de dependencias parciales o bien condiciones individuales apoyadas en el conocimiento experto sobre el problema que se esté abordando.

- **Modelos agnósticos locales**. Del mismo modo que la categoría anterior, los modelos agnósticos locales también pueden ser empleados en cualquier tipo de sistema inteligente. Igualmente su objetivo consiste en aproximar el comportamiento del sistema de caja negra con una técnica interpretable por humanos. Sin embargo dependerá exclusivamente de una instancia o de su vecindario generado a partir de la fabricación de muestras sintéticas mediante la aplicación de distintas perturbaciones o modificaciones sobre el ejemplo original [5]. Dentro de esta modalidad se ha decidido aplicar dos de las técnicas más popularmente extendidas: LIME y contrafactuales.

- **LIME**. Esta técnica tiene como propósito principal el entrenamiento de un modelo subrogado local a un conjunto de muestras con las que aproximar el comportamiento de un sistema de caja negra. Para ello genera un dataset aplicando ciertas perturbaciones sobre los ejemplos elegidos para analizar el grado de similitud entre las predicciones que produce y las asociadas a los textos originales, descubriendo el modo de razonamiento que le lleva a comportarse de una manera concreta. El modelo subrogado que fabrica es intrínsecamente interpretable y puede estar fundamentado en cualquier algoritmo, como Árboles de Decisión. Entre sus principales ventajas se encuentra la posibilidad de modificar el sistema de caja negra siendo igualmente válido el modelo interpretable si el enfoque es el mismo, la creación de explicaciones comprensibles por personas que pueden ser contrastadas utilizando conocimiento experto, así como su facilidad de uso a partir de librerías disponibles en varios lenguajes de programación. No obstante su mayor desventaja se orienta en el requisito de seleccionar un conjunto de muestras que dependiendo de su naturaleza pueden generar explicaciones muy diferentes.
- **Contrafactuales**. Una explicación contrafactual describe una hipótesis que contradice un hecho observado. Pueden ser empleadas para interpretar las predicciones que se realizan sobre un conjunto de textos limitado en el que el evento a analizar es la salida que produce el modelo, mientras que las causas que la desencadenan son las características que le son suministradas. Para llevar esta tarea a cabo basta con idear una serie de modificaciones mínimas en los valores de entrada de las variables independientes para visualizar si la predicción cambia o permanece inamovible [15].

Las últimas aportaciones más recientes a este nuevo campo consisten en la aprobación de diversas guías con principios básicos destinados a la maximización de beneficios y la mitigación de los posibles riesgos que pueden esconder la implantación de modelos de cajas negras sin el conocimiento apropiado sobre ellos. En Europa se dispone del documento **Ethics Guidelines for Trustworthy AI** en el que se detallan los tres pilares fundamentales relativos al cumplimiento de las leyes y regulaciones aplicables al ámbito concreto en el que se apliquen técnicas de Inteligencia Artificial, de los principios y valores éticos propuestos, además de una robustez formidable desde el punto de vista técnico y personal. Acompañando a esta guía existe un **framework específicamente diseñado para suministrar consejos y sugerencias** acerca de qué medidas se pueden adoptar para satisfacer los requisitos expuestos anteriormente. Las temáticas asociadas a las directrices que componen el diseño de este software hacen referencia desde el desarrollo,

despliegue y uso responsable y respetuoso con la sociedad, la fomentación de la investigación y la innovación para formalizar las necesidades y los requisitos propuestos, la trazabilidad completa del ciclo de vida del sistema inteligente hasta la adaptación de la evaluación del grado de explicabilidad de los modelos dependiendo del área considerada [3]. En relación al primer punto cabe destacar que no existía prácticamente ninguna legislatura que regulase la implantación de modelos inteligentes hasta la proposición **Artificial Intelligence Act** publicada por la Comisión de la Unión Europea en 2021. En ella se definen una variedad de estatutos y conceptos que persiguen la armonización y estandarización de los requisitos que deben alcanzar los sistemas inteligentes. Dependiendo de la criticidad del área, el volumen de reglas definidas puede aumentar en el caso de aquellos ámbitos sumamente delicados. Los principales tópicos tratados van desde la calidad esperada de los datos, la gobernanza en la toma de decisiones, las técnicas de procesamiento que se pueden emplear, la identificación y corrección de sesgos y el nivel de transparencia y explicabilidad de los modelos construidos. Hace un especial énfasis en estos dos últimos términos, declarando que la transparencia debe permitir una interacción correcta entre personas y modelos clarificando su modo de uso, además de una **interpretación humana de las salidas generadas** de una forma concisa, completa y correcta, describiendo nítidamente sus características, habilidades y limitaciones, siendo crucial dentro de los ámbitos más delicados. Asimismo, se han redactado fuertes directrices acerca del nivel de **documentación, monitorización y verificación de la transparencia** que deben demostrar los sistemas inteligentes con el fin de asegurar el cumplimiento de la regulación en las fases previas y posteriores a su integración en el mercado [4].

# Capítulo 3

# Metodología de trabajo

En este tercer capítulo se pretende aclarar el procedimiento que se ha llevado a cabo para la realización de este proyecto en todas sus facetas. Por lo tanto se detallará y justificará cada una de las etapas que han contribuido a la consecución de los objetivos establecidos al comienzo de la memoria. En primer lugar se sitúa la investigación que se efectuó acerca de los orígenes de los conjuntos de datos *EXIST* para abordar el problema de la detección de sexismo escrito, comentando en qué fuentes de información se extrayeron las conclusiones que se presentan en posteriores capítulos. A continuación se detalla y justifica cada uno de los análisis exploratorios de datos que se han planteado para conocer las peculiaridades de los datos antes de ser trabajados. Prosigue el capítulo con las distintas experimentaciones diseñadas para cada algoritmo y arquitectura considerada durante la construcción de clasificadores eficaces en la identificación de textos sexistas y no sexistas. Finaliza descubriendo las metodologías llevadas a cabo con las que intentar arrojar cierta luz acerca del funcionamiento de los mejores modelos encontrados, descubriendo sus modos de razonamiento, fortalezas y debilidades para posteriormente intentar minimizar el impacto negativo que puedan tener sobre la generalización en muestras desconocidas con la integración de técnicas adicionales. En cada una de las etapas comentadas se incluirá un manifiesto acerca de las tecnologías que se han utilizado para su ejecución.

## 3.1. Recopilación de información

En primer lugar se inició una búsqueda e investigación acerca de los orígenes de los conjuntos de datos *EXIST* [17] tanto en la página oficial como en los artículos que resumían prácticamente todos los aspectos más relevantes de las competiciones celebradas en 2021 [18] y 2022 [19]. De ellos se ha estudiado especialmente los propósitos que fundamentan el planteamiento

de esta temática y la construcción de sendos datasets, las **fuentes de información de las que proceden, así como su modo de recopilación, procesamiento, balanceado y etiquetado de datos**. La explicación que reside en porqué ha sido este paso el primero efectuado en este proyecto se fundamenta en la enorme importancia que se esconde tras la metodología de construcción de los datasets, pues influye directa y elevadamente en el futuro modelado y resolución del problema que se aborda. Conocer los conjuntos de documentos con los que se pretende trabajar resulta ser un requisito indispensable para su posterior correcto tratamiento considerando sus características, beneficios y deficiencias que puedan presentar.

## 3.2. Análisis de datos

La siguiente etapa consiste en realizar diversos estudios estadísticos acerca de las variables disponibles en los datasets y sus respectivos valores almacenados. Empieza por una **descripción profunda de las características** almacenadas en ambos conjuntos de datos, así como las variables dependientes destacando principalmente su descripción, naturaleza y propósito dentro del problema de detección de sexismo escrito. Acompañando a este estudio se han efectuado múltiples análisis univariantes en los que resalta la distribución de los valores de todas las variables, prestando especial atención al **balanceado de las clases** pertenecientes a cada una de las características nominales. Y es que resultaba particularmente relevante conocer si existía una población equitativa entre los documentos clasificados como sexistas y no sexistas, si el número de ejemplos de cada idioma también era equiparable entre sí o si existía una fuente de información mayoritaria de entre las redes sociales consideradas, puesto que cada una dispone de unos objetivos y patrones propios que pueden afectar al rendimiento de los futuros clasificadores.

La segunda fase de este estudio analítico se concentra únicamente en los textos disponibles para calcular diferentes métricas que proporcionasen información acerca de su **longitud y diversidad léxica** considerando los ejemplos originales y tras su procesamiento, realizando una comparativa acerca del volumen de caracteres y terminología útil. Adicionalmente se reflexionó acerca de las nubes de palabras y la selección de N-gramas con la posibilidad de adentrarnos más en el contenido de los documentos y descubrir los **tópicos que tratan y los conceptos más frecuentes**.

Finalmente después de descubrir en los artículos mencionados anteriormente la dificultad que podría implicar la identificación automática de ciertos documentos sexistas por su contenido no explícito como tal, se propuso la idea de intentar emplear mecanismos con los que asignar a cada texto un sentimiento y así comprobar cuán confusos podrían resultar en un análi-

sis automático. Sin embargo, nos percatamos que el planteamiento podría estar incompleto puesto que la categorización de las muestras en sentimientos positivos, neutrales o negativos no aportaría demasiada información útil que posteriormente pudiese emplearse en la construcción de clasificadores. Así continuamos investigando acerca de qué técnicas se podían usar para demostrar este fenómeno y concluimos en el empleo de un *transformer* preentrenado [20] capaz de **analizar los documentos clasificándolos en un subconjunto de seis emociones** disponibles: alegría, tristeza, miedo, enfado, amor y sorpresa. Aprovechamos este método para calcular tanto métricas a nivel global como a nivel local considerando las categorías sexistas de las muestras positivas.

Como lenguaje de programación se ha empleado Python por la gran experiencia y soltura que tengo tras trabajar con él más de tres años en el ámbito académico y empresarial, además de por enorme volumen de librerías disponibles para tareas de Ciencia de Datos. En particular, se ha hecho uso de *pandas* que facilita la manipulación, procesamiento y visualización de datos tabulares, *nltk* y *spacy* para el tratamiento y limpieza de textos por su facilidad de uso en la aplicación de multitud de técnicas con las que únicamente dejar aquellos caracteres y términos útiles para el problema, además de la librería *transformers* que permite la descarga y el uso de modelos preentrenados como ha sido el caso del *transformer* T5 empleado en la detección de emociones.

## 3.3. Modelado inteligente de datos

El principal objetivo del modelado de datos consistió en intentar igualar las métricas de validación demostradas por los mejores modelos declarados en ambas competiciones, tomando como guía los algoritmos y las arquitecturas definidos, aunque añadiendo técnicas adicionales en la experimentación por iniciativa propia. En todos los ensayos realizados se efectuó una **distinción por idioma** por diversos motivos, siendo los principales la sencillez de entrenar un clasificador por idioma en lugar de un único multilenguaje, la existencia de embeddings y modelos preentrenados en un idioma particular y las declaraciones en ambas competiciones de la existencia de una mayor representatividad por parte de los documentos españoles que provocaba un rendimiento superior en términos de las métricas de validación de los clasificadores especializados en este lenguaje. Asimismo, todos los modelos también tienen en común la política de **reducción de sobreajuste** estableciendo el mecanismo de *Early Stopping*, con un máximo de cien iteraciones suprimiendo el proceso y retornando los pesos del modelo mejor valorado tras quince iteraciones sin mostrar un avance significativo superior al 0.01 en *accuracy*.

### 3.3.1. Algoritmos clásicos de Aprendizaje Automático

Tal y como se ha anunciado en el capítulo previo, como buena práctica se inicia el modelado de datos mediante un algoritmo clásico de Aprendizaje Automático como es la **Regresión Logística en combinación con distintas técnicas de codificación de documentos**. Puesto que sus resultados son determinísticos, bastó realizar una ejecución por parámetro que se deseó ajustar para valorar su idoneidad en la maximización de las métricas de calidad. Los pasos que se establecieron para iniciar una búsqueda de la configuración más propicia se resumen a continuación:

- **Tratamiento de textos**. Empleando los valores por defecto adjudicados al algoritmo de Regresión Logística se experimentaron con la integración de distintas técnicas de procesamiento de documentos, tales como la eliminación de caracters no alfabéticos, lematización y corrección de errores ortográficos.
- **Codificación de documentos**. Se tomaron distintos métodos de conversión de documentos a representaciones vectoriales, comprobando su influencia en la bondad de los clasificadores a partir de sus evaluaciones sobre ambos conjuntos de datos.
- **Regularización**. Se trata de uno de los parámetros más relevantes de este algoritmo, que resulta ser crucial para disminuir el indeseado fenómeno del sobreajuste mayormente detectado por la sustancial diferencia entre las métricas de validación sobre el conjunto de entrenamiento y sobre el de test.

Para esta primera fase en la construcción de modelos se empleó de nuevo el lenguaje de programación Python junto con la librería *scikit-learn* con la que también dispongo de una sobrada experiencia y por la multitud de algoritmos y métodos de visualización que contiene.

### 3.3.2. Arquitecturas de Aprendizaje Profundo

Posteriormente se aumenta la complejidad de los algoritmos utilizados para producir la siguiente tanda de sistemas inteligentes basados en una tipología de red neuronal profunda como las **arquitecturas LSTM unidireccionales y bidireccionales**. Su elección radica en que, además de situarse en los primeros puestos de los rankings de ambas competiciones, se trata de una salto cualitativo importante que permite emplear una memoria limitada con la que generar un contexto que ayude a analizar el contenido de los documentos de una forma más precisa. A diferencia de la experimentación previa los resultados con arquitecturas LSTM no son determinísticos,

por lo que en ejecuciones distintas se obtienen métricas diferentes para un mismo valor en un parámetro determinado. Como consecuencia, para ajustar la configuración se ha diseñado un procedimiento en el que **entrenar y validar treinta clasificadores por cada valor de cada parámetro** considerado, calculando la media aritmética de las métricas de validación con las que valorar cuál resulta ser más beneficioso logrando una mayor capacidad de generalización. Sin embargo, con el propósito de otorgarle una continuidad a las distintas pruebas efectuadas con las diferentes aproximaciones de modelado, se integra en el punto inicial el tratamiento de documentos más ventajoso observado en el algoritmo anterior, así como la codificación con embeddings preentrenados por haber demostrado un mayor potencial de representatividad con respecto a los restantes métodos probados. Así se aprovechan los beneficios obtenidos en cada modelado reduciendo los recursos temporales y computacionales de estas investigaciones. A continuación se resumen los parámetros propios de las arquitecturas LSTM que han sido ajustados.

- **Codificación de documentos**. Aludiendo a la premisa anterior, se tomaron distintos ficheros con diversos conjuntos de embeddings preentrenados para alcanzar el objetivo de descubrir cuál de ellos resultaba ser mejor para representar los textos disponibles con la mayor precisión y volumen de términos.

- **Arquitectura**. Definiendo multitud de composiciones el propósito fijado consistía en comprobar si una arquitectura más compleja podía disponer de una mayor capacidad de predicción, ensayando tanto con capas ocultas unidireccionales como bidireccionales para también determinar si el análisis de las muestras en ambos sentidos podía proporcionar un mayor volumen de información que ayudase a los modelos en su tarea.

- **Volumen de datos**. Tanto el tamaño del lote suministrado durante cada iteración en la construcción de los clasificadores, como el número de muestras de entrenamiento, han sido puestos en tela de juicio para comprobar si por un lado un aprendizaje basado en menos muestras aunque más tiempo otorga una mejor habilidad de predicción. Mientras que por otro lado un mayor número de documentos puede ejemplificar más detalládamente los patrones que deben interiorizar para la detección de sexismo escrito.

Para los sistemas inteligentes LSTM se ha hecho uso de la librería *Keras* en Python por su facilidad de uso y mi experiencia previa con ella en otros proyectos de clasificación con imágenes y textos.

Por último, replicando el mismo procedimiento descrito anteriormente se han realizado diversos ensayos con ***transformers* BERT preentrenados** disponibles en Hugging Face, en particular *bert-base-uncased* especializado en inglés y *dccuchile/bert-base-spanish-wwm-uncased* para los documentos en español. Tras las múltiples experimentaciones y los cuantiosos puntos comunes que comparten las arquitecturas LSTM con esta nueva tipología, las lecciones aprendidas previas prácticamente pueden ser aplicadas en su totalidad, aliviando enormemente los requisitos de los ensayos producidos para aplicar ***fine tuning*** a los modelos preentrenados. Este procedimiento consiste en reentrenar los sistemas por completo sin modificar su composición para ajustar sus pesos ya establecidos a los conjuntos de datos *EXIST*.

- Volumen de datos. Se reduce el rango de valores a probar para ajustar aún más el tamaño del lote y las técnicas de generación de muestras sintéticas de entrenamiento, puesto que se intuía que dichos parámetros serían ligeramente distintos e inferiores. El motivo reside en que con arquitecturas tan complejas como *transformers*, se necesitaría disponer de un menor número de ejemplos por lote y de una combinación diferente de metodologías con las que fabricar más documentos para ejecutar *fine tuning*.
- **Tasa de aprendizaje**. Apelando a la misma teoría previa, igualmente también se reflexionó acerca de la teoría de ralentizar el aprendizaje durante la personalización de los *transformers* a los datos del conjunto *EXIST*. Este ha sido el único parámetro nuevo que se ha añadido a la experimentación y del que no se ha podido beneficiar de los ensayos procedentes de los modelados con los otros algoritmos.

Como herramientas de nuevo se ha empleado Python como lenguaje de programación y una combinación entre las librerías *transformers* con los que descargar los clasificadores preentrenados, sus tokenizadores y codificadores propios, además de *Tensorflow* y *Keras* con los que inicializar su definición, constatar los procesos de preparación de datos, entrenamiento y validación.

## 3.4. Explicabilidad de los modelos

Gracias a la meta principal de este proyecto se ha alcanzado un elevado grado de explicabilidad de los mejores modelos producidos mediante dos enfoques diferentes. Una primera aproximación desarrollada propiamente se encuentra fundamentada en tres análisis de diversa naturaleza. El primero de ellos consiste en descubrir los **intervalos de confianza** bajo los que se han cometido los falsos negativos y positivos contabilizados en cada clasificador, con el propósito de identificar con qué seguridad los sistemas han

etiquetado incorrectamente estas muestras. Así se intenta medir el grado de dificultad que conllevaría la reconducción de este comportamiento indeseado hacia su correcta categorización. La siguiente etapa consiste en cruzar los ejemplos realmente positivos con sus correspondientes **tipos sexistas** a los que pertenecen con el fin de identificar si se trata de aquellas clases tachadas de complicadas de reconocer automáticamente por no tener un contenido lo suficientemente explícito. Finalmente se intenta confirmar la teoría extraída de esta segunda fase a partir de la combinación de los errores cometidos con las **emociones** que les han sido asignadas por un *transformer* preentrenado. De este modo se puede refutar la teoría de que la terminología y las expresiones usadas son las culpables de liderar el etiquetado hacia una categorización incorrecta, desconcertando también a este modelo que lleva a la práctica un propósito diferente del de clasificación.

Posteriormente en la segunda parte sí que se emplean algunas de las metodologías de explicabilidad reconocidas dentro del ámbito *Explanaible Artifitial Intelligence* (XAI). En particular se han seleccionado dos técnicas pertenecientes a **modelos agnósticos locales** puesto que son aplicables independientemente de la tipología de los clasificadores a partir del estudio de las relaciones existentes entre las entradas que son proporcionadas y las salidas que generadas por los sistemas. Si bien son factibles en cualquier suerte de muestras, nuestro propósito se ha orientado hacia el análisis de las posibles causas que han desencadenado los errores contabilizados como falsos negativos y positivos. Dado que el volumen de sendos conjuntos es muy elevado para cada sistema engendrado, se ha realizado un **submuestreo aleatorio de diez muestras por cada intervalo de confianza** con el propósito de analizar e inspeccionar visualmente las distintas teorías explicativas de estos errores. De nuevo se ha establecido una distinción por idiomas en caso de que las conclusiones extraídas fuese diferentes, aunque al no ser así se han ejemplificado mediante el modelado español.

El primer método ejecutado ha sido **LIME** por su capacidad de crear nuevas muestras sintéticas a partir de las permutaciones automáticas que sufren un conjunto de ejemplos, evaluando posteriormente la clase asignada por cada modelo y destacando los términos en los que se fundamenta la decisión tomada. LIME es conocido por su popular uso tanto en el ámbito académico como empresarial, su facilidad de empleo gracias a la librería *lime* disponible en Python y sus representaciones tan sencillas que demuestran el grado de pertenencia a cada clase junto con los conceptos que las soportan. En la segunda parte de esta experimentación se ha hecho referencia a la producción de **contrafactuales** utilizando dos de las operaciones más extendidas a nivel general, como son la eliminación de palabras secuenciales y la negación de verbos. Para el primer caso se ha diseñado un desarrollo propio libre de bibliotecas de terceros utilizando únicamente el lenguaje de programación Python, mientras que para el segundo procedimiento se ha

optado por emplear el modelo generador de lenguaje más famoso hasta la fecha conocido como GPT-3. Debido a sus novedosas habilidades, los buenos resultados proporcionados en distintas áreas y su potencial de integración en prácticamente cualquier tipo de negocio, pensamos que era el candidato ideal para este cometido. Con estos dos procedimientos en mente, se han seleccionado de nuevo diez ejemplos de cada intervalo de confianza por cada clasificador disponible para crear tantos contrafactuales como sea el tamaño de cada texto durante la ejecución del primer método, y un contrafactual por documento en la aplicación de la segunda operación. Posteriormente se han filtrado aquellos en los que sí se ha logrado una reclasificación hacia la categoría correcta, sometiéndolos a la técnica LIME con el fin de realizar una comparativa entre el contenido original y el modificado analizando sus diferencias y el posible impacto que ha permitido su correcta identificación.

# Capítulo 4

# Análisis de datos

Como buena práctica se ha comenzado a abordar este proyecto investigando acerca de los orígenes de los conjuntos de datos *EXIST2022* [17] profundizando en las fuentes de información empleadas para su recolección, comprobando la existencia de ciertos procedimientos y tratamientos de documentos que hayan sido aplicados con diferentes propósitos, como el balanceamiento de las clases contenidas en las variables dependientes, así como la metodología llevada a cabo para etiquetar todas las muestras disponibles. A continuación se integra una segunda sección de índole propia en la que se detallan los análisis exploratorios efectuados sobre las variables disponibles desde los más genéricos hasta los más centrados en tareas de Procesamiento del Lenguaje Natural. El objetivo general que se persigue consiste en conocer las peculiaridades de las muestras de ambos datasets para posteriormente orientar su modelado maximizando el rendimiento de los clasificadores y las garantías de de su adecuación con las que conseguir una solución de calidad.

## 4.1. Conjuntos de datos *EXIST*

El proceso de recolección de datos efectuado por los organizadores de las competiciones tiene como fuentes de información primarias a **Twitter y Gab**, dos redes sociales popularmente extendidas a nivel global. Utilizando sus respectivas API establecieron una recopilación de publicaciones textuales basada en la búsqueda de un conjunto de **términos y expresiones, en inglés y español, comunmente definidos como sexistas**. Para determinar estos filtros acudieron a proyectos y artículos relacionados con la temática, cuentas y hashtags orientados a la lucha contra la discriminación por género, además de a expertos humanos en la materia que examinaron y validaron la contribución de conceptos sexistas menos frecuentes con los que intentar aumentar la representatividad de los conjuntos de datos. Debido a

la obtención de millones de muestras en inglés y español tras orquestar dos períodos de recopilación, uno por cada conjunto de datos disponible para entrenamiento y test, se aplicaron una variedad de técnicas de muestreado con las que **equilibrar el número de ejemplos en cada idioma y reducir los posibles sesgos introducidos** contemplándose desde diversos puntos de vista.

- Desde una perspectiva conceptual se han utilizado **más de cien términos distintos como semillas**, para cada idioma, con los que realizar las búsquedas y filtrados de documentos procedentes de ambos medios de comunicación.

- Considerando los intervalos temporales de recogida de muestras, afirman que el sesgo introducido es mínimo al **espaciar temporalmente** a nivel global, con una diferencia anual, la recopilación de muestras para los conjuntos de entrenamiento y test, además de a nivel local al haber ejecutado los procesos de ingesta con diferentes semillas en distintos plazos temporales.

- Desde el punto de vista de los **usuarios** detectaron y eliminaron información personal almacenada en las publicaciones textuales, verificaron que cada documento únicamente perteneciese a un usuario y excluyeron del conjunto de test a aquellos usuarios con representación en el conjunto de entrenamiento.

En el **etiquetado de datos comenzaron haciendo uso de Amazon Mechanical Turk**, un servicio que permite configurar una plataforma en la que se asigna la ejecución online de una tarea manual a un conjunto amplio de personas. En este caso se trata de la clasificación de los documentos coleccionados en función de una guía elaborada que recogía la descripción de cada categoría disponible tanto en la detección de sexismo, para conocer si un texto es o no sexista, y en caso afirmativo qué tipo de sexismo caracterizaba su contenido. Para evitar la asociación de clases aleatoriamente, adicionalmente se integraron mecanismos de calidad basados en diversas métricas, como el tiempo de respuesta o la desviación con respecto a la distribución de las clases. Tras obtener un primer etiquetado de los conjuntos de entrenamiento y test, posteriormente se organizaron varias **revisiones con un equipo de expertos de ambos géneros especialistas en la materia** para así realizar las modificaciones pertinentes y desbloquear los casos en los que había un empate entre clases. Cabe destacar que las discrepancias entre el voto mayoritario del público y la opinión de los expertos fueron extremadamente elevadas, quedando patente la dificultad y subjetividad que entraña esta tarea [18] [19].

## 4.2. Análisis exploratorios

### 4.2.1. Tipos de variables

El primer estudio confeccionado consiste en identificar los tipos de datos y el significado de cada una de las características existentes en ambos conjuntos de datos.

- *test_case*: columna nominal que indica la edición de la competición a la que pertenece cada registro.
- *id*: columna numérica que representa un identificador único por documento.
- *source*: columna nominal que define la fuente de información de la que procede un texto: Twitter o Gab.
- *language*: columna nominal que especifica el idioma en el que se encuentra escrito un documento: inglés o español.
- *text*: columna nominal que almacena el contenido de un texto.
- *task1*: columna nominal que representa la primera variable a predecir puesto que señala si un documento es o no sexista.
- *task2*: columna nominal con la segunda variable dependiente que refleja la categoría sexista a la que pertenece un texto clasificado como tal.

### 4.2.2. Distribución de datos

A continuación se procede a elaborar múltiples análisis univariantes de las propiedades originales que aportan mayor información útil para la resolución de este problema de clasificación, siendo visibles sus resultados en la Tabla 2.1 para el conjunto de entrenamiento y de test. La primera diferencia destacable es que el **conjunto de entrenamiento únicamente se compone de documentos extraídos de Twitter**, mientras que el de test sí contiene textos procedentes de ambas redes sociales aunque el grupo de muestras asociadas a Gab es muy minoritario. Un segundo aspecto sumamente influyente es el equilibrio existente entre el volumen de documentos en español e inglés, al igual que el número de textos clasificados como sexistas y no sexistas, por lo que se puede determinar que se presenta un **balanceado prácticamente total de las muestras por idiomas y por clases**, tanto relativas a la detección de textos sexistas como a la categorización de los ejemplos positivos.

**Tabla 4.1:** Tabla con los resultados numéricos de los análisis univariantes sobre los conjuntos de entrenamiento y test.

| | Entrenamiento | Test |
|---|---|---|
| Datos de Twitter | 6.977 | 3.386 |
| Datos de Gab | 0 | 982 |
| Textos ingleses | 3.436 | 2.208 |
| Textos españoles | 3.541 | 2.160 |
| Muestras no sexistas | 3.600 | 2.087 |
| Muestras sexistas | 3.377 | 2.281 |
| Total de documentos | 6.977 | 4.368 |

### 4.2.3. Frecuencia de caracteres

Proseguimos con esta sección mostrando en la Figura 2.1 la frecuencia de caracteres de los conjuntos de entrenamiento y test originales y tras aplicar las siguientes técnicas de procesamiento de textos:

1. Eliminación de las direcciones y enlaces URL puesto que no proporcionan datos de calidad para ser empleados como parte de una solución.

2. Eliminación de las menciones de usuarios puesto que no disponemos de más información sobre ellos como para que su diferenciación suponga beneficiosa.

3. Eliminación de palabras de parada o *stopwords* tanto en inglés como en español respetando la estructura de los documentos.

4. Conversión de todos los caracteres restantes a minúsculas con el fin de normalizarlos y estandarizarlos bajo el mismo estilo.

A simple vista la discrepancia más resaltable es que la longitud de los textos de entrenamiento es más diversa que la de las muestras del conjunto de test, ya que la concentración de su población se manifiesta en más intervalos. Por el contrario, en el conjunto de test prácticamente la totalidad de su población se aglutina en el primer intervalo entre 0 y 50 caracteres, por lo que la conclusión extraída determina que los **documentos para validar los modelos se caracterizan por tener un tamaño bastante inferior**. Dependiendo de su nivel de representatividad y calidad de la terminología, este fenómeno puede influir negativamente en el rendimiento de los sistemas inteligentes puesto que podrían esperar textos más enriquecidos y al proporcionarles ejemplos más escuetos podrían cometer un mayor número de fallos por falta de información. Por otro lado apenas se aprecia algunas diferencias entre las frecuencias de caracteres de los conjuntos originales y

procesados, siendo ligeramente más visibles en los textos de entrenamiento reduciendo tanto sus longitudes a menos de la mitad y sus frecuencias en más de doscientos puntos. Una teoría explicativa de este suceso radica en que la **mayoría de documentos contienen un elevado porcentaje de caracteres y términos que no son útiles** para la resolución de este problema de clasificación, ensalzando la importancia de aplicar diversos procedimientos para el tratamiento y la limpieza de textos dejando únicamente sus elementos más representativos.

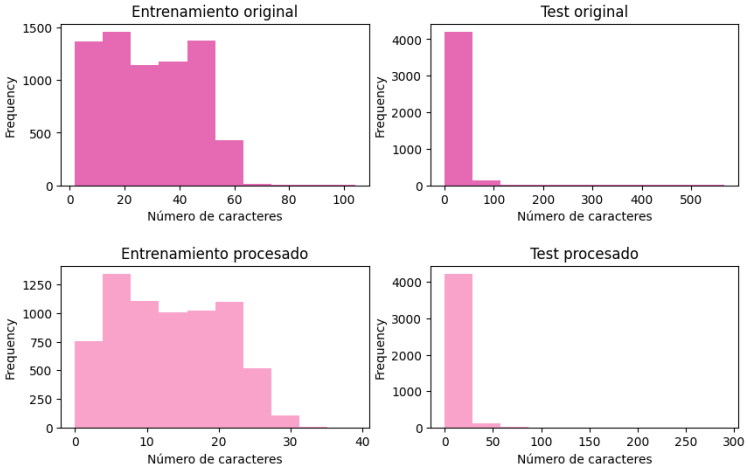


**Figura 4.1:** Análisis de la frecuencia de caracteres en los conjuntos de entrenamiento y test originales y procesados.

### 4.2.4. Diversidad léxica

Acompañando al estudio anterior basado en documentos se encuentra el siguiente análisis fundamentado en descubrir el grado de riqueza léxica asociada a los conjuntos de entrenamiento y test, también comparando los resultados entre los textos originales y los que han sufrido el tratamiento explicado anteriormente. En el objetivo consiste en elaborar una comparación entre el número de términos total y el volumen de conceptos distintos con los que determinar cuán diversas son las muestras recopiladas para este problema de clasificación y si las técnicas de procesamiento aumentan o disminuyen esta cualidad. En la Figura 2.2 se observa una gráfica de barras agrupada según la naturaleza de los textos analizados representándose en conjunto la frecuencia de términos distintos y totales. A rasgos generales se puede apreciar una ínfima variación entre ambas métricas calculadas, lo que simboliza que existe una **elevada diversidad léxica y más considerando que proceden de redes sociales**, que por lo general se tratan de medios de comunicación con escasa calidad lingüística. Este fenómeno es especial-

mente acusado en los conjuntos de datos procesados puesto que la diferencia entre los términos totales y distintos es considerablemente menor que en los datasets originales, por lo tanto se confirma que la aplicación de los debidos métodos de **tratamiento de textos puede ayudar a eliminar aquellos términos con poca información útil** de modo que únicamente se inviertan los recursos necesarios en los conceptos más enriquecedores.

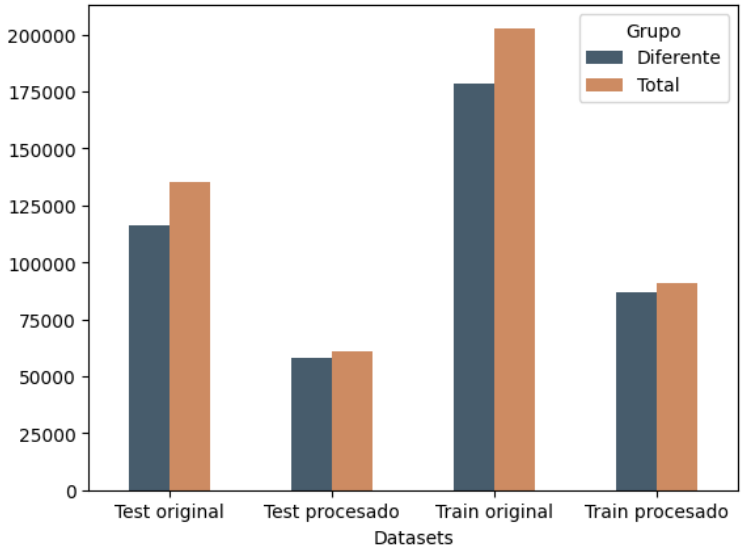


**Figura 4.2:** Análisis de la diversidad léxica en los conjuntos de entrenamiento y test originales y procesados.

### 4.2.5. Nubes de palabras

Una vez conocidas las principales peculiaridades de las muestras implicadas en sendos datasets, a continuación se presentan dos nubes de palabras considerando únicamente los documentos ya procesados con el propósito de descubrir los términos más relevantes en función de su frecuencia de aparición la cual, a su vez, determina el tamaño de su representación siendo mayor conforme más elevado es el valor calculado. Tal y como se puede observar en la Figura 2.3 y Figura 2.4, tanto en el conjunto de entrenamiento como en el de test, existe una **amplia variedad de vocablos relacionados con todo tipo de tópicos**. Gramaticalmente la inmensa mayoría de ellos son verbos que representan acciones básicas como ser, gustar, mirar o saber, propiamente utilizados para describir las capacidades y cualidades de una persona.

En la Tabla 2.2 se adjunta una recopilación de las **frecuencias absolutas de los diez términos más relevantes** y de mayor tamaño dentro


**Figura 4.3:** Nube de palabras sobre los documentos de entrenamiento ya procesados.


**Figura 4.4:** Nube de palabras sobre los documentos de test ya procesados.

de las nubes de palabras anteriores. Tal y como se aprecia existe una prominente discrepancia entre la frecuencia absoluta del verbo "gustarz el resto de terminología del conjunto de entrenamiento, por lo que esta acción se encuentra muy presente en los documentos de este dataset. Asimismo también destaca el elevado porcentaje de aparición del verbo "serçon respecto a la terminología sucesora. Adicionalmente cabe destacar el término "people"que parece sugerir la existencia de una cierta **inclinación a la generalización por parte de los autores de los documentos**, uno de los fenómenos más propensos a aparecer cuando se realizan críticas personales. Mientras que el conjunto de test los vocablos más relevantes son prácticamente los mismos, sus frecuencias absolutas son menores debido tanto al menor volumen de documentos como a su tamaño inferior, según se ha percibido en análisis anteriores. No obstante, parece compartir una tendencia similar con el dataset de entrenamiento.

**Tabla 4.2:** Tabla con los diez términos más relevantes y sus frecuencias absolutas.

| Conjunto de datos | Términos | Frecuencia absoluta |
|---|---|---|
| entrenamiento | like | 732 |
| | si | 530 |
| | dont | 503 |
| | ser | 447 |
| | feminismo | 272 |
| | one | 261 |
| | look | 256 |
| | people | 247 |
| | know | 239 |
| | get | 223 |
| test | like | 460 |
| | dont | 321 |
| | si | 292 |
| | ser | 247 |
| | get | 184 |
| | one | 164 |
| | people | 159 |
| | know | 148 |
| | would | 146 |
| | look | 144 |

### 4.2.6. Análisis de N-gramas

En la Figura 2.5 y Figura 2.6 se muestran los resultados de elaborar un estudio extendido al anterior con la pretensión de encontrar las **veinte expresiones o conjuntos de palabras con una mayor frecuencia de aparición**. Para esta tarea se han dividido los documentos de entrenamiento y test en un conjunto de sentencias o N-gramas estableciendo este parámetro a diez términos, escogiendo este valor por su balance entre el tiempo invertido en la búsqueda y la información que puede proporcionar acerca de los tópicos más populares. Tal y como se observa en el conjunto de términos de entrenamiento los mensajes parcialmente mostrados reflejan una **menospreciación hacia las capacidades y actitudes de las mujeres**, cosificando su cuerpo y ensalzando la violencia contra ellas. Sin embargo en el conjunto de los veinte N-gramas procedentes del dataset para validación las temáticas parecen ser de muy diversa índole con un contenido más difuso que no clarifica a simple vista si se trata o no de un texto sexista. Esta situación puede provocar un **elevado nivel de confusión en los modelos entrenados** al verificar su capacidad de predicción y bondad frente a muestras desconocidas.

```
['sit in front with her will he accept what if',
 'me to cook i need it for husband later in',
 'smears it on her face so she feels like a',
 'women like girls are so much prettier than men like',
 'all that people i want so say something what we',
 'tweet about it every second and if we do ya',
 'loved and he may hit the woman or use her',
 'preparate la puta que te re pari porque los viernes',
 'te pasa paleto si que eres piensa lo que te',
 'grew up in town where doles is from he is',
 'solo me lo dice a mi aunque todas la de',
 'as stupid for falling for it so we have to',
 'que lo cuides por que el es mi lo amo',
 'feministas por el mero hecho de que lo haga una',
 'just watched the long video amp no it does not',
 'el deporte pero eso no quita lo basura que fue',
 'stage in life dont feel the need to prove myself',
 'tu crush o tirndole la caa as con todas dios',
 'was handing me his receipt only to playfully yank it',
 'single and see the same women show him no interest']
```

**Figura 4.5:** Representación de los veinte 10-gramas más frecuentes en el conjunto de entrenamiento test.

```
['refugees could not enter the see the link and excerpt',
 'debate thank you donald sir bruce and think about making',
 'structure which a superficial theoretically untrained observer blinded an immediate',
 'de que la intervencin norteamericana en el conflicto hubiera sido',
 'faltado las filtraciones reveladoras sobre el particularen el seguimiento de',
 'merchant named david isaacs born in frankfurt germany in immigrated',
 'their genetic desires but what will come at a cost',
 'jews foods they say diabolical whichll be installed in afghanistan',
 'table nobody has consent to effeminated sodomize nor homosexualize not',
 'golpe de cocana en bolivia en junio de orquestado por',
 'marroqu da una paliza a una menor de aos de',
 'en directo sin disimulo su indignacin por el fallo del',
 'and white girl deadlocksim sick of seeing girls who have',
 'es que definitivamente son obras que negativamente hablando no tienen',
 'wear shirts did get a thumbs up from an elderly',
 'chica la turba sigue machacando en las calles medios argumentando',
 'you that its a biggest democracy and we have a',
 'banned from the cenotaph on certain days they have totally',
 'detenido en murcia por propinar una paliza encerrar en un',
 'me produce casi tanta repugnancia como la que me sobreviene']
```

**Figura 4.6:** Representación de los veinte 10-gramas más frecuentes en el conjunto de test.

### 4.2.7. Reconocimiento de emociones

Después de analizar las declaraciones realizadas en los resúmenes de las competiciones *EXIST* de 2021 [18] y 2022 [19], me percaté de la **dificultad que podría ocurrir en la identificación automática de ciertos**

**tipos de documentos sexistas** debido a su contenido no tan explícito como otras clases más violentas. Para intentar refutar esta teoría y tenerla en consideración para el futuro modelado, se ha planteado la integración de un mecanismo con el que reconocer la emoción intrínseca a un texto. Si realmente existe un subconjunto de muestras positivas que puede desencadenar una acumulación importante de errores, este fenómeno debería ser visible en sistemas inteligentes capaces de analizar textos aunque con objetivos diferentes. Para llevar a cabo esta tarea se ha empleado el transformer pre-entrenado T5 de Google disponible en los repositorios de Hugging Face [16]. Este estudio se divide en dos principales secciones, por un lado se calculan una serie de métricas globales acerca del número de textos pertenecientes a cada emoción disponible, mientras que por otro se realiza el mismo conteo en función de las clases sexistas que aparecen en la columna *task2*, de modo que posteriormente se puedan inspeccionar visualmente los documentos para reconocer las posibles cualidades de sus contenidos que han propiciado la asociación de una emoción u otra. Comenzando por el primer propósito en la Tabla 2.3 se encuentran el número de documentos sexistas y no sexistas pertenecientes a cada emoción dentro de los conjuntos de entrenamiento y test. Cláramente se aprecia que la **emoción con una mayor población es *joy***, lo que resulta sorprendente en ambas clases aunque especialmente en la positiva. Visualizando el contenido de los ejemplos categorizados como tal he podido observar que una posible teoría explicativa reside en el uso de **terminología positiva aunque sus significados son profundamente irónicos, negativos y sexistas**. Con este hecho podemos intuir que probablemente se trate de documentos complejos de analizar y comprender por modelos automáticos, por muy avanzadas que sean sus arquitecturas. Las tres siguientes emociones relacionadas con la ira, el miedo y la tristeza se caracterizan por disponer de textos con conceptos y expresiones sumamente violentos relativos a la reivindación de la lucha contra el sexismo, noticias sobre la represión femenina y el rechazo a las mujeres transgénero. Finalmente destaca la emoción *love* que lejos de simbolizar ningún tipo de amor, recoge a todos aquellos documentos asociados a agresiones sexuales, cosificación sexual del cuerpo de la mujer, vejaciones, etc.

La segunda subsección de este análisis consiste en replicar el mismo procedimiento de detección de emociones sobre ambos datasets aunque en esta ocasión se efectua un conteo en función de las categorías sexistas. Así en la Tabla 2.4 se aprecia que las dos emociones mayoritarias son *anger* y *joy* por lo que de nuevo se observa la tendencia descubierta anteriormente con las mismas características en los documentos clasificados como tales. En particular destaca que las categorías sexistas *objectification* y *stereotyping-dominance* situen en primera posición a una emoción que representa alegría. Tras una inspección visual del contenido de su población, se ha podido concluir que los tópicos principales tratan sobre la vestimenta femenina sexista y

**Tabla 4.3:** Tabla con el número de documentos pertenecientes a cada emoción detectada.

| Conjunto de datos | Emoción | Textos sexistas | Textos no sexistas |
|---|---|---|---|
| entrenamiento | joy | 1220 | 1468 |
| | anger | 1116 | 1081 |
| | fear | 729 | 751 |
| | sadness | 240 | 238 |
| | surprise | 12 | 31 |
| | love | 60 | 30 |
| | joy | 851 | 733 |
| test | anger | 851 | 622 |
| | fear | 466 | 427 |
| | sadness | 174 | 137 |
| | surprise | 16 | 19 |
| | love | 41 | 19 |

la cosificación sexual del cuerpo, mientras que en la segunda clase se encuentran **reivindicaciones y reproches por parte del colectivo femenino contra el masculino** por ser los causantes de sus inseguridades, desconfianzas, desigualdades laborales y salariales y acoso sexual. A pesar de que sus significados son profundamente negativos mayormente se componen de terminología y expresiones positivas que, de nuevo, confunden al modelo provocando su categorización dentro de una emoción positiva como es *joy*.

**Tabla 4.4:** Tabla con el número de documentos pertenecientes a cada emoción detectada por categoría sexista en el conjunto de entrenamiento.

| Categoría sexista | Emoción | Textos train | Textos test |
|---|---|---|---|
| ideological-inequality | anger | 325 | 258 |
| | joy | 303 | 189 |
| | fear | 168 | 126 |
| misogyny-non-sexual-violence | anger | 249 | 185 |
| | joy | 216 | 131 |
| | fear | 176 | 116 |
| objectification | joy | 214 | 133 |
| | anger | 140 | 88 |
| | fear | 104 | 65 |
| sexual-violence | anger | 215 | 196 |
| | joy | 159 | 92 |
| | fear | 85 | 75 |
| stereotyping-dominance | joy | 328 | 188 |
| | anger | 196 | 124 |
| | fear | 187 | 84 |

## 4.3. Conclusiones finales

A continuación se resumen las principales características acerca del proceso de recopilación de datos así como las deducciones más relevantes alcanzadas durante los diversos estudios exploratorios de ambos conjuntos.

- Si bien las dos fuentes de información utilizadas han sido dos redes sociales conocidas como Twitter y Gab, **el conjunto de entrenamiento únicamente se encuentra compuesto por tweets** mientras que el de test sí que contiene documentos de sendos medios. Esta discrepancia puede influir en el rendimiento de los clasificadores en tanto en cuanto el contenido disponible en cada plataforma sea considerablemente distinto por su formato, redacción y temáticas tratadas.

- Durante la recolección de información se consideraron diversas técnicas para **mitigar los posibles sesgos conceptuales, temporales y de usuarios** que podrían ser introducidos en los datasets consiguiendo igualar el número de muestras por idiomas, por clases y eliminando información personal de los autores de los documentos.

- En los análisis de frecuencias de caracteres se ha podido concluir que es **fundamental la aplicación de distintos tratamientos de textos capaces de filtrar contenido sin información útil** reduciendo tanto el tamaño de los documentos, el ruido existente en los datos así como disminuyendo el volumen de recursos necesarios para trabajar con ellos.

- A pesar de que en las nubes de palabras no se han podido detectar tópicos con suficiente nitidez como para conocer qué temas específicos son tratados en los mensajes interceptados, sí se ha descubierto que poseen una **elevada riqueza lingüística** debido a la variedad de terminología que han usado los usuarios. Asimismo han aparecido con una altísima frecuencia de aparición verbos con los que habitualmente se describen las capacidades y actitudes de las personas, como ser, gustar, mirar y saber, mientras que parece existir una **tendencia a la generalización** por la considerable frecuencia de conceptos generalistas como la palabra *people*.

- Los análisis de N-gramas y detección de emociones han conllevado multitud de inspecciones visuales del contenido de algunos documentos que han permitido la clarificación de las posibles cuestiones discutidas en los mensajes coleccionados con temas como el rechazo, la cosificación y la enalteración de la violencia y discriminación contra las mujeres. Desafortunadamente existe un amplio volumen de textos compuestos por terminología y expresiones gramaticalmente positivas pero con un

nivel elevado de **ironía y significados ocultos**. Estas muestras han generado una terrible confusión en el *transformer* preentrenado en la identificación de emociones categorizando masivamente ejemplos sexistas dentro de un sentimiento positivo. Por lo que **presumiblemente este mismo fenómeno volverá a estar presente en el modelado** del conjunto de datos. En particular, estas muestras se encuentran principalmente concentrados en tres categorías sexistas que se posicionan como las más complicadas de detectar correctamente por modelos automáticos.

# Capítulo 5

# Construcción de modelos clásicos

Una vez se conocen los aspectos más relevantes, a continuación se ha llevado a cabo la fase de modelado en la que se han utilizado, en primer lugar, **algoritmos clásicos de Aprendizaje Automático en combinación con distintas técnicas de codificación de textos**. Su razón de ser radica en aplicar el principio básico de la Navaja de Occam con el que intentar proporcionar una solución lo más sencilla posible. Si bien tras revisar los breves resúmenes disponibles de cada competición ya conocía que estos métodos no eran suficientes para aportar un modelo competitivo para este problema, pensé que podrían ayudarme a concentrar mis esfuerzos en estudiar y comprender el funcionamiento y las capacidades de diversos métodos de codificación de documentos. Por lo tanto en este primer capítulo de modelado se encuentran las métricas de validación seleccionadas para la evaluación común a los diversos sistemas inteligentes producidos, las secciones referentes a los distintos tratamientos de documentos experimentados con los que procesar y preparar los conjuntos de datos, así como los análisis de resultados generados como consecuencia del ajuste de los parámetros pertinentes de cada arquitectura probada.

## 5.1. Métricas de validación

Durante mi aprendizaje en el área de la Ciencia de Datos he podido estudiar y emplear una amplia gama de métricas de validación con las que evaluar la bondad de distintos tipos de modelos. La más popularmente utilizada es conocida como **accuracy o tasa de aciertos**, siendo definida como el porcentaje de muestras correctamente clasificadadas tomando en consideración todas las categorías disponibles dentro de un conjunto de datos

particular. Dado que el problema que se aborda para la detección de textos sexistas es de naturaleza binaria, es necesario establecer un umbral con el que definir la etiqueta que se le debe asignar a cada ejemplo dependiendo de la probabilidad de pertenencia predicha de cada clase. Si bien el valor más común es 0.5 para datasets balanceados, como es nuestro caso, el rendimiento del modelo que demuestra esta métrica es totalmente dependiente de este parámetro que es fijado por un usuario, por lo tanto se trata de una **evaluación parcialmente subjetiva**. Por esta razón, entre otras, es aconsejable utilizar una segunda medida adicional con la que verificar el comportamiento de los sistemas inteligentes. En mi caso particular he escogido como segunda métrica de validación el **área bajo la curva ROC** por las varias ventajas que entraña su uso, siendo en primer lugar totalmente **independiente del usuario** puesto que no tiene como requisito la especificación de ningún parámetro. Su explicación reside en que la propia técnica es capaz de experimentar con diferentes umbrales dentro de un intervalo [0.5, 1] sobre un mismo modelo con el propósito de generar un informe estadístico más preciso acerca de su rendimiento, considerando las tasas de verdaderos y falsos positivos. Así su interpretación puede ayudar al diseño de una estimación de la capacidad de generalización que tiene un clasificador con respeto a la característica de un modelo básico.

## 5.2. Procesamiento de textos general

Existen una amplia variedad de técnicas de tratamiento de documentos que son aplicadas prácticamente de forma generalizada a cualquier tipo de problema de clasificación basado en texto. Mientras que algunas han resultado tremendamente satisfactorias en relación a la cantidad de datos útiles considerados y al rendimiento de los modelos, otras han sido descartadas por motivos de diferente índole. Como consecuencia a continuación se detallan los procedimientos experimentados sobre los conjuntos tanto de entrenamiento como de test, el comportamiento esperado y recibido, así como si han sido integradas o no en este proyecto durante la construcción de todos los clasificadores engendrados.

- Se han eliminado los **enlaces web** puesto que no aportan ninguna información relevante para solventar el problema de detección de textos sexistas.

- Del mismo modo se suprimen las **menciones de usuarios** porque no se dispone de información sobre ellos que pueda ser combinada con los documentos que han redactado para obtener algún beneficio en el modelado de datos.

- Adicionalmente se han borrado los **caracteres especiales** y signos de puntuación ya que tampoco se caracterizan por proporcionar ninguna aportación de calidad.

- Se han transformado los caracteres restantes a **minúsculas** con el propósito de normalizar el formato de todos los elementos que componen los textos involucrados en los conjuntos de entrenamiento y test.

- Si bien la técnica stemming no parecía tener sentido dentro del contexto de este proyecto puesto que dificultaría la codificación de los términos al emplear embeddings preentrenados, sí que se han realizado diversos ensayos con el proceso de **lematización**. Si bien se especulaba con que facilitaría la conversión de los textos a representaciones numéricas al simplificar al máximo toda la terminología incrementado el volumen de información codificada que se pudiese suministrar a los modelos, apenas ha conseguido igualar el rendimiento de los experimentos en los que no se integraba la lematización a costa de incrementar considerablemente el número de recursos computacionales y temporales para llevarla a cabo. Como consecuencia de su pésimo aporte y beneficio, este tratamiento no ha sido finalmente incluido dentro del procesamiento de textos común a todos los modelos.

- Es popularmente conocido que los escritos procedentes de redes sociales son altamente caracterizados por sus faltas de ortografía y errores gramaticales que podrían provocar la pérdida de información valiosa decrementando así el rendimiento de los modelos. Como solución se han efectuado distintos ensayos con diversas librerías de Python con el propósito de **detectar y corregir aquellos términos incorrectamente redactados** tanto en español como en inglés. La hipótesis inicial de su beneficio residía en mejorar la capacidad de predicción de los modelos gracias a la codificación de un mayor volumen de palabras que así ayudasen a la extracción y estudio de los patrones requeridos para la identificación de textos sexistas y no sexistas. No obstante, apenas se ha logrado identificar un 9.43 % de conceptos erróneos de los que solamente un 7.61 % han sido corregidos, por lo que como era de esperar, no ha supuesto una diferencia suficientemente significativa como para justificar el aumento estrepitoso de recursos energéticos requeridos, por lo que finalmente no ha sido integrada dentro de este conjunto general de técnicas para la preparación y limpieza de los textos.

## 5.3. Modelos de Regresión Logística

La Regresión Logística es un algoritmo estadístico cuyo objetivo consiste en aproximar una función matemática acotada en el intervalo [0, 1] para estimar las probabilidades de pertenencia de cada muestra a las distintas clases existentes. Ha sido seleccionado como punto inicial de esta fase de modelado por su posicionamiento como **mejor algoritmo clásico en las competiciones** de 2021 y 2022, proporcionando métricas de validación más altas que el resto de métodos considerados. Puesto que la Regresión Logística no integra técnicas de codificación propias, se ha probado con el subconjunto de aquellas más popularmente extendidas como la bolsa de palabras, TF-IDF y word embeddings, detalladas en capítulos anteriores. Como consecuencia de la multilingualidad de los documentos tanto de entrenamiento como de test, y motivado por la presunción de que la población en español parece ser más representativa según las declaraciones efectuadas en las dos ediciones de las competiciones EXIST, se ha optado por **entrenar un modelo individual por idioma** en lugar de construir uno generalizado para ambos casos.

### 5.3.1. Procedimiento de experimentación

A continuación se pretende detallar cada una de las etapas diseñadas por las cuales se han podido efectuar los distintos ensayos que han permitido obtener diferentes modelos de Regresión Logística.

- **Conjuntos de datos**. Tanto los documentos pertenecientes al dataset de entrenamiento como los de test han sufrido un tratamiento compuesto por las técnicas explicadas en la primera sección de este capítulo, mientras que su codificación ha sido guiada por una variedad de transformaciones tales como la creación de bolsas de palabras, el cálculo de índices TF-IDF, el entrenamiento de word embeddings propios y el uso de embeddings preentrenados.

- **Diferenciación por idiomas**. Ya que los textos almacenados en ambos datasets se encuentran escritos en inglés o español, la aproximación recomendada en los resúmenes de las competiciones *EXIST* [18] [19] consiste en elaborar un modelado individual por idioma, puesto que parece que los documentos españoles disponen de una mejor representatividad y repercute en el rendimiento de los clasificadores. Asimismo, se trata de una práctica habitual motivada por la existencia de modelos inteligentes avanzados que se encuentran preentrenados en base a un corpus dedicado a un único lenguaje. Como consecuencia, se han entrenado dos clasificadores por cada técnica de codificación empleada compartiendo la configuración del algoritmo puesto que no se han ob-

servado diferencias significativas entre lenguajes al utilizar Regresión Logística.

- **Ajuste de parámetros**. Debido a la presencia de un fuerte fenómeno de sobreaprendizaje, el parámetro que se ha debido ajustar por cada método de codificación ha sido la **regularización**, experimentando tanto con L1, L2 como con una combinación de ambas, además de ensayar con distintos valores de penalización con el fin de encontrar un equilibrio con el que minimizar los efectos del sobreaprendizaje sin ser demasiado restrictivos en la actualización de la función interna del modelo.

- **Resultados determinísticos**. Puesto que los modelos generados con Regresión Logística proporcionan los mismos resultados en diferentes ejecuciones, únicamente se ha realizado una iteración por cada uno de los parámetros que se han deseado ajustar para maximizar su rendimiento.

### 5.3.2. Codificación mediante bolsas de palabras

El primer experimento se caracteriza por codificar los ejemplos textuales de entrenamiento y test como dos bolsas de palabras independientes para permitir su introducción a un algoritmo de Regresión Logística con el que construir un clasificador por idioma. Para esta primera tanda de pruebas ha sido vital la inclusión de una **fortísima regularización estableciendo un valor muy cercano a cero** y siendo L1 aquella que ha conseguido unas métricas de evaluación más elevadas. En caso de no aplicarla la capacidad de predicción de los clasificadores decrementa considerablemente apareciendo uno de los fenómenos más comunes cuando se dispone de un escaso volumen de datos: el sobreaprendizaje, ya que se ha observado una diferencia desmesurada entre la tasa de aciertos en entrenamiento con respecto a la validación mediante el conjunto de test. En la Tabla 5.1 se presenta la evaluación de ambos modelos en la que se observa que ningún clasificador ha alcanzado un accuracy o un área bajo la curva ROC de más del 70 % en test. Estos valores indican que sus capacidades de generalización no son suficientes para postularse como soluciones competentes en este problema de clasificación. Adicionalmente existe una **elevada concentración en la tasa de falsos negativos**, es decir, textos sexistas que no han sido detectados como tal por lo que parece que los sistemas construidos en este primer ensayo no han aprendido cuáles son las cualidades destacables de la clase positiva.

Con el propósito de comprender algo más los fenómenos que explican los pésimos resultados obtenidos en el conjunto de test, se han elaborado diversos estudios en los que se persiguen una variedad de objetivos diferentes. El primero consiste en descubrir los intervalos de confianza bajo los que se

**Tabla 5.1:** Tabla con la evaluación de modelos de Regresión Logística codificando los textos con bolsas de palabras sobre los conjuntos de entrenamiento y test.

| Datos | Idioma | Accuracy | AUC | FN | FP |
|---|---|---|---|---|---|
| entrenamiento | inglés | 0.772 | 0.769 | - | - |
| | español | 0.768 | 0.768 | - | - |
| test | inglés | 0.700 | 0.705 | 449 | 213 |
| | español | 0.682 | 0.687 | 487 | 199 |

ha sucedido la clasificación errónea de los falsos negativos y positivos. Para un mayor detalle se han establecido seis rangos de valores asigándoles una etiqueta descriptiva acerca de la confianza con la que los modelos han identificado erróneamente algunos ejemplos del conjunto de test. Si bien en la Tabla 5.2 únicamente aparecen los valores obtenidos tras aplicar este análisis sobre el clasificador inglés, se han encontrado los mismos fenómenos y teorías explicativas para el modelo español por lo que las conclusiones que se detallan a continuación son aplicables a ambos sistemas. Tal y como se aprecia en la susodicha tabla tanto los falsos negativos como los positivos se sumergen dentro de los rangos medios y altos, lo que resulta profundamente negativo puesto que simboliza que el clasificador inglés **ha cometido estos fallos con una confianza media-alta**, indicando que no dispone del conocimiento requerido para diferenciar los textos sexistas de los que no lo son.

**Tabla 5.2:** Tabla con el número de falsos negativos y positivos procedentes de modelos de Regresión Logística y bolsas de palabras por cada intervalo de confianza.

| Intervalo | FN | FP |
|---|---|---|
| Muy bajo [0.0, 0.2) | 0 | 0 |
| Bajo [0.2, 0.4) | 0 | 0 |
| Medio [0.4, 0.6) | 163 | 94 |
| Alto [0.6, 0.8) | 260 | 80 |
| Muy alto [0.8, 1.0] | 26 | 39 |

El segundo estudio confeccionado consiste en identificar la existencia de una posible relación entre los falsos negativos y las categorías sexistas a las que pertenecen. Siguiendo el mismo procedimiento anterior, en la Tabla 5.3 se muestran los resultados calculados para el clasificador inglés, siendo similares a los del modelo en español por lo que las conclusiones obtenidas son igualmente válidas. Y es que se aprecia que las tres categorías sexistas que ocupan el podio del ranking calculado son aquellas cuyos términos y expresiones no son extremadamente violentos y que por tanto **su contenido pueden estar causando la misma confusión** a los modelos de clasificación que al detector de emociones, tal y como se pudo comprobar en el

análisis exploratorio de datos.

**Tabla 5.3:** Tabla con el número de falsos negativos procedentes de modelos de Regresión Logística y bolsas de palabras por categoría sexista.

| Clase sexista | Nº de FN |
|---|---|
| ideological-inequality | 133 |
| stereotyping-dominance | 102 |
| misogyny-non-sexual-violence | 92 |
| sexual-violence | 76 |
| objectification | 46 |

### 5.3.3. Codificación mediante TF-IDF

En la segunda experimentación se ha establecido como técnica de codificación de documentos el cálculo del índice TF-IDF tanto para el conjunto de entrenamiento como para el de test. A diferencia de la situación anterior, en esta ronda de pruebas se ha aplicado una regularización L2 con un valor cercano a 1 lo que produce un efecto ligeramente menor siendo esta configuración más permisiva con la actualización de los sistemas. Con penalizaciones superiores e inferiores se ha observado un notorio sobreajuste en sendos modelos que perjudicaba considerablemente su rendimiento, particularmente en el clasificador orientado a documentos en español. No obstante este fenómeno sigue presente en los resultados que se pueden visualizar en la Tabla 5.4 debido a la **gran diferencia entre las métricas de validación del conjunto de entrenamiento en comparación con las de test** para ambos idiomas. Pese a modificar la técnica de codificación de documentos y la configuración del algoritmo de Regresión Logística, la tasa de aciertos o accuracy continúa sin alcanzar el 70 % por lo que como era de esperar, existe una importante cifra de falsos negativos y positivos. Así esta batería de pruebas tampoco ha sido capaz de proporcionar ninguna solución razonable a este problema de clasificación.

**Tabla 5.4:** Tabla con la evaluación de modelos de Regresión Logística codificando los textos con bolsas de palabras sobre los conjuntos de entrenamiento y test.

| Datos | Idioma | Accuracy | AUC | FN | FP |
|---|---|---|---|---|---|
| entrenamiento | inglés | 0.877 | 0.876 | - | - |
| | español | 0.970 | 0.970 | - | - |
| test | inglés | 0.691 | 0.694 | 427 | 255 |
| | español | 0.678 | 0.681 | 450 | 246 |

Con el mismo fin de comprender qué razones pueden esclarecer el deplorable comportamiento de los sistemas inteligentes validados sobre el conjunto

de test, se han replicado los dos análisis de resultados obteniendo para esta técnica de codificación unos valores prácticamente similares a los visualizados anteriormente. Por lo tanto la tendencia observada de nuevo se asemeja cuantiosamente al caso previo en el que el sistema inteligente parece estar **muy seguro acerca de la categorización errónea** de muestras desconocidas en la fase de validación. Este hecho significa que la segunda tanda de modelos obtenidos se caracterizan por poseer graves deficiencias en la información extraída para el reconocimiento de documentos sexistas. En referencia a la distribución de falsos negativos en función de las categorías sexistas a las que pertenecen también se ha reflejado un ranking aproximadamente igual al generado en la experimentación anterior, mostrando cómo se sitúan en los tres primeros puestos los **tópicos sexistas más complicados de detectar** por su elevado nivel de ironía que ha confundido tanto a los modelos de clasificación como al sistema de reconocimiento de emociones, puesto que existe una amplia mayoría de documentos bajo la emoción *joy* por su composición de terminología y expresiones positivas aunque con significados profundamente negativos y contrarios a la igualdad de género.

### 5.3.4. Codificación mediante embeddings propios

En la tercera ronda de experimentos se ha decidido entrenar embeddings propios de los conjuntos de datos *EXIST* **a nivel de palabra con modelos Word2Vec y de documentos empleando sistemas Doc2Vec**. Para su creación se han efectuado diversas experimentaciones con los parámetros más relevantes con el propósito de adaptar los sistemas de generación a los conjuntos de textos disponibles.

- El **tamaño de los vectores** ha sido establecido a un valor de cien elementos para incrementar el tamaño del vocabulario de los embeddings con los que, a su vez, aumentar la precisión de la codificación de los términos situados en ambas series de documentos.

- La **frecuencia mínima de aparición** que determina la inclusión de un concepto dentro del vocabulario que se produce también resulta un parámetro sumamente importante ya que si es muy elevado será muy restrictivo con su composición, y por ende disminuye el porcentaje de codificación de los textos, mientras que si es ínfimo puede introducir mucho ruido y perjudicar la calidad de las representaciones numéricas.

- El **algoritmo** que internamente especifica el procedimiento a seguir para predecir los embeddings es *CBWO* puesto que proporciona modelos con métricas de validación un 5 % mayores con respecto al rendimiento de su alternativa conocida como *Skip-Gram*, por lo tanto

la aproximación más beneficiosa consiste en predecir un término en función de un contexto determinado.

- El **número de iteraciones** es determinante puesto que a mayor número de repeticiones aumenta la presencia del sobreajuste de los sistemas inteligentes al existir una considerable diferencia entre las métricas de validación del conjunto de entrenamiento y de test. Sin embargo, si no se invierte el tiempo suficiente en su construcción el modelo resultante podría no disponer de la información debida para realizar una codificación de calidad.

- La **generación de un modelo diferente por conjunto de datos** ha demostrado ser mucho más beneficioso al suministrar unas métricas de validación considerablemente superiores a su alternativa consistente en configurar un único modelo generador orientado a los dos datasets. Una posible teoría explicativa reside en que tal y como se ha descubierto durante el análisis exploratorio de datos, el dataset de entrenamiento parece estar compuesto por un mayor número de opiniones de usuarios, mientras que el conjunto de test dispone de una población prácticamente repleta de noticias, por lo que la terminología y expresiones empleadas son notablemente diferentes.

En la Tabla 5.5 y 5.6 se muestran los cálculos procedentes de la evaluación de los modelos entrenados con Regresión Logística y word embeddings propios mediante sistemas basados en palabras (Word2Vec) y en documentos (Doc2Vec), respectivamente. Desafortunadamente, en la Tabla 5.5 los resultados son terriblemente peores que en los ensayos anteriores con razón de que **ambos clasificadores están completamente sesgados hacia la clase negativa**, categorizando aproximadamente la totalidad de los ejemplos analizados como no sexistas. Como consecuencia tanto el accuracy como el área bajo la curva ROC ni siquiera superan el 50 %, mostrando que los sistemas inteligentes creados con esta configuración no han aprendido nada útil como para poder ser soluciones de una calidad razonable para este problema. Mientras que en la Tabla 5.6 los valores visualizados no auguran ningún beneficio puesto que en este caso el **clasificador inglés resulta ser prácticamente aleatorio mientras que el español se encuentra sesgado hacia la clase positiva**, ya que todos los ejemplos han sido categorizados como sexistas. De nuevo sus comportamientos no difieren entre las dos metodologías empleadas para el entrenamiento de embeddings basados en los documentos disponibles.

Tras repetir los estudios de resultados basados en falsos negativos y positivos que se comenzaron realizando desde los primeros clasificadores más clásicos, en esta ocasión el volumen completo de **ejemplos erróneamente identificados se concentra en el intervalo de confianza moderado**.

**Tabla 5.5:** Tabla con la evaluación de modelos de Regresión Logística codificando los textos con embeddings entrenados con modelos Word2Vec sobre los conjuntos de entrenamiento y test.

| Datos | Idioma | Accuracy | AUC | FN | FP |
|---|---|---|---|---|---|
| entrenamiento | inglés | 0.551 | 0.540 | - | - |
| | español | 0.970 | 0.970 | - | - |
| test | inglés | 0.476 | 0.500 | 1158 | 0 |
| | español | 0.678 | 0.681 | 1154 | 4 |

**Tabla 5.6:** Tabla con la evaluación de modelos de Regresión Logística codificando los textos con embeddings entrenados con modelos Doc2Vec sobre los conjuntos de entrenamiento y test.

| Datos | Idioma | Accuracy | AUC | FN | FP |
|---|---|---|---|---|---|
| entrenamiento | inglés | 0.519 | 0.497 | - | - |
| | español | 0.522 | 0.516 | - | - |
| test | inglés | 0.521 | 0.520 | 517 | 540 |
| | español | 0.520 | 0.500 | 0 | 1037 |

No obstante, la interpretación de estos datos puede ser engañosa ya que tal y como se ha podido comprobar previamente, los modelos producidos no parecen haber aplicado lo que han aprendido porque sus comportamientos se han caracterizado por ser aleatorios o sesgados totalmente hacia una clase determinada. Finalmente el análisis sobre la combinación de los falsos negativos, en los modelos donde existan, con las categorías sexistas a las que pertenecen nuevamente demuestran una inclinación semejante a la conocida hasta el momento, predominando en los primeros puestos aquellas clases sexistas más difíciles de predecir debido a su contenido altamente confuso y lleno de ironía y dobles sentidos.

### 5.3.5. Codificación mediante embeddings pre-entrenados

El último conjunto de pruebas consiste en usar embeddings preentrenados basados tanto en el almacenamiento dentro de ficheros como producidos por modelos existentes que se pueden encontrar tanto online como en librerías de programación. Si bien para este proyecto se ha empleado la biblioteca denominada *gensim*, se han observado dos principales inconvenientes durante su uso. El primero de ellos reside en la imposibilidad de la aplicación de ciertos modelos debido a la generación de valores perdidos que no son tratables por el algoritmo de Regresión Logística. Mientras que en segundo lugar la totalidad de los sistemas generadores de embeddings únicamente se encuentran orientados a documentos en inglés, por lo que no ha sido posible su aplicación sobre las muestras españolas. Para trabajar con

ellas se debe acudir a modelos multilenguajes más avanzados o a la lectura y carga de embeddings en español que deben ser gestionados manualmente mediante la programación de funciones personalizadas a cada formato. Por lo tanto el último conjunto de ensayos se solamente se encuentra enfocado en los ejemplos ingleses. Como consecuencia en la Tabla 5.7 se presentan unos resultados similares a las anteriores experimentaciones con métricas de validación inferiores al 70 %. Aún utilizando embeddings prefabricados con una supuesta mayor calidad y capacidad de mejorar la codificación de los documentos, únicamente se ha conseguido **reducir el sobreajuste en los clasificadores** ya que la diferencia entre las métricas de validación de entrenamiento y test es sumamente menor a las percibidas en secciones anteriores. No obstante, se sigue mostrando una elevada concentración de falsos negativos además de un notable incremento de la cifra de falsos positivos.

**Tabla 5.7:** Tabla con la evaluación de modelos de Regresión Logística codificando los textos con embeddings preentrenados sobre los conjuntos de entrenamiento y test.

| Datos | Idioma | Accuracy | AUC | FN | FP |
|---|---|---|---|---|---|
| entrenamiento | inglés | 0.696 | 0.692 | - | - |
| test | inglés | 0.649 | 0.651 | 472 | 304 |

Replicando los análisis de resultados de secciones previas basados en intervalos de confianza y muestras erróneamente clasificadas, reiteradamente se aprecia un ranking idéntico a los observados anteriormente en el que los modelos demuestran una confianza elevada-moderada con la que han cometido los fallos contabilizados. Este fenómeno podría indicar que la corrección de su comportamiento se antoja casi inviable debido a la altísima seguridad que posee al identificar algunos documentos de sus respectivas clases contrarias. De igual modo al cruzar la información relativa con las categorías sexistas a las que pertenencen los falsos negativos también aparecen en los primeros puestos aquellas que contienen una terminología y expresiones más complicadas de detectar automáticamente y capaces de desorientar a los sistemas inteligentes generados.

### 5.3.6. Resumen y conclusiones

Esta sección le otorga el broche final a la amplia variedad de experimentaciones llevadas a cabo en torno al algoritmo de Regresión Logística, recapitulando las conclusiones alcanzadas y consideradas más relevantes así como recopilando en la Tabla 5.8 los valores de las métricas de validación que se han conseguido sobre el conjunto de test tras aplicar las mejores configuraciones encontradas para los documentos de español e inglés.

**Tabla 5.8:** Tabla resumen con las métricas de validación sobre test de las experimentaciones elaboradas con distintas técnicas de codificación y el uso de Regresión Logística.

| Codificación | Idioma | Accuracy | AUC | FN | FP |
|---|---|---|---|---|---|
| Bolsa de palabras | inglés | 0.700 | 0.705 | 449 | 213 |
| TF-IDF | inglés | 0.691 | 0.694 | 427 | 255 |
| Word2Vec | inglés | 0.476 | 0.500 | 1158 | 0 |
| Doc2Vec | inglés | 0.521 | 0.520 | 517 | 540 |
| preentrenados | inglés | 0.649 | 0.651 | 472 | 304 |
| Bolsa de palabras | español | 0.682 | 0.687 | 487 | 199 |
| TF-IDF | español | 0.678 | 0.681 | 450 | 246 |
| Word2Vec | español | 0.482 | 0.502 | 1114 | 4 |
| Doc2Vec | español | 0.520 | 0.500 | 0 | 1037 |

- A pesar de que el algoritmo de Regresión Logística es profundamente sencillo, el objetivo que se perseguía en esta sección consistía en probar diferentes métodos de codificación de textos para su estudio y comprensión. Por lo tanto, aunque los resultados no hayan sido beneficiosos para solucionar este problema de clasificación, es posible que con la combinación de otros algoritmos y configuraciones las métricas de validación sean más favorables.

- La **elevada tasa de falsos negativos y el sobreajuste son fenómenos práctiamente generalizados** a todos los sistemas inteligentes producidos con Regresión Logística. Ambos hechos demuestran la incapacidad por parte de este algoritmo de identificar y aprender los patrones característicos de los documentos sexistas, así como posibles problemas intrínsecos en los conjuntos de datos que puedan desestabilizar a los modelos entrenados. Se trata de una teoría que se deberá refutar con los análisis de las siguientes arquitecturas conforme se vaya ganando un mayor grado de complejidad.

- Tal y como se aprecia en la Tabla 5.8 han sido las técnicas de codificación más sencillas las que han conseguido fabricar los clasificadores de mayor calidad con respecto a las métricas de validación sobre test. Uno de los principales motivos es que realmente no han existido métodos competentes que pudieran superar sus resultados, puesto que el entrenamiento de embeddings personalizados a los conjuntos de datos no ha sido beneficioso por las cantidades ingentes de textos que serían necesarios para construir vocabularios más enriquecidos y capaces de codificar todo tipo de terminología. Sin embargo, sí que se ha visto el potencial existente detrás del uso de **embeddings prefabricados porque han conseguido reducir el nivel de sobreaprendizaje**.

Por lo tanto, se posiciona como una de las metodologías más popularmente empleadas en tareas de PLN en combinación con el uso de arquitecturas más complejas y orientadas a documentos.

- Si bien en los resúmenes de ambas competiciones se declara una mayor representación de los documentos en español que provocaban la obtención de sistemas inteligentes más precisos, con las configuraciones probadas y el algoritmo de Regresión Logística no se ha podido visualizar tal beneficio. De hecho en la Tabla 5.8 se aprecia un ligero empeoramiento de las métricas de validación para los clasificadores específicos de textos españoles, siendo caracterizados por una mayor tendencia a presentar un sesgo total hacia una determinada clase.

# Capítulo 6

# Construcción de modelos avanzados

Tras completar la fase anterior empleando un algoritmo clásico de Aprendizaje Automático junto con distintas técnicas de codificación de documentos, a continuación se incrementa la complejidad de los modelos adentrándome en los algoritmos de Aprendizaje Profundo cuya elección fue guiada en base a los detalles aportados por los equipos que consiguieron las mejores posiciones en el ranking de ambas competiciones. Estableciendo un hilo experimental con el que beneficiarse de los progresos logrados con los ensayos anteriores, en esta nueva etapa de modelado se emplean las métricas de validación y técnicas de tratamiento de datos anteriormente citadas. Esta tendencia será integrada en todas las rondas de pruebas con las distintas arquitecturas consideradas. Asimismo, la estructura planteada es idéntica a la del capítulo anterior introduciendo las diversas tipologías de redes neuronales, su hiperparametrización buscando las configuraciones más propicias para los objetivos planteados, ejemplos prácticos en los que se aplican dichas parametrizaciones para observar y analizar los resultados sobre modelos explícitos, además de una última sección de conclusiones y resúmenes acerca de las pesquisas encontradas para cada tipología de red.

## 6.1. Modelos LSTM

Continuamos detallando la siguiente modalidad de arquitectura empleada aumentando el nivel de complejidad con el propósito de conseguir mejores resultados que los hasta ahora conocidos tras el modelado efectuado con Regresión Logística. Para ello se ha hecho uso de la librería *Keras* por su facilidad de implementación, su diversidad y completitud de las características ofertadas, además de la experiencia fundamentada tras haber realizado

diversas prácticas dentro del propio máster. De nuevo, como este tipo de arquitecturas tampoco incluyen técnicas de codificación propias se ha experimentado con distintos conjuntos de embeddings preentrenados tratando de descubrir aquel que sea más propicio para los conjuntos de datos *EXIST*. Adicionalmente también se han ajustado los parámetros más relevantes efectuando diversos ensayos con los que comparar la bondad intrínseca a cada uno de los valores contemplados maximizando las métricas de validación detalladas al comienzo de este capítulo. Con el propósito de poder comparar el rendimiento de unos sistemas más avanzados con los clásicos analizados anteriormente y de comprobar si los documentos españoles son de mayor calidad como se afirma en las dos competiciones, de nuevo se sigue la misma estrategia de construir un clasificador diferente por idioma. Finalmente se integra un ejemplo práctico de cada lenguaje con el que analizar los resultados obtenidos en mayor profundidad apreciando las consecuencias asociadas a la aplicación de las mejores configuraciones encontradas durante la experimentación.

### 6.1.1. Lecciones aprendidas

En este apartado se pretende realzar las conclusiones extraídas durante las pruebas anteriormente elaboradas mediante el uso de la Regresión Logística y las distinas técnicas de codificación de documentos. El objetivo principal reside en **generar un punto inicial más avanzado** desde el que partir la experimentación con arquitecturas LSTM incorporando únicamente los procedimientos que han demostrado ser más ventajosos para la resolución del problema que se aborda en este proyecto. Mientras que el anterior algoritmo era de naturaleza púramente matemática y el presente representa un tipo específico de red neuronal, el único punto en común que tienen se fundamenta en el uso de **embeddings preentrenados como metodología de transformación de documentos** motivado por el enorme potencial demostrado al postularse como el único procedimiento capaz de reducir el grado de sobreaprendizaje. A diferencia de los restantes métodos de codificación, los embeddings disponen de la habilidad de respetar el orden de los términos que componen los documentos siendo estos reemplazados por sus correspondientes vectores numéricos, los cuales pueden ser similares entre sí dependiendo del grado de semejanza entre conceptos individuales.

### 6.1.2. Procedimiento de experimentación

A continuación en esta sección se introducen las configuraciones de datos, entrenamiento y validación contempladas así como la metodología relativa a la experimentación con arquitecturas LSTM.

- **Conjuntos de datos**. El procesamiento de los documentos de entrenamiento y test es idéntico al explicado en el comienzo del capítulo así como el efectuado en los ensayos de modelos de Regresión Logística. La principal diferencia reside en que únicamente se emplea una técnica de codificación, los embeddings preentrenados, aunque probando con distintos conjuntos entrenados sobre corpus diferentes con sus propias particularidades. Adicionalmente, al tratarse de una arquitectura más compleja que puede requerir su entrenamiento sobre un conjunto de datos más voluminoso, se ha investigado acerca de las tecnologías popularmente utilizadas para la generación de muestras sintéticas con las que poder ampliar el número de ejemplos de entrenamiento.

- **Diferenciación por idiomas**. De nuevo se selecciona la aproximación conocida hasta el momento consistente en modelar individualmente los textos pertenecientes a inglés y a español.

- **Ajuste de parámetros**. En primer lugar se ha considerado un rango razonable de tamaños del lote que probar con el propósito de ajustar el número de muestras que la arquitectura puede recibir intentando equilibrar la inversión de recursos necesaria para su construcción y el tiempo brindado para el estudio, extracción y aprendizaje de los patrones intrínsecos a los documentos. Por otro lado también se ha analizado el impacto producido por el incremento en el número de capas y neuronas comprobando si un modelo más complejo dispone de una capacidad de generalización mayor que un clasificador más sencillo.

- **Resultados no determinísticos**. A diferencia de la experimentación con Regresión Logística, se pueden obtener diferentes resultados en términos de las métricas de validación en distintas ejecuciones con una misma arquitectura LSTM. Como consecuencia a fin de efectuar análisis comparativos de una manera correcta y significativa entre los distintos valores considerados para cada parámetro mencionado anteriormente, se ha ideado un procedimiento de entrenamiento de treinta modelos bajo la misma configuración de datos y arquitectura con el que calcular la media aritmética tanto de accuracy como del área bajo la curva ROC y así permitir medir sus bondades eligiendo el más beneficioso para maximizar el rendimiento del sistema. De este modo las métricas de validación recopiladas en las siguientes tablas reflejan la acumulación de las evaluaciones realizadas en función de los treinta modelos generados por parámetro ajustado.

### 6.1.3. Ejemplo práctico en español

Una vez se han descubiertos los valores más propicios para los ajustes más importantes que se han ideado dentro del modelado con arquitecturas LSTM, los cuales pueden ser visualizados en el apéndice situado al final de esta memoria, a continuación se propone el siguiente ejemplo práctico en el que se aplican todas las pesquisas halladas sobre el conjunto de documentos españoles. A continuación se resumen la configuración de datos, entrenamiento y validación empleados a modo resumen de las experimentaciones detalladas previamente.

- **Codificación mediante embeddings preentrenados**. Tanto en el clasificador inglés como español el conjunto de embeddings preentrenados que se caracteriza por proveer las más elevadas métricas de validación ha sido el conocido como *Glove Twitter 27B 100d*, siendo válido en ambos idiomas debido a su contenido multilingual.
- **Tamaño del lote**. Ha sido en este parámetro en el que se ha encontrado la única discrepancia entre el modelado de los dos idiomas disponibles, donde para los textos en español se ha descubierto un aumento de aproximadamente un 1.5 % en las métricas de validación al **decrementar a dieciséis muestras** por lote en lugar de treinta y dos establecidas para los documentos ingleses.
- **Arquitectura**. La composición que ha proporcionado un mayor rendimiento está fundamentada en el uso de **tres capas ocultas unidireccionales de 128, 64 y 32 neuronas**, respectivamente. Si bien se ha escogido una composición ligeramente más compleja para este ejemplo práctico, su capacidad predictiva es bastante similar a una arquitectura básica de una sola capa oculta, aunque el número de recursos aumenta considerablemente con un impacto dependiente del entorno de ejecución elegido. Por contra no se ha integrado ningún mecanismo de *drop-out* ya que no se han apreciado beneficios algunos en la generalización de los clasificadores durante las distintas experimentaciones efectuadas.
- **Entrenamiento y validación**.
  - **Early Stopping**. Con un máximo de cien iteraciones por ensayo ha sido un requisito indispensable la adición de este mecanismo para reducir el sobreajuste más que conocido en la resolución del problema que aborda este proyecto. Para ello se ha configurado un alto en la construcción de un modelo si tras quince iteraciones la métrica de validación *accuracy* no ha mejorado en más de un 0.01 devolviendo los pesos del mejor sistema inteligente obtenido en base a dicha medida.

- **Función de pérdida**. Al tratarse de un problema de clasificación binario se ha optado por establecer la función clásica conocida como *binary crossentropy* por su adecuación a las características de las soluciones ideadas para la detección de textos sexistas.
- **Optimizador**. Al igual que en el parámetro anterior también se ha escogido el optimizador *Adam* que se encuentra por defecto debido a su perfecto equilibrio entre el rendimiento y la inversión de recursos computacionales.
- **Porcentaje de validación**. Con el propósito de conocer la evolución en la construcción del modelo se ha reservado un 20 % del total del conjunto de entrenamiento para realizar una validación tras cada iteración.

Los resultados obtenidos demuestran que sobre el conjunto de entrenamiento el modelo producido alcanza un 74.04 % de accuracy y 80.46 % de área bajo la curva ROC, mientras que la **validación sobre test responde a 71.59 % de accuracy y 77.64 % de área bajo la curva ROC**. Si bien se trata de unos valores superiores a los conocidos hasta el momento, en la Figura 6.1 se aprecia una preocupante concentración de falsos negativos también percibida en algunos sistemas de Regresión Logística con aproximadamente cuatrocientos textos sexistas no detectados correctamente. Asimismo el volumen de falsos positivos también es notable con más de doscientos documentos no sexistas categorizados como positivos. Por lo tanto, dadas estas métricas aún con la mejor configuración encontrada el clasificador resultante no dispone de una suficiente capacidad predictiva como para posicionarse como una solución competente de razonable calidad para la resolución de este problema.

Analizando los falsos negativos y positivos en la Tabla 6.1 se aprecia un cambio de tendencia, en comparación con la conocida hasta el momento, puesto que mientras la confianza con la que ha categorizado documentos sexistas como negativos es muy baja por pertenecer a los primeros intervalos considerados, los textos no sexistas clasificados como positivos sí han sido identificados bajo intervalos de confianza mayores. Este fenómeno puede indicar que el propio **modelo es capaz de reconocer sus carencias de información al intentar etiquetar ejemplos como positivos**, con lo cual podría ser más sencillo corregir este comportamiento erróneo.

No obstante al cruzar la información relativa a los falsos negativos con sus correspondientes categorías sexistas a las que pertenecen el ranking visualizado es exactamente idéntico al descubierto con los modelos de Regresión Logística. Las primeras posiciones siguen estando protagonizadas por aquellas clases más complicadas de predecir automáticamente debido a su contenido basado en terminología y expresiones positivas pero con un

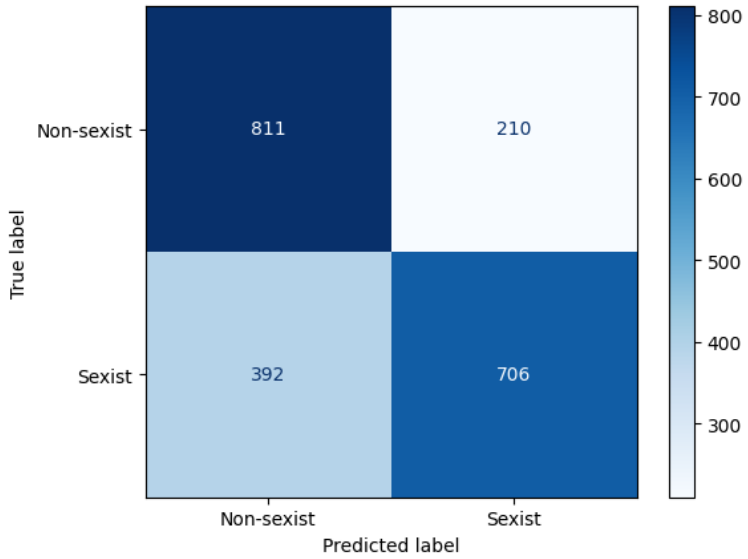


**Figura 6.1:** Matriz de confusión para la validación sobre test de un modelo LSTM específico de textos en español con la mejor configuración encontrada.

**Tabla 6.1:** Tabla con el número de falsos negativos y positivos procedentes del mejor modelo LSTM español por cada intervalo de confianza.

| Intervalo | FN | FP |
|---|---|---|
| Muy bajo [0.0, 0.2) | 78 | 0 |
| Bajo [0.2, 0.4) | 232 | 0 |
| Medio [0.4, 0.6) | 82 | 86 |
| Alto [0.6, 0.8) | 0 | 107 |
| Muy alto [0.8, 1.0] | 0 | 17 |

profundo significado irónico y negativo. De hecho más del 39 % de falsos negativos y más del 41 % de falsos positivos se les ha asignado la emoción *joy*, durante el análisis exploratorio realizado en el capítulo anterior, que expresa la existencia de un sentimiento positivo por sus características lingüísticas aunque para nada se asemeja a la realidad.

### 6.1.4. Resumen y conclusiones

Finalizando el estudio de arquitecturas LSTM unidireccionales a continuación se detallan las conclusiones conseguidas a partir de la multitud de pruebas efectuadas y narradas en secciones anteriores. En la Figura 6.2 se representa una gráfica de líneas que refleja la evolución de la media aritmética de las métricas de validación establecidas tras la selección de los valores más beneficiosos encontrados para cada uno de los parámetros ajustados.

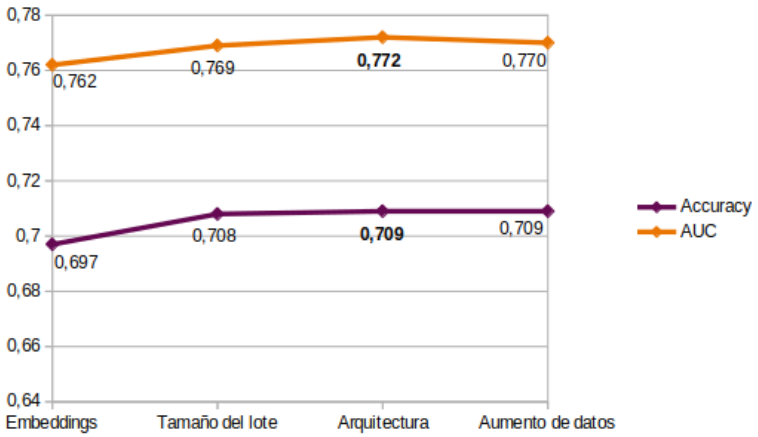


**Figura 6.2:** Gráfica de evolución de las métricas de validación durante el ajuste de los parámetros en modelos LSTM.

- Tal y como se refleja en el gráfico anterior las métricas iniciales se sitúan en el 60.7 % de accuracy y 76.2 % de AUC tras seleccionar el conjunto de embeddings más beneficioso para la codificación de los documentos almacenados en los datasets *EXIST* ya que dispone de un **vocabulario más voluminoso y enfocado en Twitter**, fuente de datos primaria de la que proceden los mencionados documentos.

- Se consigue incrementar ambas medidas aproximadamente un 1 % después del ajuste del tamaño del lote que es ligeramente mayor en el modelado de muestras inglesas mientras que ha debido de ser inferior en los ejemplos españoles para maximizar el rendimiento de los modelos específicos de este idioma. Este hecho pone de manifiesto la existencia de **diferencias significativas entre la representatividad detrás de los documentos ingleses con respecto a los españoles**. Su razón de ser parece residir en que estos últimos entrañan un mayor número de patrones a extraer y por ende es necesario reducir el volumen de muestras proporcionadas a los clasificadores para su correcto estudio y aprendizaje.

- A diferencia del parámetro anterior apenas se ha logrado una ligera mejoría al incrementar la complejidad de la arquitectura tras la integración de un mayor número de capas y neuronas. Como conclusión se puede extraer que un **modelo más complicado no siempre tiene que proporcionar un rendimiento mayor** que un clasificador simplista, por ello en este punto se enfatiza la importancia de comenzar con soluciones más sencillas antes de acudir a sistemas más avanzados que requieren una inversión de recursos temporales y computacionales bastante mayor.

- Finalmente tampoco se ha alcanzado un incremento en la capacidad de generalización del modelo LSTM al aumentar el conjunto de entrenamiento a partir de la generación de muestras sintéticas traduciendo los documentos ingleses a español y viceversa. Al contrario que el caso anterior, este fenómeno **no significa que dicha técnica no sea beneficiosa** en futuras arquitecturas con otras configuraciones, aunque claramente como hipótesis inicial se esperaba un mayor rendimiento al emplear un dataset de entrenamiento más voluminoso donde el clasificador pudiese estudiar más en detalle los patrones existentes que le permitiese una identificación más precisa.

## 6.2. Modelos BiLSTM

El objetivo de esta subsección consiste replicar la experimentación realizada en el subapartado anterior aunque integrando únicamente capas bidireccionales en lugar de unidireccionales, dando lugar a arquitecturas BiLSTM. Así se pretende investigar acerca de si resulta ventajoso la ampliación de la complejidad que se produce a nivel de capa al modificar el mecanismo que se encarga de analizar las muestras de datos proporcionadas durante la construcción del clasificador. Para esta tarea de nuevo se ha utilizado la librería *Keras* por los mismos motivos explicados anteriormente, además de los mejores valores encontrados para cada uno de los parámetros identificados en los ensayos previos. Al final de esta sección se incluye un ejemplo práctico orientado a la construcción de un sistema BiLSTM orientado a documentos ingleses donde se pueda analizar en profundidad los resultados generados como consecuencia de la adaptación de la arquitectura a capas ocultas bidireccionales.

### 6.2.1. Lecciones aprendidas

A continuación se resumen las pesquisas localizadas hasta el momento fruto de las experimentaciones que se han llevado a cabo con la idea de aplicar los beneficios a los futuros ensayos de esta modalidad, tomando en consideración el arduo trabajo realizado para únicamente explotar los puntos que difieren entre los modelados previos y los que se prueban con arquitecturas bidireccionales.

- **Codificación con embeddings**. En la construcción de sistemas inteligentes se seguirá empleando, tanto para español como para inglés, el fichero de embeddings *Glove Twitter 27B 100d* por haber demostrado ser el que mejor se adapta al vocabulario de los textos de entrenamiento y test al compartir la fuente de información de la que proceden

ambas entidades.

- **Tamaño del lote**. Si bien con las muestras en inglés es suficiente establecer treinta y dos ejemplos por lote, para los datos españoles se ha descubierto que su reducción a dieciséis proporciona una mayor capacidad de generalización al disponer de un mayor nivel de representatividad y detalle, según ha sido anunciado por las dos ediciones de las competiciones *EXIST*.

- **Arquitectura**. La principal conclusión asociada a este parámetro se fundamenta en que los rendimientos de modelos con una única capa oculta se caracterizan por ser muy cercanos a los clasificadores con hasta tres capas ocultas, aunque la inversión de recursos temporales y computacionales se eleva en mayor o menor consideración dependiendo del entorno de ejecución utilizado. Sin embargo merece la pena seguir ajustando su composición puesto que se presupone que el desenlace con sistemas bidireccionales podría ser distinto del visualizado hasta ahora.

- **Aumento del dataset de entrenamiento**. Tras generar muestras sintéticas para el conjunto de entrenamiento mediante el ensayo de una amplia gama de técnicas, se ha descubierto que con la configuración unidireccional aplicada no ha resultado ser una herramienta tan provechosa como se esperaba. Sin embargo, al igual que en el punto anterior su comportamiento se prevee que pueda diferir con capas ocultas bidireccionales.

### 6.2.2. Procedimiento de experimentación

Como el procedimiento seguido es exactamente idéntico al detallado en la sección anterior con arquitecturas LSTM unidireccionales, puede consultarlo en este enlace.

### 6.2.3. Comparación de arquitecturas undireccionales y bidireccionales

A diferencia de la tendencia percibida en la que prácticamente no existía ninguna diferencia entre el modelado de los documentos ingleses y españoles, en esta casuística sí que son plausibles algunas discrepancias que deben ser destacadas y por ende, comienza este análisis comparativo con los textos en inglés. En la Tabla 6.2 se recogen las métricas de validación realizadas sobre los conjuntos de entrenamiento y test que representan un conjunto de cuatro arquitecturas diferentes basadas en los mejores modelos obtenidos de la experimentación LSTM previa.

**Tabla 6.2:** Tabla con una comparativa de las métricas de validación de arquitecturas LSTM unidireccionales y bidireccionales durante el modelado de documentos ingleses.

| Arquitectura | Train acc | Train AUC | Test acc | Test AUC |
|---|---|---|---|---|
| UNIDIR 1 | 0.783 | 0.866 | 0.708 | 0.769 |
| BIDIR 1 | 0.783 | 0.864 | 0.705 | 0.769 |
| UNIDIR 2 | 0.786 | 0.868 | 0.708 | 0.770 |
| BIDIR 2 | 0.776 | 0.859 | 0.710 | 0.773 |
| UNIDIR 3 | 0.779 | 0.861 | 0.709 | 0.772 |
| BIDIR 3 | 0.776 | 0.856 | 0.708 | 0.771 |
| UNIDIR 4 | 0.813 | 0.879 | 0.709 | 0.770 |
| BIDIR 4 | 0.804 | 0.872 | 0.712 | 0.773 |

- Primera arquitectura. Tanto UNIDIR 1 como BIDIR 1 se componen de una **única capa oculta con 128 neuronas**, siendo unidireccional en el primer caso y bidireccional en el otro, respectivamente. El resto de la configuración se fundamenta en las lecciones aprendidas detalladas en este mismo capítulo. Tal y como se aprecia **no existen disparidades significantes entre las evaluaciones** efectuadas sobre entrenamiento y test, por lo que en este caso se concluye que no merece la inversión de recursos superior que implica la integración de mecanismos bidireccionales puesto que uno más simple es capaz de conseguir prácticamente los mismos resultados.

- Segunda arquitectura. Los modelos UNIDIR 2 y BIDIR 2 se caracterizan por disponer **dos capas ocultas de 128 neuronas cada una** unidireccionales y bidireccionales respectivamente, mientras que los restantes parámetros aplican a las lecciones aprendidas explicadas previamente. En este caso se observa una diferencia de aproximadamente el 1 % en las métricas de validación, por lo que se trata de un **comportamiento ligeramente superior por parte de la arquitectura bidireccional**. Dependiendo del objetivo que se persiga puede ser suficientemente significativo como para optar por esta composición más compleja, como podría ser dentro de un marco competitivo regido por el número de aciertos de los modelos.

- Tercera arquitectura. Los clasificadores denominados UNIDIR 3 y BIDIR 3 tienen un total de **tres capas ocultas de 128, 64 y 32 neuronas** unidireccionales y bidireccionales respectivamente. En este ensayo los valores reflejados difieren mínima aunque con una inclinación opuesta al anterior siendo la **arquitectura unidireccional la que proporciona un mayor rendimiento** con un menor coste de entrenamiento y validación.

- Cuarta arquitectura. Los modelos UNIDIR 4 y BIDIR 4 disponen de una arquitectura idéntica a la anterior con tres capas ocultas de 128, 64 y 32 neuronas con la excepción de haber **introducido muestras sintéticas en el conjunto de entrenamiento**, producidas a partir de la traducción tradicional de textos españoles a inglés, duplicando así el volumen. En este experimento si bien las métricas de entrenamiento son levemente superiores en el modelo unidireccional, es en la evaluación **sobre test donde se observa un mayor potencial por parte del sistema bidireccional** representando una cierta superioridad en la clasificación de ejemplos desconocidos.

En la Tabla 6.3 se representan los resultados suministrados tras replicar el estudio comparativo anterior aunque sobre el conjunto de documentos españoles. A excepción del primer par de arquitecturas donde los rendimientos mostrados son aproximadamente similares, en el resto se observa una **tendencia notoriamente a favor de los sistemas bidireccionales** al superar, en términos de las métricas de validación establecidas, hasta en casi un 2 % a las arquitecturas unidireccionales tanto en entrenamiento como en test. La principal teoría de este fenómeno puede residir en la confirmación de que las muestras españolas disponen de un volumen mayor de características y patrones, con respecto a los ejemplos ingleses, que requieren aumentar la complejidad de los mecanismos encargados de su análisis y extracción con el propósito de producir sistemas inteligentes más robustos y de mayor calidad para la identificación de textos sexistas y no sexistas.

**Tabla 6.3:** Tabla con una comparativa de las métricas de validación entre arquitecturas LSTM unidireccionales y bidireccionales durante el modelado de documentos españoles.

| Arquitectura | Train acc | Train AUC | Test acc | Test AUC |
|---|---|---|---|---|
| UNIDIR 1 | 0.801 | 0.874 | 0.694 | 0.760 |
| BIDIR 1 | 0.801 | 0.873 | 0.697 | 0.762 |
| UNIDIR 2 | 0.783 | 0.861 | 0.698 | 0.765 |
| BIDIR 2 | 0.807 | 0.881 | 0.691 | 0.757 |
| UNIDIR 3 | 0.774 | 0.845 | 0.701 | 0.768 |
| BIDIR 3 | 0.808 | 0.873 | 0.697 | 0.763 |
| UNIDIR 4 | 0.789 | 0.854 | 0.691 | 0.766 |
| BIDIR 4 | 0.797 | 0.864 | 0.687 | 0.763 |

### 6.2.4. Ejemplo práctico en inglés

Tras los distintos experimentos realizados con arquitecturas LSTM y su posterior comparación entre mecanismos unidireccionales y bidireccionales, a

continuación se presenta un ejemplo práctico orientado al uso de capas ocultas bidireccionales y a documentos en inglés. A continuación se establece un resumen de la configuración de datos, entrenamiento y validación aplicadas de los mejores parámetros encontrados durante este modelado concreto.

- **Codificación con embeddings**. Nuevamente se ha hecho uso del fichero con embeddings preentrenados multilenguaje conocido como *Glove Twitter 27B 100d* gracias a su interoperabilidad entre idiomas y a la aportación de las mejores métricas de validación.
- **Tamaño del lote**. Para los documentos ingleses es más que suficiente establecer este parámetro a treinta y dos ejemplos por lote para su correcta asimilación durante la construcción de un modelo unidireccional LSTM.
- **Arquitectura**. La composición que ha demostrado ser más beneficiosa para este idioma en particular se encuentran formada por **dos capas bidireccionales con 128 neuronas cada una**, ya que esta configuración ha guardado un buen equilibrio entre el rendimiento producido en los modelos y la inversión de recursos necesarios para su entrenamiento y validación.
- **Entrenamiento y validación**.
  - **Early Stopping**. Se proponen un máximo de cien iteraciones añadiendo este mecanismo con el que reducir el fenómeno del sobreajuste muy visible en las pruebas realizadas con la mayoría de algoritmos y arquitecturas. Se detiene la construcción del modelo tras quince iteraciones sin percibir una mejor de más de 0.01 en la métrica de *accuracy*, simulando el mismo enfoque competitivo del que procede el planteamiento y uso de los datos *EXIST*.
  - **Función de pérdida**. De nuevo se opta por una función de pérdida centrada en problemas de clasificación binarios como es la denominada *binary crossentropy* por su adecuación a las únicas dos clases disponibles en los datasets.
  - **Optimizador**. Este parámetro también se mantiene constante al posicionar el valor por defecto *Adam*.
  - **Porcentaje de validación**. Para conocer el rendimiento que va tomando el clasificador durante su fabricación se reserva un 20 % del total del conjunto de entrenamiento para efectuar una evaluación tras cada iteración finalizada.

Mientras que la evaluación sobre el conjunto de entrenamiento reside en un 77.77 % de accuracy y un 86.26 % de área bajo la curva ROC, en el

conjunto de **test únicamente se alcanza 70.41 % de accuracy y un 77.15 % de área bajo la curva ROC**. En la Figura 6.3 se aprecia la matriz de confusión generada sobre este último dataset que demuestra una tendencia similar a la conocida en el clasificador español, con una peligrosa tasa de falsos positivos que prácticamente roza los cuatrocientos documentos sexistas no identificados, al igual que una elevada cifra de falsos negativos.

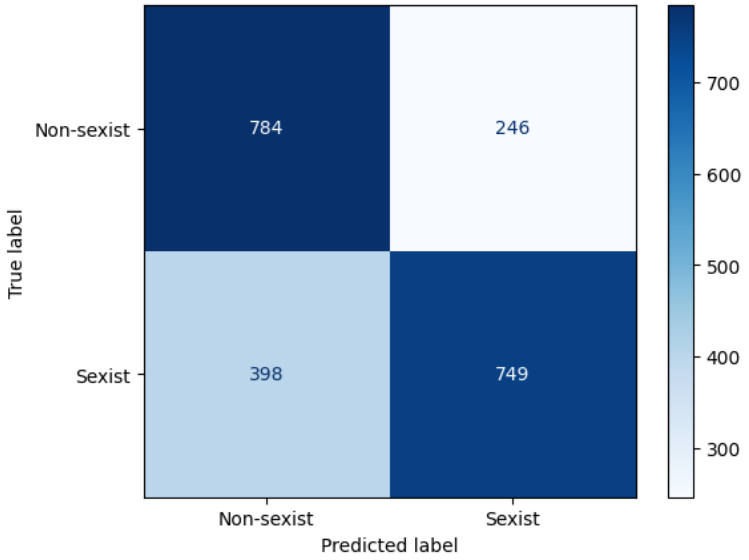


**Figura 6.3:** Matriz de confusión para la validación sobre test de un modelo BiLSTM específico de textos en inglés con la mejor configuración encontrada.

Analizando los falsos negativos y positivos se han observado unas conclusiones idénticas a las visualizadas durante el modelado unidireccional con textos españoles, ya que de nuevo el nuevo clasificador inglés demuestra una confianza ínfima al detectar erróneamente documentos no sexistas como positivos, mientras que la seguridad es considerablemente alta en el caso opuesto al identificar ejemplos positivos como no sexistas. Cruzando la información relativa a los falsos negativos y a las categorías sexistas a las que pertenecen también se ha descubierto un ranking idéntico al conocido hasta el momento, donde las principales posiciones son ocupadas por las clases sexistas que se consideran más complicadas de detectar automáticamente. Una de las razones explicativas de este fenómeno es que el contenido de más del 25 % de documentos incorrectamente categorizados han sido catalogados dentro de la emoción *joy*.

### 6.2.5. Resumen y conclusiones

Para terminar este capítulo de arquitecturas LSTM bidireccionales en la Figura 6.4 y 6.5 se representan dos gráficos de barras con las métricas de validación obtenidas en las tablas comparativas anteriores que muestran visualmente las diferencias significativas existentes entre los dos idiomas. Mientras que en la Figura 6.4 relativa al modelado inglés se observa una semejanza general entre los distintos ensayos, en términos de accuracy y AUC, existe una tendencia que se inclina hacia las capas unidireccionales por su mayor simpleza y menor consumo de recursos. Por lo tanto parece que en la **clasificación de documentos redactados en inglés los requisitos son menores** para la extracción y estudio de los patrones que llevan a los modelos a la identificación de ambas clases. Sin embargo, en la Figura 6.5 ocurre lo contrario siendo más beneficiosas las arquitecturas bidireccionales en la mayoría de los experimentos del modelado español. Ha sido gracias a esta comparativa en la que nos hemos podido percatar del anuncio que se comentaba en secciones anteriores acerca del **mayor nivel de enriquecimiento que caracteriza los textos españoles**. Si bien este fenómeno puede presumirse como una ventaja al efectuar su modelado puesto que supuestamente contienen más características que puedan aportar información más útil con la que construir su respectivo clasificador, también conlleva un nivel de complejidad más amplio, y como consecuencia un número de recursos computacionales y temporales más elevados. Únicamente existe la excepción de la composición caracterizada por una sola capa oculta en la que ambos mecanismos de estudio y análisis de muestras muestran un comportamiento similar, seguramente debido a que su simpleza elevada al máximo exponente minimiza cuantiosamente el beneficio que supone emplear una arquitectura más compleja en textos más enriquecidos.

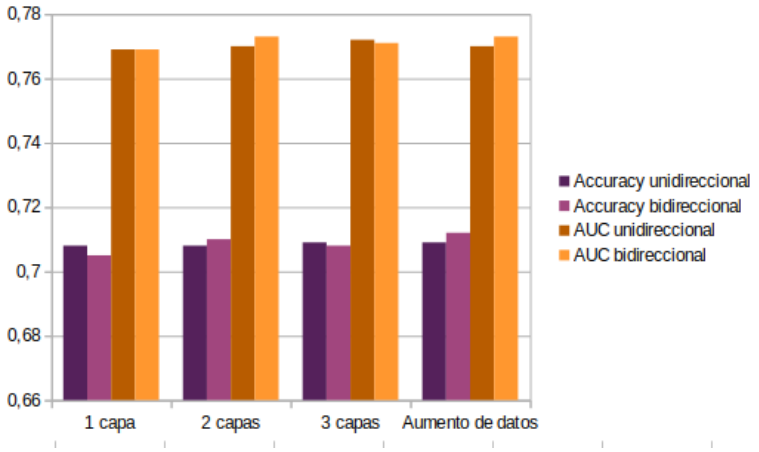


**Figura 6.4:** Evolución de las métricas de validación en la comparación de arquitecturas LSTM unidireccionales y bidireccionales en el modelado de textos en inglés.

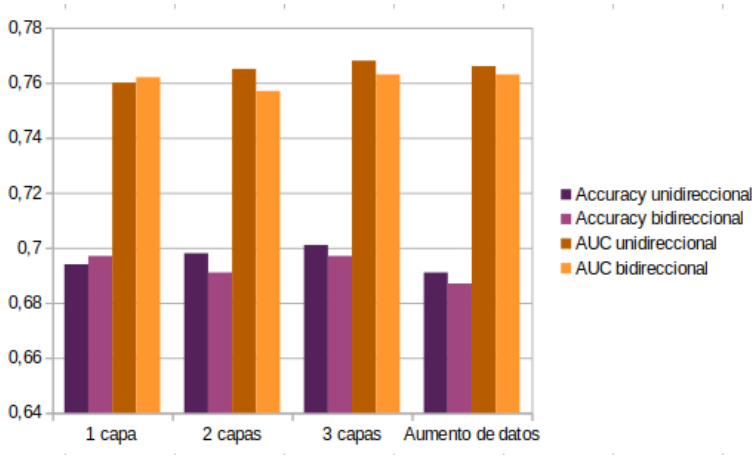


**Figura 6.5:** Evolución de las métricas de validación en la comparación de arquitecturas LSTM unidireccionales y bidireccionales en el modelado de textos en español.

## 6.3. Modelos BERT

La última arquitectura experimentada para la resolución del problema de detección y clasificación de textos sexistas consiste en elaborar un proceso de *finetuning* compuesto por diferentes configuraciones para ser aplicado en modelos *transformers* preentrenados sobre conjuntos voluminosos de documentos, con sus respectivos tokenizadores para la codificación de datos. Para ello se ha utilizado la librería *transformers* en combinación con *tensorflow* por su popularidad extendida a nivel global, su gran comunidad de usuarios y la multitud de ejemplos que facilitan la implementación de los flujos de trabajo requeridos con los que producir clasificadores más avanzados. De nuevo se mantiene la fabricación de modelos concentrados en un único lenguaje por lo que los clasificadores preentrenados seleccionados para cada uno se detallan a continuación.

- ***bert-base-uncased***. Es el *transformer* preentrenado empleado para el análisis de documentos en **inglés**. Contiene aproximadamente 110 millones de parámetros y ha sido construido persiguiendo un doble objetivo: basado en una primera tarea conocida como *Masked Language Modeling* que consiste en predecir palabras que han sido seleccionadas y reemplazadas aleatoriamente por máscaras, mientras que la segunda denominada *Next Sentence Prediction* se encuentra orientada en identificar si un par de sentencias son secuenciales dentro de un contexto determinado.

- ***dccuchile/bert-base-spanish-wwm-uncased***. Se trata de una va-

riante del sistema anterior aunque especializado sobre corpus en **español**. Comparte las mismas características descritas previamente.

Por lo tanto en este capítulo se puede encontrar, en primer lugar, la configuración procedente de los anteriores ensayos con Regresión Logística y LSTM con la que poder incorporar los mejores valores de los parámetros comunes, beneficiándonos de sus buenos rendimientos característicos y orientando las próximas experimentaciones a aquellos ajustes propios de las arquitecturas BERT. Posteriormente se detalla el diseño del proceso llevado a cabo para efectuar las distintas rondas de pruebas planteadas tanto para las muestras en inglés como en español. En este caso, a diferencia de los modelos previos, sí existen discrepancias suficientemente significativas en términos de métricas de validación entre sendos idiomas. En la última subsección se podrán encontrar un resumen de las conclusiones extraídas y una evolución de las medidas de evaluación de los modelos conforme se han adaptado los parámetros considerados más relevantes para los *transformers* empleados.

### 6.3.1. Lecciones aprendidas

A continuación se analizan las pesquisas encontradas en anteriores arquitecturas con el propósito de generar un **punto de partida más avanzado** con el que comenzar las experimentaciones con arquitecturas BERT.

- **Codificación con embeddings**. Recordamos que una de las principales investigaciones procedentes del modelado con Regresión Logística se centró en el modo de codificación de los documentos, descubriendo un magnífico potencial en el uso de **embeddings preentrenados**. Se profundizó en el uso de esta metodología en la generación de clasificadores LSTM en cuyos ensayos se demostró que el conjunto ***Glove Twitter 27B 100d*** era el que se caracterizaba por una superior capacidad de adaptación por su inmenso vocabulario y por compartir fuente de datos con los datasets *EXIST*. Como consecuencia este conjunto de embeddings será el empleado en las próximas experimentaciones con el modelado basado en *transformers* BERT.

- **Tamaño del lote**. Derivado de uno de los ajustes en la construcción de sistemas inteligentes con arquitecturas LSTM se demostró que el valor del tamaño del lote debía ser reducido para proporcionarle menos muestras al modelo durante su entrenamiento y así maximizar su rendimiento a costa de una mayor inversión de recursos. Gracias a esta investigación en las futuras pruebas se **reduce el rango de valores a ensayar, siendo los seleccionados 32, 16 y 8**. El motivo de continuar experimentando con este parámetro reside en que como teoría

propia pienso que con una arquitectura más compleja quizás se deba reducir aún más con el fin de que la extracción y estudio de patrones sea más detallado.

- **Aumento del conjunto de entrenamiento**. Si bien ninguna de las técnicas comprobadas resultó ser favorable en el modelado con Regresión Logística y LSTM, mi hipótesis es que sí pueda ser más útil en arquitecturas BERT puesto que en multitud de ocasiones los modelos más complejos necesitan un mayor volumen de datos para explotar su potencial. No obstante, de la experimentación con arquitecturas LSTM se observó que algunas de ellas proporcionaron mejores clasificadores, como la traducción simple entre idiomas y la aplicación de operaciones *Easy Data Augmentation*, mientras que la generación de muestras sintéticas por ***Contextual Embeddings*** **fue la que peores modelos proporcionó. Como consecuencia este último procedimiento queda descartado** para los ensayos con *transformers* ayudando así a reducir y enfocar el ajuste de parámetros y datos únicamente a los métodos que han resultado ser más ventajosos en anteriores modelados.

### 6.3.2. Procedimiento de experimentacion

Puesto que el diseño de las experimentaciones llevadas a cabo con las que decidir si un valor es significativamente mejor que otro dentro de los parámetros a ajustar es igual que el explicado en la sección de arquitecturas LSTM unidireccionales, si lo desea puede consultarlo en este enlace.

### 6.3.3. Ejemplo práctico en español

Tras llevar a cabo una hiperparametrización particular a esta tipología de red neuronal, ajustando los parámetros considerados más relevantes para proporcionar un buen resultado, siendo visible esta parte en el apéndice del final de esta memoria, a continuación se presentan los resultados de su aplicación. La razón de seleccionar el modelado español para ejemplificar la configuración más beneficiosa encontrada a partir de las experimentaciones anteriores reside en el énfasis que le otorga al aumento del conjunto de entrenamiento y a la ralentización de la tasa de aprendizaje que han posibilitado la obtención del clasificador de mayor capacidad de generalización durante este proyecto. Los parámetros más relevantes se detallan a continuación:

- **Aumento del conjunto de entrenamiento**. Las medidas de evaluación han sido máximas tras la aplicación del mecanismo *Easy Data*

*Augmentation* con una selección y reemplazamiento aleatorio de términos efectuada hasta tres veces por cada documento de entrenamiento, lo que conlleva un **volumen total de ejemplos cuatro veces mayor** añadiendo las nuevas muestras sintéticas a los datos originales.

- **Codificación mediante embeddings preentrenados**. Como no podría ser de otra forma en las arquitecturas BERT y en este sistema particular se ha hecho uso del conjunto de embeddings *Glove Twitter 27B 100d* para codificar los documentos disponibles en los datasets de entrenamiento y test debido a los buenos resultados proporcionados al comienzo de la construcción de modelos con LSTM.

- **Tamaño del lote**. Con un total de ocho muestras por lote se ha conseguido una mejora sustancial que ha permitido elevar las métricas de validación de las arquitecturas BERT en comparación con los anteriores modelos. A diferencia de la sección anterior, para los *transformers* el tamaño del lote se ha mantenido invariable para ambos idiomas.

- **Tasa de aprendizaje**. De forma similar a la reducción del tamaño del lote producida sobre el modelado de documentos en español con arquitecturas LSTM, en este caso ha ocurrido una situación semejante aunque con el parámetro que marca la velocidad de aprendizaje de un clasificador BERT también centrado en el mismo idioma. Mientras que para **textos ingleses fue suficiente fijar el valor 2e-5, en los ejemplos españoles ha debido ser reducido hasta 1e-5** para conseguir las mejores métricas de validación conocidas en todas las experimentaciones llevadas a cabo. Parece ser que una vez más se desmuestra la necesidad de proporcionar más tiempo a la fabricación de sistemas detectores de textos sexistas y no sexistas españoles gracias a su mayor representatividad y riqueza que permite alcanzar umbrales de generalización más elevados.

- **Arquitectura**. Se ha utilizado un modelo preentrenado conocido como *dccuchile/bert-base-spanish-wwm-uncased* basado únicamente en corpus españoles, compuesto por un total de 110 millones de parámetros y cuya fase de construcción perseguía dos objetivos: la predicción de palabras aleatoriamente enmascaradas (*Masked Language Modeling*) y la estimación de si un par de sentencias son secuenciales dentro de un mismo documento (*Next Sentence Prediction*).

- **Entrenamiento y validación**.

  - **Early Stopping**. Este mecanismo se lleva integrando en la fabricación de todos los modelos con el propósito de reducir el fenómeno del sobreaprendizaje tan común en la detección de textos sexistas. Para ello se otorga un margen de cien iteraciones

que se verán interrumpidas si tras quince no se ha producido una mejora de más del 0.01 en la métrica de accuracy, retornando los pesos del modelo con mayor rendimiento encontrado.

- **Función de pérdida**. De nuevo se ha escogido la función denominada *binary crossentropy* con la que calcular la pérdida provocada tras cada iteración con la que medir la bondad del modelo durante su entrenamiento y validación.
- **Optimizador**. En este caso también se ha hecho uso del optimizador más popularmente utilizado *Adam* que se encuentra establecido por defecto en la configuración de entrenamiento debido a su magnífico rendimiento dentro de un gasto de recursos razonable.
- **Porcentaje de validación**. En cada iteración se reserva un 20 % del total de muestras de entrenamiento para efectuar una validación sobre la marcha con la que conocer la evolución del clasificador en construcción.

Mientras que la verificación en el conjunto de entrenamiento da como resultado más del 99 % de accuracy y AUC, la **evaluación sobre las muestras de test se reducen al 77.15 % y 77.32 %, respectivamente**. Cabe destacar que la máxima puntuación alcanzada en ambas competiciones por los equipos ganadores únicamente alcanzó un 80 % en accuracy por lo que los resultados conseguidos en este proyecto se encuentran muy cercanos. Como consecuencia en la Figura 6.6 se aprecia una **drástica reducción de las tasas de falsos negativos y positivos** aunque el valor de estos últimos sigue siendo considerablemente elevado dado el volumen total del conjunto de test.

Tras analizar las muestras clasificadas erróneamente en cada intervalo de confianza se ha podido observar un ranking idéntico al visualizado actualmente en el que se representa una confianza de muy alta a moderada con el que el clasificador español ha cometido estos fallos. Sin embargo, cruzando los falsos negativos con las mencionadas categorías sexistas sí que se han podido notar discrepancias elevadamente importantes puesto que la distribución cambia completamente. En la Tabla 6.4 se aprecia un cambio de tendencia en el que en **primera posición se situa una de las categorías con contenido más violento** de entre las disponibles, que prácticamente en todos los modelos se ha localizado en la parte baja de la tabla. Después de una inspección visual con la que hacerse una idea explicativa de este fenómeno se ha podido concluir que el contenido de los falsos negativos pertenecientes a esta clase se caracteriza por hacer alusión a **protestas y reivindicaciones a favor de la igualdad de género**, que al haber utilizado términos con connotación sexista según el modelo han sido incorrectamente clasificados, así como la **composición de hashtags y la**

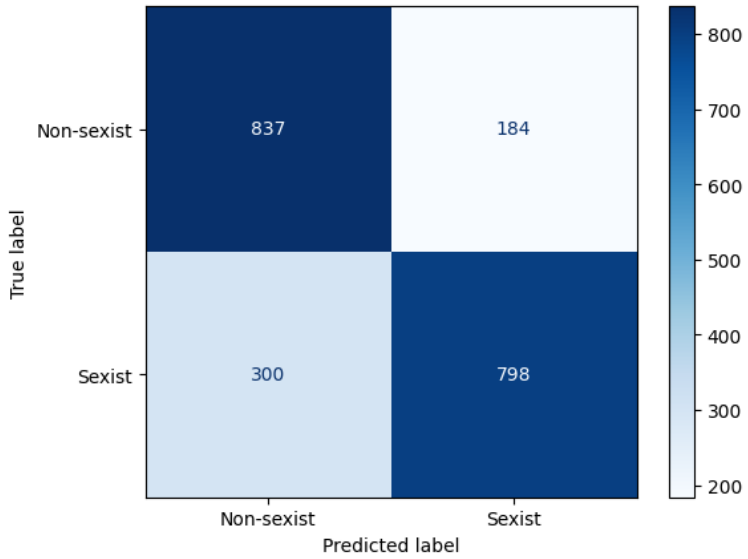


**Figura 6.6:** Matriz de confusión para la validación sobre test de un modelo BERT específico de textos en español con la mejor configuración encontrada.

**aparición de errores gramaticales** que han dificultado su codificación, análisis e identificación.

**Tabla 6.4:** Tabla con el número de falsos negativos procedentes de un modelo BERT español por cada intervalo de confianza.

| Clase sexista | FN |
|---|---|
| sexual-violence | 72 |
| misogyny-non-sexual-violence | 69 |
| ideological-inequality | 61 |
| stereotyping-dominance | 54 |
| objectification | 44 |

### 6.3.4. Resumen y conclusiones

Por último se procede a explicar las pesquisas descubiertas tras las diversas investigaciones realizadas sobre arquitecturas BERT durante el modelado de documentos en inglés y español. A continuación en la Figura 6.7 se representa una gráfica de líneas que refleja la evolución de la media aritmética tanto de la métrica accuracy como AUC tras establecer el valor más propicio encontrado para cada uno de los parámetros considerados.

- Ajustando el tamaño del lote a ocho muestras ya se visualiza una destacable mejora en las métricas de validación con respecto a los re-

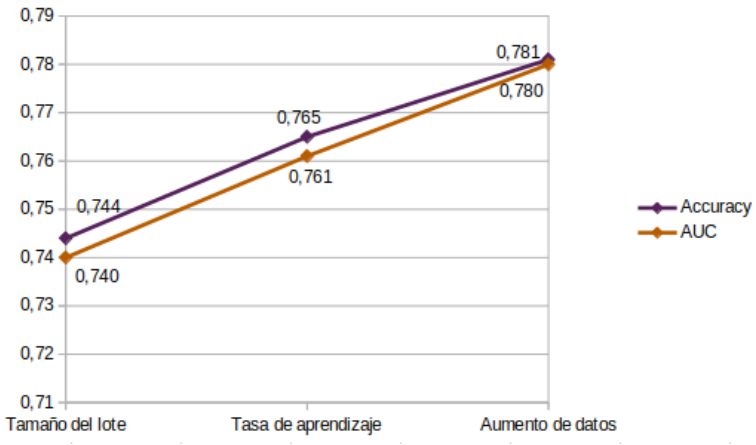


**Figura 6.7:** Gráfica de evolución de las métricas de validación durante el ajuste de los parámetros en modelos BERT.

sultados proporcionados por anteriores arquitecturas que le ayudan a comenzar su trayectoria en más del 74 % de accuracy y AUC. Y es que tal y como se predijo al comienzo de los ensayos con *transformers* BERT era de esperar que su **elevada complejidad funcional necesitase una reducción aún más contundente del número de muestras por lote** a suministrar durante la fabricación de los modelos. Si bien el hecho de limitar más estrictamente la cifra de ejemplos a extraer y estudiar sus patrones por lote ha provocado un efecto beneficioso traducido en el incremento de la capacidad de generalización, también produce un considerable aumento del volumen de recursos computacionales necesarios para su ejecución.

- Con una analogía similar a la previa **reduciendo la tasa de aprendizaje se ha conseguido** una ralentización en el procedimiento mencionado anteriormente de análisis de las muestras de entrenamiento que ha elevado sumamente las métricas de validación en más de un 2 %, consiguiendo por primera vez un accuracy y AUC de más del 76 %. De nuevo a costa de un gasto superior de recursos se le proporciona al clasificador más tiempo con el propósito de profundizar aún más en los vecindarios de búsqueda de cada uno de los datos de entrenamiento para encontrar una mayor cantidad de información útil con la que identificar las muestras de cada clase.

- Finalmente en esta arquitectura sí se ponen de manifiesto las ventajas que supone la disposición de un **volumen elevado de ejemplos de entrenamiento con el que fabricar un clasificador más robusto** y capacitado para conseguir unas métricas de validación acumuladas cercanas al 80 %. Cabe destacar que la teoría que se planteó al comien-

zo de esta sección también apostaba porque serían los modelos BERT quienes sacarían mayor provecho de las técnicas de aumento de datos, a pesar de los resultados encontrados en arquitecturas previas.

# Capítulo 7

# Análisis y deducciones del modelado

En este breve capítulo se pretende poner en valor las deducciones alcanzadas durante el modelado de los documentos *EXIST* en inglés y español a través del uso de distintas configuraciones de datos, arquitecturas construidas desde cero y modelos preentrenados ajustados a la problemática que aborda este proyecto. En él se podrán visualizar dos primeras gráficas que resumen la evolución del proceso de modelado efectuado destacando únicamente los puntos en los que se han alcanzado mejoras importantes, desde el comienzo con algoritmos clásicos como Regresión Logística hasta su etapa final con sistemas inteligentes avanzados y especializados en tareas del Procesamiento del Lenguaje Natural. Asimismo, se le desea otorgar una especial relevancia representativa a las muestras de test incorrectamente identificadas con el propósito de observar la variación de su volumen en relación a la producción de modelos conforme se incorporaban nuevos ajustes que mejoraban la capacidad de generalización frente a ejemplos desconocidos. Por último se han calculado algunos datos estadísticos que serán tremendamente útiles en el capítulo siguiente para la aplicación de técnicas de explicabilidad. Consisten en comparar los falsos negativos y positivos de los mejores clasificadores generados en cada idioma con el propósito de identificar ejemplos canónicos que reflejen los textos categorizados erróneamente de forma común a todos los modelos o a parejas de ellos. Así se pretende analizar las cifras de documentos complicados de reconocer automáticamente con los que intentar descubrir los posibles factores causantes de estos fallos.

## 7.1. Evolución del modelado según las métricas de validación

En esta primera sección se pretende cubrir el estudio del progreso realizado con las experimentaciones detalladas en el capítulo anterior gracias a las cuales se han conseguido cuatro modelos de distintas arquitecturas. En la Figura 7.1 se observa la evolución asociada a la clasificación de las muestras en inglés ordenada de forma ascendente en función de las medias aritméticas de accuracy y el área bajo la curva ROC (AUC). Tal y como era de esperar, el sistema entrenado con el algoritmo clásico de Regresión Logística en combinación con la codificación en forma de bolsas de palabras es el que ocupa la primera posición por su menor tasa de aciertos y capacidad de generalización. Al ser una **técnica de global aplicación es totalmente plausible que no proporcione un buen comportamiento frente a tareas más específicas**, como es la detección de textos sexistas por su contenido. Asimismo la transformación de textos utilizada también se caracteriza por contener deficiencias importantes como no respetar el orden de los términos dentro de una frase concreta ni considerar la similitud entre conceptos.

A continuación se situa una arquitectura LSTM undireccional y una segunda bidireccional con la misma composición de dos capas ocultas de 128 neuronas cada una, un tamaño del lote de treinta y dos muestras y empleando el conjunto de embeddings *Glove Twitter 27B 100d* para la codificación de documentos. Como se puede apreciar ambas suministran unas medidas de evaluación prácticamente similares, aunque cabe destacar que en el segundo modelo la inversión de recursos era muy superior al sistema más sencillo. Dados los bajos porcentajes de las tasas de aciertos y capacidades de predicción he de decir que **mi teoría suponía un mejor rendimiento** por parte de las arquitecturas LSTM, incluso apostando a que podían posturlarse como soluciones viables y de cierta calidad para resolver el problema abordado en este proyecto. Sin embargo, pese a que dispone de uno de los valores de AUC máximos con respecto a los otros clasificadores, **apenas superan el 70 % de accuracy**, por lo que ninguno de ellos podría plantearse como recurso para la detección de sexismo basado en textos. Quizás la adición de algún complemento, como un mecanismo de atención, podría haber mejorado las cualidades de los modelos otorgándoles una mayor capacidad de competición frente a arquitecturas más complejas.

Finalmente se localiza el *transformer* BERT con el que se ha conseguido la **mayor tasa de aciertos para el modelado de documentos en inglés** gracias a la combinación entre el aumento del volumen de ejemplos

de entrenamiento utilizando la técnica *Easy Data Augmentation*, la disminución del tamaño del lote a únicamente ocho datos y la reducción de la tasa de aprendizaje para permitirle una extracción de características más detallada con un menor porcentaje de ejemplos en cada ronda. Tal y como se comentó en el capítulo previo, era de esperar que un modelo tan complejo necesitase un mayor número de datos para su construcción y validación, siendo esta arquitectura la única que ha podido aprovechar la aplicación de procedimientos de generación de muestras sintéticas. Por otro lado también disponía de la hipótesis relacionada con la ralentización que supondría el uso de estas arquitecturas, especialmente si se le proporcionaba más ejemplos de entrenamiento. Sin embargo, el punto más discrepante reside en que su área bajo la curva ROC, la medida que supuestamente representa su capacidad de predicción, es aproximadamente un 3 % menor que en los sistemas LSTM, a pesar de que su tasa de aciertos es más cuantiosa.

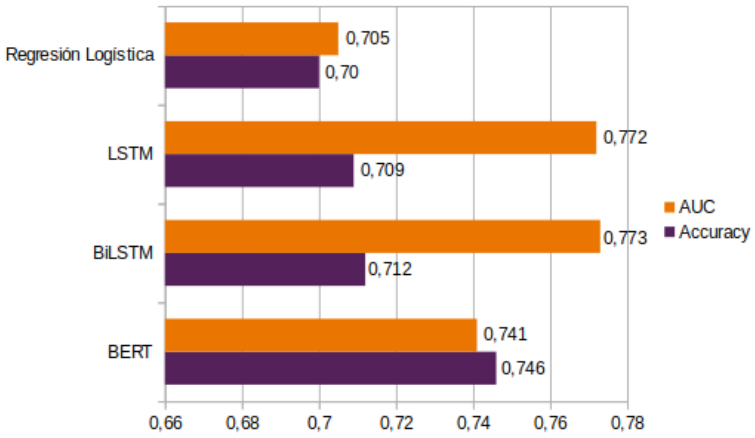


**Figura 7.1:** Evolución del modelado de documentos ingleses en función de los mejores modelos obtenidos y las métricas de validación.

En la Figura 7.2 se representa de igual forma la acumulación de las métricas de validación en las configuraciones más beneficiosas durante el análisis de textos españoles. En las tres primeras arquitecturas las conclusiones son equitativas con respecto a las detalladas en la clasificación de documentos ingleses. No obstante, se ha incluido esta gráfica para resaltar dos aspectos importantes. Por un lado el sistema BERT bajo una configuración que reside en una menor tasa de aprendizaje y un mayor volumen de muestras, responde con un rendimiento bastante superior a la misma arquitectura en el idioma inglés. Por lo tanto, como se mencionaba en el capítulo anterior, se confirma la teoría anunciada en las competiciones acerca del **mayor valor y enriquecimiento que demuestran los documentos en español permitiendo así la construcción de sistemas automáticos con una capacidad superior de generalización**. Mientras que por otro lado, si

bien En la Figura 7.1 se aprecia una mayor área bajo al curva ROC en los modelos LSTM con respecto al sistema BERT pese a que este indica una tasa de aciertos más elevada, En la Figura 7.2 el protagonista claramente es el **clasificador BERT puesto que supera en ambas métricas de validación a los restantes modelos**. Como consecuencia, en este lenguaje no hay ninguna duda de que el sistema BERT es el mejor preparado para abordar la detección y clasificación de documentos sexistas y no sexistas.

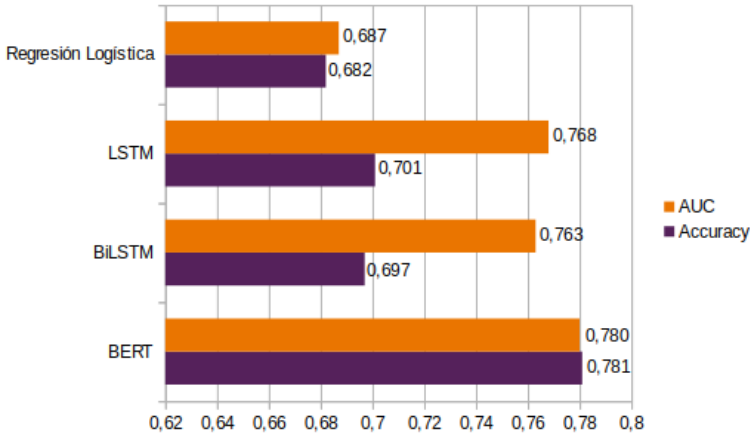


**Figura 7.2:** Evolución del modelado de documentos españoles en función de los mejores modelos obtenidos y las métricas de validación

## 7.2. Evolución del modelado según los falsos negativos y positivos

En este nuevo apartado el propósito es doble, primeramente consiste en apoyar las conclusiones resumidas de la sección anterior con unas métricas de evaluación adicionales como son la tasa de falsos negativos y positivos, mientras que en segundo lugar se pretende analizar si existe una clase particular donde se cometan un mayor número de fallos de manera generalizada a todos los modelos o de manera parcial. En la Figura 7.3 los cálculos realizados se corresponden con la identificación de ejemplos en inglés y tal y como se aprecia se encuentran ordenados ascendentemente, al igual que las representaciones anteriores. En ella se observa que en la más alta posición del ranking se encuentra el **modelo BERT previamente descrito, que se postula como la mejor solución hallada para este idioma. Sin embargo, al contrario que las demás arquitecturas su punto débil es la clase negativa** puesto que su tasa de falsos positivos es superior a la de los negativos. De hecho es el porcentaje de muestras negativas incorrectamente clasificadas más alto de todas las arquitecturas consideradas. En las

restantes posiciones se refleja el orden inverso visualizado en las métricas de validación, teniendo como punto común la elevada tasa de falsos negativos con la que demuestran no haber aprendido demasiado bien las características de los textos sexistas.

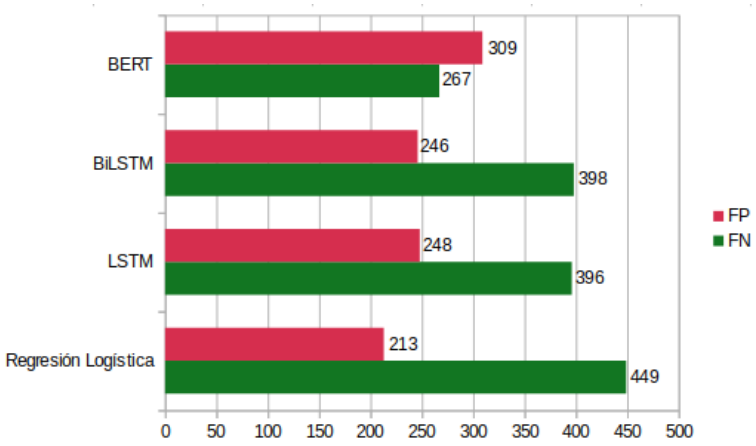


**Figura 7.3:** Evolución del modelado de documentos ingleses en función de los mejores modelos obtenidos y las muestras erróneamente clasificadas.

Por último En la Figura 7.4 que recopila el conteo de falsos negativos y positivos durante el modelado de documentos en español se visualiza una tendencia más homogénea en la que todos los clasificadores engendrados disponen de una tasa de muestras positivas erróneamente categorizadas más elevada con respecto a la clase contraria. El hecho más destacable de esta representación reside en que la **diferencia entre el número de falsos negativos y positivos es mucho más notable** que en el análisis de textos ingleses, siendo una discrepancia hasta aproximadamente el doble superior en la mayoría de sistemas. Como conclusión se puede afirmar que pese a que los documentos españoles parezcan estar caracterizados por una mayor información útil que ha liderado a obtener los clasificadores con mejores métricas de validación, sus ejemplos positivos siguen sin ser suficientemente nítidos como para ser estudiados e identificados de forma automática y correcta.

## 7.3. Análisis estadístico de errores comunes

En este nuevo apartado se ha contabilizado el número de falsos negativos y positivos que son comunes a los tres mejores modelos generados durante el modelado de los documentos en inglés y español, considerando el algoritmo de Regresión Logística y las arquitecturas LSTM (español) / BiLSTM (inglés) y BERT. Estos clasificadores han sido profundamente explicados en el capítulo anterior detallando las experimentaciones que han desembocado

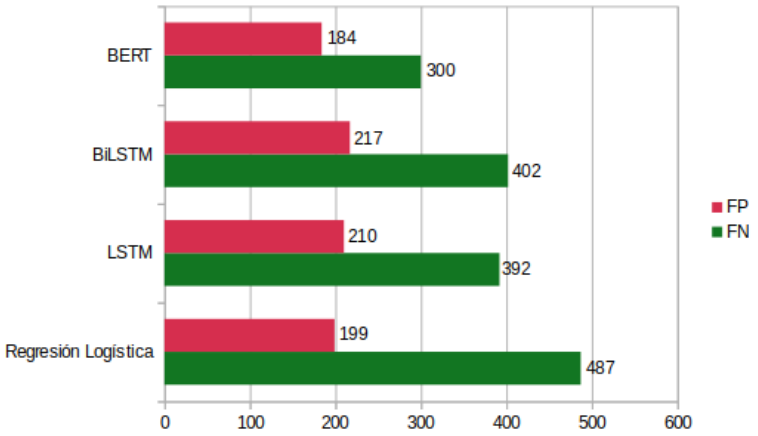


**Figura 7.4:** Evolución del modelado de documentos españoles en función de los mejores modelos obtenidos y las muestras erróneamente clasificadas.

en sus configuraciones de datos, entrenamiento y validación. El propósito de esta sección consiste en descubrir si existen ciertas muestras que han resultado difíciles de identificar para los tres sistemas tan variopintos o para algunos pares en particular. En el capítulo posterior de explicabilidad se le dedicará una sección particular a dichos ejemplos comunes para tratar de reconocer cuáles han podido ser las causas que expliquen los fenómenos calculados. En la Tabla 7.1 se recopilan los valores generados por idioma y tipo de sistema. El primer hecho destacable es que **no existe ningún falso negativo ni positivo común a los tres mejores clasificadores**, y ciertamente parecer ser lógico puesto que sus composiciones son de naturaleza completamente distinta. A continuación se observa un volumen bajo de errores comunes al algoritmo clásico de Regresión Logística y a las arquitecturas más avanzadas, siendo menor en la última pareja en la que se encuentra con el modelo BERT seguramente motivado porque este duo tiene cualidades notablemente dispares. Una situación opuesta se encuentra en el par **LSTM/BiLSTM y BERT que disponen de una elevada tasa de fallos comunes en ambos idiomas**, superando formidablemente el número de falsos negativos y positivos contabilizados en los dos anteriores pareados. Al contrario que ocurría en la situación previa, esta dos arquitecturas sí se las puede visualizar como más similares en relación a su procedencia, ya que ambas son tipos de redes neuronales complejas, y a su modo de funcionamiento al utilizar codificación por embeddings preentrenados para preparar los documentos con los que trabajar. Una posible consecuencia que se puede traducir de este hecho se fundamenta en la existencia de un número importante de ejemplos que han sido incorrectamente clasificados por sendos modelos.

Cruzando los falsos negativos característicos de las tres parejas confec-

**Tabla 7.1:** Tabla con el número de falsos negativos y positivos comunes a los mejores modelos encontrados.

| Idioma | Error | Todos | LR-LSTM | LR-BERT | LSTM-BERT |
|---|---|---|---|---|---|
| inglés | Falsos negativos | 0 | 66 | 27 | 161 |
| | Falsos positivos | 0 | 20 | 13 | 150 |
| español | Falsos negativos | 0 | 61 | 20 | 188 |
| | Falsos positivos | 0 | 25 | 14 | 99 |

cionadas con respecto a las categorías sexistas a las que pertenecen, se ha podido apreciar una tendencia común a todas ellas y es que la **mayoría pertenecen a los tipos de sexismo identificados como los más complicados de reconocer automáticamente**. Recordemos que del análisis exploratorio de datos que se efectuó al comienzo del proyecto las categorías *ideological-inequality*, *misogyny-non-sexual-violence* y *stereotyping-dominance* se conocían porque su contenido era capaz de crear confusión hasta en el *transformer* preentrenado utilizado para detectar emociones, además de poseer documentos con considerables niveles de ironía y doble significado aunque repletos de terminología y expresiones positivas o no sexistas.

# Capítulo 8

# Técnicas de explicabilidad

Como se ha estado anunciando durante la memoria actual en este quinto capítulo se podrá disfrutar de la aplicación de dos de las técnicas locales de explicabilidad más popularmente utilizadas dentro del nuevo ámbito conocido como *Explainable Artificial Intelligence*, que tanto peso ha ganado en los últimos años. En particular se trata de LIME (*Local Interpretable Model Agnostic Explanation*) que como se ha explicado anteriormente se concentra en destacar qué palabras contribuyen a qué clases. Mientras que la generación de contrafactuales consiste en la producción de nuevas muestras con las que probar la bondad de los modelos visualizando qué modificaciones contribuyen hacia la categorización de cada clase. Para la fabricación de estos ejemplos sintéticos existe una amplia gama de operaciones [23] que suelen ser ejecutadas sobre los propios documentos de test, de las que se han considerado y seleccionado aquellas con un mayor potencial:

- **Eliminar términos**. Este proceso se fundamenta en la eliminación, de forma secuencial o aleatoria, de uno o más conceptos dado un texto de ejemplo para producir un nuevo documento por iteración. En el planteamiento integrado en este proyecto se ha optado por minimizar el número de cambios necesarios para producir una nueva muestra y por ende se suprimen todos los términos secuencialmente a razón de uno por ejecución por cada texto de test. Por lo tanto el volumen de contrafactuales descendientes de esta operación dependerá del tamaño de cada ejemplo escrito.
- **Negar verbos**. En esta segunda metodología el propósito reside en emplear el modelo generador de lenguaje tan popularmente extendido como es GPT-3 para enviarle tareas basadas en la transformación de los verbos contenidos en los documentos de positivos a negativos y viceversa, consiguiendo así un nuevo dato por cada ejemplo de test cuyo significado sea opuesto al original realizando el menor número de

cambios. A diferencia de la operación anterior, esta tiene un carácter semiautomático puesto que necesita supervisión ya que en ocasiones GPT-3 elabora reformas excesivas que hacen perder prácticamente toda relación con los ejemplos originales.

Gracias al uso y desarrollo de las acciones comentadas previamente con las que arrojar algunas teorías explicativas acerca del comportamiento de los tres mejores modelos encontrados tras las múltiples experimentaciones con distintas configuraciones de datos, arquitecturas, entrenamientos y validaciones se han podido extraer una variedad de **conclusiones que son comunes a los modelos específicos de cada idioma**. Como consecuencia, para no prolongar demasiado este capítulo su enfoque se hará en torno a las muestras en español por ser nuestro idioma materno y por haber obtenido los sistemas con las más elevadas tasas de accuracy y área bajo la curva ROC.

## 8.1. LIME

Tal y como se ha explicado en el segundo capítulo de esta memoria, LIME es una técnica local capaz de generar explicaciones que resaltan los términos en los que el modelo se ha fundamentado para asignar una determinada clase a una muestra particular. Con el propósito de llevar a cabo los siguientes análisis se ha hecho uso de la librería *lime* de Python que destaca por su facilidad de uso y por su empleo generalizado a nivel académico y empresarial. Si bien esta metodología es válida para su aplicación a cualquier tipo de ejemplos, en este caso se ha deseado apoyar en los **datos erróneamente clasificados descubiertos en secciones previas, tanto falsos negativos como positivos**. Así se pretende encontrar las causas que justifican tal comportamiento en cada uno de los clasificadores engendrados y dependiendo de los resultados se podrán plantear soluciones adicionales que permitan mejorar sus capacidades de predicción frente a ejemplos desconocidos. Dentro de esta sección se da a conocer el procedimiento diseñado para la selección de los documentos a estudiar con LIME, al igual que los resultados obtenidos y representaciones gráficas de los patrones extraídos.

### 8.1.1. Procedimiento de experimentación

Dado el elevado volumen de documentos incorrectamente identificados se ha optado por seleccionar aleatoriamente **diez ejemplos de los conjuntos de falsos negativos y positivos del conjunto de test asociados a cada intervalo de confianza** establecido dentro de los tres mejores modelos producidos por arquitectura. Se les aplicará la técnica LIME a su totalidad

buscando causas comunes que resuman detalladamente las teorías explicativas de los fallos cometidos durante la asignación de etiquetas a los documentos españoles pertenecientes al conjunto de test. Finalmente se efectuarán diversas inspecciones visuales individuales y colectivas con las que reconocer posibles patrones en los contenidos de los documentos, temáticas y terminologías coincidentes que permitan su agrupación en diferentes casuísticas.

### 8.1.2. Mejor modelo de Regresión Logística español

Recordamos que en este clasificador se contabilizaron un total de 462 falsos negativos y 199 falsos positivos distribuidos en intervalos de confianza desde muy altos a moderados. Iniciamos el análisis de explicabilidad empleando el primer conjunto que almacena los falsos negativos en el que destacan tres principales aspectos. En el intervalo de confianza máximo los errores inducidos se deben a la asignación de un **valor no sexista a terminología neutra**, como determinantes, preposiciones, pronombres y conceptos varios, que contrarrestan el sentido sexista del contenido. Se trata de un fenómeno severo en el que parece introducirse ruido con la presencia de estos tipos de vocablos produciendo una alteración significativa en la identificación y clasificación de las muestras. A continuación en la Figura 8.1 se representa un ejemplo ilustrativo de un documento en el que se visualiza cómo determinantes y preposiciones son consideradas no sexistas, además del vocablo "*embarazada*", produciendo una suma suficiente como para desviar la etiqueta hacia la clase negativa.

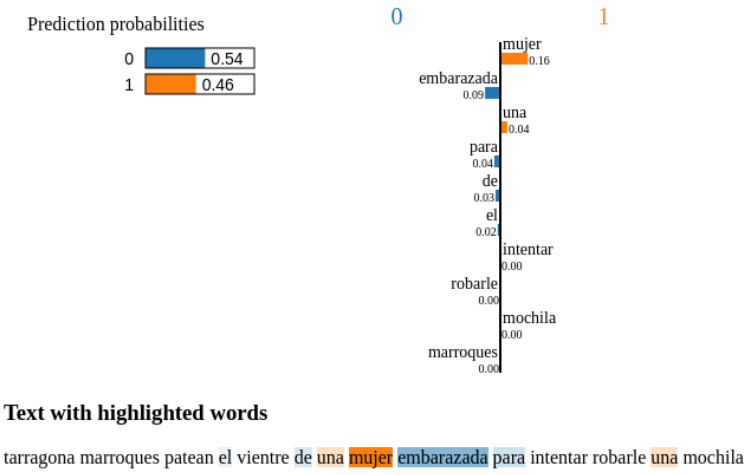


**Figura 8.1:** Análisis de explicabilidad por LIME en torno a un falso negativo producido por asignar una connotación no sexista a términos neutros.

Por otro lado, en el segundo rango de confianza más elevado la tendencia es ligeramente diferente aunque relacionada con lo expuesto. En esta ocasión

la problemática reside en que un amplio volumen de documentos se encuentran compuestos por palabras que a priori son positivas pero que dentro de un contexto su significado es profundamente sexista. A esta cualidad se la conoce como **ironía** y se ha postulado, desde hace décadas, como uno de los retos más complejos de resolver en el área del Aprendizaje Automático y Profundo. En la Figura 8.2 se puede observar un ejemplo ilustrativo acerca de un falso negativo relacionado con esta casuística en el que pocos términos son negativos o sexistas pero en conjunto su intención es denigrar las habilidades del colectivo femenino.


**Figura 8.2:** Análisis de explicabilidad por LIME en torno a un falso negativo producido por ironía.

Finalmente el intervalo de confianza moderado se caracteriza por poseer muestras con una frecuencia elevada de aparición de **hashtags y errores ortográficos** que dificultan la codificación, análisis y clasificación de documentos. Si bien ya se intuyó que esta deficiencia podría tener un impacto negativo en el rendimiento de los modelos, a pesar de experimentar con diversas librerías de Python para la detección y reemplazamiento por sus correspondientes términos correctos, no se obtuvieron resultados positivos en términos de las métricas de validación. En la Figura 8.3 aparece una salida de la librería LIME que refleja esta situación en la que el verbo "*abuso*"de estar correctamente escrito se podría haber interpretado como sexista reconduciendo la categorización de este texto hacia su debida clase.

Ejecutando LIME sobre el conjunto de falsos positivos se han confirmado dos tendencias evidentes. En los tres intervalos de confianza establecidos sus documentos contenían noticias y reivindicaciones de usuarios contra los ataques y la violencia contra la mujer. Al estar su contenido repleto de **palabras típicamente usadas para mencionar y caracterizar a ambos géneros** que han sido tachadas como sexistas han superando el umbral de clasificación y por lo tanto han desembocado en una categorización dentro de la etiqueta positiva. En la Figura 8.4 se adjunta un texto que refleja acontecimiento en el que términos comunes como *"mujeres"* y *"marido"* son

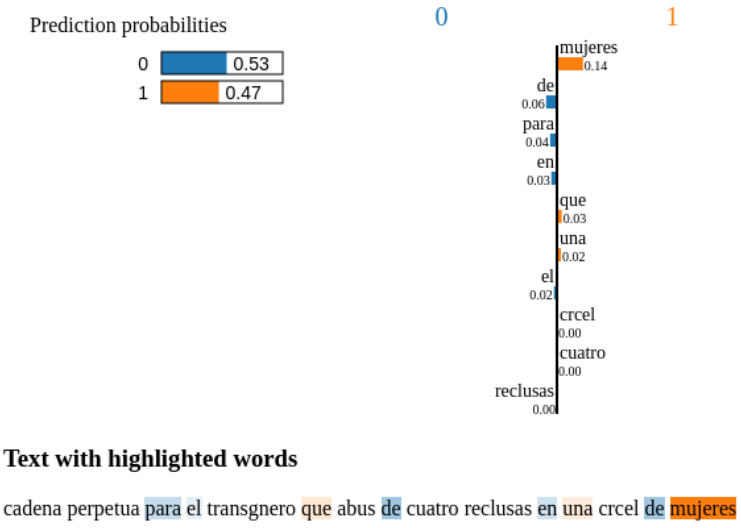


**Figura 8.3:** Análisis de explicabilidad por LIME en torno a un falso negativo producido por errores ortográficos.

identificados como sexistas cuando en realidad su contexto no puede ser considerado como tal. A pesar de los resultados este comportamiento se podría tomar como razonable ya que algoritmos clásicos como la Regresión Logística no alcanzan a formular un contexto en el que situar todos los vocablos antes de agruparlos en una u otra clase.

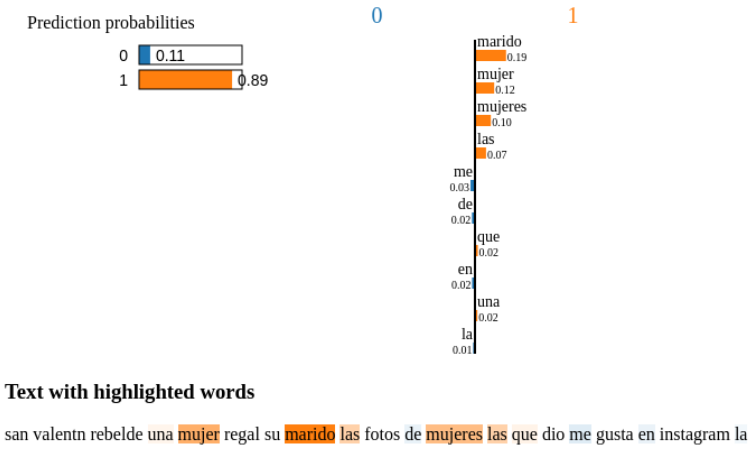


**Figura 8.4:** Análisis de explicabilidad por LIME en torno a un falso negativo producido por asignar una connotación sexista a menciones de los dos géneros.

### 8.1.3. Mejor modelo LSTM español

Este segundo modelo mejor considerado y más avanzado que el anterior se caracterizaba por haber cometido 392 falsos negativos y 210 falsos positivos. Después de aplicar LIME sobre el primer conjunto y estudiar sus características se confirma que las conclusiones extraídas son idénticas a los patrones vistos previamente, por lo que no existen nuevas aportaciones para este primer caso. Bien diferente ha sido el análisis de falsos positivos en los que, además de encontrar ejemplos coincidentes con los fenómenos descritos anteriormente, se han avistado nuevos que continuan desenmascarando las debilidades de una arquitectura LSTM. En la Figura 8.5 se sitúa un ejemplo complicado de corregir puesto que "*florero*"siendo uno de los conceptos evaluados como más sexista solamente alude al apellido de una mujer. Esta **falta de información** sobre la terminología que aparece en el documento produce como consecuencia un contexto incorrecto en el que el modelo auotmáticamente interpreta un mensaje contra la población femenina cuando no es cierto.

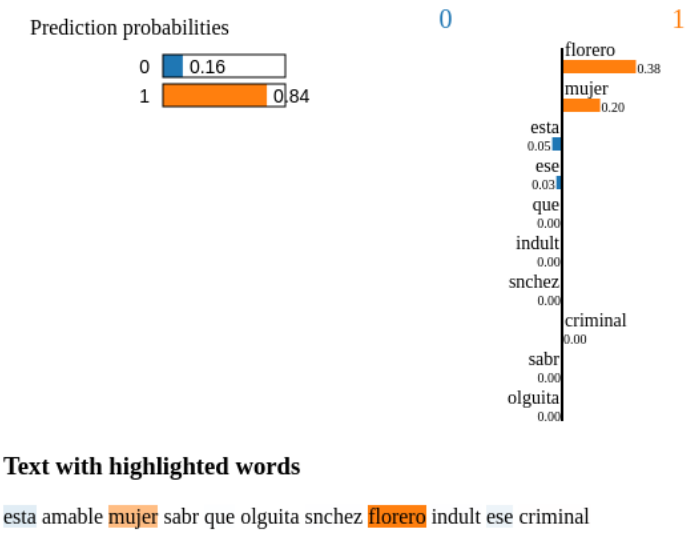


**Figura 8.5:** Análisis de explicabilidad por LIME en torno a un falso positivo producido por confundir un apellido con un adjetivo sexista.

Relacionado con el anterior ejemplo en la Figura 8.6 se demuestra cómo un sistema inteligente tampoco es capaz aún de identificar **expresiones hechas** compuestas por vocablos negativas y sexistas aunque el significado verdadero sea positivo. Se trata de otra muestra más de un tipo distinto de ironía que se postula como una debilidad incurable en una arquitectura LSTM con la mejor configuración encontrada de entre los experimentos realizados.

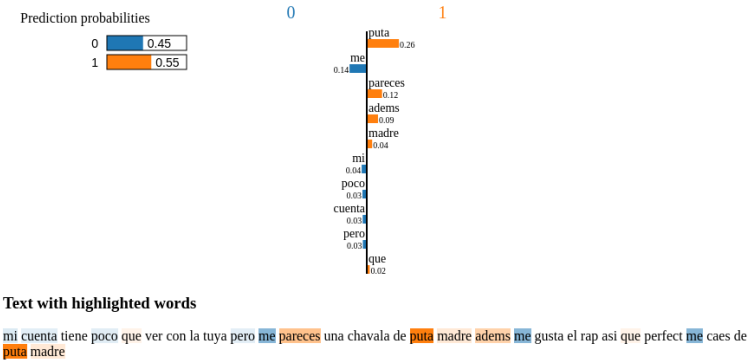


**Figura 8.6:** Análisis de explicabilidad por LIME en torno a un falso positivo producido por la no comprensión de una expresión hecha.

### 8.1.4. Mejor modelo BERT español

En esta última arquitectura se ha producido la mejor configuración de datos, entrenamiento y validación que ha resultado en un modelo con las mayores tasas de accuracy y área bajo la curva ROC, aunque con un total de 300 falsos negativos y 184 falsos positivos. Durante el análisis del primer subconjunto de falsos negativos se ha podido observar cómo este sistema es capaz de identificar **un mismo término como sexista en unos documentos y como no sexista en otros en base al contexto** de los mismos, por lo tanto su capacidad de generalización es superior a la de los otros modelos. No obstante, en todos los intervalos de confianza se sigue padeciendo las deficiencias destacadas en secciones anteriores tales como la **ironía**, que se visualiza en la Figura 8.7 en la que el clasificador percibe la negación de un término sexista de manera literal aunque en realidad la intención del usuario es menospreciar a una mujer.

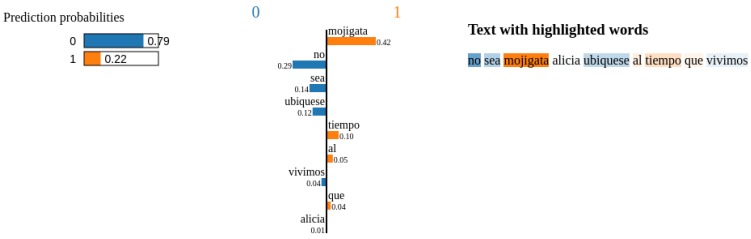


**Figura 8.7:** Análisis de explicabilidad por LIME en torno a un falso negativo producido por falta de comprensión de la ironía.

En aquellos rangos de confianza más moderados se aglutinan un conjunto de documentos que son realmente **complicados de entender por seres humanos** ya que su redacción e hilo conversacional son considerablemente pobres. Un ejemplo gráfico de este fenómeno se encuentra en la Figura 8.8

en la que particularmente he tenido que releer el contenido al menos un par de veces para hacerme una idea, aunque difusa, de lo que ha querido comentar un usuario. Asimismo también se aprecia cómo la terminología típica de ambos géneros como la palabra *"mujeres"* continua apareciendo con un elevado grado de sexismo que contribuye a la errónea clasificación de este documento. Mientras que por otro lado aparecen como no sexistas conceptos que deberían ser neutros como artículos y preposiciones.

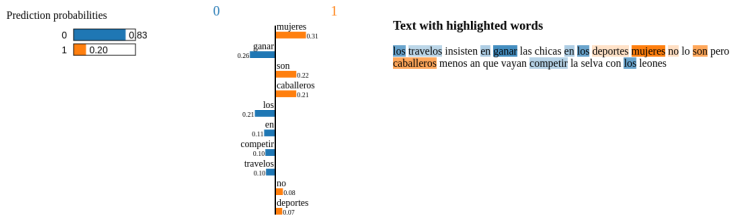


**Figura 8.8:** Análisis de explicabilidad por LIME en torno a un falso negativo producido por la cosificación sexista de terminología neutra y de género.

Por último en el estudio de falsos positivos se han descubierto unas pesquisas idénticas a las ya conocidas que son características de este subconjunto de muestras erróneamente clasificadas. Como ejemplos representativos se localiza la Figura 8.9 en la que se cosifica el vocablo *"guapa"* asignándole un altísimo grado sexista cuando a partir del contexto se demuestra que no hace referencia a ninguna mujer, es simplemente una **expresión hecha** con la que el usuario manifiesta su gusto por el modo de jugar que tiene una segunda persona a un videojuego particular. En este mismo texto también se presenta unas de las ventajas asociadas a este sistema BERT puesto que en anteriores documentos la palabra "*no*"ha sido interpretada como sexista, mientras que el mismo término en este texto ha sido ignorado presumiblemente porque no lo ha encontrado relevante para el análisis de su contenido.

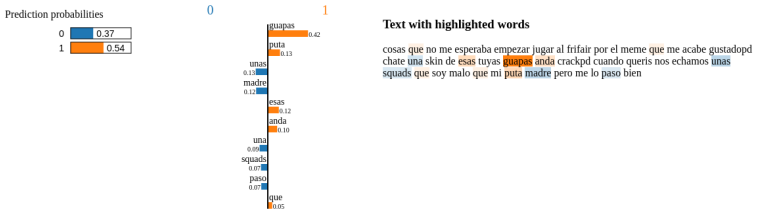


**Figura 8.9:** Análisis de explicabilidad por LIME en torno a un falso positivo producido por falta de comprensión de una expresión hecha.

Mientras que en la Figura 8.10 se puede observar un ejemplo de un intervalo moderado de confianza que ejemplifica perfectamente el **pobre grado de redacción que tienen ciertos textos** en los que hasta para una persona resulta complicado conocer qué significado tiene. No obstante, el modelo se ha dejado influenciar por las palabras que aluden al género masculino,

como "*hombre*", tachándolas de sexistas y produciendo una identificación incorrecta de la muestra de test. Al igual que en el caso anterior en este texto también se evidencia la capacidad del modelo BERT de asignar un valor diferente al mismo vocablo previamente destacado "*no*"modificando su rol a no sexista.

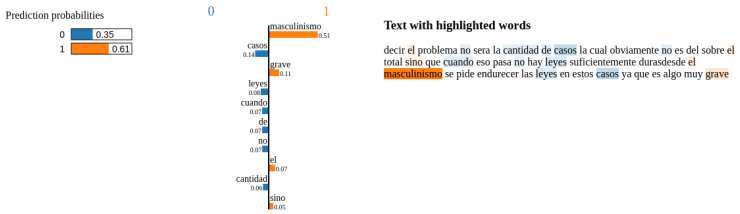


**Figura 8.10:** Análisis de explicabilidad por LIME en torno a un falso positivo producido por la cosificación de terminología de género y con pobre nivel de redacción.

### 8.1.5. Errores comunes entre modelos

Efectuando el procedimiento de experimentación comentado al comienzo de esta sección se han muestreado los falsos negativos y positivos comunes a la pareja de modelos de **Regresión Logística y LSTM** con el propósito de conocer si existen diferencias en las interpretaciones que han hecho ambos sistemas de los errores cometidos en común. Mientras que con los falsos positivos se ha observado un comportamiento semejante en el que prácticamente resulta general la asignación de un grado de sexismo a terminología de género, como las palabras *"mujeres"* y *"hombres"*, en los falsos negativos sí que se han visualizado discrepancias en el modo de interpretación con el que detectan si un concepto es o no sexista. Una muestra de este fenómeno se ejemplifica en la Figura 8.11 y Figura 8.12 en la que se aprecia que mientras el sistema clásico únicamente fundamenta su decisión en palabras irrelevantes como conjunciones y artículos, el clasificador LSTM considera especialmente sexista el término *"fea"* con el que se denigra el aspecto de una mujer. Por lo tanto como consecuencia se puede determinar que con una **arquitectura más avanzada y una codificación mediante embeddings se ha conseguido formular un contexto** con el que comprender el significado de cada uno de los vocablos para detectar las vejaciones que incluye un usuario en este comentario.

En el siguiente par de sistemas representado por **Regresión Logística y BERT** se debe considerar su inmensa disparidad al provenir de diferentes tipos de algoritmos, arquitecturas y configuraciones de datos. Por lo tanto, como no podía ser de otra forma tanto en los falsos negativos como positivos se han encontrado numerosas discrepancias entre los modos de funcionamiento e interpretación de estos dos sistemas con respecto a los

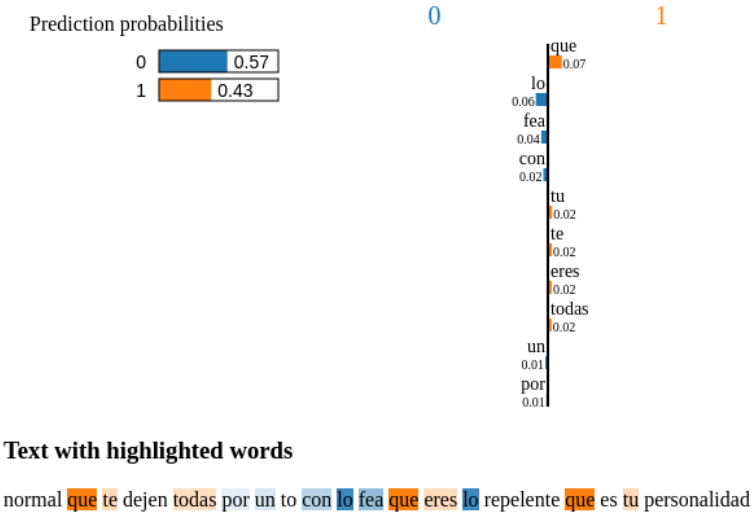


**Figura 8.11:** Comparación de la explicabilidad generada por LIME en torno a un falso negativo común desde la perspectiva del modelo de Regresión Logística producido por la cosificación de terminología neutra como preposiciones y artículos.

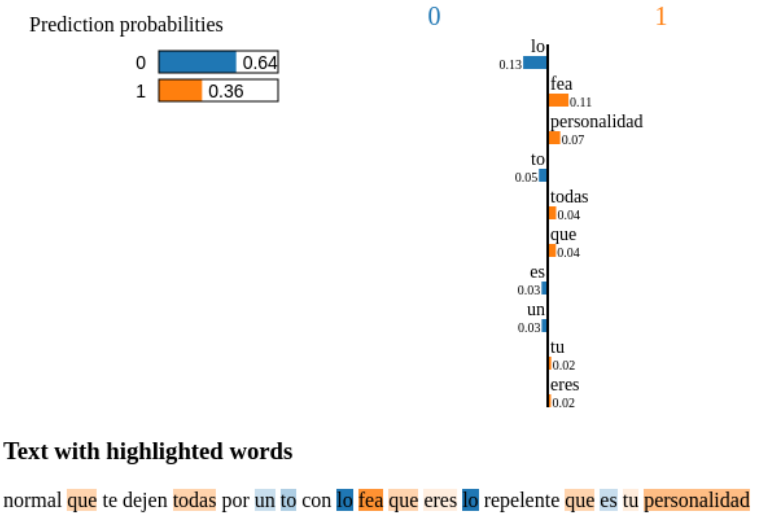


**Figura 8.12:** Comparación de la explicabilidad generada por LIME en torno a un falso negativo común desde la perspectiva del modelo LSTM producido por la cosificación de terminología neutra como preposiciones y artículos.

ejemplos seleccionados aleatoriamente del conjunto de test. Mientras que el clasificador de Regresión Logística generalmente se ha apoyado en **términos irrelevantes que no deberían tener ningún grado sexista o no sexista** asignado, tales como preposiciones, artículos, conjunciones, entre otros, el *transformer* ha fundamentado sus decisiones en entidades individuales que han sido analizadas presumiblemente gracias al contexto generado en base

a la totalidad de cada documento. A pesar de posicionarse como el sistema más complejo y avanzado de los empleados en este proyecto, persiste en los fallos destacados anteriormente consistentes en la incorrecta identificación de conceptos sexistas cuando han sido referenciados con otros motivos. Un claro ejemplo de este fenómeno se puede observar en la Figura 8.13 que muestra cómo el modelo de Regresión Logística ha resaltado vocablos que pertenecen a los tipos de palabras mencionados anteriormente además de a lenguaje referente al género femenino como es la palabra *"mujer"*. Por otro lado en la Figura 8.14 se observa una tendencia completamente distinta en la que el clasificador BERT se ha fijado mayormente en adjetivos y sustantivos, aunque no parece haber comprendido el significado del documento al tachar como sexistas algunos de ellos puesto que no tienen un sentido peyorativo hacia el colectivo femenino.

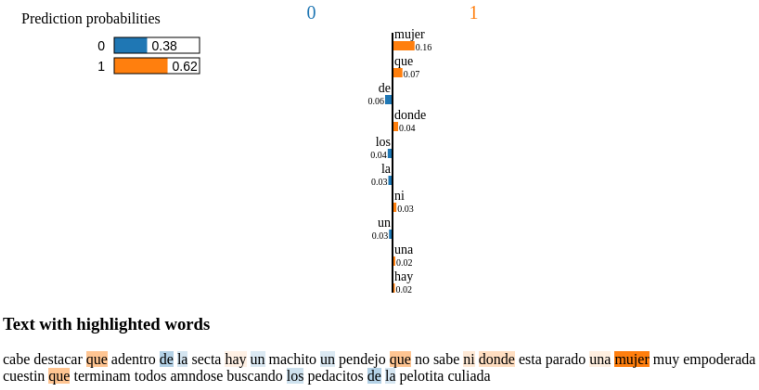


**Figura 8.13:** Comparación de la explicabilidad generada por LIME en torno a un falso positivo común desde la perspectiva del modelo de Regresión Logística producido por la cosificación de terminología neutra y de género.

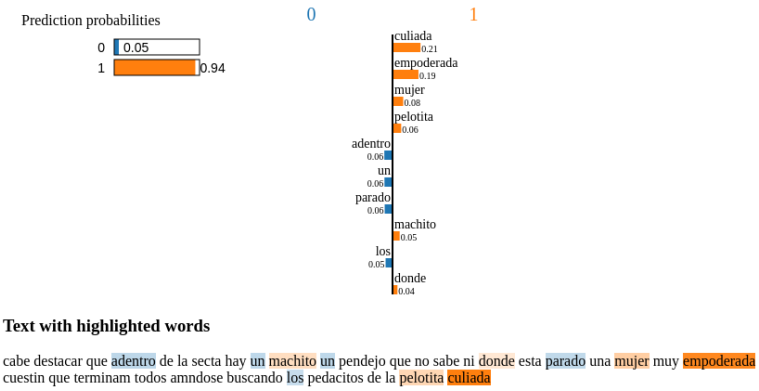


**Figura 8.14:** Comparación de la explicabilidad generada por LIME en torno a un falso positivo común desde la perspectiva del modelo BERT producido por la incorrecta clasificación de terminología de género y adjetivos como sexista.

Finaliza esta subsección con el análisis de los falsos negativos y positivos comunes a la pareja de modelos **LSTM y BERT** durante el modelado de documentos en español. Tras observar detalladamente las muestras seleccionadas de cada conjunto se ha podido determinar que la principal conclusión que resume la diferencia fundamental entre el comportamiento de ambos sistemas. Esta reside en que el primer clasificador considera demasiada terminología neutra, como determinantes, artículos, conjunciones, especialmente cuando no existe información que considere relevante en el documento, mientras que el *transformer* no le importa no destacar demasiados vocablos si los restantes no le parecen importantes. Por lo tanto parece que la **arquitectura LSTM cae en la tentación de sobrejustificar sus decisiones tomando conceptos instrascendentes** sin ningún grado sexista mientras que el sistema BERT solo resalta aquellas entidades que piensa son útiles para la asignación de una etiqueta aunque su volumen sea muy inferior. En la Figura 8.15 y 8.16 se encuentra un ejemplo gráfico de las cualidades explicadas de cada arquitectura. Mientras que el modelo LSTM colorea la mayor parte de palabras contenidas en el documento, el clasificador BERT apenas destaca algunas de ellas. La segunda discrepancia que pone de manifiesto cuán distintos son sus respectivos modos de razonamiento se apoya en que el sistema LSTM cosifica terminología de género como "*chica*", *"hombre"*, cuando por el contexto conocemos que no se está cometiendo ninguna agresión verbal contra la mujer. Sin embargo esta situación no ocurre en el modelo BERT aunque sí se observa una tasa moderada de categorización no sexista asociada a terminología que no debería ser analizada para la toma de decisiones.

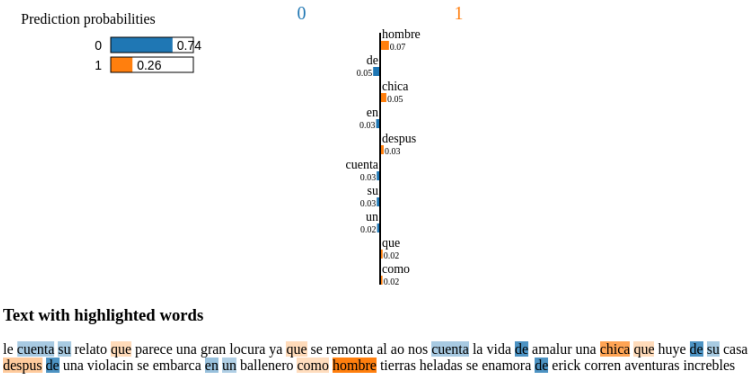


**Figura 8.15:** Comparación de la explicabilidad generada por LIME en torno a un falso negativo común desde la perspectiva del modelo de LSTM producido por la sobrevaloración de términos.

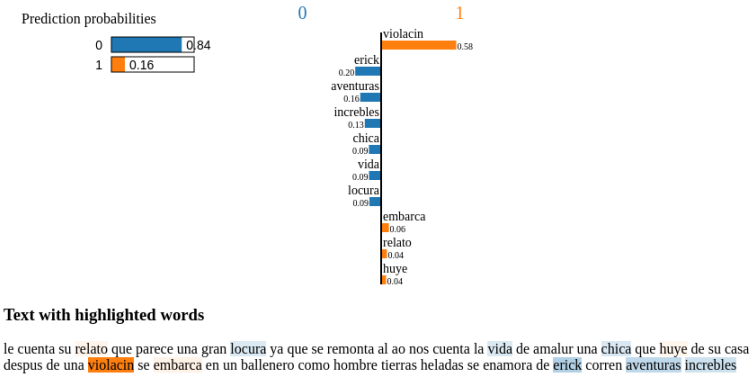


**Figura 8.16:** Comparación de la explicabilidad generada por LIME en torno a un falso negativo común desde la perspectiva del modelo BERT producido por la falta de valoración de términos relevantes.

### 8.1.6. Resumen y conclusiones

Por último en este apartado se pretende resumir y destacar las conclusiones extraídas de la aplicación de la técnica de explicabilidad local LIME sobre los tres mejores modelos encontrados considerando un subconjunto de falsos negativos y positivos con el propósito de conocer las posibles causas y teorías que explican tales fallos cometidos de manera individual y común a ciertas parejas de clasificadores. Estas pesquisas se encuentran basadas en el modelado de documentos en español aunque son idénticas a los textos ingleses por lo que son totalmente válidas para ambos idiomas.

- El primer fenómeno generalizado a todas las arquitecturas se concentra en la asignación de un **grado no sexista a terminología que debería considerarse neutra**, como preposiciones, artículos, determinantes, sustantivos y verbos sin relación con la temática. Su consecuencia directa desemboca en la categorización errónea de documentos sexistas dentro de la clase negativa causando los falsos negativos contabilizados. Es especialmente acusado en el modelo de Regresión Logística y el clasificador LSTM presumiblemente debido a su menor capacidad de generar un contexto preciso que represente la idea que refleja cada documento. Sin embargo, se debe considerar la ventaja con la que cuenta el *transformer* BERT al estar preentrenado mientras que los dos sistemas anteriores han sido construidos desde cero. Por lo tanto gracias a su preparación previa y a su arquitectura más compleja, su habilidad de análisis y predicción es superior a la de sus competidores.

- Una segunda situación común a todos los modelos consiste en **etiquetar como sexistas sustantivos y adjentivos típicos de ambos**

**géneros** aunque el significado del contenido no conlleve ninguna connotación sexista. En este caso se producen los falsos positivos, estando su inmensa mayoría relacionados con reivindicaciones, protestas y noticias a favor de la igualdad de género. No obstante el uso de conceptos de género, insultos, palabras mal sonantes, entre otros términos provoca el aumento de la probabilidad de pertenencia a la clase positiva desviando su correcta clasificación. En muchos de los documentos tomados como ejemplos el principal problema radia en la **falta de comprensión de los documentos especialmente por la introducción de ironía, dobles significados y expresiones hechas** que confunden a todos los modelos generados, si bien el *transformer* BERT es el que consigue unos mejores resultados por una contextualización superior a sus modelos compañeros.

- En los intervalos moderados e inferiores de confianza se ha visualizado una tendencia que aumenta el número de textos caracterizados por una **pésima calidad ortográfica, gramatical, semántica y la existencia de hashtags** que incluso dificulta su comprensión por seres humanos. Estos ejemplos son obviamente complicadísimos de analizar e identificar automáticamente, por lo tanto se podrían establecer unos **estándares mínimos de calidad** para no considerar documentos de tales cualidades con los que medir la bondad de los clasificadores.

- En las comparaciones de parejas de sistemas inteligentes se ha concluido que los clasificadores de **Regresión Logística y LSTM disponen de un comportamiento semejante en el tratamiento de falsos positivos**. Por el contrario en los falsos negativos el modelo con arquitectura LSTM es capaz de etiquetar terminología más relevante para la asignación de una clase u otra, mientras que el sistema más clásico parece realzar conceptos sin ningún tipo de lógica. Me ha sorprendido malamente esta similitud entre ambos modelos ya que **mis expectativas eran mayores con las arquitecturas LSTM** y sin embargo, gracias a los análisis de explicabilidad nos hemos percatado que coincide en demasiados puntos con un algoritmo clásico de Aprendizaje Automático. Como se anunció anteriormente, puede que necesite complementos adicionales como **mecanismos de atención** que ayuden a concentrar sus esfuerzos en el análisis de términos relevantes reduciendo su desviación hacia conceptos de naturaleza irrelevantes o fuera de la temática sexista que se aborda en este proyecto.

- Por otro lado el modelo LSTM también comparte ciertas tendencias localizadas en el *transformer* BERT que le otorga el potencial suficiente como para situarse en una posición intermedia entre un algoritmo clásico de Regresión Logística y una arquitectura avanzada y especializada de tareas PLN como es el modelo BERT. A pesar de ello este últi-

mo clasificador generalmente solo destaca entidades individuales tales como sustantivos y adjetivos que suelen aportar más información útil para el estudio de una muestra, mientras que el **clasificador LSTM en ocasiones resalta demasiados términos** independientemente de su naturaleza pecando de sobreanalizar el contenido introduciendo demasiado **ruido** en su etiquetado. De nuevo en esta pesquisa se pone de manifiesto la necesidad de complementar su conducta con algún mecanismo que le ayude a focalizarse únicamente sobre los conceptos más apropiados.

## 8.2. Contrafactuales

El segundo procedimiento de explicabilidad que se ha escogido para su aplicación en este proyecto consiste en la generación de contrafactuales, muestras sintéticas producidas en base a determinados documentos tomados como ejemplos a los que se les efectúa ciertas transformaciones. Estas ya han sido mencionadas al comienzo del mismo capítulo y mientras que la eliminación secuencial de términos es una implementación propia, en la negación de verbos se ha hecho uso del modelo de generación de lenguaje GPT-3. Nuevamente, esta técnica puede ser elaborada sobre cualquier tipo de documentos aunque el objetivo que se persigue es el de comprender qué posibles causas pueden entrañar los fallos cometidos por los modelos. Por lo tanto, la experimentación diseñada es idéntica a la anterior realizada en la que se han utilizado un subconjunto de falsos negativos y positivos de cada intervalo de confianza disponible en cada uno de los tres mejores clasificadores encontrados. Asimismo la estructura de esta nueva sección es compartida por el apartado previo en el que se realizará un análisis acerca de los contrafactuales fabricados, los resultados obtenidos y las conclusiones y mejoras que surjan dependiendo de la comparación entre los resultados esperados y los reales.

### 8.2.1. Procedimiento de experimentación

Tras seleccionar aleatoriamente **diez documentos de entre los falsos negativos y otros diez de los positivos por intervalo de confianza** en cada una de las arquitecturas cntempladas, cada ejemplo será sometido a las dos operaciones especificadas al comienzo de este capítulo para generar dos conjuntos de contrafactuales con los que estudiar la clasificación que realizan los sistemas a nivel individual. Mediante la supresión secuencial de términos individuales el objetivo que se persigue consiste en analizar la ausencia de qué conceptos ha sido vital para la reclasificación de la muestra en su etiqueta correcta. Por otro lado con la negación de verbos el propósito que se

desea obtener trata de estudiar el nivel de modificación que ha alcanzado cada muestra con respecto al significado original así como el grado de contribución que aporta cada verbo negado en la identificación de la clase correcta. Una vez se encuentren todos los contrafactuales disponibles, en la etapa final cada modelo se ha enfrentado a la **clasificación de los contrafactuales fabricados** destacando solamente aquellos ejemplos sintéticos que hayan conseguido reconducir el comportamiento erróneo realizado con las muestras originales. A continuación se han realizado inspecciones visuales con las que intentar descubrir si existe un determinado tipo de terminología cuya desaparación/negación haya resultado ser beneficiosa. Para ello se pretende aplicar una vez más la técnica *LIME* con la que apreciar la categorización de los conceptos en sexistas o no sexistas que hayan ayudado a determinar la etiqueta resultante.

### 8.2.2. Mejor modelo de Regresión Logística español

Recordemos que los falsos negativos y positivos del mejor modelo producido mediante Regresión Logística estaban distribuidos en los tres máximos intervalos de confianza. El primer fenómeno destacable es que los contrafactuales procedentes de los falsos negativos son menos voluminosos puesto que únicamente se sitúan en el rango moderado de confianza. Mientras que los ejemplos sintéticos de los falsos positivos son más numerosos al localizarse en todos los intervalos de confianza, aunque siendo mayoritario el inferior. Por lo tanto por un lado se concluye que la **reconducción de los textos sexistas no identificados como tal se intuye más complicada** con la fabricación de contrafactuales en comparación con los falsos positivos, en los que existe una mayor población con distintos niveles de confianza. Una de las posibles teorías explicativas de este hecho puede residir en la categorización sexista que siempre realiza de la terminología asociada a ambos géneros, cuya frecuencia de aparición es prácticamente generalizada y por lo tanto dificulta el estudio y reconocimiento de patrones.

Durante el procedimiento efectuado sobre la primera operación de supresión secuencial de términos individuales, se ha concluido que la **eliminación de la terminología resaltada en la sección previa con LIME** ha resultado ser muy beneficiosa para la reclasificación correcta de varias muestras. En los falsos negativos se ha frenado la contrarrestación de la probabilidad sexista por el borrado de conceptos neutros, como artículos, preposiciones y conjunciones, mal identificados como no sexistas. Mientras que en los falsos positivos ha sido la ausencia de palabras referentes a ambos géneros, como las muchas veces mencionadas *"mujer"* y *"hombre"*, así como insultos, adjetivos y partes del cuerpo las que han propiciado la corrección de ciertos documentos no sexistas. Si bien una solución a los falsos negativos podría plantear la introducción de una nueva etapa en el procesamiento de textos

que eliminase las palabras de parada conocidas por su composición mayoritaria de la terminología que provoca estos errores, a continuación se expone un ejemplo con el que no se puede lidiar. En la Figura 8.17 se encuentra el texto original mientras que en la Figura 8.18 se localiza un contrafactual generado a partir de él en el que se ha eliminado el concepto "*embarazada*", un cambio que ha sido suficiente como para realizar un correcto etiquetado. No obstante su clasificación en la categoría positiva parece fundamentarse solamente en el término "*mujer*", por lo tanto este comportamiento tampoco sería adecuado para que el modelo de Regresión Logística pudiese posicionarse como una solución en la detección de sexismo escrito.

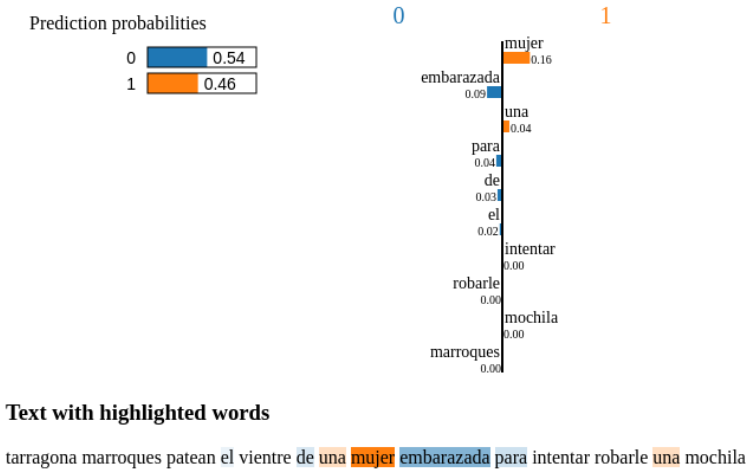


**Figura 8.17:** Explicabilidad generada por LIME acerca de un falso negativo producido por el mejor modelo de Regresión Logística por su falta de comprensión del documento.

Desafortunadamente no se han podido obtener resultados positivos en la generación de contrafactuales mediante la negación de verbos ya que en esta operación no se ha conseguido reclasificar correctamente ninguna muestra.

### 8.2.3. Mejor modelo LSTM español

Con una distribución de intervalos de confianza idéntica a la explicada en la sección previa, las conclusiones detalladas son de igual modo aplicables a los contrafactuales y los resultados obtenidos por el mejor modelo LSTM encontrado en el modelado de documentos españoles. En los falsos negativos la supresión de terminología neutra, como artículos, preposiciones y conjunciones, ha sido beneficiosa para interrumpir la reducción de la probabilidad de pertenencia a la clase positiva. En oposición a los falsos positivos en los que el borrado de palabras asociadas a ambos géneros, insultos, partes del cuerpo y prendas femeninas ha propiciado la reconducción

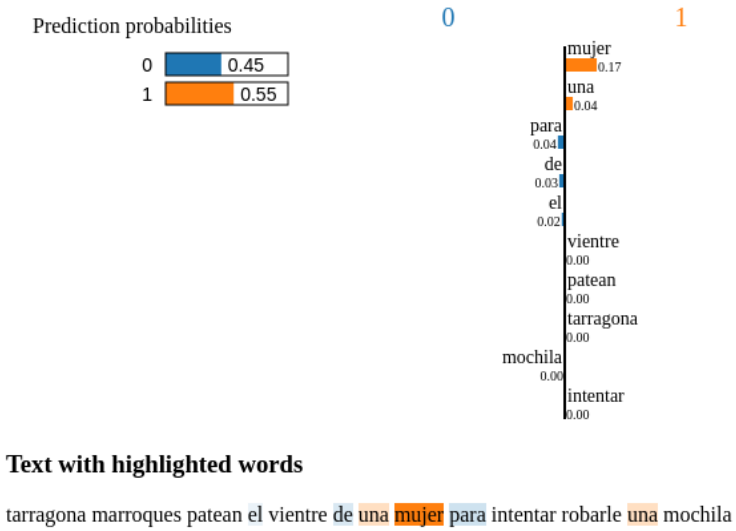


**Figura 8.18:** Explicabilidad generada por LIME acerca del contrafactual creado en base al falso negativo de la Figura 7.17 cuya clase ha sido asignada correctamente por el mejor modelo de Regresión Logística gracias a la supresión de una palabra concreta.

de estas muestras hacia su verdadera clase no sexista. Sin embargo, debido a que el planteamiento ideado en el uso de GPT-3 para la negación de verbos ha sido establecido como semiautomático, me he podido percatar de que al enviar esta tarea, en ocasiones además **parafrasea el documento realizando demasiados cambios** hasta que pierde su esencia original. Un ejemplo ilustrativo de esta situación se puede visualizar en la Figura 8.19, en la que se localiza el texto original cuyo único concepto resaltado ha sido tachado de sexista aunque por el contexto del contenido no parece ser esa la intención del usuario. Mientras que en la Figura 8.20 se encuentra el contrafactual fabricado por GPT-3 durante la negación de verbos en el que se aprecia la modificación prácticamente completa de la terminología del documento por otras palabras que poco tienen que ver con el ejemplo original. Por lo tanto aunque se ha conseguido su reclasificación correcta bajo mi criterio este contrafactual no es válido puesto que apenas guarda relación con la muestra proporcionada.

### 8.2.4. Mejor modelo BERT español

Como en los análisis anteriores los contrafactuales fundamentados en los falsos negativos y positivos de la propia arquitectura BERT demuestran unas inclinaciones semejantes, habiendose corregido un pequeño volumen de ellos pertenecientes también a todos los intervalos de confianza involucrados, desde el máximo al moderado. En el primer subconjunto de errores se han

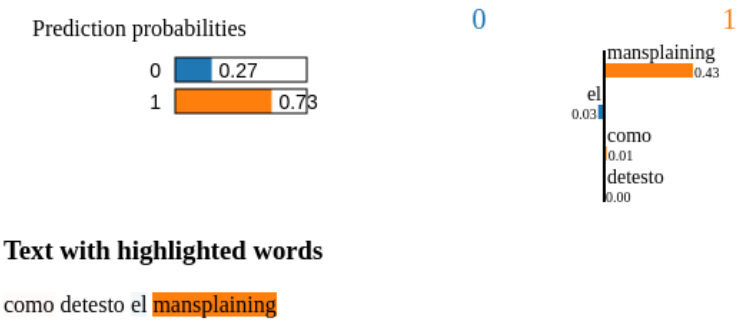


**Figura 8.19:** Explicabilidad generada por LIME acerca de un falso positivo producido por el mejor modelo LSTM por su falta de comprensión del contexto del documento.

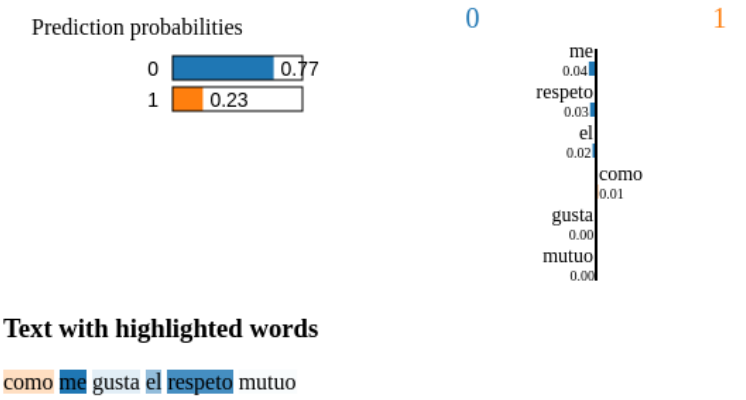


**Figura 8.20:** Explicabilidad generada por LIME acerca del contrafactual basado en un falso positivo que ha sido correctamente clasificado por el mejor modelo LSTM gracias al parafraseo efectuado por GPT-3.

visualizado unas prácticas similares a las descubiertas previamente, destacando especialmente la **eliminación de adjetivos y la transformación de verbos negativos en positivos** al suprimir la palabra *"no"* como las principales técnicas más propicias para este modelado. Por lo tanto como conclusión se puede determinar que la arquitectura BERT le asocia una mayor relevancia a este tipo de terminología y su aparición o desaparición puede causar efectos adversos en el etiquetado de muestras. Este hecho le diferencia aún más de los otros dos sistemas inteligentes por ser capaz de centrarse en los conceptos que más información pueden aportar en un documento para generar un contexto y una idea nítida de la temática que tratan. Por otro lado continúa la escasa tasa de resultados proporcionados por el uso de GPT-3 en la conversión de verbos, aplicando un procedimiento de parafraseo más intenso aunque beneficioso con el que le aporta más sentido a los documentos facilitando su etiquetado. A diferencia de los contrafactuales evaluados por

el modelo LSTM, en este caso **no se han obtenido ejemplos sintéticos tan alterados** como el mostrado en la sección previa. En los falsos positivos se han concluido unas pesquisas similares a las conocidas hasta el momento, destacando de nuevo la supresión de adjetivos tradicionamente utilizados como "piropos"hacia el colectivo femenino aunque realmente empleados como expresiones hechas con un significado nada sexista. Con el uso de GPT-3 para contraponer los verbos existentes se han apreciado elevados niveles de correcciones gramaticales, división de hashtags en sus palabras individuales y el parafraseo ejecutado sobre las muestras proporcionadas como ejemplo que de nuevo han conseguido otorgarle una mayor claridad a los contenidos siendo más sencillo el análisis automático por parte del mejor modelo BERT encontrado. En la Figura 8.21 se muestra el contenido original de un falso positivo en el que el concepto mayormente destacado como sexista también aparece como tal en el contrafactual generado y visualizado en la Figura 8.22, aunque en combinación con una serie de términos clasificados como no sexistas que contrarrestan el fallo que comete el modelo. Si bien se aprecia que ambos textos son muy similares entre sí, mientras que con el primero existe cierta dificultad de comprensión con el segundo es más sencillo determinar su significado, por lo que una identificación automática es más plausible.

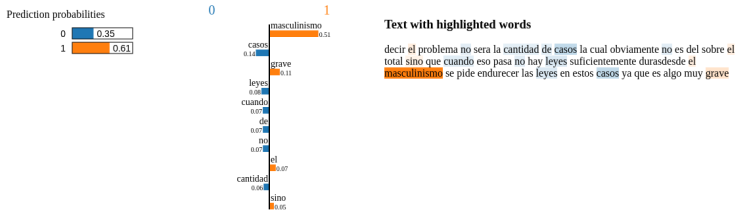


**Figura 8.21:** Explicabilidad generada por LIME acerca de un falso positivo producido por el mejor modelo BERT por su falta de comprensión del contexto del documento.

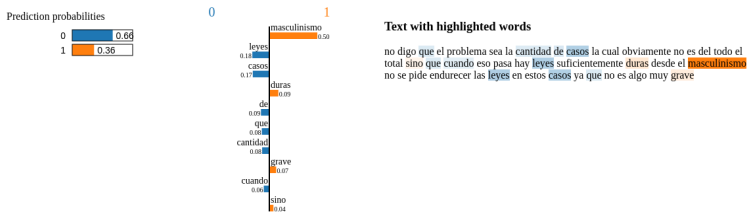


**Figura 8.22:** Explicabilidad generada por LIME acerca del contrafactual basado en un falso positivo que ha sido correctamente clasificado por el mejor modelo BERT gracias al parafraseo y a la negación de verbos efectuada por GPT-3.

### 8.2.5. Lecciones aprendidas y posibles mejoras

Antes de finalizar el capítulo se procede a resumir las deducciones fundamentadas en la generación y análisis de contrafactuales intentando relacionarlas con las conclusiones extraídas durante la aplicación exclusiva de LIME en la sección previa. De esta forma, gracias a la generación de muestras sintéticas y al estudio acerca de qué modificaciones conllevaron una reclasificación correcta, se pueden proponer algunas mejoras adicionales para aumentar la capacidad de predicción de los modelos.

- Los clasificadores de Regresión Logística y LSTM se han visto beneficiados por la supresión de preposiciones, conjunciones y artículos, entre otra terminología neutra debido a su incorrecta categorización como no sexistas, especialmente presente en los falsos negativos. Por lo tanto **eliminando estos conceptos popularmente conocidos como palabras de parada** se puede reducir la probabilidad de desviación hacia la clase errónea. Este fenómeno fue el primero expuesto durante las resoluciones detalladas anteriormente como uno de los principales problemas generales aunque particularmente notorios en sendos algoritmos. No obstante, tras los múltiples análisis realizados sobre sus conductas cabe destacar que esta técnica **no se plantea como una solución definitiva** puesto que también se han visualizado diversos ejemplos en los que se identificaban como no sexistas entidadesde distinto estilo sin relación con la temática abordada. Asimismo con esta operación se resquebrajaría la estructura de los documentos pudiendo alterar la intención y el significado que deseó proporcionarle el usuario originalmente. Estas deficiencias parecen ser incorregibles en ambos sistemas puesto que la raíz que las causa se fundamenta en la falta de comprensión del texto y del significado que el usuario ha pretendido otorgar con su redacción.

- El segundo inconveniente avistado fue la cosificación de terminología de género como sexista, siendo palabras tales como *"mujer"* y *"hombre"* las más frecuentes. Si bien su desaparición ha sido ventajosa produciendo contrafactuales capaces de ser correctamente etiquetados, la erradicación de esta conducta parece no tener una solución viable. De nuevo su incapacidad de generar un contexto que le ayude a identificar tales conceptos dependiendo de la intención del usuario pone de manifiesto la imposibilidad de utilizar algoritmos y arquitecturas tan sencillas como la Regresión Logística y la LSTM simple. Este fenómeno se encuentra presente aunque de forma más reducida en el *transformer* BERT que si bien falla en ocasiones, sí que se le ha visto hábil en la adaptación de cada vocablo a su contexto, mejorando así las predicciones realizadas y demostrando una capacidad de generaliza-

ción superior. Por lo tanto para solventar esta problemática la única recomendación consiste en utilizar **modelos más complejos y especializados en tareas NLP**.

- La operación con un mayor porcentaje de resultados ha sido la eliminación de términos secuencialmente debido a que la producción de contrafactuales es mucho mayor que la segunda planteada en la negación de verbos, que solamente proporcionaba un contrafactual por muestra. Si bien la primera técnica se caracteriza por tener la ventaja de poder ser automática y como consecuencia más veloz en términos de aplicación y análisis, el uso de GPT-3 también ha proporcionado pesquisas muy importantes. Aquella más fundamental se apoya en el impacto efectuado sobre la habilidad de **clasificación por parte de los modelos después de transformar los documentos** otorgándoles un contenido más legible, comprensible y correctamente redactado en términos ortográficos, sintácticos y semánticos. Como consecuencia el uso de este tipo de sistemas generadores de lenguajes se postula como una solución a la pobreza característica de los mensajes procedentes de redes sociales que tanto dificultan su reconocimiento automático. No obstante, o bien se le introducen **límites con los que intentar controlar el número y el tipo de modificaciones** que realizan sobre los ejemplos proporcionados, o sus resultados deberán ser comprobados manualmente cualificando esta fase como semiautomática con la correspondiente inversión de recursos temporales y personales que conlleva.

- Debido a que su naturaleza y composición se encuentra sumamente alejada de sus modelos compañeros, el sistema BERT requiere de otro tipo de soluciones adicionales con las que mejorar su habilidad de construcción de contextos, precisando aún más el propósito con el que cada autor ha utilizado los conceptos pertenecientes a los documentos. Este hecho se justifica con la influencia que ha ejercido la eliminación de adjetivos, la transformación de verbos y el parafraseo de textos en la reconducción hacia la categorización correcta de un volumen considerable de falsos negativos y positivos. Por lo tanto para mejorar su rendimiento apostaría por **aumentar la representatividad de los ejemplos, generando nuevas muestras sintéticas que reflejasen los distintos significados** que pueden asociarse a los vocablos que más frecuentemente aparecen en la temática sexista que aborda este proyecto. Así dispondrá de datos suficientes con los que extraer los patrones necesarios para identificar en cada documento el verdadero significado que tiene cada palabra en lugar de únicamente considerar el conocido por defecto, siendo así menos vulnerable a la ironía y las expresiones hechas.

# Capítulo 9

# Conclusiones y trabajo futuro

Para finalizar la memoria de este proyecto en este último capítulo se presentan las conclusiones finales que se han podido extraer a lo largo de su desarrollo, aportando también tanto reflexiones personales como futuras líneas de investigación con las que continuar mejorando este trabajo. En primer lugar se confirma una teoría acerca de los beneficios que conllevaría el uso de **embeddings preentrenados** en la codificación de documentos. Y es que tal y como se ha percibido en las primeras experimentaciones, con su uso se ha logrado reducir el fenómeno del sobreaprendizaje como ninguna otra técnica lo ha hecho. Sin embargo, un buen componente de manera aislada no es sinónimo de un buen comportamiento, sino que necesita estar integrado en un entorno con competencias adicionales para lograr un buen rendimiento. Este hecho se ha podido observar al introducir esta metodología en arquitecturas más avanzadas como LSTM y BERT, en las que sí ha resultado ser más beneficiosa que con un algoritmo clásico como la Regresión Logística.

A continuación se confirma otra hipótesis planteada acerca del buen funcionamiento que aportan las arquitecturas de Aprendizaje Profundo, siendo especialmente notable el rendimiento de los *transformers* BERT. Se podría determinar que han sido los únicos sistemas inteligentes capaces de **obtener provecho de las técnicas de procesamiento y aumento de textos** utilizadas para proporcionar las métricas de validación más elevadas. Para ello ha sido un requisito indispensable la etapa de hiperparametrización y *fine tuning* que destacan por la ralentización máxima del aprendizaje y la disminución del número de datos por lote. Aunque haya conllevado un aumento importante de recursos computacionales y temporales, gracias a su configuración minuciosa sus múltiples **mecanismos de atención** han podido extraer un mayor número de patrones con los que enfrentar la clasificación

de muestras desconocidas con garantías de calidad. Estos mecanismos quizás han sido los que han marcado la diferencia con respecto a las arquitecturas LSTM, de las que personalmente esperaba mejores resultados. Sin embargo, dadas las pesquisas halladas durante el capítulo de explicabilidad, se han podido observar deficiencias claras en su comportamiento, como el sobreanálisis que realizan destacando demasiados términos cuando en realidad son únicamente unos pocos los que suministran información útil. No obstante, una lección importante que se ha aprendido de esta tipología de redes se apoya en el lema de que una arquitectura más compleja no siempre causa un mejor rendimiento. Por lo tanto cuando en el modelado no se obtengan soluciones de calidad, quizás es preferible investigar acerca de componentes adicionales más avanzados que puedan añadir nuevas funcionalidades al clasificador, en lugar de simplemente insertar más capas y más neuronas.

Por otro lado gracias al análisis exploratorio de datos hemos podido adelantarnos a un suceso que ha influenciado negativamente en la búsqueda de soluciones eficaces. Tras conocer la dificultad que entraña la identificación automática de ciertas categorías sexistas por sus **contenidos irónicos, repletos de expresiones hechas y dobles intenciones**, se ha podido confirmar su futuro impacto en el modelado gracias a la introducción de un modelo preentrenado especializado en detectar emociones. Cuando la mayoría de los documentos pertenecientes a estas complicadas clases han sido amparados bajo **emociones positivas** como *joy* o *love*, hemos comprendido que probablemente la confusión causada en este clasificador también estaría presente en nuestros modelos generados. Adicionalmente a los análisis de explicabilidad propios, parece que el enfoque de la aplicación de técnicas de explicabilidad con una distinción entre intervalos de confianza ha sido de mucha utilidad. Por un lado se ha demostrado que los errores cometidos bajo un paraguas de alta seguridad apenas han podido ser corregidos, a diferencia de aquellos pertenecientes en rangos moderados y bajos de confianza en los que sí se ha conseguido una reclasificación correcta más voluminosa. Gracias a LIME se ha identificado que una posible causa de la existencia de falsos negativos es la consideración de **terminología neutra como no sexista**, un fenómeno que se ha ido reduciendo conforme se aumentaba la complejidad de la arquitectura hasta llegar a los *transformers* donde este comportamiento prácticamente había desaparecido. Mientras que por otro lado los falsos positivos han sido originados como consecuencia de la cosificación de conceptos referentes a ambos géneros, partes del cuerpo y vestimenta que no eran empleados con una connotación sexista aunque los modelos lo creyesen así. En este caso la **falta de contexto o la comprensión parcial del significado** de los documentos ha podido ser el detonante generalizado a todas las arquitecturas planteadas, con un impacto menor en los sistemas BERT. Por el contrario, en los intervalos moderados y bajos de confianza las principales teorías se fundamentan en el elevado número de errores

ortográficos y hashtags que han dificultado su codificación, perdiendo una información valiosa que no ha podido ser empleada durante la clasificación. Con la extracción de estas conclusiones se ha quedado demostrado el notorio potencial que esconde la técnica LIME en generar explicabilidad resaltando la terminología en la que se fundamentan las respuestas proporcionadas por los clasificadores, quedando patentes los defectos que desembocan en la creación de falsos negativos y positivos.

Una situación similar ha ocurrido con la producción de contrafactuales, siendo los falsos negativos y positivos reconducidos hacia sus correctas clases mayoritariamente pertenecientes a intervalos moderados y bajos de confianza. Principalmente el foco se ha originado en la primera operación de supresión secuencial de términos debido al enorme volumen de muestras sintéticas que ha generado, en comparación con la segunda consistente en la negación de verbos que únicamente producía un contrafactual por documento empleando GPT-3 para la tarea. Mientras que en el primer caso se ha observado la reclasificación de contrafactuales fabricados por la ausencia de esta terminología neutra considerada como no sexista y de palabras referentes a los géneros, con la segunda operación también se han entrevisto conclusiones muy interesantes. En primer lugar se ha observado el potencial que tiene GPT-3 de corregir las faltas de ortografía, dividir los hashtags en términos individuales y parafrasear los documentos otorgándole un mejor hilo semántico que ha favorecido la comprensión y reconocimiento por los modelos generados. Adicionalmente gracias a la negación de los verbos se ha podido clarificar el sentido que le han brindado los usuarios a sus mensajes facilitando también la identificación de su correspondiente clase. Independientemente de los resultados proporcionados, con esta experimentación se ha alcanzado un nivel de interpretabilidad de los modelos generados bastante preciso puesto que hemos sido capaces de identificar sus modos de razonamiento, discrepancias y similitudes entre ellos, además de posibles soluciones a priori para intentar paliar su impacto negativo en el rendimiento.

Como trabajo futuro se plantean diversas líneas de investigación relacionadas tanto con el procesamiento de datos, el modelado y la explicabilidad. En el primer caso, dadas las cuantiosas habilidades del modelo generador de lenguaje GPT-3 se podría intentar llevar a cabo un tratamiento de los documentos más exhaustivo e inteligente, intentando eliminar las imperfecciones que contienen mejorando así su nivel de legibilidad semántica con el fin de aumentar la calidad y cantidad del conjunto de entrenamiento. Por otro lado en el modelado queda pendiente la introducción de mecanismos de atención en las arquitecturas LSTM para comprobar si estos componentes pueden ayudar a mejorar su capacidad de predicción como ha quedado patente en los *transformers* BERT. Mientras que también se puede idear la aplicación de otro tipo de técnicas de explicabilidad, como por ejemplo globales, con las que proporcionar enfoques adicionales acerca del funcionamiento de mode-

los tan complejos como los producidos en este proyecto. Toda la orientación realizada y mencionada igualmente podría ser aplicada al segundo problema que aborda el conjunto de datos *EXIST* en el reconocimiento automático de las distintas categorías sexistas, un paso más allá acerca de identificar qué tipología sexista aparece un texto clasificado como positivo.

# Apéndice A

# Hiperparametrización de modelos avanzados

En este apéndice se encuentran las secciones relativas a la búsqueda de las mejores configuraciones relativas a las experimentaciones con arquitecturas LSTM y *transformers* BERT. Mientras que en el primer caso se han considerado aspectos tales como la codificación de documentos en word embeddings, la composición de la arquitectura y el volumen de datos de entrenamiento y de muestras por lote. Para los modelos BERT además se ha ensayado con diversos ritmos de aprendizaje con el que ralentizar aún más su procedimiento otorgándole más tiempo para el análisis de muestras y la extracción de patrones y características.

## A.1. Modelos LSTM

### A.1.1. Selección de embeddings

El primer ajuste que se pretende realizar en el modelado con arquitecturas LSTM consiste en seleccionar el conjunto de embeddings que mejor representa a los documentos disponibles para la detección de textos sexistas y no sexistas. Para ello se ha dispuesto una arquitectura básica de una única capa oculta con 128 neuronas en su interior. En la Tabla A.1 se aprecian los distintos ficheros considerados, para el modelado en inglés y español, caracterizados por fundamentarse en datos procedentes de *Wikipedia+Gigaword* y de *Twitter* con distintas proporciones de *tokens* y tamaños, estando ordenados de forma ascendente. Observando las métricas de validación generadas se visualiza un primer aspecto que demuestra un aumento de la capacidad de predicción de los modelos al incrementar la complejidad del conjunto de embeddings empleado para la codificación de documentos. Mientras que

una segunda conclusión relevante se concentra en los **mejores resultados proporcionados por los embeddings procedentes de Twitter** con respecto a la combinación de Wikipedia y Gigaword. Una de las posibles teorías explicativas de este fenómeno puede apoyarse en que **su fuente de datos coincide con la procedencia de los datasets *EXIST***, por lo tanto aumenta el número de posibilidades de encontrar la terminología de los textos en los conceptos almacenados en los embeddings. Como consecuencia el fichero integrado en la solución es el Twitter 27B 100d puesto que apenas existen diferencias significativas con el inmediatamente posterior en términos de las métricas de validación, aunque la inversión computacional y temporal sí que es considerablemente superior.

**Tabla A.1:** Tabla con las métricas de validación de arquitecturas LSTM para elegir un conjunto de embeddings preentrenados.

| Embeddings | Conjunto | Accuracy | AUC |
|---|---|---|---|
| (W+G) 6B 100d | entrenamiento | 0.740 | 0.817 |
| (W+G) 6B 200d | entrenamiento | 0.757 | 0.835 |
| (W+G) 6B 300d | entrenamiento | 0.766 | 0.846 |
| Twitter 27B 100d | entrenamiento | 0.762 | 0.843 |
| Twitter 27B 200d | entrenamiento | 0.785 | 0.865 |
| (W+G) 6B 100d | test | 0.667 | 0.726 |
| (W+G) 6B 200d | test | 0.684 | 0.749 |
| (W+G) 6B 300d | test | 0.681 | 0.748 |
| Twitter 27B 100d | test | 0.697 | 0.762 |
| Twitter 27B 200d | test | 0.698 | 0.759 |

### A.1.2. Selección del tamaño del lote

A continuación la siguiente característica que se ha valorado como relevante para su estudio es el tamaño del lote con el que se proporciona el conjunto de entrada durante la construcción del modelo. Nuevamente se ha empleado una arquitectura sencilla de una única capa oculta con 128 neuronas, en combinación con el conjunto de embeddings que ha resultado ganador del torneo anterior. Así en cada parámetro se aplican las mejoras descubiertas hasta el momento para continuar enriquenciendo el modelado de los documentos ingleses y españoles. En la Tabla A.2 se demuestra que es necesario **decrementar el tamaño del lote para conseguir una mayor capacidad de acierto y predicción** aunque a costa de asignar más recursos computacionales y temporales para su entrenamiento y validación. Así se intenta fijar un valor que refleje el mejor equilibrio posible entre rendimiento y gasto, siendo suficiente un volumen de treinta y dos muestras puesto que apenas existen diferencias con el siguiente valor del ranking, que

si bien proporciona métricas ligeramente superiores, prácticamente duplica el tiempo de ejecución.

**Tabla A.2:** Tabla con las métricas de validación de arquitecturas LSTM para elegir el tamaño del lote en el modelado de documentos.

| Tamaño del lote | Conjunto | Accuracy | AUC |
|---|---|---|---|
| 128 | entrenamiento | 0.762 | 0.843 |
| 64 | entrenamiento | 0.771 | 0.853 |
| 32 | entrenamiento | 0.783 | 0.866 |
| 16 | entrenamiento | 0.785 | 0.835 |
| 128 | test | 0.697 | 0.762 |
| 64 | test | 0.699 | 0.766 |
| 32 | test | 0.708 | 0.769 |
| 16 | test | 0.716 | 0.774 |

### A.1.3. Selección de la arquitectura

La composición de la arquitectura y el nivel de complejidad también son partes fundamentales de un clasificador y como consecuencia se ha postulado como tercer parámetro a ajustar. Si bien la discrepancia entre las medidas de validación en entrenamiento en comparación con las de test no son tan significativas como con el algoritmo anterior, sí que se aprecia un leve sobreajuste que se ha intentado minimizar añadiendo capas ocultas adicionales, diferentes combinaciones de neuronas e incluso capas *drop-out*. La primera conclusión extraída reside en que el modelo **no es capaz de incrementar su capacidad predictiva al añadir más de tres capas ocultas**, por lo tanto esta cifra ha sido el máximo número de capas generadas en un sistema. En relación al número de neuronas se ha comprobado que para un diseño con dos capas ocultas es más beneficioso que estén constituidas por 128 neuronas cada una, sin añadir más puesto que no mejoran las métricas de validación pero sí aumentan los costes temporales y computacionales. Sin embargo para una composición de tres capas ocultas parece ser más ventajoso fijar un número de neuronas distinto en cada capa de manera decreciente, consiguiendo los mejores resultados con 128, 64 y 32 neuronas. Por último se ha probado a incorporar niveles de *drop-out* con los que desactivar aleatoriamente algunas neuronas en las capas ocultas elegidas. Para ello se han ensayado distintas configuraciones mostrándose en la tabla las más propicias consistentes en agregar este mecanismo a la arquitectura completa o únicamente a las dos últimas capas desactivando un 20 % de neuronas. No obstante, en la Tabla A.3 **no parece que exista ninguna mejora en la capacidad de generalización del clasificador puesto que apenas llega a igualar las métricas de validación de los modelos sin *drop-out***.

De igual modo, en comparación con la arquitectura simplista de una única capa que ha proporcionado el resultado más alto de accuracy 70.8 % y AUC 76.9 % en test, tampoco ayuda demasiado la adición de un mayor número de capas y/o neuronas para la resolución de este problema, ya que los valores de la evaluación realizada son prácticamente idénticos.

**Tabla A.3:** Tabla con las métricas de validación de arquitecturas LSTM para experimentar con distintas composiciones de capas y neuronas.

| Arquitectura | Conjunto | Accuracy | AUC |
|---|---|---|---|
| 2 capas | entrenamiento | 0.786 | 0.868 |
| 2 capas + AllDropOut | entrenamiento | 0.798 | 0.881 |
| 2 capas + LastDropOut | entrenamiento | 0.787 | 0.868 |
| 3 capas | entrenamiento | 0.779 | 0.861 |
| 3 capas + AllDropOut | entrenamiento | 0.796 | 0.871 |
| 3 capas + LastDropOut | entrenamiento | 0.778 | 0.859 |
| 2 capas | test | 0.708 | 0.770 |
| 2 capas + AllDropOut | test | 0.708 | 0.773 |
| 2 capas + LastDropOut | test | 0.709 | 0.771 |
| 3 capas | test | 0.709 | 0.772 |
| 3 capas + AllDropOut | test | 0.707 | 0.770 |
| 3 capas + LastDropOut | test | 0.707 | 0.769 |

### A.1.4. Aumento del conjunto de entrenamiento

En esta última subsección se ha reflexionado que una posible teoría explicativa acerca de los pésimos rendimientos de los modelos encontrados hasta el momento podría ser la escasa cifra de muestras de entrenamiento con las que se realiza la fabricación de sistemas inteligentes para la resolución de un problema particular. Este fenómeno se encuentra más acusado aún al dividir el conjunto de entrenamiento por idiomas para generar modelos específicos de un lenguaje concreto. Como consecuencia se han llevado a cabo diversas investigaciones sobre las distintas metodologías disponibles con las que generar ejemplos sintéticos adicionales a los ya existentes. A continuación se exponen aquellas que han sido valoradas positivamente por su potencial y modo de funcionamiento, ordenadas de menor a mayor complejidad procedural.

- **Traducción de textos**. Se trata de convertir los documentos ingleses a español y viceversa con el fin de duplicar el número de ejemplos de entrenamiento para cada idioma. Para esta tarea basta con utilizar cualquier API o librería de traducción como ha sido *deep-translator* empleada en este proyecto.

- **Easy Data Augmentation**. Es un conjunto de técnicas sencillas de tratamiento de textos que integran cuatro operaciones básicas: el reemplazamiento o inserción de sinónimos procedentes de términos escogidos aleatoriamente, el intercambio de posiciones de dos palabras seleccionadas al azar o la eliminación de conceptos dentro de una determinada frase también en base a una probabilidad aleatoria [21]. Pueden ser aplicadas en su totalidad, en combinaciones de subconjuntos o individualmente.
- **Round-trip Translation**. Conocido también como traducción bidireccional este método transforma una palabra, una frase o un documento de un idioma de origen a otro destino para posteriormente realizar el proceso inverso aprovechándose de las modificaciones que surgen tras traducir el contenido de vuelta al primer lenguaje [22].
- **Contextual Word Embeddings**. Esta última metodología trata de reemplazar algunos términos seleccionados aleatoriamente por los más similares semánticamente utilizando los embeddings preentrenados de modelos bidireccionales y transformers disponibles en Hugging Face.

En la Tabla A.4 se proyectan las métricas de validación de los ensayos más relevantes para cada una de las técnicas anteriormente comentadas. Tal y como se aprecia la primera consistente en **traducir los textos de español a inglés ha sido la más provechosa** por proporcionar un modelo con un 70.9 % de accuracy y 77 % de AUC, aunque la inversión de recursos computacionales y temporales ha sido notable y podría incrementarse exponencialmente a mayor número de muestras a traducir. En el resto de procedimientos se observa un fortísimo sobreaprendizaje debido a las diferencias desmesuradas entre las evaluaciones basadas en entrenamiento y en test. En la segunda posición se encuentra el método EDA en el que se han aplicado las cuatro operaciones disponibles de manera secuencial, aunque el modelo conseguido ha cometido un 3 % más de fallos. Sorpredentemente el modo de traducción RTT, pese a ser un método más completo en la generación de nuevos ejemplos sintéticos, ha suministrado un clasificador de peor calidad que la traducción tradicional de un idioma a otro. En el experimento *RTT 1* se han transformado los documentos ingleses a alemán, por ser el lenguaje más similar según su gramática, mientras que en el ensayo *RTT 2* se ha incluido el francés como tercer lenguaje con el que triplicar el conjunto de entrenamiento. No obstante, no existen apenas diferencias entre ambos casos pese a que se pensó en un primer momento que se postularía como uno de los mejores generadores por su modo de funcionamiento tan particular. Tampoco ha supuesto un beneficio la búsqueda y el reemplazamiento de sinónimos basados en transformers con embeddings preentrenados. Se ha experimentado con una arquitectura BERT y otra RoBERTa sin ninguna diferencia aparente en la calidad de las muestras generadas.

**Tabla A.4:** Tabla con las métricas de validación de arquitecturas LSTM para experimentar con técnicas de generación de muestras sintéticas adicionales al conjunto de entrenamiento

| Data Augmentation | Conjunto | Accuracy | AUC |
|---|---|---|---|
| Traducción | entrenamiento | 0.813 | 0.879 |
| EDA | entrenamiento | 0.976 | 0.995 |
| RTT 1 | entrenamiento | 0.930 | 0.970 |
| RTT 2 | entrenamiento | 0.939 | 0.980 |
| CWE (BERT) | entrenamiento | 0.925 | 0.970 |
| CWE (ROB) | entrenamiento | 0.917 | 0.967 |
| Traducción | test | 0.709 | 0.770 |
| EDA | test | 0.678 | 0.713 |
| RTT 1 | test | 0.657 | 0.704 |
| RTT 2 | test | 0.658 | 0.700 |
| CWE (BERT) | test | 0.653 | 0.692 |
| CWE (ROB) | test | 0.651 | 0.685 |

## A.2. Modelos BERT

### A.2.1. Selección del tamaño del lote

Tal y como se observa en la Tabla A.5 la teoría mencionada anteriormente queda confirmada puesto que **reduciendo aún más el tamaño del lote se aprecia una mejora sustancial del rendimiento** de los modelos hasta en más de un 1 % en las métricas de validación sobre el conjunto de test. Dada la dificultad que presenta el problema de detección de textos sexistas, me parece que merece la pena el incremento de la inversión de recursos que supone el entrenamiento de sistemas BERT con solamente ocho muestras por lote si a cambio se pueden rascar algunos puntos que aumenten la capacidad de generalización.

**Tabla A.5:** Tabla con las métricas de validación de arquitecturas BERT para elegir el tamaño del lote en el modelado de documentos.

| Tamaño del lote | Conjunto | Accuracy | AUC |
|---|---|---|---|
| 32 | entrenamiento | 0.763 | 0.771 |
| 16 | entrenamiento | 0.833 | 0.836 |
| 8 | entrenamiento | 0.822 | 0.824 |
| 32 | test | 0.730 | 0.720 |
| 16 | test | 0.739 | 0.733 |
| 8 | test | 0.741 | 0.739 |

### A.2.2. Selección de la tasa de aprendizaje

En base a mi experiencia uno de los parámetros que suelen impactar más en el comportamiento de un sistema automático es la velocidad con la que se le impone la el descubrimiento y asimilación de los patrones almacenados en los datos proporcionados como entrada. En la Tabla A.6 se visualizan los distintos valores de evaluación sobre los conjuntos de entrenamiento y test a partir de la aplicación de un amplio rango de tasas de aprendizaje. Con un valor inferior se consigue incrementar el rendimiento medio y su capacidad de predicción reflejándose en unas métricas de validación ligeramente superiores, aunque si la tasa de aprendizaje es demasiado baja, como ocurre en los últimos valores, puede resultar perjudicial puesto que el modelo demora demasiado tiempo en estudiar los ejemplos de entrenamiento sin extraer las características suficientes para realizar una clasificación de calidad. Por lo tanto, este parámetro se establece a **2e-5 para el modelado de documentos en inglés y en 1e-5 para los textos españoles**, debido a que una tasa de aprendizaje más pausada ha resultado suministrar más de un 2 % de accuracy y AUC. De nuevo en las arquitecturas BERT se aprecia el fenómeno que diferencia el nivel de representatividad y riqueza de los ejemplos españoles con respecto a los redactados en inglés.

**Tabla A.6:** Tabla con las métricas de validación de arquitecturas BERT para experimentar con distintas tasas de aprendizaje en el modelado de documentos.

| Tasa aprend. | Conjunto | Accuracy | AUC |
|---|---|---|---|
| 5e-5 | entrenamiento | 0.802 | 0.806 |
| 4e-5 | entrenamiento | 0.770 | 0.765 |
| 3e-5 | entrenamiento | 0.824 | 0.827 |
| 2e-5 | entrenamiento | 0.822 | 0.824 |
| 1e-5 | entrenamiento | 0.814 | 0.813727 |
| 5e-5 | test | 0.729 | 0.724 |
| 4e-5 | test | 0.676 | 0.683 |
| 3e-5 | test | 0.740 | 0.736 |
| 2e-5 | test | 0.741 | 0.739 |
| 1e-5 | test | 0.725 | 0.727 |

### A.2.3. Aumento del conjunto de entrenamiento

Es en esta subsección en la que verdaderamente se demuestra la fabulosa acogida de la generación de muestras sintéticas de entrenamiento reflejada en el rendimiento del clasificador especializado en ejemplos españoles, aunque también se aprecia un cierto beneficio en el sistema específico del lenguaje inglés. En la Tabla A.7 dentro de las filas correspondientes a la experimentación en español se aprecia una fuerte discrepancia del 6 % y 4 % comparando

la primera técnica consistente en traducir los textos de inglés a español, con respecto a aquellas más avanzadas como ***Easy Data Augmentation*** **y** ***Round-Trip Translation***. Si bien estas últimas incrementan considerablemente el rendimiento de los modelos producidos en términos de las métricas de validación, produciendo los dos **mejores clasificadores españoles** conocidos hasta el momento, también un cierto sobreaprendizaje debido a la enorme diferencia de más de treinta puntos entre las evaluaciones sobre entrenamiento y sobre el conjunto de test. Analizando las posibles explicaciones de este fenómeno su razón de ser parece proceder de los propios datos *EXIST*, puesto que no se ha apreciado tal *overfitting* en el ajuste de parámetros anteriores hasta este punto en el que se han producido más ejemplos sintéticos a partir de la combinaión de diversas operaciones de transformación de documentos y de traducciones cíclicas entre el español y el portugués. Previamente en los análisis exploratorios se observaron importantes diferencias en el contenido de los textos del conjunto de entrenamiento, siendo todos procedentes de Twitter y por tanto compuestos únicamente por *tweets* de usuarios, mientras que las muestras del dataset de test también incluían documentos procedentes de Gab y tras inspecciones visuales sus contenidos eran de más diversa índole, siendo *tweets* de usuarios o noticias que incluso trataban otros tópicos como la homofobia, el racismo, entre otros.

**Tabla A.7:** Tabla con las métricas de validación de arquitecturas BERT para experimentar con distintas tasas de aprendizaje en el modelado de documentos.

| Data Augmentation | Idioma | Conjunto | Accuracy | AUC |
|---|---|---|---|---|
| Traducción | entrenamiento | inglés | 0.709 | 0.719 |
| EDA | entrenamiento | inglés | 0.964 | 0.964 |
| RTT | entrenamiento | inglés | 0.976 | 0.976 |
| Traducción | test | inglés | 0.692 | 0.677 |
| EDA | test | inglés | 0.746 | 0.741 |
| RTT | test | inglés | 0.739 | 0.738 |
| Traducción | entrenamiento | español | 0.806 | 0.806 |
| EDA | entrenamiento | español | 0.980 | 0.980 |
| RTT | entrenamiento | español | 0.951 | 0.951 |
| Traducción | test | español | 0.729 | 0.734 |
| EDA | test | español | 0.781 | 0.780 |
| RTT | test | español | 0.765 | 0.765 |

# Bibliografía


[1] World Economic Forum. *Global Gender Gap Report*. 2022. URL: https://www3.weforum.org/docs/WEF_GGGR_2022.pdf.

[2] Fan-Osuala O. Humanit Soc Sci Commun. "Women's online opinions are still not as influential as those of their male peers in buying decisions." En: (2023). URL: https://doi.org/10.1057/s41599-023-01504-5.

[3] Brussels European Commission. *Independent High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI*. 2019. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[4] Nick Ferguson Balint Gyevnar. *Aligning Explainable AI and the Law: The European Perspective*. 2023. URL: https://arxiv.org/abs/2302.10766.

[5] W.J. von Eschenbach. "Transparency and the Black Box Problem: Why We Do Not Trust AI". En: (2021). URL: https://link.springer.com/article/10.1007/s13347-021-00477-0.

[6] Jens Allmer Malik Yousef. *MicroRNA Biology and Computational Analysis, Chapter 7: Introduction to Machine Learning*. 2014. URL: https://link.springer.com/protocol/10.1007/978-1-62703-748-8_7.

[7] James Whitaker Uday Kamath John Liu. *Deep Learning for NLP and Speech Recognition*. 2019. URL: https://link.springer.com/content/pdf/10.1007/978-3-030-14596-5.pdf.

[8] Ajay Shrestha University of Bridgeport y Ausif Mahmood. "Review of Deep Learning Algorithms and Architectures". En: (2019). URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8694781.

[9] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks, Chapter 4 Long Short-Term Memory*. 2012. URL: https://link.springer.com/chapter/10.1007/978-3-642-24797-2_4.

[10] Elizabeth D. Liddy Syracuse University. "Natural Language Processing". En: (2001). URL: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub.

[11] Attila Kovari Zsolt Krutilla. "The origin and primary areas of application of natural language processing". En: (2022). URL: https://ieeexplore.ieee.org/abstract/document/10029432?casa_token=RL5p3W3UDfYAAAAA:B7mmIoGphRktPbsI2cAOblrmitG3GN3w59YIVD3GMF8kIz01Pm7

[12] J. R. Medina D. W. Otter y J. K. Kalita. "A Survey of the Usages of Deep Learning for Natural Language Processing". En: (2021). URL: https://ieeexplore.ieee.org/abstract/document/9075398.

[13] Anna Rumshisky Anna Rogers Olga Kovaleva. "A Primer in BERTology: What We Know About How BERT Works". En: (2021). URL: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00349/96482/A-Primer-in-BERTology-What-We-Know-About-How-BERT.

[14] Nature. "Can we open the black box of AI?" En: (2016). URL: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[15] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2023. URL: https://christophm.github.io/interpretable-ml-book/.

[16] Manuel Romero. *T5-base fine-tuned for Emotion Recognition*. URL: https://huggingface.co/mrm8488/t5-base-finetuned-emotion.

[17] Adrián Mendieta Aragón UNED et al. *EXIST: sEXism Identification in Social neTworks*. 2022. URL: http://nlp.uned.es/exist2022/.

[18] Francisco Rodríguez Sánchez UNED et al. "Overview of EXIST 2021: sEXism Identification in Social neTworks". En: (2021). URL: https://diposit.ub.edu/dspace/bitstream/2445/181257/1/715155.pdf.

[19] Adrián Mendieta Aragón UNED et al. "Overview of EXIST 2022: sEXism Identification in Social neTworks". En: (2022). URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[20] Manuel Romero. "Hugging Face: mrm8488/t5-base-finetuned-emotion". En: (2019). URL: https://huggingface.co/mrm8488/t5-base-finetuned-emotion.

[21] Jason Wei Protago Labs Research Dartmouth College y Kai Zou Georgetown University. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". En: (2019). URL: https://arxiv.org/abs/1901.11196.

[22] Vukosi Marivate University of Pretoria y Tshephisho Sefara. "Improving short text classification through global augmentation methods". En: (2019). URL: https://arxiv.org/abs/1907.03752.

[23] Allen Institute for Artificial Intelligence Jeffrey Heer et Daniel S. Weld University of Washington Tongshuang Wu Microsoft Research Marco Tulio Ribeiro. "Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models". En: (2019). URL: https://www.cs.cmu.edu/~sherryw/assets/pubs/2021-polyjuice.pdf.