UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Identificación y clasificación de contenido sexista en memes

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Maeso Olmos, Alba

Tutor/a: Rosso, Paolo

Director/a Experimental: Chulvi Ferriols, María Alberta

CURSO ACADÉMICO: 2023/2024

Agradecimientos

Me gustaría agradecer especialmente a mi tutor, Paolo Rosso, por creer en mí y haberme dado la oportunidad de trabajar con él en este Trabajo Final de Grado, lo que ha sido un verdadero placer. Agradezco su apoyo y dedicación a lo largo de todo el trabajo, lo que me ha permitido conocer a un gran profesional, pero sobre todo a una gran persona.

Agradecer también el apoyo y ayuda de mi familia y amigos, en especial a mi pareja, él ha sido mi apoyo incondicional y el mejor compañero que podría tener en esta montaña rusa de emociones, triunfos y, a veces, desesperación que ha sido la carrera. Sencillamente él ha sido lo que he necesitado que fuera en cada momento, gracias por tanto.

De la carrera quiero agradecer el apoyo y ayuda a mis compañeros y amigos, con una especial mención a mi gran amigo Francisco Tomás García-Ruiz, una maravillosa persona con un gran talento al que quiero agradecerle todo su apoyo emocional y la ayuda que me ha brindado en este Trabajo Final de Grado y a lo largo de la carrera.

Por último, a mi compañero David Gimeno-Gómez, a quien ha sido un placer conocer y compartir este año con él, quiero agradecerle su ayuda y consejos para este Trabajo Final de Grado.

Resumen

El elevado número de contenido sexista dirigido hacia las mujeres en la web y su amplia difusión en las redes sociales, constituye un serio problema para las mujeres y la sociedad en general. Este fenómeno no se limita únicamente a la transmisión de contenido textual sexista, sino que también se manifiesta a través de recursos visuales, e incluso mediante su combinación, resaltando así la importancia de abordar su detección desde una perspectiva multimodal como son los memes.

Los memes, aunque aparentemente concebidos para fines humorísticos e irónicos, también se utilizan para expresar y transmitir ideologías y creencias con efectos negativos. Concretamente, en ocasiones son creados con el propósito de atacar y difamar a ciertos grupos, como es el caso de las mujeres, lo cual representa una forma de expresión de contenido sexista. La automatización del reconocimiento del sexismo en los memes presenta un desafío considerable debido a la naturaleza subjetiva de este fenómeno, la cual varía según la percepción y las características individuales de las personas.

Este Trabajo Final de Grado se enfoca en la construcción de un conjunto de datos formado por memes relacionados con esta temática, para posteriormente llevar a cabo dos tareas fundamentales: la primera busca automatizar la detección del sexismo en los memes, mientras que la segunda se centra en categorizar los diferentes tipos de sexismo presentes en dicho contenido: descrédito ideológico, estereotipos y dominancia, cosificación, misoginia y violencia sexual. Este enfoque integral busca contribuir al análisis y la comprensión de la problemática del sexismo en los memes desde una perspectiva de la Ciencia de Datos, específicamente haciendo uso de técnicas de Procesamiento del Lenguaje Natural y Visión por Computador con modelos de *Deep Learning*, como los Transformers.

Palabras clave: memes, sexismo, Deep Learning, Procesamiento del Lenguaje Natural, Visión por Computador, modelos unimodales, modelos multimodales, Transformers.

Abstract

The high number of sexist contents directed towards women on the web and its wide dissemination on social networks, constitutes a severe problem for women and society in general. Although, most of studies related to this phenomenon are based on the analysis of textual content, this problem also manifests itself through visual resources, and even through their combination, thus highlighting the importance of approaching its detection from a multimodal perspective.

Memes, although originally conceived for humorous and ironic purposes, also they are used to express, and transmit ideologies and beliefs with negative effects. Specifically, sometimes they are created to the purpose of attacking and defaming certain groups, such as women, which represents a form of expression of sexist content. Automating the recognition of sexism in memes presents a considerable challenge due to the subjective

nature of this phenomenon, which varies according to people's perceptions and individual characteristics.

This Final Degree Project focuses on the construction of a dataset formed by memes related to this topic, to subsequently carry out two fundamental tasks: the first one focuses on the detection of sexism in memes, while the second one focuses on categorizing the different types of sexism present in such content: ideological inequality, stereotyping and dominance, objectification, misogyny and sexual violence. This comprehensive approach seeks to contribute to the analysis and understanding of the issue of sexism in memes from a Data Science perspective, specifically making use of Natural Language Processing and Computer Vision techniques using Deep Learning models from both a unimodal and multimodal perspective with different Transformer models.

Keywords: memes, sexism, Deep Learning, Natural Language Processing, Computer Vision, uni-modal models, multi-modal models, Transformer.

Índice de contenidos

Índice de figuras

Índice de tablas

1 Introducción

1.1 Sexismo y memes

El sexismo supone un gran problema social en la actualidad principalmente afectando a las mujeres. Podemos definirlo como una actitud discriminatoria dirigida a las personas en función del sexo[1]. Dependiendo de la expresión y la motivación se puede diferenciar entre sexismo hostil, que se entiende como una actitud de prejuicio o discriminación basada en la supuesta inferioridad o diferencia de las mujeres, o el sexismo benevolente que en cambio utiliza formas sutiles de expresión, que pasan más inadvertidas y que se siguen caracterizando por un tratamiento desigual y perjudicial hacia las mujeres [1]. Como que las mujeres necesitan de protección masculina o se caracterizan por su habilidad para el cuidado doméstico y la crianza de los hijos entre otras [2]. Estudios previos han demostrado que tanto el sexismo hostil como el sexismo benevolente influyen en la violencia de género y en la aceptación de abusos en las relaciones adolescentes [3].

La importancia en investigaciones para frenar el sexismo se ve respaldada con la persistencia de actitudes, comportamientos sexistas y estudios estadísticos que demuestran la gravedad ante este fenómeno. Según el Instituto Europeo de la Igualdad de Género (EIGE) en España un 50 % de las mujeres han sufrido acoso sexual[2]. El Instituto Nacional de Estadística (INE), revela que el 10,8% de mujeres residentes en España han sufrido algún tipo de violencia de género en los últimos 12 meses. Además, el número de feminicidios en 2023 ha sufrido un incremento de 5 víctimas más respecto a los dos años anteriores [4]. A pesar de todo esto, España es el cuarto país de la Unión Europea en cuanto a progreso para alcanzar la igualdad de género, aun así, según una estimación del EIGE harían falta casi tres generaciones para alcanzarlo.

La lucha de las mujeres por acabar con la desigualdad y la violencia de género no ha cesado en las últimas décadas. Las redes sociales han sido una potente herramienta para dar voz a esta lucha y su gran capacidad de difusión ha permitido denunciar y reivindicarse con manifestaciones globales como *#NiUnaMenos o #8M 2017* que pudieron coordinarse por los distintos países gracias al activismo digital feminista impulsado por las redes sociales [5]. Sin embargo, estas plataformas también representan herramientas poderosas para la difusión y normalización de actitudes y comportamientos sexistas, especialmente cuando se manifiestan desde el anonimato y se disfrazan bajo la apariencia de humor, como es el caso de los memes.

En el mundo digital, los memes han emergido como una forma de expresión única y poderosa, especialmente entre los jóvenes que son los principales consumidores y creadores. Su éxito reside en que se entienden como bromas, y en su capacidad para divertir y viralizarse rápidamente [6]. Sin embargo, los memes también pueden ser un vehículo para la transmisión de mensajes sexistas. Esto es particularmente preocupante en fechas significativas como el 8 de marzo, cuando la circulación de memes machistas aumenta debido a una reacción de sectores que se sienten amenazados por los avances en igualdad de género [7]. Y con movimientos meméticos antifeministas en las redes sociales a través de los memes en el contexto de la misoginia y el machismo. Así como

[1] https://dle.rae.es/sexismo
[2] https://eige.europa.eu/gender-equality-index/game/ES/W

la deslegitimación de la defensa de los derechos de la mujer y la perpetuación de estereotipos sexistas reforzando la ideología heteropatriarcal [8].

El problema radica en que los memes sexistas no solo reflejan, sino que también refuerzan y normalizan actitudes discriminatorias, lo que afecta negativamente trivializando, perpetuando el sexismo y dando un paso atrás para alcanzar una sociedad en igualdad. El evidente problema que esto supone en la actualidad plantea la necesidad de realizar cambios con el objetivo de ejercer un mayor control sobre esta situación. Las redes sociales requieren la implementación de soluciones eficaces que sean capaces de automatizar su detección y poner barreras ante la difusión de este tipo de contenido.

En este contexto la Inteligencia Artificial (IA) juega un papel muy importante, ya que puede ayudar a crear mecanismos para su identificación. Concretamente el área del Procesamiento del Lenguaje Natural (PLN) y la Visión por Computador pueden proporcionar las técnicas necesarias para para abordar la identificación y categorización de contenido sexista en los memes como se propone llevar a cabo a lo largo de este Trabajo Final de Grado (TFG).

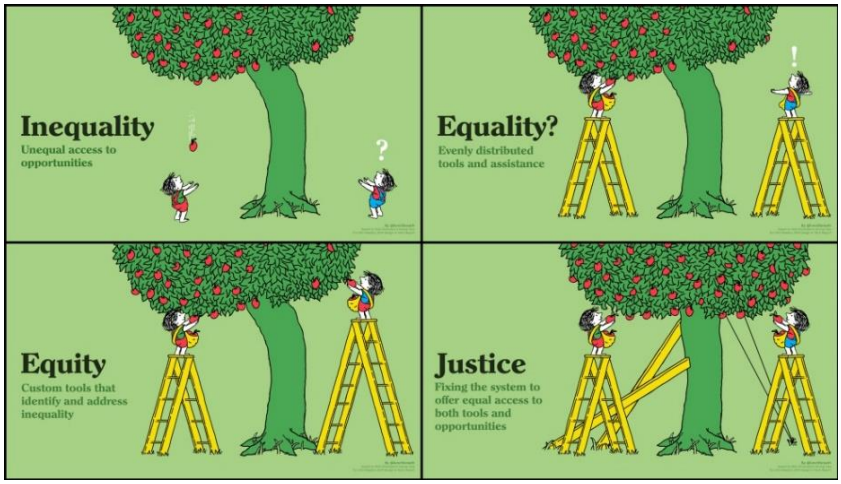

Figura 1.1: Igualdad de oportunidades [3]

1.2 Motivación

El evidente problema que supone el sexismo en la actualidad y la falta de soluciones eficaces para frenarlo motiva fuertemente a investigar en esta área. Concretamente, el sexismo en los memes es un tipo de contenido muy difundido y, a la vez, poco estudiado en la actualidad, ya que los estudios anteriores sobre sexismo se han centrado exclusivamente en el análisis de contenido textual. Quizás esto se deba a la naturaleza multimodal de los memes, lo que añade complejidad a los sistemas de análisis para extraer un contexto a través de dos canales de comunicación distintos, como el texto y las imágenes. Además, de tratarse de un tema subjetivo que puede

[3] https://verne.elpais.com/verne/2020/06/10/articulo/1591799815_274864.html

percibirse de manera distinta según el género, la cultura y otros factores que influyen en las interpretaciones individuales.

Todo esto genera un especial interés y se plantea como un gran desafío que motiva profundamente a abordar. Asimismo, resulta muy motivador la posibilidad de colaborar en proyectos como *Sexism Identification in Social Networks (EXIST)*[4] , que desde hace años se dedica a la investigación del sexismo en redes sociales. Así como abordar en este proyecto los nuevos retos que propone EXIST en su última edición para la identificación y categorización del sexismo en los memes.

Llevar a cabo un trabajo como este es muy motivador, ya que, además de poder aplicar los conocimientos adquiridos a lo largo de la carrera, permite explorar nuevas técnicas de IA, aprender y profundizar en aspectos más concretos de PNL y Visión por Computador. Pero, sobre todo, es gratificante poder contribuir a investigaciones que den voz a problemas como el sexismo y puedan aportar, de alguna forma, a conseguir una sociedad más igualitaria.

1.3 Objetivos

Dado el estado actual del fenómeno de los memes sexistas, en este TFG se proponen varios enfoques para identificar el contenido sexista en los memes, así como para determinar cuáles son los factores más influyentes y los canales que más información aportan para su identificación. Por consiguiente, se plantean los siguientes objetivos:

1. Crear un *dataset* de memes en español relacionados con la temática del sexismo para abordar el problema de identificación y categorización del sexismo.
2. Identificar el sexismo en los memes desde una perspectiva unimodal y multimodal, y analizar cuál de estas perspectivas se adapta mejor a este tipo de contenido.
3. Clasificar los memes en las distintas categorías del sexismo propuestas en este trabajo, utilizando tanto modelos unimodales como multimodales.
4. Analizar texto e imagen como elementos de comunicación: determinar cuál de estos canales aporta más información para la identificación del sexismo y sus distintas categorías.

1.4 Estructura del TFG

Este documento se organiza de la siguiente forma: El capítulo 2 es una recopilación de algunos conceptos, métricas y modelos más relevantes. El capítulo 3 recoge el estado del arte y se incluye un apartado más extenso dedicado a EXIST debido a que el *dataset* de memes se ha creado en el marco de este TFG y se ha utilizado en su tarea de evaluación 2024. En el capítulo 4 se aborda cómo se ha creado el *dataset*. En el capítulo 5, se explica la metodología y la fase de experimentación que se ha llevado a cabo para abordar las tareas. En el capítulo 6 se explica el proceso de evaluación de los sistemas planteados en el capítulo anterior y se realiza un posterior análisis de los errores. Por último, en el capítulo 7 se explica la conclusión de los hallazgos y posibles futuros cambios o ampliaciones. Así como la relación de este TFG con el grado en Ciencia de Datos y posibles aplicaciones. Finalmente, se lleva a cabo un análisis de los aspectos legales y éticos.

[4] http://nlp.uned.es/exist2024/

2 Modelos y métricas

2.1 Machine Learning

El *Machine Learning (ML)* o Aprendizaje Automático o es una rama de la Inteligencia Artificial (IA) que se centra en el desarrollo de algoritmos y técnicas que permiten a los computadores aprender a partir de datos. En lugar de ser programados de manera explícita para realizar una tarea, las máquinas utilizan estos algoritmos para identificar patrones en los datos y mejorar su desempeño con el tiempo. Esto se logra mediante la creación de modelos matemáticos que pueden hacer predicciones o tomar decisiones basadas en nuevos datos. El ML se puede clasificar en tres tipos principales:

- **Aprendizaje supervisado**: El algoritmo aprende de un conjunto de datos etiquetados y utiliza esta información para hacer predicciones sobre los nuevos datos.
- **Aprendizaje no supervisado**: El algoritmo trabaja con datos no etiquetados y buscando patrones o estructuras ocultas.
- **Aprendizaje por refuerzo**: El algoritmo aprende a través de la interacción con un entorno y aprende a base de sus errores y recompensas.

2.1.1 Modelos clásicos

En este apartado se van a definir de forma muy breve algunos de los modelos clásicos más conocidos de ML y que aparecen en el estado del arte para una mayor comprensión cuando se citen a lo largo de este trabajo.

- ***Support Vector Machine*** **(SVM),** es un modelo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. Su objetivo principal es encontrar un hiperplano en un espacio de alta dimensión que maximice el margen entre las diferentes clases de datos. Este modelo es especialmente efectivo en espacios de alta dimensión y cuando el número de dimensiones es mayor que el número de muestras.

- ***Random Forest*** **(RF),** es un modelo de aprendizaje supervisado que utiliza un conjunto de árboles de decisión durante el entrenamiento para obtener el promedio de predicciones si se trata de un modelo de regresión o de clases en un modelo de clasificación. Los modelos de RF son robustos y pueden manejar grandes conjuntos de datos con alta dimensionalidad.

- ***Naïve Bayes*** **(NB),** es un modelo de aprendizaje supervisado basado en el teorema de Bayes, que asume la independencia entre los predictores. A pesar de esta suposición tan simple, este modelo ha demostrado ser muy eficiente y eficaz para problemas de clasificación, especialmente en aplicaciones de filtrado de spam y análisis de sentimientos. La simplicidad y velocidad de NB lo hacen adecuado para problemas con grandes volúmenes de datos.

2.2 Deep Learning

El *Deep Learning* o Aprendizaje Profundo es una subrama del ML que utiliza redes neuronales artificiales con múltiples capas para modelar y entender datos con una alta complejidad. Estas capas adicionales permiten al modelo capturar características complejas en los datos, lo que lo hace especialmente útil para tareas como el reconocimiento de imágenes y procesamiento del lenguaje natural.

A diferencia de otros métodos tradicionales de aprendizaje automático, los modelos de *deep learning*, son capaces de identificar características de mayor complejidad intrínsecas en los datos. Estos modelos son muy adecuados cuando se requiere realizan tareas de inferencia en las que se necesitan predicciones muy precisas y no tanta capacidad de explicabilidad.

2.2.1 Redes neuronales

Las redes neuronales son un tipo de red en la que se basan las arquitecturas de los modelos de *deep learning*. Estas redes están inspiradas en la estructura y funcionamiento del cerebro humano, y consisten en capas de neuronas artificiales, también conocidas como nodos o neuronas. Cada neurona recibe una o varias entradas, las procesa y transmite una salida a las neuronas de la siguiente capa. Podemos distinguir entre los tres tipos de capa siguientes (véase figura 2.1):

- **Capa de entrada:** Recibe una representación interpretable de los datos iniciales.
- **Capas ocultas:** Realizan la mayor parte del procesamiento mediante la aplicación de funciones de activación no lineales.
- **Capa de salida:** Produce la predicción o el resultado final del modelo.

Cada conexión entre las neuronas tiene un peso asociado, que se ajusta durante el proceso de entrenamiento para minimizar el error del modelo. Las redes neuronales profundas *(Deep Neural Networks)* pueden tener decenas o incluso cientos de capas ocultas, lo que les permite aprender representaciones muy complejas de los datos.

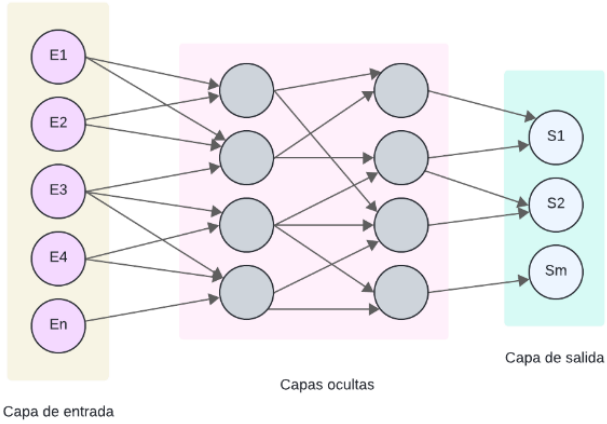

Figura 2.1: Esquema de red neuronal artificial

2.3 Text Transformers

Los Transformers desde su aparición en 2017 [9] se han convertido en un grupo de modelos muy relevantes en el área del PLN. Esto se debe a que estos modelos poseen una arquitectura de red neuronal con mecanismos de autoatención que consigue relacionar cada término con el resto extrayendo el contexto global más allá de su significado semántico. Además, previamente entrenados con grandes volúmenes de datos de diversos dominios. A diferencia de otros modelos, los Transformers parten de un conocimiento aprendido que junto con posteriores ajustes adaptados a cada tarea específica son capaces de conseguir un rendimiento excepcional con menos tiempo de cómputo que otros modelos de *deep learning* en tareas de PLN. Para comprender mejor su funcionamiento es importante definir algunos conceptos relacionados, los cuales explicaremos en los siguientes apartados.

2.3.1 Tokenización

Cuando trabajamos con este tipo de modelos el preprocesamiento previo es mínimo o en algunos casos innecesario. Sin embargo, nos enfrentamos al problema de que los ordenadores no son capaces de procesar los textos directamente. Para ello, necesitamos hacer una transformación que convierta los textos en una representación vectorial de *embeddings* interpretable para el modelo, lo que se denomina **tokenización.**

El **tokenizador** lleva a cabo un proceso de segmentación de las entradas de texto a unidades de texto más pequeñas, lo que se denominan **tokens**. Los tokens se representan en un vector de *embeddings* mediante los ***input-ID*** que son una secuencia de números enteros que representan los índices únicos de cada token en el vocabulario del modelo. También se crea otro vector llamado ***token_type_ids*** que es una lista de enteros que indican el tipo de cada token en la secuencia de entrada. Los vectores de *embeddings* que esperan los modelos de PNL suelen ser de longitud fija. Para solucionar esto el tokenizador crea un vector binario de 0 y 1 llamado ***attention mask*** donde los 1 ocupan las posiciones de los *tokens* y los 0 son elementos de relleno hasta completar la longitud máxima del vector.

2.3.2 Estructura y tipos

- ***Encoder*:** Su finalidad es transformar los vectores de *embeddings* de los tokens denominados *hidden states*, o *embedding* de tipo semántico-contextual. De esta forma se introduce un contexto a la palabra más allá del mero significado semántico de la palabra en cuestión, lo que se consigue con mecanismos de atención.

- ***Decoder*:** Este mecanismo es el encargado de recibir los *hidden states* extraídos de los tokens y generar la secuencia de salida a partir de este contexto.

Entre ellos podemos diferenciar los siguientes grupos:

- ***Encoder Only***: En este tipo de estructura, solo se utiliza la parte de codificación. El *encoder* procesa la entrada y genera una representación útil de la misma. Esto puede ser útil para tareas como la clasificación de texto o la generación de *embeddings*. Algunos de los modelos más conocidos son los de la familia BERT [10].

- ***Decoder Only***: Toma una representación de entrada y genera su salida correspondiente. Esto es útil en tareas de generación de texto, donde se desea generar una secuencia a partir de una representación interna, como los modelos de la familia GPT [11].

- ***Encoder-Decoder***: En este esquema el *encoder* procesa la entrada y la convierte en una representación interna, que luego se pasa al *decoder* para generar una salida. Esta estructura es común en tareas de traducción automática, donde el *encoder* procesa la oración en un idioma y el *decoder* genera la traducción en otro idioma [12].

2.3.3 Mecanismo de atención

Los mecanismos de atención o *self attention* se aplican a cada capa de los bloques de *encoder* y *decoder.* Cuando se procede a generar una representación o *embedding,* los *self attention* asignan unos pesos que indican el nivel de importancia que tiene un token dado en la capa anterior. De esta forma durante el proceso de entrenamiento los tokens van aprendiendo el contexto según su importancia en las capas anteriores.

2.3.4 Componentes de un modelo Transformer

- **Cuerpo o *Body*:** el cuerpo de un modelo Transformer es la parte principal del modelo e incluye múltiples capas de atención, como las *multi-head self attention* y las capas de *feed-forward*, que componen la arquitectura básica del Transformer. En esta parte del modelo está contenido su *encoder* y/o *decoder*, el cual proporciona su *hidden state* (véase figura 2.2).

- **Cabeza o *Head*:** la cabeza está conectada al cuerpo del modelo y se refiere a la capa o capas finales que se utilizan para realizar tareas específicas, como la clasificación de texto, el etiquetado de secuencias, la generación de texto, etc.

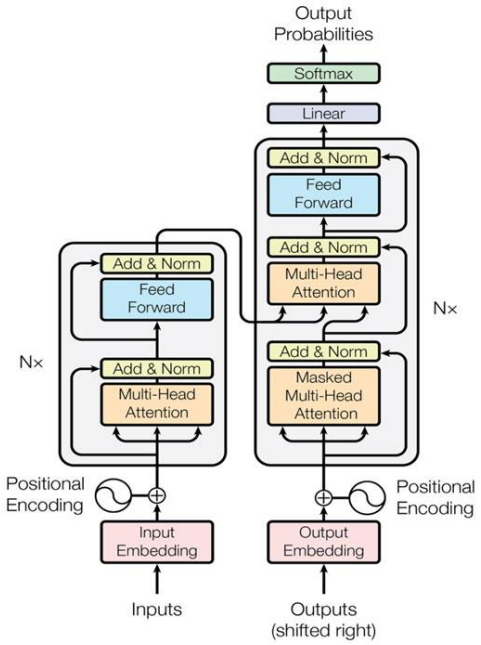

Figura 2.2: Esquema de arquitectura modelo Transformer [5]

2.3.5 Fine-tuning en modelos Transformers

El *fine-tuning* es una técnica de entrenamiento muy importante en los modelos Transformers. Se refiere a la adaptación o ajuste fino de un modelo pre-entrenado a una nueva tarea o dominio de interés.

Durante esta fase, se pueden llevar a cabo algunas modificaciones en la arquitectura del modelo pre-entrenado, como congelar algunas de sus capas y realizar ajustes en su estructura para adaptarlo a la nueva tarea. También se puede optar por ajustar todo el modelo durante el entrenamiento, ajustando todos los parámetros del modelo para adaptarlo a los nuevos datos.

Para optimizar los parámetros del modelo, y dado que no se puede acceder a ellos de forma directa, se utilizan algunos hiperparámetros y entre los más comunes podemos encontrar los que definimos a continuación.

- La **función de pérdida** es un elemento muy importante en el entrenamiento de un modelo, ya que lo ayuda a determinar lo equivocado que está para poder corregirse. Lo que hace es calcular la diferencia entre los valores predichos y los valores reales y su *loss* o medida de error determina lo bien que el modelo se ajusta a los nuevos datos para que el modelo se actualice con los parámetros óptimos.

[5] https://arxiv.org/abs/1706.03762

El error entre dos distribuciones de probabilidad se suele medir utilizando una función de pérdida de entropía cruzada. Para dos variables aleatorias discretas p y q podemos definirla de la siguiente forma:

$$H(p,q) = -\sum_{x} p(x) \log q(x)$$

Esta definición no es simétrica ya que *p* se entiende como la distribución "verdadera", solo parcialmente observada, mientras que *q* se entiende como la distribución "no natural" obtenida a partir del modelo [13] [14].

- El **optimizador,** es un algoritmo encargado de ajustar los parámetros o pesos del modelo para minimizar la diferencia de la función de pérdida. Utiliza el gradiente de la función para actualizar los pesos del modelo y reducir el error.

- Los ***epochs***, son las pasadas completas de todos los datos de entrenamiento a través del modelo. En el proceso de *fine-tuning*, generalmente se repiten múltiples *epochs* para mejorar gradualmente el rendimiento del modelo a medida que se ajustan sus parámetros.

- Los ***batchs*** o lotes son conjuntos de datos que se procesan juntos en paralelo durante una única iteración del modelo en el entrenamiento. Permiten acelerar el proceso de entrenamiento y mejorar su estabilidad. En lugar de actualizar los pesos del modelo después de cada muestra individual, se calcula el gradiente promedio por *batch* y se utiliza para actualizar los pesos.

- El ***Learning Rate (LR)*** o tasa de aprendizaje controla la magnitud de los ajustes que se realizan a los pesos del modelo durante el proceso de entrenamiento. Es un factor de escala que determina cuánto cambiará el modelo en respuesta al gradiente calculado durante la retropropagación. Un *LR* más alto significa que los pesos se ajustarán más en cada paso, lo que puede llevar a un entrenamiento más rápido, pero también puede provocar oscilaciones o divergencia. Un *LR* más bajo permite un entrenamiento más estable, pero puede llevar más tiempo converger hacia una solución óptima.

2.4 Visión por computadora

La Visión por Computadora es una disciplina que busca replicar la capacidad humana para comprender y analizar el contenido de las imágenes y videos mediante modelos matemáticos.

Entre las metodologías más recientes y revolucionarias en el campo se encuentran los Visual Transformers (ViTs), que han mostrado un rendimiento excepcional en diversas tareas de reconocimiento y clasificación de imágenes.

2.4.1 Visual Transformers

Inspirándose en el éxito de los modelos Transformers, y después de años en los que los modelos más utilizados en tareas como clasificación o tratamiento de imágenes fueron las redes neuronales convolucionales [15], aparecieron los Visual Transformers [16].

Los Visual Transformers cambiaron el paradigma de la visión por computadora al incorporar mecanismos de atención, arquitecturas y procesamiento similares a las de los Transformers, pero adaptados a las imágenes. Podemos distinguir entre las siguientes fases que permiten procesar las imágenes de manera eficiente y extraer las características significativas:

- **División de la imagen en parches:** La imagen de entrada se divide en parches más pequeños y manejables. Cada uno de estos parches se trata como un "token", similar a las palabras en una secuencia de texto en NLP.

- ***Embedding* de parches:** Cada parche se transforma en un vector de características *(embedding)* a través de una capa lineal. Además, se añade información de posición a estos *embeddings* para mantener la estructura espacial de la imagen.

- **Procesador:** La secuencia de *embeddings* de parches se alimenta a una pila de bloques de un procesador estándar. Cada bloque incluye mecanismos de autoatención y capas *feed-forward*, que permiten capturar relaciones entre diferentes parches de la imagen.

- **Clasificación:** La salida del procesador se pasa por una capa de clasificación, que genera las predicciones finales.

En la imagen de la figura 2.3 podemos ver a la izquierda el procesamiento de las imágenes con el procesador de Vision Transformer. Y en la parte de la derecha la arquitectura del *encoder* de un modelo ViT que recibe las imágenes procesadas como un vector de *embedding*.

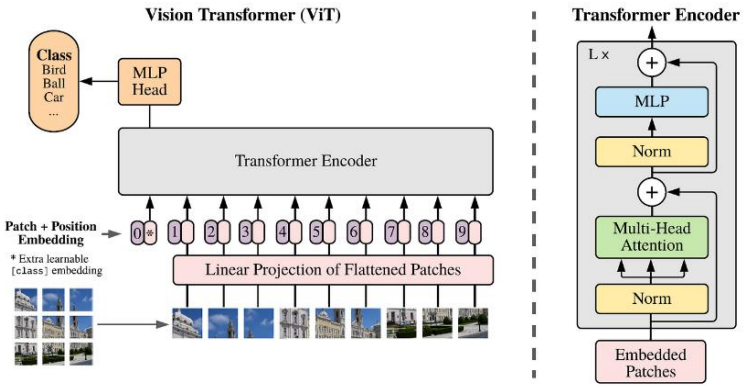

Figura 2.3: Esquema de arquitectura Visual Transformer

2.5 Métricas de evaluación

Terminologías utilizadas en las métricas de evaluación:

- **Verdaderos Positivos (TP):** Muestras positivas correctamente identificadas por el modelo.
- **Verdaderos Negativos (TN):** Muestras negativas correctamente identificadas por el modelo.
- **Falsos Negativos (FN):** Muestras negativas incorrectamente identificadas por el modelo.
- **Falsos Positivos (FP):** Muestras positivas incorrectamente identificadas por el modelo.

La **matriz de confusión** muestra la relación entre las predicciones del modelo y los valores reales de los datos. Además, es especialmente útil porque permite identificar no solo cuántas predicciones se hicieron correctamente, sino también cómo se distribuyen los errores. Podemos ver un esquema en la figura 5.

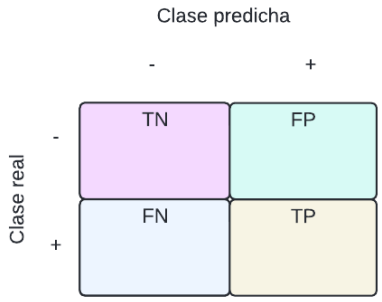

Figura 2.4: Matriz de confusión

El ***accuracy*** indica la proporción total de elementos clasificados correctamente. Esta métrica solo es adecuada cuando las clases están balanceadas. Se calcula de la siguiente forma:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

La ***precision*** es una métrica que mide la capacidad del modelo para clasificar correctamente las muestras que son positivas. Se calcula de la siguiente forma:

$$Precision = \frac{TP}{TP + FP}$$

El ***recall*** es una métrica que indica la proporción total de elementos clasificados correctamente en base al número total de muestras positivas en el conjunto de datos. Se calcula de la siguiente forma:

$$Recall = \frac{TP}{TP + FN}$$

El ***F1 score*** es una métrica muy utilizada en problemas de clasificación ya que combina la *precision* y el *recall* en una sola puntuación. Indica como de bien predice el modelo teniendo en cuenta tanto los casos positivos como los negativos para cada clase. Es muy útil cuando las clases están desbalanceadas. Se calcula de la siguiente forma:

$$F1_i = 2 * \frac{precision_i + recall_i}{precision_j + recall_j}$$

El ***F1 score macro*** es la media aritmética de cada *F1 score* obtenido para cada clase. Es decir, se suman todos los *F1 score* de cada clase y se divide por el número total de clases. De esta forma todas las clases contribuyen de manera equitativa en el cálculo del *F1 score macro*, sin tener en cuenta el tamaño y distribución de las mismas.

El ***F1 score weighted*** es la media ponderada por el soporte (número de datos por clase) de cada *F1 score* de cada clase a la que pertenece. Es la más robusta cuando tenemos datos desbalanceados y queremos tener en cuenta el peso que tiene cada clase.

3 Estado del arte

En los últimos años, la Inteligencia Artificial, y concretamente en las áreas de PLN y la Visión por Computador, se ha experimentado un crecimiento significativo reflejándose en el estado del arte actual.

En este capítulo, se pretende hacer una recapitulación del estado del arte de los estudios más recientes relacionados con sexismo, e incluir estudios sobre lenguaje ofensivo y discurso de odio que a menudo se relacionan con el sexismo. Así como incluir tanto estudios que involucren el análisis de texto e imagen como la multimodalidad, describiendo el proceso de extracción de los datos y los procedimientos llevados a cabo para obtener conjuntos de datos utilizados en este tipo de investigaciones.

3.1 Hate speech y lenguaje ofensivo en modelos multimodales

La detección y clasificación del discurso de odio en memes multimodales todavía supone un gran desafío en la actualidad. El discurso de odio o más conocido como *hate speech* según la Recomendación núm. 15 de política general de la ECRI, se basa en la suposición injustificada de que una persona o un grupo de personas son superiores a otras; incita a cometer actos de violencia o discriminación, socavando así el respeto de los grupos minoritarios y perjudicando la cohesión social. Esta discriminación se basa en una lista no exhaustiva de características o estatus personales, como la raza, el color, la lengua, la religión o las creencias, la nacionalidad o el origen nacional o étnico, así como la ascendencia, la edad, la discapacidad, el sexo, el género, la identidad de género y la orientación sexual[6].

En esta línea, se propuso *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal* Memes [17]. El conjunto de datos que se utilizó para llevarlo a cabo se generó a partir de imágenes en las que posteriormente se añadió texto hasta construir un meme. De esta forma los autores incluyen una característica que llaman “factores de confusión benignos” en la que, partiendo de una misma imagen, generan textos con discurso de odio, así como textos inofensivos para superponer en una misma imagen construyendo memes que aportan una información completamente opuesta. Del mismo modo utilizan un mismo texto para distintas imágenes. El proceso de anotación fue llevado a cabo por 3 personas distintas capacitadas para reconocer el discurso de odio e indicar su grado de odio con un conjunto final de 10.000 memes.

El propósito de este estudio no era entrenar modelos desde cero sino ajustar y probar modelos multimodales previamente entrenados. Los autores evaluaron una variedad de modelos unimodales pre-entrenados como BERT [10] para la modalidad textual. Modelos multimodales que fueron entrenados unimodalmente con la media o concatenación de las puntuaciones de salida como ResNet-152 [18] y BERT, posteriormente entrenados con una MLP. Y modelos multimodales pre-entrenados de forma multimodal como ViLBERT [19] y Visual BERT [20].

[6] https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15

En los resultados observaron cómo los modelos unimodales aportaron los peores resultados, y concretamente los modelos de imagen peor que los textuales. Los mejores resultados fueron de 62.80% de *accuracy* con BERT. En cuanto a los modelos multimodales los resultados fueron mejores en los modelos preentrenados multimodalmente que los preentrenados de forma unimodal y posteriormente concatenados. Obteniendo un mejor resultado de 69.47 % de *accuracy* con un modelo Visual BERT.

Los memes en el contexto de las redes sociales también representan una forma de expresión de ideas y emociones ofensivas para algunos grupos, además de expresar odio pueden convertirse en un medio para difundir lenguaje y transmitir comportamientos ofensivos hacia un individuo o un grupo ya sea su origen étnico, orientación sexual y religión entre otros.

En este sentido, para clasificar si un meme era ofensivo o no en función de su texto e imagen, los autores de *Multimodal Offensive Meme Classification with Natural Language Inference* [21] propusieron tratarlo como una tarea unimodal para la cual extrajeron el texto del meme y obtuvieron el título y la transcripción con ClipCap basado en el codificador de imágenes CLIP [22]. Para evaluar los distintos enfoques hicieron un análisis comparativo de los distintos conjuntos de datos: *Memotion* [23]*, Hateful memes* [17] *y conjuntos de datos MultiOFF* [24]. Para llevarlo a cabo siguieron un proceso de *NLI-fication,* que consistía en convertir de imagen-texto-etiqueta a premisa-hipótesis-etiqueta. La premisa se obtenía a partir de la transcripción, el título y texto del meme. La etiqueta *"OFF"* se convirtió en una hipótesis con palabras clave como "ofensivo", "ataque", "deshumanizante", "burlas", "odio" y "crimen". Y sugirieron tres niveles: el primario solo utilizaba la etiqueta "ofensivo" como hipótesis, el nivel extendido incluía una oración que describía la etiqueta y el nivel de definición, agregaba una definición detallada del contenido. Para medir su efectividad utilizaron RoBERTa [25], obteniendo los mejores resultados en el nivel de definición, con lo que resaltan la importancia de las palabras clave específicas.

Como *baselines* se utilizaron los siguientes modelos: *twitter-roberta-base-sentiment1*[7]*, twitter-roberta-base-emotion2*[8] y *roberta-base-offensive3*[9] de RoBERTa afinados específicamente a análisis de emociones, sentimientos y tweets ofensivos. Y para comparar si había una diferencia significativa entre ellos se realizó una prueba de significancia 5X2cv [26] con los tres modelos y el conjunto de datos *Memotion*, lo que mostró que todas las parejas de modelos eran significativamente diferentes entre sí.

Por otro lado, se realizaron tres *ablations* que consistían en los enfoques siguientes: la primera utilizaba solo la transcripción del meme, la segunda el texto del meme y la tercera eliminaron la *NLI-fication* descartando la hipótesis y teniendo en cuenta solo la premisa. Las *ablations* mostraron que el texto del meme era más importante que la leyenda para identificar los memes ofensivos. Además, los modelos *Emotion RoBERTa* y *Sentiment RoBERTa* mejoraron su rendimiento en configuraciones NLI para los conjuntos de datos *MultiOFF y Hateful Memes*, respectivamente. Y en general, los mejores resultados se obtuvieron con la configuración NLI-fication y RoBERTa.

[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion
[9] https://huggingface.co/cardiffnlp/roberta-base-offensive

3.2 Misoginia en modelos de texto, imagen y multimodales

La misoginia constituye una faceta del sexismo, que implica sentimientos de odio, repulsión hacia las mujeres e incluso violencia. Estudios como el de *Automatic Identification and Classification of Misogynistic Language on Twitter* [27] representan una contribución significativa al campo de la detección de misoginia en redes sociales. Los autores propusieron llevar a cabo un estudio para distinguir entre contenido misógino del no misógino y en el caso de serlo siguieron una taxonomía que distingue entre las siguientes categorías: desacreditación, acoso sexual y amenazas de violencia, estereotipo y cosificación, dominación y desviación, a partir de un corpus representativo de comentarios en Twitter obtenido por la búsqueda de palabras clave y hashtags relacionados. El *gold standard* se etiquetó por dos anotadores, cuyos casos de desacuerdo fueron resueltos por un tercer anotador experimentado. En la siguiente, el resto de los tweets se decidieron por el voto mayoritario de anotadores de una plataforma de *crowdsourcing*. Finalmente consiguieron un conjunto de datos compuesto de 4.454 tweets, equilibrados entre misóginos y no misóginos.

Los autores abordaron este estudio desde una perspectiva de caracterización lingüística de los textos considerando más representativas las siguientes clasificaciones: *N-gramas*, tanto a nivel de carácter como de token; Lingüística cuantitativa para la caracterización de las distintas categorías de la misoginia, donde tuvieron en cuenta la longitud del tweet, la presencia de URL, el número de adjetivos y el número de menciones de usuario. A nivel sintáctico consideraron *Bag-Of-POS (Part of Speach)* y n-gramas para las etiquetas de partes del discurso. Y, por último, utilizaron una representación de *embeddings*.

La fase de experimentación se llevó a cabo con una validación cruzada de 10-folds para cada representación de textos mencionada en el párrafo anterior y con cada uno de los modelos supervisados de ML siguientes: SVM, RF, NB y *Message-passing neural network* (MPNN) [28]. Las métricas que utilizaron para la evaluación de los modelos fueron *accuracy* para la primera experimentación puesto que la clase misógina y la no misógina estaban balanceadas y F1 macro para la categorización de la misoginia debido al desbalanceo entre las distintas categorías.

Los mejores resultados tanto para la identificación de un tweet misógino como no misógino y para la categorización de la misoginia se obtuvieron utilizando una representación de los tokens con *n-gramas* y un modelo de SVM. Para la identificación de la misoginia los resultados fueron muy cercanos entre sí, obteniendo el mejor resultado de 0,799 de *accuracy*. En cambio, para la categorización de la misoginia, los resultados diferían sustancialmente entre sí, de lo que se obtuvo como mejor resultado un 0,382 de *F1 macro*. De lo que los autores destacaron la dificultad de reconocer las diferentes categorías de la misoginia.

Dado el aumento de ataques en las redes hacia las mujeres y lo que supone este preocupante problema social se propuso la tarea *Automatic Identification and Classification of Misogynistic Language on Twitter (AMI*) [29] [30] organizada por IberEval[10] en español e inglés y posteriormente Evalita[11] en inglés e italiano con las siguientes dos subtareas: La subtarea A consistió en la discriminación de contenido misógino del no misógino. La subtarea B, en el reconocimiento de los distintos grupos

[10] https://sites.google.com/view/ibereval-2018
[11] https://www.evalita.it/campaigns/evalita-2018/

que se pueden encontrar en el contenido misógino, para ello siguieron la misma taxonomía que se utiliza en el artículo anterior [27].

El proceso de extracción de datos y etiquetado para IberEval fue llevado a cabo siguiendo la misma metodología que en el artículo anterior. En cuanto al corpus en entrenamiento fue de 3.307 tweets, mientras que el inglés por 3.251 tweets. Los datos de prueba por 831 tweets en español y 726 en inglés. En Evalita, tanto el corpus italiano como el inglés se compuso de 4.000 tweets para entrenamiento y 1.000 para evaluación.

El mejor resultado para la subtarea A de IberEval fue en inglés fue de 0,913 *accuracy* frente a de 0,814 para el español. El modelo que consiguió los mejores resultados fue un SVM entrenado con una combinación de características estilísticas, estructurales y léxicas con presencia de hashtags, de enlaces, recuento de malas palabras, de insultos sexistas y palabras relacionadas con mujeres.

En la subtarea B el mejor resultado fue de 0,339 *F1 score* para el idioma español también con un modelo SVM entrenado con *Bag of Word*[12], *Bag of Hashtags, Bag of Emojis* e insultos sexistas hacia la mujer. Es interesante destacar que hubo una gran dificultad para reconocer la categoría de misoginia desviación debido a los pocos ejemplos disponibles en el conjunto de entrenamiento.

En cuanto a Evalita el mejor resultado se obtuvo para el idioma italiano con un *accuracy* de 0,844 para la primera subtarea y 0,501 de *F1 macro* para la de categorización de misoginia, con una representación de TF-IDF[13] y descomposición de valores singulares y el clasificador *Boosting*. De nuevo los peores resultados fueron para la categorización de la misoginia y los autores consideraron que podía deberse a que hubiera mucha superposición entre expresiones textuales de las diferentes categorías y pudiera resultar muy subjetivo para los anotadores y en consecuencia para que los sistemas seleccionaran una categoría u otra.

Centrándonos ahora en contenido multimedia, el primer estudio que se llevó a cabo fue la tarea compartida de SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification [31]. A diferencia de otros estudios, este da un paso más estudiando la misoginia desde una perspectiva multimodal en los memes. Esta competición se subdividió en dos subtareas distintas: la subtarea A, es una tarea de clasificación binaria que consiste en identificar el meme como misógino o no misógino. Y la subtarea B, es una clasificación multiclase y multi-etiqueta que consiste en distinguir entre las siguientes categorías que proponen como misóginas: *avergonzar, estereotipo, cosificación y violencia*.

Para esta tarea los autores formaron un *dataset* a partir de memes en inglés que extrajeron de redes sociales como Twitter[14] y Reddit[15] e hilos dedicados a memes sobre mujeres. El proceso de etiquetado se llevó a cabo por tres expertos de *crowdsourcing* y

[12] *Bag of Word*: método que se utiliza en PNL para representar documentos ignorando el orden de las palabras.

[13] El TF-IDF: mide la importancia de los términos t en un documento d y se calcula de la siguiente forma: $\text{TF-IDF}(\text{t}, \text{d}, \text{D}) = TF(t,d) * IDF(t,D)$, donde $TF = \frac{f_{t,d}}{N_d}$, e $IDF = \log(\frac{N}{|\{d \in D : t \in d\}|})$.

[14] https://www.twitter.com

[15] https://www.reddit.com/

la etiqueta final se decidió por el voto mayoritario de los anotadores. La extracción de texto de los memes se transcribió utilizando la plataforma de Google Cloud Vision[16].

Para la subtarea A, el mejor resultado lo obtuvieron con un modelo de ensamblado de características multimodales profundas con una *Multilayer Perceptron (MLP)* con el que obtuvieron un 0,834 de *F1 score*. Para la subtarea B, un 0,731 de *F1 score* por un lado con una MLP como en la subtarea anterior. Y también con un modelo CLIP [32] basado en características de imagen y texto combinado con una *Long Short-Term Memory (LSTM).*

Por último, los autores realizaron un análisis de error en el que observaron que la mayoría de los memes estaban bien clasificados por la mayoría de los equipos, pero en los memes mal clasificados había más errores en los 'no misóginos' que en los 'misóginos'. Además, se obtuvo mejor valor en el *recall* que *precision* por lo que se dedujo que la mayoría de los sistemas tendían a estar sesgados hacia la misoginia. A raíz de la problemática del sesgo que se produce en muchos modelos de PLN, los autores de este artículo [33] estudiaron si una perspectiva unimodal o multimodal contribuía más a la detección de la misógina en los memes y qué perspectiva tendía a producir más sesgo.

Los autores observaron que los modelos unimodales basados en texto y el modelo BERT [10] de Transformers, funcionan mejor que los basados en imágenes. Lo cual indica, que el componente textual es más informativo que las otras fuentes. Sin embargo, los enfoques multimodales como *Multimodal Text and Tags* (MTT) que concatena la representación de texto y etiquetas del meme, *Multimodal Text and Caption* (MTC) que concatena la representación imagen y texto y *visual-BERT* [20] superaron a los unimodales.

Con el enfoque multimodal algunos modelos tuvieron alta *precision* pero bajo *recall* en una clase y baja *precision* pero alto *recall* en la otra, mientras que los *F1 score* fueron más altos en la clase misógina que en la no misógina, obteniendo los mejores resultados con un modelo MTT. Según los autores este desequilibrio podía deberse al sesgo de selección y que este se produjera por términos o elementos visuales específicos que se asociaran directamente con estas dos clases lo cual impidiera reflejar la distribución de los datos en un entorno real.

Todos los estudios previos en enfoques multimodales y en relación con los memes se han centrado en la identificación y categorización de la misoginia. Sin embargo, la misoginia es solo una faceta del sexismo y también se puede encontrar mucho contenido sexista no misógino en las redes sociales. Por esta razón, es importante abordar investigaciones que se centren también en el estudio y detección de otros tipos de sexismo.

[16] https://cloud.google.com/vision/docs/ocr?hl=es-419

3.3 Sexismo en modelos de texto

Uno de los estudios más recientes en la identificación del sexismo es la tarea *Task 10: Explainable Detection of Online Sexism* [34]. Dos de las subtareas que plantean son las siguientes: la subtarea A es una clasificación binaria que indica si el comentario es sexista o no sexista. La subtarea B es una clasificación de las siguientes categorías: amenazas, planes de daño e incitación, descrédito, animosidad y discusión prejuiciosa.

La mayoría de los participantes utilizó modelos pre-entrenados (90%) como RoBERTa [25]. Los mejores resultados se obtuvieron en la subtarea A, alcanzando un 0,874 de *F1 macro*, seguido de 0,732 *F1 macro* en la subtarea B atribuyendo estos resultados a la mayor complejidad de la segunda tarea.

En el análisis de error se observó que en la subtarea A un 38,3% de errores fueron FP y el 61,7% FN. De los falsos negativos, el 41% pertenecía a la clase *animosidad,* 34% a *descrédito*, 16% a discusión prejuiciosa y 9% a amenazas, planes de daño e incitación. Esto se puede explicar porque la categoría con más margen de error suele contener un lenguaje implícito más difícil de detectar mientras que las amenazas suelen expresarse de forma explícita.

La falta de barreras en las redes sociales para frenar la transmisión de contenido sexista insta a crear nuevos mecanismos que sean capaces de identificar el contenido sexista y se pongan medios para su censura. Los autores de este artículo [35] platearon un sistema de alerta anti-sexista capaz de detectar el contenido como sexista, no sexista o dependiendo del contexto. Para ello, definieron unos umbrales donde dado un texto de una noticia o un post en una red social, si la proporción de comentarios recibidos sexistas era superior al 5%, se consideraba sexista. Si la proporción de comentarios sexistas era superior al 2,5% e inferior al 5% se consideraba potencialmente sexista. Y si era inferior al 2,5% se consideraba no sexista.

Las fuentes que utilizaron para crear el *dataset* fueron publicaciones en español obtenidas de periódicos en la web, Twitter y Youtube, donde tuvieron en cuenta si los comentarios iban dirigidos a una persona, a un colectivo o ambos. En el proceso de etiquetado utilizaron el voto mayoritario para la anotación final y en la fase de entrenamiento tuvieron en cuenta si la fuente era de una mujer o un hombre y también el género al que iba dirigido.

Para crear el sistema anti-sexista los autores optaron por un modelo *Hate-speech-CNERG/dehatebert-mono-span*[17] previamente entrenado para detectar discursos de odio en español ya que consideraron que existían fuertes similitudes entre las estructuras lingüísticas del discurso de odio y los comentarios sexistas ofensivos. Por consiguiente, consiguieron clasificar correctamente el 84,6% de los comentarios y entre los mal clasificados la mayor confusión fue entre las clases de los umbrales bajo y medio. Por otro lado, identificaron que cuando la protagonista era una mujer, el 28,6% de las fuentes de contenidos se encontraban en el umbral bajo (no sexista), el 14,3% fueron asignados al umbral medio (potencialmente sexista) y el 57,1% fueron asignados como sexistas. Sin embargo, cuando un hombre era el protagonista, el 100% de las fuentes de contenido se consideraban como no sexista. Lo que corrobora una vez más que hay una alta difusión de contenido sexista en la red y la mayoría de los ataques van dirigidos hacia las mujeres.

[17] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-spanish

3.4 Identificación de sexismo en redes sociales

A diferencia de otros estudios sEXism Identification in Social neTworks[18] (EXIST) organizada en la conferencia CLEF[19], es una organización dedicada exclusivamente a la detección de sexismo hacia las mujeres. EXIST lleva tres ediciones celebrando una competición de PNL para identificar y clasificar los comentarios de contenido sexista en las redes sociales. A los participantes de la competición se les proporciona un corpus con el que a partir de técnicas de PNL y modelos de lenguaje tiene que ser capaces de clasificar el contenido de tweets como sexista o no sexista y categorizarlo en distintas categorías del sexismo.

3.4.1 Conjunto de datos y etiquetado

Para la primera edición de EXIST 2021[20] los autores formaron un *dataset* a partir de comentarios de Twitter y Gab[21] en inglés y español. Para evitar el sesgo temporal y de usuario, la descarga de datos se realizó en distintos periodos de tiempo durante un año utilizando un solo comentario por usuario. Además, se recopilaron semillas que obtuvieran contenido tanto sexista como no sexista. La estrategia que utilizaron para el etiquetado fue la del voto mayoritario por 5 expertos de *crowdsourcing* a través de la plataforma *Amazon Mechanical Turk*[22]. El *dataset* definitivo se formó de 5.644 comentarios para el inglés y 5.701 para el español [36].

En la segunda edición de EXIST 2022[23] se utilizó el *dataset* obtenido en la edición anterior para la fase de entrenamiento y se creó un nuevo conjunto de datos para evaluación. Se siguió el mismo procedimiento para descargar los tweets, y en este caso no se descargó contenido de Gab. En la fase de etiquetado se incluyó un anotador más y se tuvo en cuenta que hubiera el mismo número de anotadores mujeres y hombres para evitar un posible sesgo de género y en caso de empate el tweet se descartaba. Al *dataset* anterior se le añadieron 1.058 tweets para test resultando un total de 12.403 comentarios [37].

Para la edición de EXIST 2023[24], los autores siguen la misma metodología para la extracción de datos. En la fase de etiquetado en este caso utilizaron la plataforma Prolific[25] ya que permite elegir el perfil de los anotadores. Tuvieron en cuenta el género, contratando el mismo número de hombres que de mujeres, diferentes rangos de edad, procedencia y grado de comprensión del idioma para mitigar posibles sesgos de género, edad y culturales. Finalmente, el *dataset* se compuso de 3.200 tweets por idioma para el conjunto de entrenamiento, alrededor de 500 por idioma para el conjunto de validación y casi 1.000 tweets por idioma para el conjunto de prueba.

[18] http://nlp.uned.es/exist2024/
[19] https://clef2024.imag.fr/index.php?page=Pages/conference.html
[20] http://nlp.uned.es/exist2021/
[21] https://gab.com/
[22] https://www.mturk.com/
[23] http://nlp.uned.es/exist2022/
[24] http://nlp.uned.es/exist2023/
[25] https://www.prolific.com/

3.4.2 Tareas de EXIST

Para llevar a cabo la identificación y clasificación del sexismo EXIST propuso las siguientes tres tareas:

Identificación del sexismo. En todas las ediciones de EXIST la primera tarea ha consistido en una clasificación binaria para identificar el contenido de comentarios como sexistas o no sexistas.

- **Sexista:** *"Mujer al volante, tenga cuidado!"*
- **No sexista***: "Alguien me explica que zorra hace la gente en el cajero que se demora tanto."*

Categorización del sexismo. La siguiente tarea ha consistido en la categorización de los comentarios sexistas. Es una tarea jerárquica, multiclase y multietiqueta en la que el objetivo es clasificar los comentarios sexistas en las siguientes categorías:

- **Descrédito ideológico, negación de la desigualdad y narrativa invertida:** desacredita el movimiento feminista, rechaza la existencia de desigualdad entre hombres y mujeres, o presenta a los hombres como víctimas de la opresión de género, como en:

  "Mi hermana y mi madre se burlan de mí por defender todo el tiempo los derechos de todos y me acaban de decir feminazi, la completaron".

- **Estereotipos y dominancia:** el contenido expresa ideas falsas sobre las mujeres y sugieren que son más adecuadas o inapropiadas para ciertas tareas o afirma que los hombres son de alguna manera superiores a las mujeres, como en:

  "@Paula2R @faber_acuria A las mujeres hay que amarlas…solo eso… Nunca las entenderás.".

- **Cosificación:** el contenido presenta a las mujeres como objetos al margen de su dignidad y personalidad, o asume o describe ciertas cualidades físicas que las mujeres deben tener para cumplir con los roles tradicionales de género, como en:

  "Pareces una puta con ese pantalón" - Mi hermano de 13 cuando me vio con un pantalón de cuero".

- **Violencia sexual:** se realizan sugerencias sexuales, solicitudes de favores sexuales o acoso de carácter sexual (violación o agresión sexual), como en:

  "#MeToo Estas 4 no han conseguido su objetivo. El juez estima que se abrieron de patas ***https://t.co/GSHiiwgY6A****ánta lagartona hay en este \metoo"! 👇🏻 👇🏻 🤔 🤔 🤔* ***https://t.co/8t5VmFIUFn"***

- **Misoginia y violencia no sexual**: el contenido expresa odio y violencia hacia la mujer, como en:
- *"Las mujeres de hoy en dia te enseñar a querer… estar soltero"*

Intención de la fuente. En la edición de EXIST 2023 se incorporó una nueva tarea jerárquica y multiclase para identificar la intención de la fuente. El objetivo fue distinguir entre las siguientes categorías:

- **DIRECT:** La intención es difundir un mensaje que es sexista en sí mismo, como en:

  "Una mujer necesita amor, llenar la nevera, si un hombre puede darle esto a cambio de sus servicios (tareas domésticas, cocinar, etc), no veo que más necesita"

- **REPORTED**: la intención es denunciar y compartir una situación sexista sufrida por una mujer o mujeres en primera o tercera persona, como en:

  "Hoy, uno de mis alumnos de primer curso no podía creer que hubiera perdido una carrera contra una chica"

- **JUDGEMENTAL**: la intención es juzgar o reivindicar, ya que el meme describe situaciones o comportamientos sexistas con el objetivo de condenarlos, como en:

  "Como de costumbre, la mujer fue la que dejó su trabajo por el bienestar de la familia..."

3.4.3 Resultados

Para la evaluación de los modelos se utilizaron las métricas propuestas en la plataforma *Marco de Evaluación EvALL*[26], concretamente la métrica de *accuracy* para evaluar la primera subtarea, puesto que las clases estaban balanceadas, y *F1 macro* para la subtarea de categorización del sexismo de acuerdo con las diferentes categorías.

En la tarea de EXIST 2021 la mejor puntuación para el español fue para la subtarea 1 obteniendo un 0,794 de *accuracy*, con un conjunto de modelos *Transformers* para diferentes configuraciones: multilingüe, específico del idioma y con aumento de datos. Para la subtarea 2 la mejor puntuación fue 0,607 de *F1 macro* utilizando DeBERTa [38], una versión de los modelos BERT [10] y RoBERTa [25].

En la siguiente edición EXIST 2022 siguieron el mismo paradigma de evaluación y el mejor resultado para el español de nuevo es para la subtarea 1 que fue 0,780 de *accuracy* utilizando un conjunto de modelos con RoBERTuito [39] y BERT. Para la subtarea 2 se alcanzó un 0,487 de F1 macro, para la cual se usó un conjunto de 5 modelos Transformers distintos para el inglés y para el español.

[26] https://evall.uned.es/

En la última edición de EXIST 2023, los mejores resultados de nuevo se obtienen con modelos que utilizan la transferencia de aprendizaje, aumento de datos y modelos específicos de dominio de Twitter. Para la subtarea 1 el mejor resultado alcanza un 0,811 de *F1 macro* global (español e inglés). Para la subtarea 2 de categorización se alcanza un mejor resultado de 0,629 *F1 macro* global. Por último, en la subtarea de clasificación de la intención de la fuente se obtiene como mejor resultado un 0,571 *F1 macro* global.

En general, los modelos Transformers resultaron ser los más eficaces para capturar el contenido sexista, superando a los modelos de ML tradicionales en cada una de las ediciones de EXIST. Sin embargo, la complejidad que conlleva la naturaleza jerárquica multiclase y multietiqueta de la categorización del sexismo sigue planteando un desafío en el que seguir investigando para mejorar los resultados que se han conseguido hasta la actualidad.

4 Dataset de memes

En este capítulo se van a detallar cuáles han sido los pasos para crear un *dataset* de memes en español representativo de la temática del sexismo para poder llevar a cabo las tareas propuestas en EXIST 2024[27] y alcanzar los objetivos de este TFG.

4.1 Descarga de datos

Previamente a la descarga de los datos se ha realizado una búsqueda manual en Google Imágenes [28] a partir de una lista de términos y *hashtags* relacionada directamente con la temática de género, para evaluar cómo de productivas son las semillas que vamos a utilizar para obtener los memes. Estas expresiones han sido extraídas de diferentes fuentes: (a) trabajos previos en el área; (b) cuentas de Twitter (de periodistas, de adolescentes, etc.) o hashtags utilizados para denunciar situaciones sexistas; (c) expresiones extraídas de *The Every Day Sexism Project* [29] y (d) un compendio de diccionarios feministas. Además, han sido validadas por una psicóloga social experta en estudios de género y con la organización de EXIST 2024 puesto que este *dataset* será el utilizado para las nuevas tareas de su última edición.

Una vez validada la lista definitiva de semillas que nos devuelven memes representativos de la temática de género, tanto sexistas como no sexistas y en los que se identifican distintas tipologías del sexismo se realiza la descarga de datos en Google Imágenes. Se utiliza este mecanismo ya que nos permite extraer contenido de una amplia variedad de fuentes distintas de difusión de memes.

Posteriormente a la descarga se depuran los memes siguiendo los siguientes criterios: descartar imágenes que no se consideran memes porque no se ajustan directamente con la definición, como imágenes promocionales, anuncios, recortes de periódico etc.; los memes deben de estar directamente relacionados con la temática de la semilla; su contenido textual debe estar en español; se descartan recortes de memes o varios memes juntos; la calidad de la imagen debe de ser lo suficientemente buena para que el algoritmo de reconocimiento óptico pueda detectar el texto; por último, se descarta el contenido explícito considerado potencialmente delicado para algunos grupos.

Seguidamente se utiliza el algoritmo de reconocimiento óptico (OCR) de Google Vision[30] para extraer el texto de los memes. Esto nos servirá para incluir los textos en el *dataset* para posteriores análisis y para descartar memes repetidos. Dado que en algunos casos los mismos memes pueden aparecer en la búsqueda de semillas distintas, podemos descartar los memes repetidos tanto con la misma imagen y texto como los memes con el mismo texto, pero imágenes distintas ya que observamos que las pequeñas variaciones en la imagen no aportan ninguna información extra, por el contrario, podrían introducir sesgo en los modelos.

[27] http://nlp.uned.es/exist2024/
[28] https://images.google.com/
[29] https://everydaysexism.com/
[30] https://cloud.google.com/use-cases/ocr?hl=es

4.2 Anotación de los memes

Hemos utilizado la plataforma de *crowdsourcing* Prolific para llevar a cabo el proceso de etiquetado. Esta plataforma permite elegir distintas características del perfil de los anotadores. Para mitigar el sesgo de los anotadores hemos seleccionado el mismo número de hombres y mujeres, distintos rangos de edades, nacionalidades distintas y además hemos tenido en cuenta su nivel de estudios. Previamente a la anotación del *dataset* completo se ha probado con un subconjunto de datos para valorar la calidad de las anotaciones. Una vez llevada a cabo la anotación final se ha contado con 887 anotadores de 53 países diferentes y de los cuales cada meme ha sido etiquetado por 3 hombres y 3 mujeres.

Debido a la importancia que supone la correcta comprensión de la tarea y dado que tratamos con un tema de cierta subjetividad, se ha proporcionado a los anotadores una guía muy completa con definiciones y ejemplos:

- **Definición del sexismo**: en este apartado se incluye la definición de sexismo y una descripción clara del problema que supone en la actualidad en el contexto de las redes sociales y su dificultad para identificarlo.
- **Definición de meme**: en este apartamos se incluye la definición de meme y cómo se puede representar el sexismo en este tipo de contenido.
- **Formas de expresar el sexismo**: en este apartado se incluye las distintas formas de expresión que podemos encontrarnos, como amistoso, irónico, ofensivo, odio y violencia, etc. En forma de opresión, discriminación o prejuicio. Una discriminación basada en estereotipos, machismo, misoginia etc.
- **Objetivos de la tarea**: en este apartado se explican en detalle los dos objetivos de la tarea: identificación y categorización del sexismo.
- **Contenido sensible**: en este apartado advertimos que se va a visualizar contenido potencialmente sexista y de odio por si esto puede herir la sensibilidad del anotador y puede decidir no continuar con la tarea.
- **Instrucciones de etiquetado:** mediante un cuestionario de *Google Forms*[31] mostramos 30 memes, uno por uno. Hay que observar cuidadosamente la imagen y leer cualquier texto adjunto para responder a las siguientes preguntas:

  1. ¿El meme muestra sexismo de alguna manera? Esto incluye presentar contenido sexista, describir situaciones que involucren discriminación hacia las mujeres o criticar comportamientos sexistas.

  - **SI**, el meme es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista. Algunos ejemplos de memes en esta categoría son:

[31] https://docs.google.com/forms/

Figura 4.1: Memes sexistas

- **NO,** el meme no perjudica, discrimina o menosprecia a las mujeres ni se refiere a contenidos o situaciones en las que sí se haga. Algunos ejemplos de memes en esta categoría son:

Figura 4.2: Memes no sexistas

2. De acuerdo con que la faceta de la mujer está siendo atacada, ¿qué tipo de sexismo encontramos en el meme?

El sexismo puede afectar a las mujeres en muchas facetas de sus vidas, incluidas las funciones domésticas y de crianza, las oportunidades profesionales, la imagen sexual y las expectativas de vida, entre otras. De acuerdo con la faceta de la mujer que se ve afectada, se debe asignar al meme una o varias de las siguientes categorías que se definieron en el apartado 3.4.2 de las cuales mostramos algunos ejemplos a continuación:

- **Descrédito ideológico, negación de la desigualdad y narrativa invertida**

Figura 4.3: Memes descrédito ideológico

- **Estereotipos y dominancia**

Figura 4.4: Memes de estereotipo y dominancia

- **Cosificación**

Figura 4.5: Memes de cosificación

- **Violencia-sexual**

Figura 4.6: Memes de violencia-sexual

- **Misoginia y violencia no sexual**

Figura 4.7: Meme de misoginia y violencia

4.3 Dataset EXIST memes 2024

El *dataset* de EXIST 2024 y también el que se va a utilizar para llevar a cabo este trabajo, contiene un total de 2034 memes en español. Este conjunto de datos se proporciona en formato JSON y cada meme se representa en un objeto con los siguientes atributos:

1. **"id_EXIST":** un identificador único para cada meme.
2. **"lang":** lenguaje del meme (en este caso en "ES" para español).
3. **"text":** texto automático extraído del meme.
4. **"meme":** nombre del archivo que contiene al meme.
5. **"path_memes":** ruta que contiene al meme.
6. **"number_annotators":** número de personas que anotaron el meme.
7. **"annotators":** identificador único de cada anotador del meme.
8. **"gender_annotators":** género de los diferentes anotadores. Los posibles valores son: "F" and "M", para mujer y hombre respectivamente.
9. **"age_annotators":** el grupo de edad de los diferentes anotadores. Los posibles valores son: 18-22, 23-45 y 46+.
10. **"ethnicity_annotators":** la etnia declarada por los distintos anotadores. Los posibles valores son: *"Black or African America", "Hispano or Latino", "White or Caucasian", "Multiracial", "Asian", "Asian Indian"* y *"Middle Eastern".*
11. **"study_level_annotators":** el nivel de estudios alcanzado declarado por los distintos anotadores. Los posibles valores son: *"Less than high school diploma", "High school degree or equivalent", "Bachelor's degree", "Master's degree"* y *"Doctorate".*
12. **"country_annotators":** el país donde viven declarado por los distintos anotadores.
13. **"labels_task4":** el conjunto de etiquetas (una por cada uno de los anotadores) que indica si el meme contiene expresiones sexistas o hace referencia a comportamientos sexistas o no. Los posibles valores son: "YES" y "NO".
14. **"labels_task5":** un conjunto de etiquetas (una para cada uno de los anotadores) que registran la intención de la persona que creó el meme. Las posibles etiquetas son: "DIRECT", "JUDGEMENTAL", y "UNKNOWN".

15. **"labels_task6":** un conjunto de listas de etiquetas (una lista para cada uno de los anotadores) que indican el tipo o tipos de sexismo que se encuentran en el meme. Las posibles etiquetas son: “IDEOLOGICAL”, “INEQUALITY”, “STEREOTYPING-DOMINANCE”, “OBJECTIFICATION”, “SEXUAL-VIOLENCE”, “MISOGYNY-NON-SEXUAL-VIOLENCE”, “-”, y “UNKNOWN”.
16. **“split”:** subconjunto del *dataset* al que pertenece el meme (“TRAIN-MEME”, “TEST-MEME” + “EN” /”ES”, en este caso ES).

4.4 Estrategias de anotación

La estrategia que llevamos a cabo para la anotación final de la tarea binaria es la del voto mayoritario con el sistema de mayoría absoluta, ya que nos parece el criterio más justo para esta tarea sin forzar a que la etiqueta final tenga que pertenecer a una clase obligatoriamente. Sin embargo, tomando esta estrategia nos encontramos con 322 memes con empate, en los que ha habido un desacuerdo entre los anotadores de los cuales tres han votado como sexista y los otros tres como no sexista. Más adelante se seguirán distintas estrategias para intentar solucionar este problema.

En la segunda tarea de categorización del sexismo utilizamos la estrategia del voto simple en la que con mayor o igual a tres votos se considere perteneciente a esa categoría. Utilizamos esta estrategia menos restrictiva para poder incluir más elementos ya que si no el *dataset* se vería muy reducido para la segunda tarea y parece un criterio justo dado que estos memes ya se han considerado sexistas previamente.

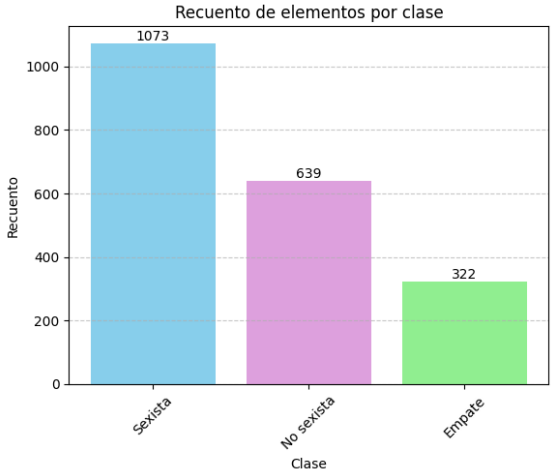

Figura 4.8: Distribución de muestras tarea 1

El *dataset* resultante es de 1.712 memes de los cuales hacemos una partición del 10% para entrenamiento y 10% para prueba. Y de nuevo otra partición del 10% de entrenamiento para evaluación de lo que resulta lo siguiente: 1.386 para entrenamiento, 154 para evaluación y 172 para prueba.

Como podemos observar en la figura 4.8, 1.073 de los memes son sexistas y 639 no lo son, lo que implica un desbalanceo notable en las distintas clases de la tarea

binaria. También se puede observar en la figura 4.9 un desbalanceo incluso mayor para las categorías del sexismo. Si observamos por clases incluso encontramos casos muy extremos, como por ejemplo la clase de misoginia y violencia no sexual solo contiene 88 elementos sexistas frente a 985 no sexistas. Este desbalanceo es esperable porque hay menos casos de sexismo de las conductas extremas. Por el contrario, hay más casos de las manifestaciones de sexismo que están más aceptadas socialmente.

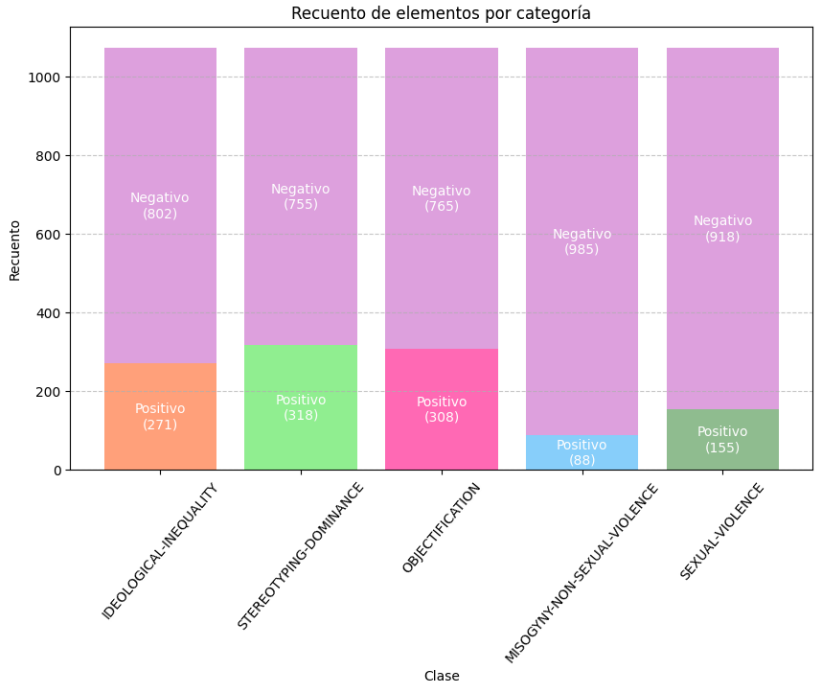

Figura 4.9: Distribución de muestras tarea 2

4.5 Desbalanceo de clases

Como hemos podido observar en el apartado anterior, existe un notable desbalanceo en general en todas las clases del conjunto de datos, tanto para la tarea binaria como para la tarea multiclase. Para ello, probaremos con las siguientes estrategias en el entrenamiento y valoraremos cuál es la más adecuada para nuestros datos.

Balance de pesos en la función de pérdida

Esta estrategia consiste en calcular los pesos de clase inversamente proporcionales a la frecuencia de cada clase y ajustar esos pesos a la función de pérdida en la fase de entrenamiento. Definimos el proceso paso a paso a continuación:

1. **Frecuencia de cada clase:** Se obtiene el número de muestras para cada clase de los datos.

2. **Frecuencia relativa por clase:** Se divide el número total de muestras por el número de muestras para cada clase. Posteriormente se normalizan estos pesos dividiéndolos por el máximo peso de clase para evitar que los pesos sean demasiado grandes.

3. **Tensor de pesos para la función de pérdida**: Se usan los pesos calculados para construir un tensor que pueda ser utilizado como argumento '*weight'* en la función de pérdida '*CrossEntropyLoss()'.*

Desempate

Una estrategia que podemos seguir, dado que conocemos el *dataset* y el proceso de anotación, es añadir un voto extra a los casos en los que los anotadores no se han puesto de acuerdo para conseguir el desempate.

A priori, esto no nos garantiza que los datos se vayan a balancear ya que se puede intuir que los memes son algo confusos de identificar. Sin embargo, aumentaríamos el número de datos del *dataset* y una vez llevado a cabo se puede valorar con qué conjunto de datos funciona mejor el proceso de experimentación.

Métodos de balanceo de datos

El ***oversampling*** **o sobremuestreo** es una técnica que replica el número de instancias de la clase minoritaria para aumentar su representación. Aunque de esta forma se puede equilibrar el conjunto de datos de forma sencilla, en los casos en los que el desequilibrio es elevado, al introducir muestras redundantes, pueden introducir sobreajuste al modelo. No obstante, lo probaremos y analizaremos cómo funciona con nuestros datos.

El ***undersampling*** **o submuestreo** es una técnica utilizada para equilibrar conjuntos de datos desbalanceados reduciendo el número de instancias de la clase mayoritaria. En nuestro caso, descartamos su aplicación porque reduciría demasiado el número de muestras disponibles.

5 Metodología y experimentación

En este capítulo, vamos a explicar la metodología y experimentación para llevar a cabo las tareas de identificación y categorización del sexismo en los memes mediante un aprendizaje supervisado con modelos Transformers.

Cuando tratamos con un tipo de contenido que fusiona varios canales de información, como es el caso de los memes, que combinan texto e imagen, podemos optar por separar estos canales y tratarlos de forma unimodal, o bien plantear un enfoque multimodal que combine ambos canales para obtener información adicional. En este apartado, se plantearán ambos enfoques y se describirá el proceso de experimentación y optimización de los modelos para conseguir el mejor sistema para llevar a cabo las dos tareas planteadas.

5.1 Modelos de texto e imagen

Para nuestro primer sistema unimodal vamos a usar dos modelos de tipo *text* Transformers de arquitectura encoder. Concretamente la versión específica de BERT (Bidirectional Encoder Representations from Transformers) [10] para el español, desarrollada por el equipo de investigación de la Universidad de Chile, que se conoce como BETO (Bidirectional Encoder Representations from Transformers for Spanish) [40]. Y *roberta-base-bne* de arquitectura RoBERTa [25] que fue desarrollado por PlanTL (Plataforma de Tecnologías del Lenguaje), una iniciativa del gobierno de España (GOB-ES) entrenado con corpus en español. Elegimos estos modelos por su buen rendimiento observado en la revisión del estado del arte y disponer de versiones específicamente entrenadas para el español.

Por otro lado, para el sistema de imagen vamos a utilizar un modelo pre-entrenado de visión por computador ViT [16]. Concretamente *google/vit-base-patch16-224-in21k* que es un modelo pre-entrenado en ImageNet-21k con 14 millones de imágenes y 21.843 clases en la resolución 224x224.

5.2 Preprocesado de texto e imagen

El preprocesado en los modelos Transformers es muy importante y a la vez muy fácil de implementar ya que no se requiere en la mayoría de los casos de un preprocesado previo, en su lugar llevamos a cabo un proceso de **tokenización** como se ha explicado en más detalle en el apartado 2.3.1.

Los tokenizadores que utilizamos son ***BertTokenizer*** y ***RobertaTokenizer*** de la librería *transformers*[32] ya que están optimizados específicamente para trabajar con estos modelos y proporcionan una mejor integración de los datos.

En el caso de ***BertTokenizer***, se utiliza el algoritmo de tokenización *WordPiece* [41] que descompone las palabras en subpalabras o tokens más pequeños. Esto método

[32] https://pypi.org/project/transformers/

es muy efectivo ya que permite captar los signos de puntuación. Los fragmentos de subpalabras se prefijan con "##" si no son el inicio de una palabra.

RobertaTokenizer**,** en cambio utiliza el algoritmo de codificación de pares de bytes *(Byte-Pair Encoding, BPE)* [42]. Este método es similar a *WordPiece* pero para la codificación combina los pares de bytes (caracteres o subpalabras) más frecuentes en nuevas subpalabras, lo que también permite captar los signos de puntuación.

Ambos tokenizadores tienen configurada una longitud máxima de secuencia de 512 tokens, que en nuestro caso equivale a una longitud máxima de 512 palabras por cada texto del meme. Dado que la configuración de los memes suele contener textos breves analizamos la longitud de nuestros textos y encontramos que el texto más largo contiene 207 términos, por lo que ajustamos este parámetro para no consumir tanta memoria de forma innecesaria.

En el caso de las imágenes hacemos un preprocesamiento en el que redimensionamos todas las imágenes a un único tamaño, en nuestro caso de 224x224 píxeles. Por otro lado, simplificamos el canal de colores para reducir su complejidad a RGB, ya que proporciona información sobre la intensidad de los colores primarios en cada píxel de la imagen.

Una vez hecho esto, las imágenes se pasan al procesador del modelo VIT y se hace una normalización de los vectores de los píxeles para que todos los valores estén en un rango de [-1,1]. Con esto se genera un tensor de *embeddings* que representará las características más importantes y será la entrada del modelo de imagen VIT. Este proceso se ha descrito más detallado en el apartado 2.4.1 del capítulo 2.

5.3 Tarea 1: Identificación del sexismo en memes

La primera tarea se centra en identificar el contenido sexista en los memes. Es una tarea de clasificación binaria en la que los modelos tienen que ser capaces de clasificar cada meme como sexista o no sexista.

5.3.1 Arquitectura unimodal de texto

Para aprovechar el conocimiento del que parte un modelo pre-entrenado como BETO y RoBERTa podemos congelar la mayoría de las capas de su red neuronal interna e incluir algunas nuevas para que se adapten a las tareas específicas y consigan extraer lo más importante de estos nuevos datos. Para ello, probamos con distintas arquitecturas hasta que damos con la que mejor se adapta a cada modelo y tarea.

El procedimiento en ambos modelos es similar, modificamos la última capa del modelo pre-entrenado (*last hidden state*), añadimos una capa que recibe el modelo y otra capa encargada de recibir las secuencias de texto tokenizadas, *(input_ids)* y máscaras de atención *(mask)*. Posteriormente, añadimos una capa *BatchNorm* con 768 neuronas (dimensión oculta o salida de BETO y RoBERTa) que aplica normalización por *batchs* a la salida del modelo, esto ayuda a estabilizar y acelerar el entrenamiento al normalizar las activaciones de cada capa.

Por otro lado, para evitar que la red se vuelva demasiado dependiente de algunas características específicas de los datos de entrenamiento añadimos una capa de regularización *Dropout* que en nuestro caso apaga aleatoriamente el 30% de las neuronas de la red. Posteriormente, añadimos una capa lineal que transforma la salida de un espacio de características del tamaño del modelo de 768 a 256 neuronas. Esto introduce no linealidades adicionales al modelo y permite aprender representaciones más complejas.

Seguidamente, añadimos una capa con una función activación *ReLU (Rectified Linear Unit)* [43] que se aplica después de la capa lineal para no introducir linealidades al modelo. Por último, añadimos otra capa lineal para la clasificación que reduce la dimensionalidad de la salida de 256 a 2 neuronas (correspondientes al número de clases), lo que obtendrá una salida con la probabilidad de pertenencia predicha para cada una de las clases. Podemos observar esta arquitectura en la figura 5.1.

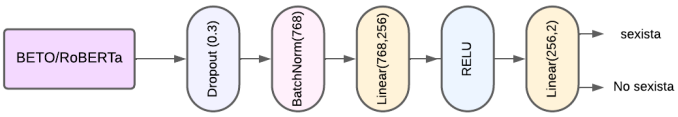

Figura 5.1: Esquema arquitectura de texto en tarea 1

5.3.2 Arquitectura unimodal de imagen

En la misma línea que con el modelo de texto congelamos la mayoría de las capas del modelo y hacemos una modificación en la última capa de la red para que el modelo sea capaz de extraer las características más importantes de las nuevas imágenes. La primera capa de esta subred es la que recibe el vector de *embeddings* de las imágenes. La siguiente es una capa *BatchNorm* con 768 neuronas (dimensión oculta o salida de ViT) y añadimos una capa *Dropout* de 0,3.

Para este modelo disminuimos el número de neuronas en un paso más suave. Añadimos una capa lineal que pasa de 768 a 512 neuronas, otra de 512 a 256 y añadimos una capa de función no lineal *ReLU*. Por último, una capa lineal para la clasificación que reduce la dimensión de 256 a 2 neuronas. Se puede ver el esquema de la arquitectura en la figura 5.2.

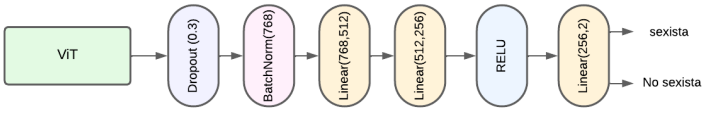

Figura 5.2: Esquema arquitectura de imagen tarea 1

5.3.3 Arquitectura multimodal

El enfoque multimodal combina los distintos canales de texto e imagen en un mismo sistema. Entre las distintas técnicas de fusión que podemos encontrar, hacemos uso de la concatenación ya que suele ser la que mejor funciona al conservar la

información original y a su vez permite la interacción entre las características de cada canal durante el entrenamiento del modelo.

Para llevar a cabo este proceso podemos distinguir entre dos arquitecturas. En *early fusion* o fusión temprana, una vez que se han extraído los vectores de *embeddings* de cada modalidad se fusionan en un único vector y se sigue con el procesamiento. En *late fusion* o fusión tardía se mantienen las representaciones de los *embeddings* hasta que todas las capas han sido procesadas y posteriormente se fusionan para generar la salida final. Podemos ver estos esquemas en la figura 5.3.

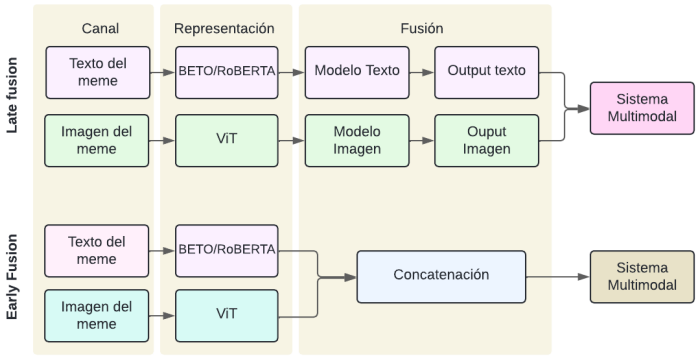

Figura 5.3: Esquema 'late fusion' y 'early fusion'

Probamos ambos esquemas y en nuestro caso *late fusion* es el que mejor funciona. Por lo tanto, utilizaremos esta técnica para crear la arquitectura del sistema multimodal. Además, esta técnica suele funcionar mejor cuando una de las modalidades aporta resultados significativamente mejores. Esto se debe a que el sistema captura una gama más amplia de características a diferentes niveles de abstracción y evita la pérdida de información en etapas tempranas, ya que las representaciones en capas posteriores contienen información más contextual y refinada [44].

El sistema multimodal final recibe los dos sistemas unimodales de texto e imagen con las arquitecturas descritas en los apartados anteriores e incluye una capa de concatenación que recibe la suma del número de neuronas del modelo de texto e imagen, 256 en cada caso (512), y reduce su dimensión a 256. Se añade una capa de función no lineal ReLU para no introducir linealidades. Por último, se añade la capa lineal para la clasificación que recibe 256 y tiene una salida de 2 neuronas, correspondientes al número de clases. Véase la representación de esta arquitectura en la figura 5.4.

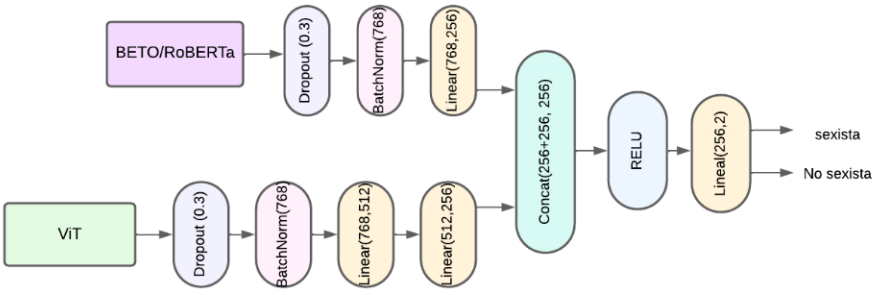

Figura 5.4: Esquema arquitectura multimodal tarea 1

5.4 Tarea 2: Categorización del sexismo en los memes

La segunda tarea consiste en clasificar los memes en las siguientes categorías del sexismo: descrédito ideológico, estereotipos y dominancia, cosificación, misoginia y violencia no sexual, las cuales están definidas en detalle en el apartado 3.4.2 del capítulo 3.

Esta tarea tiene una mayor complejidad, debido a su naturaleza multiclase y multietiqueta, donde las distintas categorías no son excluyentes entre ellas. Además, al quedarnos solo con los memes sexistas (1.703 memes) reducimos significativamente el conjunto de datos respecto a la tarea anterior. Además, si recordamos la distribución de las categorías en el capítulo anterior, nos encontramos categorías con muy pocas muestras, las cuales van a resultar muy difíciles de clasificar.

5.4.1 Arquitectura unimodal de texto

Para el modelo BETO multiclase generamos una arquitectura en la que la primera capa recibe el modelo pre-entrenado y una segunda capa el vector de *embeddings* que representa el texto. Añadimos una capa *Dropout* de 0,3 con 768 neuronas, seguida de una capa lineal de clasificación con una entrada de 768 neuronas y una salida de 5, que se corresponde con el número de clases de la tarea. Por último, una capa no lineal *Sigmoide*, que permite que cada clase sea predicha de forma independiente y las probabilidades de pertenencia a las clases no necesiten sumar 1.

En el caso de RoBERTa multiclase la arquitectura es similar, sin embargo, se añade una capa lineal extra que reduce el número de neuronas de 768 a 256 neuronas antes de la otra capa lineal de clasificación de 256 a 5 neuronas. Podemos ver los esquemas de las arquitecturas de los modelos de texto multiclase en la figura 5.5 y 5.6.

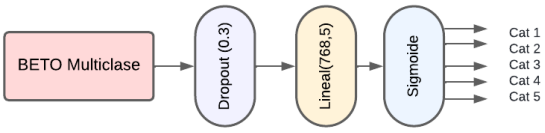

Figura 5.5: Esquema arquitectura de BETO tarea 2

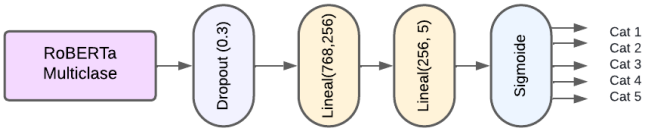

Figura 5.6: Esquema arquitectura de RoBERTa tarea 2

5.4.2 Arquitectura unimodal de imagen

Para el modelo VIT multiclase generamos una arquitectura en la que la primera capa recibe el modelo pre-entrenado y el vector de *embeddings* con la representación de las imágenes y una capa *Dropout* de probabilidad 0.3 con 768 neuronas.

Para este modelo añadimos una capa lineal que reduce el número de neuronas de 768 a 512, una segunda capa lineal que reduce de 512 a 256 y una capa lineal de clasificación con una entrada de 256 y una salida de 5 neuronas. Por último, una capa de función no lineal *Sigmoide*. Podemos ver el esquema de la arquitectura en la figura 5.7.

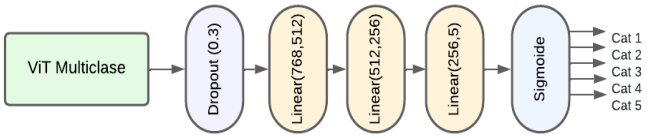

Figura 5.7: Esquema arquitectura Visual Transformer tarea 2

5.4.3 Arquitectura multimodal

Para crear los sistemas multimodales se sigue un procedimiento similar de concatenación como en la primera tarea. En este caso hay una variación respecto a fusionar directamente los sistemas unimodales y en el caso de BETO funciona mejor si se sigue la misma arquitectura que con RoBERTa, añadiendo una capa lineal más de 768 a 256 neuronas.

Para la representación de imágenes, utilizamos la misma arquitectura descrita en el enfoque unimodal. Posteriormente añadimos una capa de concatenación que recibe la suma de 256 neuronas de cada canal (512) y una salida de 5, correspondiente al número de clases. Por último, una capa con la función de activación no lineal *Sigmoide*. Podemos ver la arquitectura representada en el esquema de la figura 5.8.

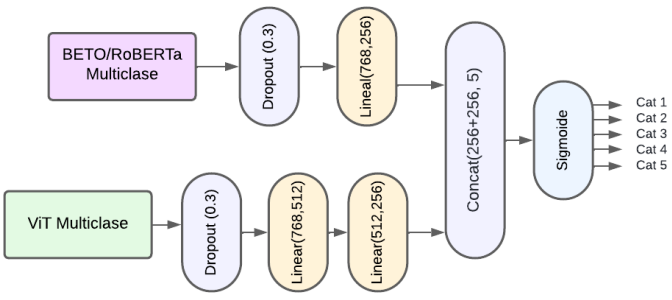

Figura 5.8: Esquema arquitectura multimodal tarea 2

5.5 Experimentación

El proceso de *fine-tuning* o 'ajuste fino' de un modelo Transformer es muy importante para adaptar y sacar el mayor partido al modelo con el conjunto de datos específico para la tarea, como hemos explicado con más detalle en el apartado 2.3.5 del capítulo 2.

En este tipo de modelos en los que existe cierta aleatoriedad en la partición de datos en algunas fases del entrenamiento es recomendable utilizar la misma semilla para que los experimentos sean reproducibles y comparables entre sí. Para ello, probamos con varias semillas y finalmente fijamos la *seed(42)* al inicio de la función de entrenamiento.

También incluimos mecanismos de regularización y normalización en algunas capas de los sistemas, como aplicar capas *LayerNorm* y capas *Dropout*, como se ha comentado en los apartados anteriores. En estas últimas se hicieron distintas pruebas con probabilidades de 0,1, 0,2, 0,3 y 0,4 en cada modelo durante el proceso de *fine-tuning* hasta dar con la que mejor funciona.

A continuación, especificamos algunos de los hiperparámetros que se han utilizado durante el proceso de *fine-tuning* y su funcionamiento fue descrito con detalle en el capítulo 2. Para calcular la diferencia entre la distribución de probabilidad predicha por el modelo y la distribución de probabilidad verdadera, utilizamos ***Cross Entropy Loss*** en la tarea binaria y ***Binary Cross Entropy Loss*** en la tarea multiclase y multietiqueta. Esta última es más adecuada cuando las categorías no son mutuamente excluyentes, puesto que calcula la pérdida como el promedio de las pérdidas binarias individuales y cada etiqueta es tratada como una clasificación binaria independiente.

Para ajustar los pesos y minimizar el *loss* de la función de pérdida probamos con dos algoritmos de **optimización:**

Por un lado, **Adam** [45] que combina las ventajas de dos otros métodos de optimización: *AdaGrad y RMSProp*. Además, mantiene un promedio de los gradientes y sus cuadrados y utiliza estos promedios para adaptar la tasa de aprendizaje de cada parámetro. Y **AdamW** [46] que es una variante de Adam e incorpora la regularización por decaimiento del peso directamente en la actualización de los parámetros. Además, modifica la forma en que se aplican los decaimientos de los pesos, corrigiendo el sesgo en la implementación de Adam.

También incluimos un **planificador** que sirve para ajustar la tasa de aprendizaje del optimizador durante el entrenamiento siguiendo la estrategia del *"One Cycle Policy"* [47]. Concretamente, esta estrategia se basa en ir variando cíclicamente el *LR* de forma que al inicio del entrenamiento parte de un valor muy pequeño y aumenta gradualmente hasta alcanzar el valor máximo *(max_lr).* Una vez alcanzado este punto, el *LR* disminuye gradualmente hasta un valor muy pequeño. Este proceso puede repetirse varias veces para intentar mejorar el mínimo global y acelerar la convergencia.

Por último, se llevaron a cabo varias pruebas con diferentes ***LR***, como 1e−3 y 1e−4 y también con 4e−3 y 4e−4, y finalmente con 5e−3, 5e−4 y 5e−5. Así como distintos números de ***batch*** en múltiplos de 3 entre 8 y 32. Y también con distintos números de ***epoch*** de 5, 10, 15 y 20.

5.6 Discusión

En el apartado 4.5 del capítulo anterior se comentaron distintas técnicas para intentar mejorar el problema del desbalanceo de las clases. En el proceso de experimentación se probaron todas ellas tal y como se detallaron y ninguna consiguió mejorar los resultados, por lo que descartamos su aplicación.

Por otro lado, en cuanto a la *loss* de la función de pérdida en la fase de experimentación, los valores son más pequeños al usar el optimizador AdamW respecto a Adam y mejoran al incluir el planificador, por lo que la evaluación se llevará a cabo con la aplicación de estos.

Algunas de las observaciones en la fase de experimentación han sido que en general todos los modelos tienden a sobreajustarse durante el entrenamiento, aun utilizando mecanismos para que esto no ocurra. Esta observación viene dada porque mientras que el valor *loss* en el entrenamiento disminuye de forma constante en cada *epoch*, el valor de *loss* en la validación aumenta considerablemente, lo que puede estar indicando que los datos estén sobreaprendiendo en exceso de los datos de entrenamiento. Podemos ver la representación gráfica del modelo multimodal de BETO + ViT en la tarea binaria en la figura 5.9.

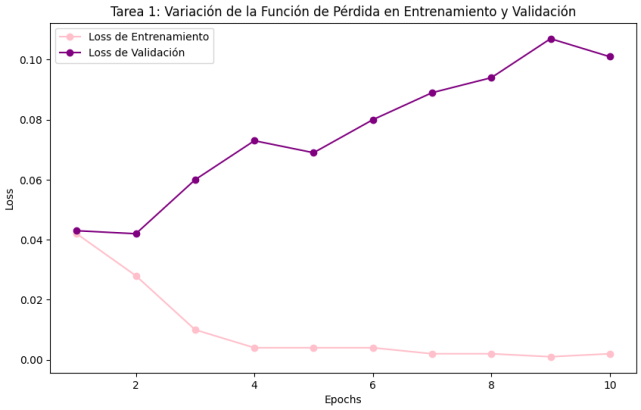

Figura 5.9: Variación de función de pérdida tarea 1

Sin embargo, cuando observamos la variación de la *loss* en la tarea multiclase se puede apreciar como el incremento de la *loss* en validación es mucho más estable, por lo que parece que los sistemas en esta tarea tienden a sobreajustarse menos. Podemos observar esa variación con el modelo multimodal de BETO + ViT de la tarea 2 en la figura 5.10.

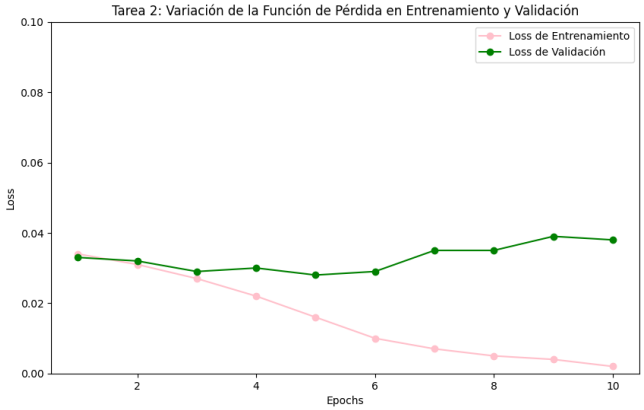

Figura 5.10: Variación de función de pérdida tarea 2

En la fase de experimentación, también se ha podido observar que el hiperparámetro más cambiante dependiendo del modelo y la tarea ha sido el número de *batchs*. En líneas generales, si aumentamos el número de *batchs* hasta 32, el *loss* en el entrenamiento disminuye de la misma forma y el aumento del *loss* en validación es mucho menor. Aunque esto no ocurre de la misma forma en la tarea multiclase, ya que con un número de *batchs* mayor a 16 los resultados empeoran.

Por otro lado, si aumentamos el número de *epochs* hasta 10, el valor de *loss* en entrenamiento disminuye, pero a partir de 10 *epochs* no consigue mejorar. En cambio, el *loss* en validación con más de 10 *epochs* sigue aumentando, y lógicamente su tiempo de computación también. Por lo que parece que 10 *epochs* es el valor óptimo para alcanzar el punto de convergencia en ambas tareas. Podemos observar los hiperparámetros óptimos dependiendo del modelo y la tarea en la tabla 5.1.

Modelo	Tipo	Epochs	Batchs	LR
BETO	**binario**	**10**	**16**	**5e-5**
RoBERTa	**binario**	**10**	**32**	**5e-5**
ViT	**binario**	**10**	**32**	**5e-4**
Multimod BETO_ViT	**binario**	**10**	**16**	**5e-4**
Multimod RoB_ViT	**binario**	**10**	**32**	**5e-4**
BETO	**multiclase**	**10**	**8**	**5e-4**
RoBERTa	**multiclase**	**10**	**16**	**5e-4**
ViT	**multiclase**	**10**	**8**	**5e-4**
Multimod BETO_ViT	**multiclase**	**10**	**16**	**5e-4**
Multimod RoB_ViT	**multiclase**	**10**	**16**	**5e-4**

Tabla 5.1: Hiperparámetros óptimos tareas 1 y 2

6 Evaluación y resultados

6.1 Resultados Tarea 1: Identificación del sexismo

Como métrica de evaluación y debido al desbalanceo que se ha observado en las clases utilizamos métricas ponderadas por el soporte de cada clase. Aunque las observaciones durante el proceso de *fine-tuning* parecen indicar que el modelo se está sobreajustando, los resultados en líneas generales indican que los sistemas han conseguido generalizarse pudiendo clasificar correctamente con más del 0.70 de F1 score de forma global en la mayoría de los casos. Incluso si lo observamos por clases, conseguimos hasta 0.82 en la clase positiva. En la tabla 6.1 incluimos el F1 score por cada clase y el global de cada modelo.

Modelo	Input	F1 global	F1 (+)	F1 (-)
BETO	**Texto**	**0.76**	**0.81**	**0.67**
RoBERTa	**Texto**	**0.71**	**0.78**	**0.63**
ViT	**Imagen**	**0.64**	**0.67**	**0.48**
Multimodal BETO ViT	**Imagen+texto**	**0.77**	**0.82**	**0.68**
Multimodal RoB ViT	**Imagen+texto**	**0.73**	**0.78**	**0.59**

Tabla 6.1: Resultados tarea 1

Es apreciable también que existe una diferencia significativa entre los dos enfoques unimodales donde el canal de texto es el que más información aporta al modelo lo cual se puede ver reflejado claramente en la clasificación. Concretamente, BETO es el modelo de texto que mejor funciona.

Cuando los distintos canales de forma unimodal consiguen diferencias significativas entre ellos, como es en este caso, el enfoque multimodal no siempre logra superar los resultados que consigue el canal más potente. Sin embargo, el sistema multimodal consigue una ligera mejora respecto al modelo de texto, lo que parece indicar que en estos casos las imágenes sí que están aportando algo de información extra al modelo.

6.1.2 Análisis estadístico

En el apartado anterior se ha evaluado la capacidad de predicción de los modelos, pero también es interesante analizar estadísticamente si hay una diferencia significativa entre el modelo de texto y el sistema multimodal puesto que los resultados son muy similares.

El estadístico de McNemar [48] comparándolo con una distribución chi-cuadrado puede ser útil para determinar si hay una diferencia significativa entre dos modelos cuando utilizamos la misma muestra para evaluarlos. La idea es que solamente los pares cuyos miembros se comportan de forma diferente en los dos modelos son los que contribuyen en la diferencia de comportamiento del modelo.

Por consiguiente, calculamos una tabla de contingencia y para calcular el estadístico solo tendremos en cuenta los resultados donde un modelo predice positivo y el otro negativo y viceversa. El resultado que obtenemos con el estadístico de McNemar es de 15.00 y el p-valor de 1,00. Tomando como referencia un nivel de significancia de α = 0,05 y 1 grado de libertad, ya que la tabla de contingencia tiene unas dimensiones de (2 filas - 1) * (2 columnas - 1), el valor correspondiente en la tabla de distribución Chi-Cuadrado es de 3,841[33].

Dado que el valor de McNemar es mayor que el valor de la chi-cuadrado parece haber una diferencia estadística entre ambos modelos. Sin embargo, el p-valor es mayor que el nivel de significancia y por lo tanto no podemos rechazar la hipótesis nula de que no hay diferencia significativa entre los dos modelos de clasificación.

6.2 Resultados Tarea 2: Categorización del sexismo

En general, los sistemas no parecen tener la capacidad suficiente para clasificar correctamente todas las categorías del sexismo. Esto es razonable si consideramos el número de elementos, ya que en algunas clases puede ser insuficiente para generalizar.

Las métricas utilizadas en este caso son el F1 score para evaluar cada clase y el F1 score ponderado por el soporte de la clase, para tener una visión global de la tarea. El mejor resultado en términos globales ha sido un F1 score ponderado de 0,54 con el sistema multimodal de BETO [10] y VIT [16], lo que sugiere una clasificación un tanto aleatoria. Sin embargo, al observar los resultados por clases, se obtiene un F1 score de 0,72 para la clase de cosificación, lo que representa un resultado significativamente mejor. Le sigue un F1 score de 0,67 con el modelo de texto BETO, y un F1 score de 0,68 con el modelo RoBERTa [25] en la clase de estereotipo y dominancia.

Las clases mejor clasificadas coinciden con las que tienen más muestras positivas, lo que parece indicar que el resto de las clases no se clasifican correctamente debido a la falta de muestras y, por lo tanto, a la insuficiente información para los modelos. Esto parece evidente en la clase de misoginia y violencia no sexual, que contienen un número significativamente menor de muestras y ningún modelo ha logrado clasificar correctamente estos elementos. Es importante señalar que, debido a la falta de muestras en esta clase, solo seis muestras misóginas aparecieron en los datos de evaluación, y en ningún caso los modelos las clasificaron correctamente, resultando todas falsos negativos. Todos estos resultados, junto con los del resto de las clases, se pueden observar en la tabla 6.2.

Modelo	Des-Ideolog.	Ester-Dom.	Cosific.	Misog.	Viol-sex.	F1 Global
BETO	**0.51**	**0.50**	**0.67**	**0.0**	**0.48**	**0.52**
RoBERTa	**0.54**	**0.68**	**0.62**	**0.0**	**0.26**	**0.53**
ViT	**0.50**	**0.38**	**0.53**	**0.0**	**0.22**	**0.41**
BETO+ViT	**0.51**	**0.50**	**0.72**	**0.0**	**0.48**	**0.54**
RoB+ViT	**0.62**	**0.58**	**0.64**	**0.0**	**0.11**	**0.52**

Tabla 6.2: Resultados tarea 2

[33] https://es.slideshare.net/slideshow/tabla-chi-cuadrado-pspp/81846451

6.3 Análisis de errores

En este apartado vamos a centrarnos en hacer una observación de los errores en la que analizaremos cada uno de los memes clasificados incorrectamente (FP y FN). Concretamente lo haremos con el modelo multimodal de BETO + ViT ya que ha sido el mejor sistema que hemos obtenido en ambas tareas. En este análisis se pretende identificar en la medida de lo posible qué tipo de errores son los más habituales y si los errores guardan algún tipo de relación para cada categoría. También si algunos elementos en los memes son más confusos o problemáticos a la hora de conseguir una correcta clasificación.

6.3.1 Errores en identificación del sexismo

En la primera tarea, si observamos la matriz de confusión del sistema multimodal de BETO y VIT vemos que, a pesar de haber una gran diferencia entre el número de elementos en la clase positiva y negativa, la diferencia entre los distintos errores es mínima. Únicamente de dos elementos más en la clase negativa clasificados incorrectamente (FP), como podemos observar en la matriz de confusión de la figura 6.1.

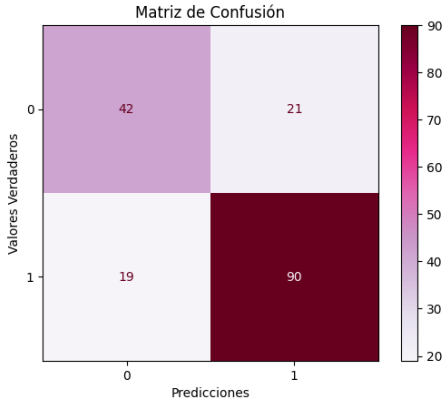

Figura 6.1: Matriz de confusión tarea 1

Observándolo por clases y analizando los errores se puede observar como la mayoría de los errores se deben a que son elementos algo confusos y con cierta dificultad para clasificar.

Por otro lado, también nos encontramos con algunos casos en los que el elemento textual tiene más peso que el de la imagen o su conjunto, y esto se refleja al cometer el error, como es el caso de la figura 6.2 que es un FP donde parece que el modelo lo clasifica como sexista por contener el término ‘zorra’ pero en este caso no se refiere al término despectivo y malsonante con el que se refieren en algunos casos a las mujeres, si no al animal.

Figura 6.2: Meme FP tarea 1

En cuanto a los casos en los que el modelo no los ha identificado como sexistas y si lo son (FN), los errores más comunes han sido casos en los que se refiere a alguna crítica de situación o conducta sexista y que en este estudio también se incluyen como contenido sexista. O casos en los que parece que los memes no contienen evidentes elementos textuales, y al contrario todo el peso sexista recae en la parte de la imagen, lo que para los modelos es difícil de identificar. Como es el caso del meme de la figura 6.3, donde si nos fijáramos solo en los elementos textuales no se consideraría sexista.

Figura 6.3: Meme FN tarea 1

6.3.2 Errores en las categorías del sexismo

Obtenemos una matriz de confusión para cada categoría del sexismo y de esta forma podemos identificar de forma sencilla cuáles han sido las clasificaciones correctas (TP y TN) y las incorrectas (FP y FN) para cada clase. Si hablamos en términos globales en cuanto a clasificaciones correctas un 72,59% han sido TN y un 10,55% TP, lo que no es sorprendente debido al desbalanceo que hay entre las distintas categorías y que ya se ha comentado con anterioridad. Respecto a las clasificaciones incorrectas nos encontramos con un 6,48% de FP y un 10,37% de FN en términos globales.

En las matrices de confusión por categorías, se puede observar la gran dificultad que ha habido para clasificar correctamente algunas de las clases. La clase de cosificación ha sido la clase mejor clasificada, posiblemente debido a ser la que más muestras tiene de la clase positiva. Además, en cuanto a las clasificaciones erróneas están bastante equilibradas. Sin embargo, con respecto a la clase de misoginia, el modelo no ha conseguido clasificar ningún elemento de la clase positiva correctamente. Posiblemente pudiendo deberse a contener un número insuficiente de muestras, como ya se comentó anteriormente. Véase las matrices de confusión en la figura 6.4.

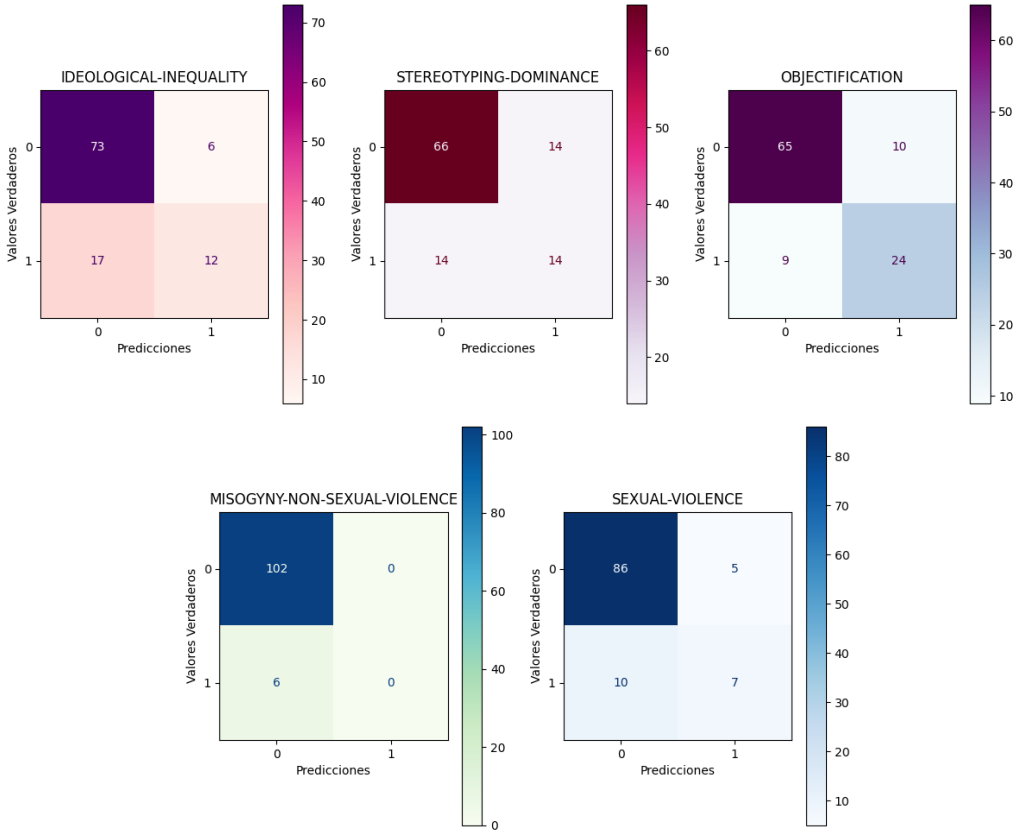

Figura 6.4: Matrices de confusión por categorías

A continuación, vamos a analizar con más detalle los tipos de errores más comunes en cada categoría:

▪ **Descrédito ideológico**: en esta categoría el error más cometido es el de FN con un 15,74% de los casos. Se puede observar cierta tendencia a clasificar erróneamente cuando el meme contiene algún término o imagen muy relacionado con la otra categoría. Por ejemplo, el caso de atributos físicos sobre las mujeres con la categoría de cosificación, o si hay alguna mención a algún término de tipo sexual con la categoría de violencia-sexual. Podemos ver un ejemplo de un FN en la figura 6.5, clasificado como cosificación posiblemente debido a detectar el elemento de Barbie en el meme y por ello relacionarlo directamente en esta categoría.

En cuanto a los FP nos encontramos con un 5,55% de los casos, y aquí ocurre lo mismo, por ejemplo, encontramos casos en los que el meme se clasifica como descrédito-ideológico cuando en el meme se identifica algún elemento relacionado directamente con esta categoría. Por ejemplo, cuando en el meme se reconocen a mujeres políticas defensoras de la igualdad.

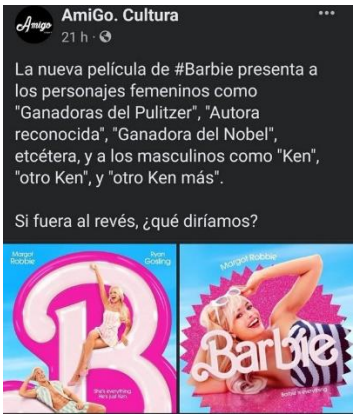

Figura 6.5: FN descrédito ideológico como cosificación

▪ **Estereotipo y dominancia**: en esta categoría nos encontramos con el mismo porcentaje de errores de 12,66 % tanto para FN como FP. Podemos observar como en el caso de los FN en la mayoría de los casos el modelo tiende a clasificarlos con la categoría de cosificación, tanto si el meme pertenece a ambas o solo a estereotipo-dominancia. En algunos casos se observan ciertos elementos textuales más relacionados con esta clase, por ejemplo, cuando el meme menciona algo sobre 'las rubias', como es el caso del meme de la figura 6.6.

Por otro lado, se suele observar cierta tendencia en los FP con referencias textuales de generalización de las mujeres en el tiempo como, por ejemplo, 'a veces las mujeres…'. Pero no necesariamente en todos los casos el meme menciona un estereotipo.

Figura 6.6: FN estereotipo y dominancia como cosificación

▪ **Cosificación**: en esta clase nos encontramos que la mayoría de los errores son FP con un porcentaje del 9,25% frente a los FN 8,33%. Por un lado, observamos que los memes que se clasifican en esta categoría en algunos casos también se clasifican

en la categoría de violencia sexual. En cuanto a los FN ocurre lo mismo que en la categoría anterior: el error más común es clasificar el meme en la categoría de estereotipo-dominancia y suele ser porque detecta algún tipo de expresión de generalización como, por ejemplo, 'la mayoría de las mujeres…' se puede ver un ejemplo en la figura 6.7.

En el caso de los FP se puede observar algo similar ya que en algunos casos el modelo relaciona ciertos términos con esta categoría, por ejemplo, cuando se menciona algo relacionado con el aspecto físico o gordofobia.

Figura 6.7: FN de cosificación como estereotipo y dominancia

▪ **Misoginia y violencia no sexual**: nos encontramos con una categoría que solo contiene un 5,55% de errores correspondientes a FN, a pesar de eso no se ha conseguido clasificar ningún meme de la clase positiva correctamente, como ya se ha comentado con anterioridad. Lo cual tiene aún más sentido al realizar el análisis de errores ya que nos encontramos con memes algo confusos y difíciles de clasificar. Podemos ver un ejemplo en la figura 6.8, donde el meme debería clasificarse en esta categoría y sin embargo se ha clasificado como estereotipo y dominancia.

Figura 6.8: FN de misoginia y violencia no sexual como estereotipo y dominancia

- **Violencia sexual**: en esta categoría encontramos un 9,25 % de FN frente a 4,62% de FP. Se puede observar cierta tendencia por parte del modelo a cometer errores de tipo FN cuando contextualmente el meme hace referencia a algo de carácter sexual pero no de forma explícita textualmente. En esos casos el modelo tiende a clasificarlos erróneamente en la categoría de cosificación, como es el caso del meme de la figura 6.9.

De la misma forma, pero al contrario ocurre con los FP, en los que en muchos casos al detectar algún término de carácter sexual lo clasifica directamente con esta categoría.

Figura 6.9: FN violencia y sexual como cosificación

En resumen, como hemos podido observar algunas categorías son más difíciles de clasificar tanto por contener muy pocas muestras de la clase positiva como por contener memes que pueden considerarse algo confusos para su clasificación, como es el caso de la categoría de misoginia y violencia no sexual.

Por otro lado, los elementos textuales además de aportar más información al modelo, como se ha visto a lo largo del trabajo, también tienden a tener más peso para confundirlo e impedir su correcta clasificación.

7 Conclusiones y trabajos futuros

7.1 Conclusiones

En el marco de este TFG se ha abordado el problema de la identificación del sexismo en memes. Se ha creado un *dataset* representativo de la temática de género en el que se incluyen memes sexistas, no sexistas y de las distintas categorías del sexismo: desigualdad ideológica, estereotipo-dominancia, cosificación, misoginia y violencia sexual. Esto ha permitido poder llevar a cabo las tareas que ha propuesto EXIST[34] en su última edición. Además, el *dataset* podrá servir a otros investigadores para poder seguir investigando sobre el sexismo y los memes puesto que anteriormente no existía un *dataset* específico para poder llevar a cabo este tipo de tarea.

Para llevar a cabo estas tareas se han empleado modelos Transformers y Visual Transformers para procesar los textos e imágenes en los memes. Se han planteado tanto enfoques unimodales como multimodales y aunque el canal de texto es el que aporta la mayor parte de información a los modelos, en líneas generales, el canal de imagen también aporta un extra de información y, al fusionar ambos canales en un sistema multimodal, se consigue una mejor identificación y clasificación de contenido sexista en los memes. Por otro lado, el tiempo de cómputo de un sistema multimodal es prácticamente el doble, por lo que en líneas generales es interesante valorar si la diferencia es significativa para emplear estos sistemas.

En cuanto a la primera tarea de identificación de sexismo en los memes, se ha conseguido clasificar correctamente la mayoría de los casos obteniendo un F1 score de 0,77 con el sistema multimodal de BETO + ViT. Además, la diferencia de los errores de clasificación entre la clase positiva y negativa ha sido mínima pese a la diferencia de muestras entre las dos clases.

Sin embargo, la clasificación de los memes sexistas en las distintas categorías ha resultado una tarea muy complicada. Esto se debe, por un lado, a la complejidad de la tarea al tratarse de una tarea multiclase y multietiqueta, pero también porque al quedarnos solo con los datos sexistas reducimos significativamente el número de datos con respecto a la otra tarea. Además, el reparto entre las muestras de las distintas clases es desigual, y nos encontramos clases con muy pocas muestras. De lo que se observa una mejor clasificación para las categorías con más muestras, como es la clase de cosificación con la que se consigue un 0.72 de F1 score con el sistema multimodal de BETO + ViT. Por el contrario, la categoría con menos muestras de misoginia y violencia no sexual ha demostrado una gran dificultad para la identificación de estos memes, ya que ninguno de los sistemas ha conseguido su clasificación.

7.2 Análisis de problemas, legal y ética

En relación con los aspectos legales y las posibles problemáticas asociadas a ello, especialmente cuando tratamos con datos, en este TFG se emplean memes descargados de Google Imágenes, para los cuales no se requiere obtener consentimiento debido a su naturaleza pública. Aunque estos memes se originan en

[34] http://nlp.uned.es/exist2024/

otras plataformas de difusión de contenido multimedia, en ningún caso se tiene acceso a datos personales del autor responsable de la creación o manipulación del contenido, descartando la necesidad de un tratamiento específico en términos de protección de datos.

En cuanto a la dimensión ética, este estudio trata con contenido delicado para ciertos grupos, especialmente para las mujeres. Por esta razón, se mantiene una consciente consideración de este aspecto a lo largo de todo el proyecto. En este sentido se descarta el material explícito que pueda herir la sensibilidad y tener un impacto negativo. Además, en el proceso de anotación se advierte de que se va a tratar con contenido sensible y se da la posibilidad de abandonar este proceso si de alguna forma puede afectar a la sensibilidad de las personas que llevan a cabo esta tarea.

Por otro lado, y también en referencia a la parte ética, es importante tener especial cuidado cuando se llevan a cabo tareas en las que se tratan temas con cierta subjetividad, como es el sexismo. O cuando hay más muestras de algunas clases y los modelos pueden sesgarse hacia alguna de ellas impidiendo representar la realidad. En este sentido, desde la fase inicial de creación del conjunto de datos hemos tratado de evitar posibles sesgos de género, de edad y culturales eligiendo un perfil equitativo en los anotadores, además de utilizar las estrategias más justas para decidir el voto final de la clase. Así como llevar a cabo distintas estrategias para solucionar el desbalanceo entre las clases y en la fase de evaluación controlar que los modelos no estuvieran sesgados hacia las clases mayoritarias.

7.3 Mejoras y trabajos futuros

Como ampliación de este trabajo se propone incrementar el *dataset* desarrollado, lo que parece bastante viable debido a la naturaleza tan cambiante y creciente que tienen los memes. Además, sería importante identificar nuevas semillas para extraer sobre todo más elementos de las clases que tienen menos muestras.

En esta línea y de forma complementaria, también se propone utilizar técnicas de *Data Augmentation* para generar nuevas muestras sintéticas, dado que en este proyecto no ha sido posible debido a la limitación del tiempo. Una propuesta para ello sería utilizar los memes en inglés, traducir sus textos e incluirlos como nuevos memes en el *dataset*.

Por otro lado, se propone utilizar otras técnicas no tan rígidas como la del voto mayoritario para evaluar las anotaciones de las clasificaciones, como por ejemplo utilizar el paradigma de *Learning with Disagreement* [49] que se propone en el marco de EXIST.

Dado que en este TFG se ha estudiado el sexismo tanto desde perspectivas unimodales como multimodales y en líneas generales la multimodalidad ha arrojado mejores resultados, se propone profundizar en el estudio de la multimodalidad y utilizar modelos Transformers multimodales [50].

7.4 Legado

Por un lado, la creación de un *dataset* de memes representativos del sexismo, hasta ahora inexistente, proporciona a los investigadores una herramienta valiosa para continuar trabajando en esta área. Este *dataset* puede ser utilizado en estudios futuros

relacionados con el análisis del sexismo en los memes, así como en investigaciones complementarias.

Por otro lado, este trabajo ofrece una base sólida para arrojar más conocimiento sobre cómo el sexismo se manifiesta y se difunde a través de los memes y aporta algunas ideas para seguir trabajando en ello.

7.5 Relación del trabajo con la carrera Ciencia de Datos

Este TFG ha sido en parte una recapitulación de diversas ramas de la Ciencia de Datos, donde hemos podido aplicar muchos de los conocimientos aprendidos a lo largo de la carrera, así como explorar nuevas técnicas no vistas y desarrollar una mayor capacidad de investigación y aprendizaje autodidacta.

Se ha abordado la recopilación y descarga de datos utilizando técnicas como *web scraping* estudiada en la asignatura de Adquisición y transmisión de datos. Seguido de procesos de limpieza y formateo de los datos para crear un buen *dataset,* con el que posteriormente se ha llevado a cabo un análisis explorario de los datos, procesamiento del lenguaje, despliegue y evaluación de modelos de ML con técnicas y metodología estudiadas en asignaturas como Modelos descriptivos y predictivos I y II, Lenguaje natural y recuperación de la información y Evaluación, despliegue y monitorización de modelos. Gran parte de este trabajo se ha puesto en práctica anteriormente en otros proyectos como los que se han realizado en las asignaturas de Proyecto I, II y III a lo largo de la carrera. Por último, este TFG se ha llevado a cabo desde un punto de vista ético teniendo en cuenta el conocimiento aprendido en la asignatura de Marco profesional, legal y deontológico.

A su vez es importante destacar que todo el trabajo realizado se ha llevado a cabo utilizando el lenguaje de programación Python, que es ampliamente reconocido y utilizado en la Ciencia de Datos. Todo esto ha sido posible tras llevar a cabo un constante proceso de aprendizaje en programación, desde el inicio hasta el final de la carrera y aplicado en la mayoría de las asignaturas. Además, entre las librerías de Python más utilizadas en este TFG podemos destacar: Pytorch[35], Transformers[36], Sklearn[37] y Matplotlib[38] entre otras.

7.6 Posibles aplicaciones

Es sumamente importante abordar este tipo de estudios para dar visibilidad al sexismo en cualquier forma de manifestación, y en especial en los memes, ya que no se había abordado hasta la fecha. La detección automática del sexismo en los memes puede proporcionar mecanismos para frenar e impedir su difusión masiva, la cual se está realizando prácticamente sin censura hasta ahora.

[35] https://pytorch.org/
[36] https://pypi.org/project/transformers/
[37] https://scikit-learn.org/stable/
[38] https://matplotlib.org/

Una posible aplicación sería automatizar la detección de memes sexistas en las redes sociales, permitiendo bloquear y frenar su difusión. Esta herramienta podría integrarse en plataformas sociales para identificar y detener la propagación de posible contenido sexista antes de que alcance una amplia difusión.

Otra aplicación podría ser en las páginas web dedicadas a la creación de memes. En este caso, los memes tendrían que pasar por un filtro antes de poder ser descargados, bloqueando aquellos que sean potencialmente sexistas. Esto aseguraría que los usuarios no puedan acceder ni difundir contenido sexista desde su origen.

En resumen, la implementación de estas aplicaciones no solo visibilizaría el problema del sexismo en los memes, sino que también proporcionaría herramientas efectivas para mitigar su impacto negativo en la sociedad.

Referencias

[1] P. Glick y S. T. Fiske, «Ambivalent sexism» *Advance in experimental social psychology*, vol. 33,115-188, 2001.

[2] M. Arteaga-Barba, K. Escamilla, K. Sánchez, J. León, G. Guzman y J. Herrera, «Aproximación socio-histórica y psicoanalítica del machismo y sexismo» *Boletín Científico de la Escuela Superior Atotonilco de Tula,* vol. 8, pp. 45-50, 2021.

[3] M. Zabalgoitia Herrera, «Masculinidad/es y violencia sexista/sexual en las relaciones cotidianas en la universidad» *Instituto de Ciencias Nucleares UNAM,* 2023.

[4] Delegación del Gobierno contra la Violencia de Género, «Mujeres víctimas mortales por violencia de género en España. Disponible en: https://violenciagenero.igualdad.gob.es/wp-content/uploads/VMujeres_2023_act_11_04_2024.pdf» 2024.

[5] C. Laudano, «Acerca de la apropiación feminista» *Memoria Acádemica. Compartimos lo que sabemos,* 2018.

[6] M. Nuevo Espín, «El éxito de los memes, ¿por qué triunfa su lenguaje entre los jóvenes?» *Hacer familia,* 2024.

[7] M. Z. Herrera, «Retóricas del meme masculinista» *Mitologías hoy. Revista de pensamiento, crítica y estudios literarios latinoamericanos,* 2022.

[8] L. A. García- González y O. Bailey Guedes, «Memes de Internet y violencia de género a partir de la protesta feminista #UnViloladorEnTuCamino» *Virtualis. Revista de cultura digital,* 2020.

[9] A. Vaswan, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser y I. Polosukhin, «Attention is All you Need» *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA,* 2017.

[10] F. A. Acheampong, H. Nunoo-Mensah y C. Wenyu, «Transformer models for text-based emotion detection: a review of BERT-based approaches» *Artificial Intelligence Review,* vol. 54, p. 5789–5829, 2021.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child y A. Ramesh, «Language Models are Few-Shot Learners» de *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877--1901.

[12] K. C. Bart van Merriënboer Caglar Gulcehre, D. Bahdanau, F. B. Holger Schwenk y Y. Bengio, «Learning Phrase Representations using RNN Encoder–Decoder» *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 1724--1734, 2014.

[13] J. Vermorel, «Cross-entropy» Lokad Quantitative Supply Chain, 2018. [En línea].

[14] K. Pykes, «Cross-Entropy Loss Function in Machine Learning: Enhancing Model Accuracy» *Radar AI Edition,* 2024.

[15] K. O'Shea y R. Nash, «An Introduction to Convolutional Neural Networks» *ArXiv e-prints,* 2015.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghan, . M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit y N. Houlsby, «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale» *ICLR Conference 2021.*

[17] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia y D. Testuggine, «The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes» *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada..*

[18] K. He, X. Zhang, S. Ren y J. Sun, «Deep Residual Learning for Image Recognition» *Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. ,* pp. 770-778, 2016.

[19] J. Lu, D. Batra, D. Parikh y S. Lee, «ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks» *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.,* pp. 13-23, 2019.

[20] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh y K.-W. Chang, «VisualBERT: A Simple and Performant Baseline for Vision and Language» *CoRR,* 2019.

[21] S. Shardul, A. Mihael, L. Suzanne y B. Paul, «Multimodal Offensive Meme Classification with Natural Language Inference» *Proceedings of the 4th Conference on Language, Data and Knowledge, 2023,* pp. 134-145.

[22] R. Mokady, A. Hertz y A. H. Bermano, «ClipCap: CLIP Prefix for Image Captioning» *CoRR,* vol. abs/2111.09734, 2021.

[23] D. Yuhao, M. Muhammad Aamir y K. Joseph, «Understanding Visual Memes: An EmpiricalAnalysis of Text Superimposed on Memes Shared on Twitter» *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020),* pp. 153-164.

[24] S. Suryawanshi, B. R. Chakravarthi, M. Arcan y P. Buitelaar, «Multimodal Meme Dataset Multi-OFF for Identifying Offensive Content in Image and Text» *European Language Resources Association (ELRA). Marseille, France,* 2020.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer y V. Stoyanov, «RoBERTa: A Robustly Optimized BERT Pretraining Approach» *CoRR,* 2019.

[26] T. G. Diettrich, «Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms» *Department of Computer Science. Oregon State University Corvallis,* 1997.

[27] M. Anzovino, E. Fersini y P. Rosso, «Automatic Identification and Classification of Misogynistic Language on Twitter» *Proc. 23rd Int. Conf. on Applications of Natural Language to Information Systems (NLDB 2018),* pp. 57-64, 2018.

[28] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals y G. E. Dahl, «Neural Message Passing for Quantum Chemistry» *Proceedings of the 34 th International Conference on Machine,* 2017.

[29] M. Anzovino, E. Fersini y P. Rosso, «Overview of the Task on Automatic Misogyny Identification at IberEval 2018» *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).*

[30] E. Fersini, D. Nozza y P. Rosso, «Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)» *EVALITA Evaluation of NLP and Speech Tools for Italian,* 2018.

[31] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees y S. Jeffrey, «Multimedia Automatic Misogyny Identification» *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022),* p. 533–549, 2022.

[32] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwa, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger y I. Sutskever, «Learning Transferable Visual Models From Natural Language Supervision» *Proceedings of the 38 th International Conference on Machine,* 2021.

[33] G. Rizzi, F. Gasparini, . A. Saibene, P. Rosso y E. Fersini, «Recognizing misogynous memes: Biased models and tricky» *Information Processing and Management,* vol. 60, 2023.

[34] H. Kirk, W. Yin, B. Vidgen y P. Röttg, «SemEval-2023 Task 10: Explainable Detection of Online Sexism» *17th International Workshop on Semantic Evaluation,* pp. 2193-2210, 2023.

[35] R. P. Díaz, A. Fernández, R. M. Mateo, S. Valladares, S. Torres, M. Hafez, «Anti-Sexism Alert System: Identification of Sexist Comments» *Applied Sciences (Switzerland),* vol. 13, 2023.

[36] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet y T. Donoso, «Overview of EXIST 2021: sEXism Identification in Social neTworks» *Sociedad Española para el Procesamiento del Lenguaje Natural,* vol. 67, pp. 195-207, 2021.

[37] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina y P. Rosso, «Overview of EXIST 2022: sEXism Identification in Social neTworks» *Sociedad Española para el Procesamiento del Lenguaje Natural,* vol. 69, pp. 229-240, 2022.

[38] P. He, X. Liu, J. Gao y W. Chen, «DeBERTa: Decoding-enhanced BERT with Disentangled Attention» *Published as a conference paper at ICLR 2021.*

[39] N. Calzolari, F. Béchet, L. Alonso Alemany y F. M. Luque, «RoBERTuito: a pre-trained language model for social media text in Spanish» *Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association,* pp. 7235--7243, 2022.

[40] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. Pio Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre y M. Villegas, «MarIA: Spanish Language Models» *Sociedad Española para el Procesamiento del Lenguaje Natural,* vol. 68, pp. 39-60, 2022.

[41] M. Schuster y K. Nakajima, «Japanese and Korean voice search» *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5149--5152, 2012.

[42] A. Dwi Suarjaya, «A New Algorithm for Data Compression» *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 3, 2012.

[43] K. Hara, D. Saito y H. Shouno, «Analysis of Function of Rectified Linear Unit» *International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland.,* pp. 1-8, 2015.

[44] M. Pawłowski, A. Wróblewska y S. Sysko-Romańczuk, «Effective Techniques for Multimodal Data Fusion: A Comparative Analysis» *Sensors,* 2023.

[45] D. P. Kingma y J. Ba, «Adam: A Method for Stochastic Optimization» *Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego,* 2015.

[46] I. Loshchilov y F. Hutter, «Decoupled Weight Decay Regularization» *Published as a conference paper at ICLR,* 2019.

[47] L. N. Smith y N. Topin, «Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates» *CoRR,* 2018.

[48] M. Smith y G. Ruxton, «Effective use of the McNemar test» *Behavioral Ecology and Sociobiology,* 2020.

[49] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank y M. Poesio, «Learning from Disagreement: A Survey» *Journal of Artificial Intelligence Research,* vol. 72, pp. 1385-1470, 2021.

[50] P. Xu, X. Zhu y D. A. Clifton, «Multimodal Learning with Transformers: A Survey» de *Transactions on Pattern Analysis & Machine Intelligence*, 2023, pp. 12113-12132.

Apéndice

Apéndice A: Publicaciones

Cabe mencionar que parte del trabajo presentado en este TFG ha sido publicado o se va a publicar en las actas de varias conferencias:

Carrillo-De-Albornoz J., Plaza L., Amigó E., Gonzalo J., Spina D., Rosso P., Morante R., Chulvi B., Maeso A., Ruiz V. EXIST 2024: «sEXism Identification in Social neTworks and Memes». *In: Proc. 46th European Conf. on Information Retrieval, ECIR-2024, Springer-Verlag,* LNCS (14612), pp. 498-504, 2024

Plaza L., Carrillo-De-Albornoz J., Amigó E., Gonzalo J., Spina D., Rosso P., Morante R., Chulvi B., Maeso A., Ruiz V. «Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental {IR} Meets Multilinguality, Multimodality, and Interaction». *Proceedings of the Fifteenth International Conference of the CLEF Association,* 2024.

Plaza L., Carrillo-De-Albornoz J., Amigó E., Gonzalo J., Spina D., Rosso P., Morante R., Chulvi B., Maeso A., Ruiz V. «Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes» (Extended Overview, Working Notes of {CLEF} 2024 - *Conference and Labs of the Evaluation Forum*, Guglielmo Faggioli and Nicola Ferro and Petra Galuvsvacakova and Alba García Seco de Herrera, 2024

Apéndice B: Objetivos de Desarrollo Sostenible

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**				X
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**	X			
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				X
ODS 9. **Industria, innovación e infraestructuras.**				X
ODS 10. **Reducción de las desigualdades**.	X			
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**		X		
ODS 17. **Alianzas para lograr objetivos.**				X

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados

Este Trabajo de Final de Grado (TFG) está estrechamente relacionado con varios de los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030. Por un lado, con el de ODS 5. Igualdad de género y ODS 10. Reducción de las desigualdades y de forma no tan directa con el ODS 16. Paz, justicia e instituciones sólidas. Estos objetivos están interconectados y se refuerzan mutuamente para lograr una sociedad en igualdad de derechos.

El ODS 5, Igualdad de género, busca eliminar todas las formas de discriminación contra las mujeres y niñas. En el contexto de este TFG, los memes pueden ser un vehículo para la transmisión de mensajes sexistas, perpetuando estereotipos de género, actitudes misóginas y violencia, como se ha podido ver a lo largo de este trabajo. Identificar y clasificar las formas de sexismo en los memes, como el descrédito ideológico, estereotipos y dominancia, cosificación, misoginia, y violencia es esencial para visibilizar estas manifestaciones de desigualdad. Este TFG contribuye a una comprensión más profunda de cómo se perpetúan las desigualdades de género en la cultura digital y ayuda a desarrollar estrategias para combatirlas.

El ODS 10, Reducción de las desigualdades, se centra en disminuir las brechas de desigualdad. Analizar los memes sexistas permite revelar cómo se contribuye a la perpetuación de roles de género y estereotipos donde las personas según su género son supuestamente más apropiadas o inapropiadas para desempeñar ciertas funciones, lo que contribuye a mantener las desigualdades de género. Este TFG aporta un conocimiento valioso sobre las manifestaciones del sexismo en los memes, proporcionando un *dataset* para que otros investigadores pueden abordar el problema desde diferentes perspectivas. Esta investigación no solo destaca la necesidad de mayor atención y recursos para estudiar estas formas de comunicación, sino que también sugiere la implementación de mecanismos en las redes sociales para frenar la difusión de contenido sexista.

El ODS 16, Paz, justicia e instituciones sólidas, promueve sociedades pacíficas e inclusivas para el desarrollo sostenible, asegurando el acceso a la justicia para todos y construyendo instituciones eficaces, responsables e inclusivas. La lucha contra el sexismo y la violencia de género es fundamental para alcanzar este objetivo. Las instituciones deben ser capaces de reconocer y abordar tanto las formas explícitas como sutiles de discriminación y violencia. Este TFG, al identificar y clasificar el sexismo en los memes, proporciona una herramienta valiosa para que estas instituciones desarrollen mejores políticas de prevención y respuesta al sensibilizar sobre cómo los medios digitales contribuyen a la normalización de la violencia y la discriminación, y promoviendo la igualdad y justicia.