Trabajo de Fin de Grado

Análisis del sesgo de género en los modelos NLP

Junio 2021

Resumen

Este proyecto explora la existencia del sesgo de género en el ámbito del Procesamiento del Lenguaje Natural o NLP (*Natural Language Processing*). Se trata de un fenómeno de extrema importancia, dada la importancia exponencialmente creciente del aprendizaje automático dentro de la sociedad, y cómo la extrema dependencia de las técnicas NLP en datos históricos, de probado sesgo, puede contribuir a que el mismo no solo no desaparezca de la sociedad, sino que contribuya a mantenerlo o incluso ampliarlo.

Por un lado se ha hecho una recopilación de diferentes estudios que ya han analizado, medido y mitigado el mencionado sesgo en base a estudios publicados de las técnicas del estado del arte actuales, lo cual ayuda a conocer de primera mano el estado actual.

Posteriormente se procede a aplicar de forma práctica una de las formas de medición del sesgo de género en los modelos de lenguaje, en este caso concreto en los BERT (Bidirectional Encoder Representations from Transformers). Esto nos permite conocer cómo funciona y cómo se aplica esta técnica y, adicionalmente, lo que nos dice sobre el nivel del sesgo de género en una de las herramientas del estado del arte más utilizadas en el ámbito de NLP. Además se ha extendido la técnica a los idiomas castellano y euskera, aparte del inglés originalmente estudiado, y para versiones de BERT tanto monolingües como multilingües.

El alumno partía con escasos conocimientos previos sobre el procesamiento del lenguaje natural, por lo que para la realización del proyecto ha sido necesaria una primera fase de puesta al día sobre plataformas, técnicas y herramientas disponibles para trabajar en el ámbito de NLP, especialmente en aquellas disponibles de forma gratuita y online. Entre ellas se encuentran Google Cloud, GitHub, Google Colaboratory y Overleaf.

En el aspecto formativo y personal, gracias a la realización de este trabajo se ha reforzado la capacidad de trabajar de forma más autónoma, solucionando dudas y problemas según iban surgiendo. También ha sido útil para aprender a gestionar mejor la realización de

tareas y consecución de objetivos en base a un plan, incluyendo la fase de documentación del proyecto realizado.

Índice general

Índice de figuras

Índice de tablas

1. CAPÍTULO

Introducción

Podemos definir el sesgo de género como la preferencia o prejuicio hacia un determinado género, respecto de otros.

A medida que el Procesamiento del Lenguaje Natural o NLP (*Natural Language Processing*) y sus consiguientes herramientas han ido creciendo en popularidad, se hace vital el reconocer el rol que pueden jugar las mismas en dar forma (o trasladar) a determinados sesgos y estereotipos sociales.

Centrándonos específicamente en los prejuicios de género, se puede afirmar que, aunque se han creado modelos de lenguaje muy satisfactorios para aplicaciones específicas, está demostrado a estas alturas que dichos modelos propagan (y en algunos casos amplían) el sesgo de género que se encuentra en los corpus de texto que sirven de base a los mismos. El sesgo de género se manifiesta en múltiples áreas del NLP, incluidos datos de entrenamiento (*training data*), recursos disponibles, modelos preentrenados (por ejemplo, los llamados *Word embeddings*, mediante los cuales una palabra se mapea a un vector n-dimensional, y se parte del supuesto de que palabras que se encuentran en un espacio semejante deben tener algún tipo de relación), e incluso en los propios algoritmos (Zhao et al., 2018a[1]; Bolukbasi et al., 2016[2]; Caliskan et al., 2017[3]; Garg et al., 2018[4]). Las soluciones NLP que contienen sesgo en cualquiera de sus componentes pueden llegar a producir predicciones basadas en prejuicios y, en algunos casos, amplificar dicho sesgo (Zhao et al., 2017[5]). Dicho fenómeno de propagación no ha de considerarse como banal, ya que puede afectar de forma dramática en casos reales específicos como, por ejemplo, en el filtrado automático de currículos, dando preferencia a los candidatos masculinos

cuando el único factor de distinción es el género.

En este trabajo vamos a revisar esta problemática desde dos perspectivas diferentes. Por un lado, vamos a resumir cuál es el estado actual de las técnicas estado del arte existentes para detectar, medir y/o mitigar los comentados sesgos de género en el ámbito de modelos de NLP. En segundo lugar, vamos a analizar y medir, usando una de las mencionadas técnicas, el sesgo existente en una serie de modelos BERT (Bidirectional Encoder Representations from Transformers) para los idiomas inglés, español y euskera.

1.1. Estructura de la memoria

La memoria del TFG consta de cinco capítulos, además de dos anexos:

1. El presente capítulo es donde se realiza una **Introducción** al TFG.
2. En el **Documento de objetivos del proyecto** se detallan los objetivos principales a cumplir para la finalización del trabajo.
3. En el **Resumen de técnicas del estado del arte** se recopila en profundidad el estado del arte relacionado con la detección, medición y mitigación del sesgo de género en los modelos de lenguaje.
4. En el capítulo de **Medición del sesgo de género en modelos BERT** se detalla lo que son los modelos BERT, y se explica de forma detallada una técnica de medición de sesgo sobre los mismos, y cómo se ha aplicado a una serie de modelos BERT para varios idiomas.
5. En las **Conclusiones** se plantean el impacto del trabajo, sus discusiones y el trabajo futuro a realizar, junto con una reflexión de lo que ha supuesto a nivel personal.
6. En el **Anexo A** se proporcionan detalles sobre los modelos BERT que han sido objeto de estudio.
7. Finalmente, en el **Anexo B** se da una explicación detallada de la composición de los corpus de evaluación utilizados en la técnica de medición del sesgo de género seleccionada.

2. CAPÍTULO

Documento de objetivos del proyecto

2.1. Descripción y objetivos del proyecto

El objetivo de este trabajo es, por un lado, hacer una recopilación pormenorizada de los últimos estudios publicados que cubren la detección, medición y posibles métodos de reducción del sesgo de género en el ámbito del aprendizaje automático y, más concretamente, en el área del Procesamiento de lenguaje natural (PLN) o NLP en inglés.

En la segunda parte del proyecto, se busca el elegir una de las técnicas de medición del sesgo en los modelos de lenguaje, y aplicarla de forma práctica a diferentes BERT (Bidirectional Encoder Representations from Transformers), extendiéndola al español y euskera, idiomas para los que todavía no ha sido aplicado en el momento del inicio dle presente trabajo.

2.2. Planificación del proyecto

Una parte importante de la planificación es identificar las diferentes tareas que formarán parte del proyecto. En la estructura de descomposición del trabajo (EDT) de la figura 2.1, se puede ver la estructura jerárquica de las tareas necesarias para cumplir con los objetivos establecidos.

Adicionalmente, la tabla 2.1 muestra la estimación en horas para cada una de las fases y la desviación final en cada una de ellas.

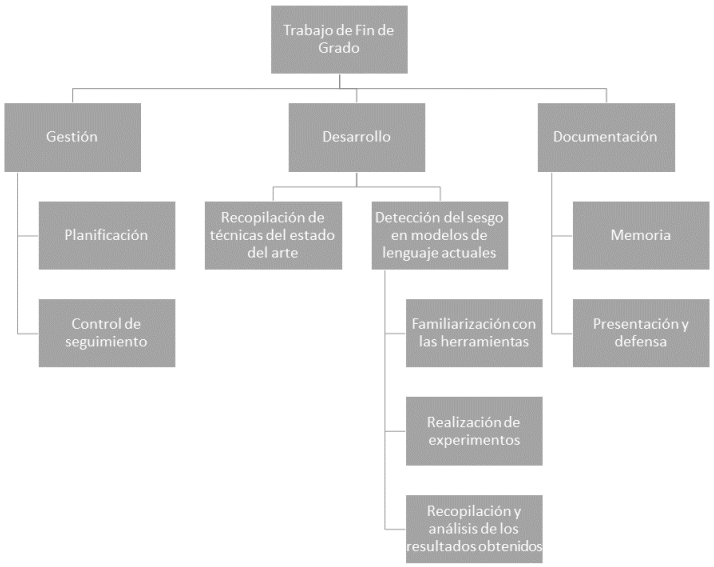

Figura 2.1: Estructura de la Descomposición del Trabajo.

Fase	Horas estimadas	Horas empleadas	Desviación
Gestión	15	15	0
Planificación	5	5	0
Control de seguimiento	10	10	0
Desarrollo	200	204	4
Recopilación de técnicas estado del arte	75	72	-3
Detección del sesgo en modelos de lenguaje actuales	125	132	7
Documentación	100	105	5
Memoria	75	80	5
Presentación y defensa	25	25	0
Total dedicación	**315**	**324**	**9**

Tabla 2.1: Tiempo estimado e invertido, y desviación de la dedicación

2.3. Tareas

Como se puede observar, el trabajo se divide en tres partes principales, la gestión, el desarrollo y la documentación. A continuación, procedemos a detallarlas.

2.3.1. Gestión

Esta fase que está presente durante todo el ciclo de vida del proyecto, pero precisa de más tiempo al principio. Se divide en dos partes diferentes, la planificación y el control de

seguimiento.

Planificación

Es en esta parte de la gestión donde se identifican los objetivos y alcance del proyecto. Además es cuando se determina la metodología de trabajo y el momento en el que se estiman las horas necesarias para cada fase.
Adicionalmente, es donde se lleva a cabo un análisis de los peligros y riesgos del proyecto.

Control de seguimiento

Esta parte del proyecto está activa durante toda la duración del mismo, y es la manera de garantizar que la planificación principal y las fechas límite establecidas se cumplan, y asegurar de que el proyecto seguirá adelante aún produciéndose cualquier desviación (documentando estas últimas).
Como parte de esta tarea, se controlan las horas que se hayan necesitado para cada uno de los paquetes de trabajo definidos, y al final, se compara con en el tiempo estimado al inicio. También se tienen en cuenta las reuniones realizadas con el tutor, para actualizar el estado del proyecto.

2.3.2. Desarrollo

Esta fase contiene la mayor carga de trabajo, ya que es la que posibilita la consecución de los objetivos definidos para el trabajo.

Recopilación de técnicas estado del arte

Es en esta fase donde se procede a realizar una investigación exhaustiva de las diferentes técnicas estado el arte que han investigado la existencia del sesgo de género en el NLP, incluso para aquellos modelos más reconocidos y utilizados.
La cantidad de material disponible online, incluido aquel originalmente proporcionado por las tutoras del trabajo, requiere de un esfuerzo extra para sintetizar la información, intentando visibilizar muchos de los ámbitos dentro de NLP en los que el mencionado sesgo tiene presencia.

Detección del sesgo en modelos de lenguaje actuales

En esta fase, con ayuda de las tutoras, se elige una de las técnicas de detección y medición del sesgo recientemente publicadas, aplicable sobre modelos BERT, modelos de

lenguaje que tienen su origen en Google para su motor de búsqueda, y que actualmente son extensivamente utilizados. Para aplicar la mencionada técnica se optará por hacerla extensiva a los idiomas español y euskera, aplicándola sobre modelos BERT multilingüe y/o específicos para esos idiomas.

Familiarización con las herramientas

El escaso conocimiento inicial del ámbito del NLP por parte del alumno hace preciso que en una de las fase iniciales del proyecto sea necesario añadir una fase de formación en conceptos NLP teóricos, y familiarización con herramientas necesarias para poder ejecutar los experimentos imprescindibles para la aplicación del método de medición elegido.

Realización de experimentos

Para la realización de experimentos se utilizarán plataformas online para generación y ejecución de código (tales como GitHub y Google Colaboratory), así como portales de recursos y modelos de lenguaje (Huggingface).
Se estima que el proceso de diseño y ejecución de experimentos puede precisar en ocasiones de métodos de prueba y fallo, debido a la potencialmente necesaria adaptación de métodos ya existentes para obtener los datos necesarios, y poder así documentar los resultados obtenidos.

Recopilación y análisis de los resultados obtenidos

En esta fase trabajará en recopilar de forma ordenada y precisa los datos necesarios para poder analizar, con los resultados obtenidos y en base a la metodología descrita elegida, el fenómeno del sesgo de género en los diferentes modelos y para los idiomas objeto de estudio.

2.3.3. Documentación

Esta fase también esta presente durante todo el desarrollo del proyecto, ya que si bien el resultado final son la memoria y presentación utilizada en la defensa, lo cual solo es preciso al final de la ejecución del proyecto, es necesario documentar todo lo ocurrido durante toda la duración del mismo. Se plasmarán así la información recopilada y analizada, las decisiones tomadas, y los resultados y conclusiones obtenidos en cada momento, consiguiendo al mismo tiempo adelantar trabajo y suavizar el impacto final de la finalización de los entregables.

Memoria

En la memoria del proyecto es donde se plasma toda la información mencionada anteriormente. Además, tiene que ser un documento detallado, para que otras personas que la lean obtengan toda la información posible sobre los procesos realizados, y ha de usar un formato estándar que cumpla con unos requerimientos mínimos definidos por la universidad.

Presentación y defensa

Finalmente, en la presentación de la defensa se tiene que resumir la memoria documentada a través de diapositivas. Esas diapositivas servirán de soporte para la presentación oral de no más de 20 a 25 minutos, con todos los detalles del proyecto creado.
Se considera parte de esta fase también la preparación de la propia defensa, con los ensayos realizados como parte de la propia preparación.

2.4. Metodología de trabajo

Como ya se ha indicado con anterioridad, para la realización de este trabajo se utilizarán herramientas online, consiguiendo de esta manera el no depender del buen funcionamiento de un equipamiento local, eliminando así uno de los mayores riesgos detectados para la realización de las tareas definidas, la posible pérdida de datos y, consecuentemente, del trabajo realizado.

2.4.1. Reuniones

A lo largo de la ejecución del proyecto se mantendrán una serie de reuniones de seguimiento con las tutoras del trabajo, con el objetivo de establecer las pautas a seguir para el cumplimiento de los objetivos, así como la adaptación de los mismos en caso de que los objetivos originales no se fueran a cumplir.
Debido a la distancia geográfica de la vivienda del alumno, y a la situación sanitaria de pandemia, todas las reuniones se realizarán de forma virtual mediante la herramienta Blackboard Collaborate.

2.4.2. Horarios planificados

Con el objeto de mantener una rutina de evolución continua del proyecto, se han planificado un mínimo de horas semanales con una proyección final de unas 300 horas a cumplir antes de Junio de 2021.

2.5. Riesgos y medidas correctivas

Es habitual que durante el desarrollo de un proyecto haya riesgos o inconvenientes que lleguen a cambiar totalmente el rumbo del mismo. Debido a esto, es muy importante reconocer los riesgos antes de empezar el proyecto y tomar las medidas preventivas o correctivas necesarias para afrontarlos, ya sea eliminándolos o mitigándolos. A continuación se muestran los principales riesgos detectados, pero se ha de tener en cuenta que probablemente haya algún riesgo más que no se haya documentado:

- Pérdida de código o datos, o del material informático:
  - Se solventa con duplicidad de entorno en PC y portátil, y con backup a sitio online (Google Drive).
  - Además se plantea el uso de una solución online tanto para la implementación (Google Colaboratory + GitHub) como para la documentación (Overleaf).
- Dificultad para reunirse con las directoras del TFG:
  - Debido a la distancia geográfica de mi vivienda actual.
  - Unido a las dificultades adicional propiciadas por el Covid-19 y las posibles limitaciones de movilidad.
  - Solucionado con la realización de reuniones remotas virtuales usando herramientas como Blackboard Collaborate o Google Meet, junto con comunicación complementaria vía email.
- Retrasos sobre la planificación:
  - Debido a circunstancias laborales del alumno que puedan afectar a la disponibilidad diaria para progresar en las tareas u otros posibles inconvenientes (inesperados).
  - Se pueden aplicar 2 medidas:

- Adaptar la planificación a las situaciones cambiantes que puedan afectar al plan inicial de las tareas.
- Distribuir la planificación semanalmente para dar opción a recuperar los posibles retrasos diarios con sesiones intensivas durante días no laborales. 12 horas semanales para una duración aproximada de 6 meses (1 Dic 2020 - 31 Mayo 2021).

2.6. Dedicación y análisis de la desviación

Tal y como se puede observar en la tabla 2.1, la estimación inicial de horas dedicadas se ha cumplido con una deseable precisión. Las mayores desviaciones se han encontrado a la hora de la realización de la memoria, ya que la decisión final de realizarla en LaTeX ha precisado de una re-adaptación y formateo de las tablas de resultados de los experimentos de medición del sesgo y, especialmente, a la hora de realizar los propios experimentos, ya que la fase de familiarización con las herramientas y modelos BERT ha precisado de mayor número de horas, lo cual puede haberse dado por la falta de conocimiento inicial por parte del alumno.

3. CAPÍTULO

Resumen de técnicas del estado del arte para la medición del sesgo de género en modelos NLP

En este apartado se muestra un resumen detallado del estado actual en lo que se refiere a los estudios sobre la existencia del sesgo de género en los modelos de NLP ampliamente utilizados, ya que, si bien el estudio de los prejuicios/sesgos en el entorno de la Inteligencia Artificial (IA) no es nuevo, el desarrollo de métodos que disminuyan o hagan desaparecer los sesgos de género tiene aún mucho camino por delante.
Una de las maneras de categorizar el sesgo es en términos de adjudicación y de representación de género. El sesgo en adjudicación se podría formular como problema económico donde un sistema distribuye de forma injusta medios a determinados grupos sobre los otros, y en el entorno de NLP dicho sesgo se refleja cuando los modelos tienen mejor rendimiento en datos correspondientes al género mayoritario. Por otro lado, el sesgo en representación puede considerarse como el hecho de que los sistemas restan valor a la identidad social y representación de algunos grupos (Crawford, 2017[6]), y en el entorno de NLP es un sesgo que se da cuando se capturan asociaciones entre género y determinados conceptos en procesos de *word embedding* o parámetros de los modelos (Crawford, 2017[6]).
Profundizando en el sesgo de representación en el entorno de NLP, podríamos clasificar el mismo en diferentes categorías (Crawford, 2017[6]):

1. Denigración, que se refiere al hecho de usar términos históricamente y culturalmente despectivos.

2. Estereotipado, donde se refuerzan los estereotipos sociales existentes.

3. De reconocimiento, el cual implica la falta de precisión de los algoritmos en tareas de reconocimiento.

4. Infrarrepresentación, es decir, la desproporcionadamente baja representación de un determinado grupo o grupos.

La Tabla 3.1 presenta ejemplos de la mencionada categorización del sesgo de representación en el contexto del sesgo dentro de NLP:
(D)enigración, (E)stereotipado, (R)econocimiento e (I)nfrarrepresentación.

Tarea	**Ejemplo del sesgo de representación en el contexto de género**	**D**	**E**	**R**	**I**
Traducción automática	Traducir "He is a nurse. She is a doctor." de inglés a húngaro y de vuelta a inglés da como resultado "She is a nurse. He is a doctor." (Prates et al., 2017[7]).		X	X	
Generación de leyendas	Un modelo generador de leyendas predice de forma incorrecta que el agente es un hombre porque hay un ordenador cerca (Hendricks et al., 2018[8]).		X	X	
Reconocimiento de voz	El reconocimiento de voz automático funciona mejor con voces masculinas que con las femeninas (Tatman, 2017[9]).			X	X
Análisis de sentimientos	Los sistemas de análisis de sentimiento clasifican las frases de sustantivo femenino como indicativas de enfado con mayor frecuencia que las frases de sustantivo masculino (Park et al., 2018[10]).		X		
Modelado de lenguaje	"He is doctor" tiene una mayor probabilidad condicional que "She is doctor" (Lu et al., 2018[11]).		X	X	X
Word embedding	Analogías del tipo "man : woman :: computer programmer : homemaker" son generadas de forma automática por modelos entrenados en *word embeddings* sesgados (Bolukbasi et al., 2016[3]).	X	X	X	X

Tabla 3.1: Ejemplos del sesgo de género en diferentes tareas NLP

Mediante los diferentes estudios en el campo, es reconocido que los perjuicios de adjudicación y representación anteriormente mencionados aparecen a menudo en los sistemas NLP debido a patrones estadísticos en los corpus de entrenamiento, los cuales se integran en las representaciones semánticas y en el propio modelo.
El sesgo de género en NLP es, por lo tanto, un problema complejo que requiere de un trabajo interdisciplinar. Además, como los sistemas NLP se están integrando incrementalmente en nuestro día a día, gracias a los desarrollos modernos de inteligencia artificial,

es preciso desarrollar soluciones inmediatas y enfoques elementales que “reparen” los sistemas actuales.
A continuación se procede a detallar los estudios y técnicas más actuales, y que se pueden considerar de referencia, que se enfrentan a la problemática del sesgo de género, cómo detectarlo y cómo reducirlo, en el entorno del NLP, intentando mostrar al mismo tiempo las ventajas e inconvenientes de cada uno de ellos. La Figura 3.1 a continuación muestra los diferentes ámbitos dentro del proceso de NLP en los que los mencionados métodos son aplicables:

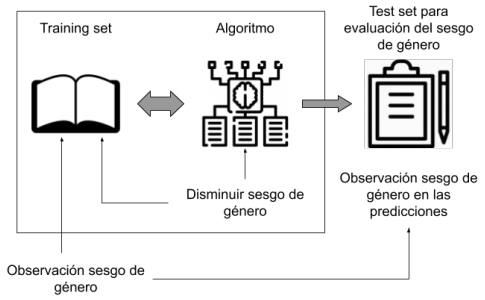

Figura 3.1: Ámbitos donde aplicar métodos de observación y disminución del sesgo.

3.1. Métodos de detección del sesgo de género

Los recientes trabajos de análisis del sesgo de género en NLP se han enfocado en cuantificar el mencionado sesgo a través de test psicológicos, la geometría de los espacios vectoriales en *word embeddings*, y las diferencias en rendimiento de los modelos entre los diferentes géneros para determinadas tareas. Seguidamente se detallan los estudios para cada uno de los aspectos indicados.

3.1.1. Test psicológicos

En Psicología, el test de asociación implícita o IAT (*Implicit Association Test*) es una medida usada para evaluar el sesgo de género subconsciente en los humanos, el cual se cuantifica como la diferencia en tiempo y precisión para que los humanos asocien palabras con dos conceptos que consideran similares frente a dos conceptos considerados diferentes (Bolukbasi et al., 2016[2]; Caliskan et al., 2017[3]). Por ejemplo, para medir

las asociaciones subconscientes de los géneros con las artes y ciencias, se les pide a los participantes en el test que creen categorías que agrupen hombres y ciencias por un lado, y mujeres y artes por otro (Caliskan et al., 2017[3]). A los participantes se les pide a continuación que agrupen a hombres y artes por un lado, y mujeres y ciencias por el otro. El hecho de que los participantes contesten antes y de forma más acertada al primer ejercicio indica que los humanos, de forma subconsciente, asocian a los hombres con las ciencias y a las mujeres con las artes.
En esta línea, Caliskan et al., (2017)[3] adoptan el concepto del IAT, midiendo el sesgo de género mediante la diferencia en asociación de conceptos en los *word embeddings* incluidos dentro del Word Embedding Association Test (WEAT). Usando esta metodología confirman que los sesgos de género en los humanos existen también en los *word embeddings* GloVe y Word2Vec. Por último, los autores del estudio demuestran, usando porcentajes de ocupación disponibles en el Bureau of Labor Statistics de EEUU, la positiva correlación entre la fuerza asociativa de una profesión y el género femenino en los *word embedding* mencionados, y el porcentaje de mujeres que desarrollan ese trabajo en el mundo real. Es digno de reseñar también cómo Garg et al., (2018)[4] demuestran que el sesgo en los *word embeddings* puede utilizarse para monitorizar a lo largo del tiempo la mencionada participación femenina en los diferentes oficios.
Posteriormente, May et al. (2019)[12] ampliaron el WEAT y crearon el Sentence Encoder Association Test (SEAT), el cual es capaz de buscar sesgos humanos descubiertos en los IAT en codificadores de frases como el ELMo (Peters et al., 2018[13]).

3.1.2. Análisis del subespacio vectorial de género en los *word embeddings*

Bolukbasi et al., (2016)[2] definen el sesgo de género como la correlación de la magnitud de la proyección en el subespacio de género de un *word embedding* que representa a una palabra neutral en cuanto al género, y la valoración en cuanto a sesgo de esa misma palabra por parte de un grupo significativo de trabajadores.
Para identificar el mencionado subespacio crearon una máquina de vectores de soporte (SVM - Support Vector Machine) lineal para clasificar palabras entre específicas o neutrales en cuanto al género, haciendo uso de un *training set* y un conjunto de palabras de género específico seleccionadas a mano. Los autores determinan en su estudio la dirección correspondiente al género, agregando diez parejas de palabras de género del tipo *she-he*, *her-his*, *woman-man*, etc., y usan Principal Component Analysis (PCA) para encontrar el eigenvector que muestra una varianza mayor que el resto (llegando incluso a definir

métricas para cuantificar sesgos directo e indirecto en los *word embeddings*).
Sin embargo, Gonen and Goldberg (2019)[14] apuntan que el método anteriormente descrito reduce el sesgo de género de forma superficial, y que falla en su intento de capturar de forma completa el sesgo de género en los espacios vectoriales. Concretamente muestran cómo, incluso después de haber eliminado las proyecciones de los *word embeddings* de las palabras neutrales de género en el subespacio de género, los *word embeddings* de aquellas palabras con sesgos similares acaban agrupándose.
Introducen así la noción de *cluster bias*: el *cluster bias* de una palabra *w* puede medirse como el porcentaje de palabras estereotipadas masculinas o femeninas que se encuentran entre los *k-nearest neighbours* del *word embedding* de *w*, donde las palabras estereotipadas masculinas o femeninas se obtienen a través de anotación humana. Concluyen además que la dirección del género dentro del espacio vectorial proporciona una forma de medir la asociación de género de una palabra concreta, pero que no la determina.

3.1.3. Medición de la diferencia en rendimiento de los modelos entre los diferentes géneros

En la mayor parte de las tareas de NLP la predicción realizada por un modelo no debería verse influenciada por el género de la entidad o el contexto de la entrada. Para evaluar si este es el caso, podemos considerar 2 frases que actúen como entrada de un modelo, las cuales se diferencian únicamente en las palabras que hacen referencia al género del sujeto (por ejemplo, "'Él ha ido al parque" y "Ella ha ido al parque").
Si hablamos del intercambio de género (*gender-swapping*), hacemos referencia a cambiar los sustantivos de género (sujetos) en una frase (cambiar "él" por "ella" en la frase anterior, por ejemplo), técnica utilizada por Zhao et al., (2018a)[1]; Lu et al., (2018)[12]; Kiritchenko y Mohammad, (2018)[15]. Hipotéticamente, si el modelo utilizado no toma decisiones basadas o influenciadas por el género del sujeto, sus resultados deberían ser igual de efectivos para ambos casos o géneros (sujeto masculino y femenino). Si se diera el caso contrario, la diferencia en los resultados de evaluación reflejaría la dimensión del sesgo de género existente en el modelo evaluado.
Por ejemplo, Dixon et al. (2017)[16] demuestran en su trabajo cómo el desequilibrio en los datos de entrenamiento puede llevar al sesgo involuntario en los modelos y, por lo tanto, a herramientas de aprendizaje automático potencialmente injustas. Introducen igualmente dos métricas para indicar la diferencia entre rendimientos (*False Positive Equality Difference* o FPED, y *False Negative Equality Difference* o FNED), las cuáles se definen como las diferencias en las ratios de falsos positivos y falsos negativos, respectivamente, de las

predicciones de un determinado modelo para las entradas originales y para con los datos de testeo con *gender-swapping* aplicado. Las mencionadas métricas son usadas para la detección de lenguaje ofensivo (Park et al., 2018[10]), pero se sugiere que podrían ser utilizadas para otras tareas de NLP.

Profundizando en la línea de medición de diferencias de rendimiento, cuando se diseñan los materiales de testeo en el ámbito del aprendizaje automático, el medir las diferencias de rendimiento entre géneros saca a la luz el sesgo de género de representación en los ámbitos de reconocimiento, estereotipado e infrarrepresentación detallados al inicio de este capítulo. Si, por ejemplo, un modelo de generación de leyendas se emplea de forma menos acertada cuando ha de reconocer a una mujer en lugar de a un hombre, estando ambos sentados delante de un ordenador (Hendricks et al., 2018[8]), este supone un fenómeno claramente indicador de existencia del sesgo en reconocimiento. Si esta falta de precisión es debida a un algoritmo que asocia "hombre" y "ordenador", entonces nos encontramos también delante de un ejemplo de estereotipado. Se podría por lo tanto imaginar que, si en el mencionado modelo no se aplica un proceso de eliminación del sesgo y, por lo tanto, los errores se propagan en una muestra importante de imágenes, el propio modelo contribuirá también de forma relevante a la infrarrepresentación de la minoría (en este caso el género femenino).

Cuando Bhardwaj et al., (2020)[17], por otro lado, realizan su estudio para analizar el sesgo de género que BERT (modelo de NLP desarrollado por Google) induce en cinco tareas relacionadas con el análisis de sentimientos, plantean la hipótesis de que los *word embeddings* con mucha dependencia del contexto pueden ser evaluados mediante el sesgo detectado en las tareas derivadas del modelo, dado que para dicho tipo de modelos es muy complicado analizar el sesgo implícito al mismo.

En base a diversos estudios publicados podemos derivar que los *datasets* de evaluación estándar nos son adecuados para medir el sesgo de género, dado que ya contienen sesgo por sí mismos (siendo un claro ejemplo el hecho de que tengan por defecto más referencias masculinas que femeninas). Por lo tanto, podemos concluir que la evaluación mediante el uso del mencionado material de evaluación puede no revelar adecuadamente el sesgo de género. Es más, dado que las predicciones hechas por los sistemas que realizan tareas complejas de NLP dependen de numerosos factores, los *datasets* utilizados han de ser diseñados cuidadosamente para aislar el efecto del género en la salida, para ser capaces así de probar el sesgo de género en el modelo objeto del estudio.

Podemos denominar los *datasets* así creados (en referencia a aquellos que buscan aislar el efecto del género) como *Gender Bias Evaluation Testsets* (GBETs). El objetivo de su creación es el proporcionar formas de verificación de que los sistemas NLP sean capaces

de evitar realizar errores en su salida debido al sesgo de género. Aunque se pueda argumentar que el diseño artificial de los GBETs no refleja la verdadera distribución de los datos disponibles (y extenderlo al mundo real), lo cual implicaría que las propias evaluaciones sean igualmente artificiales, esta afirmación puede ser rebatida afirmando que, si los humanos pueden evitar errar debido a sus prejuicios, lo mismo deberían poder hacer las máquinas. Asimismo, debería considerarse que los sistemas que hacen predicciones sesgadas podrían disuadir a las minorías de usarlos, lo cual disminuiría su utilización y eficacia, empeorando de esta manera la disparidad en los *datasets* y acrecentando aún más el fenómeno.
La tabla 3.2 incluye un extracto de los GBETs disponibles de forma pública en el momento de la realización del presente trabajo.

Dataset	**Tarea**	**Ámbito de sondeo**	**Tamaño**
Winogender Schemas (Rudinger et al., 2018[18])	Resolución de correferencia	Profesiones	720 frases en inglés
WinoBias (Zhao et al., 2018a[1])	Resolución de correferencia	Profesiones	3160 frases en inglés
GAP (Kocijan et al., 2020[19])	Resolución de correferencia	Sustantivos	4454 frases en inglés
EEC (Kiritchenko y Mohammad, 2018[15])	Análisis de sentimientos	Emociones	8640 frases en inglés

Tabla 3.2: Listado de GBETs

A continuación se detallan los GBETs listados anteriormente:

- Para la tarea de resolución de correferencia, Rudinger et al., (2018)[18] y Zhao et al., (2018a)[1] diseñaron, de forma paralela e independiente, GBETs basados en los modelos Winograd. Los corpus generados están construidos con frases en inglés que contienen una profesión de género neutral (p.e., *doctor*), un segundo participante (p.e., *patient*), y un pronombre indicativo de género que hace referencia o bien a la profesión o bien al participante. Es preciso recordar que el sistema de resolución de correferencia precisa que se identifique al antecedente del pronombre. En este sentido, para cada frase Rudinger et al., (2018)[18] consideran 3 tipos de pronombres; masculino, femenino y neutro; mientras que Zhao et al., (2018a)[1], por su lado, consideran únicamente pronombres masculinos y femeninos. Resulta,

por lo tanto, evidente que ambos *datasets* son notablemente diferentes.

En este caso, el medir únicamente las diferencias globales en acierto entre los diferentes pronombres con género es insuficiente. Por ejemplo, un modelo podría predecir que mujeres y hombres son correferentes con "secretaria" con un 60 % y 20 % de acierto respectivamente y, al mismo tiempo, predecir que mujeres y hombres son correferentes con "doctor" en un 20 % y 60 %. En el mencionado ejemplo la media global de acierto sería equivalente y, sin embargo, estaríamos hablando de un caso evidente de sesgo de género. Por lo tanto, y con el objetivo de evitar una forma de medición con tan escaso rendimiento, tanto Rudinger et al., (2018)[18] como Zhao et al., (2018a)[1] diseñaron sus métricas para analizar el sesgo de género examinando cómo la diferencia de rendimiento de cada género con respecto de cada profesión se correlaciona con las estadísticas ocupacionales del U.S. Bureau of Labor Statistics.

- Otro GBET para evaluar la tarea de resolución de correferencia es el llamado GAP, el cual contiene frases en inglés minadas de Wikipedia y, por lo tanto, permite realizar la evaluación en contextos reales en lugar de contextos generados de forma artificial (Kocijan et al., 2020[19]). GAP no incluye sustantivos estereotípicos. En su lugar, los pronombres hacen referencia únicamente a los nombres. En este caso, el sesgo de género puede medirse como la ratio de puntuación $F1$ en las entradas para las que el pronombre es femenino, respecto de aquellas en las que el pronombre es masculino. Es necesario destacar que, en este caso, no se realiza un *gender-swapping* con las frases, luego podría haber diferencias de dificultad para las frases masculinas y femeninas en los *datasets* de evaluación.

- Para el caso concreto de la tarea de análisis de sentimientos, Kiritchenko y Mohammad, (2018)[15] diseñan un GBET llamado Equity Evaluation Corpus (EEC). Cada frase que forma parte del EEC incluye una palabra emocional (p.e., enfado, miedo), con intensidades asociadas de uno a cinco para cada emoción, y una palabra específica de género. En este caso el sesgo de género se mide como la diferencia en intensidad emocional entre las predicciones para las frases en las que se ha realizado *gender-swapping*.

3.2. Métodos de manipulación de datos para reducción del sesgo

En el ámbito de NLP, se han propuesto diversos enfoques para reducir los estereotipos de género, centrándose principalmente en dos aspectos:

- Los corpus de texto y sus representaciones.
- Los algoritmos de predicción.

Seguidamente se abordan técnicas para eliminar el sesgo de género en los corpus de texto y los *word embeddings*, las cuales se pueden clasificar en reentrenamiento e inferencia. El reentrenamiento requiere que, tal y como su nombre indica, los modelos vuelvan a ser entrenados, mientras que la inferencia utiliza el *set* de entrenamiento original. La tabla 3.3 muestra ejemplos de técnicas clasificadas de esta forma.

Métodos	**Tipo de método**
Incremento de datos mediante *gender-swapping*	Reentrenamiento
Etiquetado del género	Reentrenamiento
Fine-Tuning del sesgo	Reentrenamiento
Eliminación del subespacio vectorial del género	Inferencia
Aprendizaje de *word embeddings* de género neutral	Reentrenamiento
Predicciones restrictivas/limitadas	Inferencia
Ajuste del *Adversarial Network Discriminator*	Reentrenamiento

Tabla 3.3: Listado de métodos de manipulación de datos

3.2.1. Métodos de reentrenamiento

Los métodos de reentrenamiento tienden a abordar el sesgo de género en las fases iniciales de NLP o incluso en la fuente de los datos. Sin embargo, volver a entrenar un modelo en una nueva muestra de datos puede resultar muy costoso tanto en recursos como en tiempo. En las secciones posteriores se detallan diferentes métodos de reentrenamiento.

Incremento de datos

En numerosas ocasiones el *dataset* utilizado en el entrenamiento de un modelo tiene un número desproporcionado de referencias a un género (p.e., OneNotes 5.0) (Zhao et al., 2018a[1]). Para mitigar la mencionada desproporción, Zhao et al., (2018a)[1] proponen la creación de una muestra de datos ampliada, idéntica a la original, pero partidista hacia el género opuesto del mismo, y entrenar la unión de ambos *datasets* (el original y el *gender-swapped*). Es un método similar al de los GBETs citados anteriormente en este trabajo, pero en este caso el objetivo es reducir el sesgo mediante el entrenamiento de un *dataset* balanceado a nivel de género, mientras que con los GBETs se busca únicamente evaluar el propio sesgo, antes y después del proceso de reducción (o eliminación) del mismo.
El funcionamiento del mencionado incremento de datos es el siguiente:

1. Para cada frase en la muestra de datos original se crea una frase equivalente, pero cambiándole el género.

2. Seguidamente se procede a la anonimización de ambos *datasets* (el original y el de género opuesto). En este caso se entiende por anonimización el proceso de intercambio de entidades con nombre propio por entidades anónimas del tipo $E1$. P.e. si aplicamos anonimización y cambio de género, la frase *"Mary likes her mother Jan"* se convierte en *"E1 likes his father E2"*. Este paso tiene un doble objetivo: el eliminar las asociaciones de género y, al mismo tiempo, entidades con nombre propio.

3. Por último, el modelo se entrena con la unión de muestra de datos original anonimizada y la muestra aumentada mediante cambio de género.

La técnica de incremento de datos ha demostrado ser muy flexible, y uno de sus grandes beneficios es que permite mitigar el sesgo de género en modelos y tareas de diferente tipo. Por ejemplo:

- Aplicada sobre un modelo de resolución de correferencia basado en redes neuronales (Lee et al., 2017[20], 2018[21]), entrenado originalmente en OntoNotes 5.0 y testado en WinoBias, la técnica de aumento de datos reduce la diferencia en valores de $F1$ para muestras de datos pro-estereotipado y anti-estereotipadas. Esto indica que el modelo estaba menos inclinado a realizar predicciones sesgadas en cuanto al género (Zhao et al., 2018a[1], 2019[22]).

- Aplicada sobre detección de discursos de odio en una CNN (*Convolutional Neural*

Network), el incremento de datos redujo notablemente las diferencias en FNED y FPED entre predicciones masculinas y femeninas (Park et al., 2018[10]).

El aumento de la muestra de datos es fácil de implementar, pero tiene los siguientes inconvenientes:

- El paso de creación de la lista de términos relevantes (indicadores de género) puede acabar siendo extremadamente caro a nivel de recursos y tiempo en aquellos casos en los que hay mucha variedad en los datos o la muestra de datos original es amplia. Una de las consideraciones a tener en cuenta a la hora de utilizar métodos que hacen uso de la identificación de las palabras específicas de género y su correspondiente opuesto, es que requiere habitualmente de intervención de grupos masivos de personas.

- Asimismo, esta técnica duplica el *training set*, lo cual puede llegar a aumentar el tiempo de entrenamiento por el factor específico de la tarea a realizar.

- Por último, el hacer ciegamente intercambio de género en las frases da lugar a la creación de frases sin sentido (p.e. cambiar *"ella dio a luz"* a *"él dio a luz"*).

Etiquetado del género

En algunas tareas NLP, como la traducción automática o MT (*Machine Translation*), confundir el género en el lenguaje de origen puede conllevar predicciones o traducciones imprecisas. Los modelos actuales de MT predicen la fuente como masculina de forma desproporcionada (Prates et al., 2018[7]; Vanmassenhove et al., 2018[23]). Esto sucede porque las muestras de entrenamiento son prominentemente masculinas. Como consecuencia los modelos que se entrenan con las mismas aprenden relaciones estadísticas distorsionadas y, por lo tanto, es más probable que su predicción sea que el género de la fuente sea masculino, independientemente de la realidad de la fuente.
La técnica de etiquetado del género atenúa este fenómeno, al etiquetar el género de la frase origen al inicio del propio punto de información (p.e. *"I'm happy"* se convertiría en *"MALE I'm happy"*). Teóricamente, el codificar la información sobre el género en las propias frases podría mejorar aquellas traducciones en las que el género del interlocutor afecta a la traducción (p.e. al traducir *"I am happy"* del inglés al francés podría hacerse como *"Je suis heureux"* [Masculino] o *"Je suis heureuse"* [Femenino], dado que en este caso el inglés original no señala el género, mientras que el idioma francés destino de la traducción sí que se diferencia según el sujeto de la frase).

Posteriormente, la etiqueta originada se analiza de forma independiente por el modelo, con el objetivo de preservar el género de la frase original, y que el modelo realice predicciones más acertadas (Vanmassenhove et al., 2018*23*).
Esta técnica demuestra ser efectiva al aplicarla sobre una red neuronal *Sequence-to-Sequence* entrenada sobre el corpus Europarl, ya que incrementa de forma significativa los resultados de BLEU en las traducciones automáticas de Inglés a Francés en aquellas frases en las que el primer interlocutor era una mujer. Asimismo, las frases en las que el mismo primer interlocutor es un hombre también mejoran de forma sustancial.
Sin embargo, el etiquetado del género puede resultar costoso, dado que el conocer el género de la fuente requiere el uso de meta-información, y su obtención incrementa el uso de memoria y tiempo. Es más, los modelos de traducción automática que hagan uso de la misma pueden necesitar ser rediseñados para ser capaces de tratar correctamente las etiquetas generadas.

Fine-Tuning del género

Las muestras de datos sin sesgo pueden ser muy escasas, pero sin embargo pueden existir de forma específica para determinadas tareas. El *Fine-Tuning* del sesgo es una técnica que hace uso de un *dataset* no sesgado para intentar asegurar, mediante aprendizaje por transferencia, que un modelo contiene el menor sesgo posible antes de afinarlo con *datasets* específicos para la tarea a ejecutar. Dichos *datasets* pueden estar sesgados de forma implícita (Park et al., 2018[10]).
Esta técnica, por lo tanto, permite a los modelos evitar aprender de muestras de datos sesgadas, al mismo tiempo que son entrenados de forma adecuada para la tarea que han de realizar.
Es una técnica de relativa eficiencia. Park et al., (2018)[10] entrenaron una *Convolutional Neural Network* (CNN), usando una muestra de datos no sesgada de *tweets* ofensivos, para realizar el mencionado aprendizaje por transferencia, y posteriormente realizando el afinamiento del modelo mediante el uso de un *dataset* sesgado de *tweets* sexistas. Evaluaron la CNN resultante usando un GBET, y testaron el modelo tras entrenarlo en muestras de datos con *gender-swapping*. Al realizar la comparación de ambas técnicas mostraron que la técnica de *gender-swapping* era más eficiente a la hora de reducir el sesgo de género y mantener el rendimiento del modelo.
Sin embargo, existe la duda de si el aprendizaje de transferencia pudo ser inefectivo en este caso en concreto debido a que ambos *datasets* utilizados en el experimento (el de detección de *tweets* ofensivos y el de detección de *tweets* sexistas) eran significativamente

diferentes. Queda la duda de si esta técnica puede ser más efectiva si los *datasets* usados, tanto para la transferencia de aprendizaje como para el afinado, son similares.

3.2.2. Métodos de inferencia

A diferencia de los métodos de reentrenamiento anteriormente descritos, los métodos de inferencia no requieren que los modelos vuelvan a ser entrenados, sino que, en su lugar, parchean los modelos existentes para ajustar la salida, proporcionando la reducción del sesgo durante el proceso de *testing*.
A continuación se detallan diferentes métodos de inferencia para eliminación del sesgo de género.

Eliminación del subespacio vectorial del género en los *word embeddings*

Schmidt, (2015)[30] eliminó la afinidad con el subespacio de género de los *word embeddings*, creando un *framework* sin género mediante la utilización de la similitud del coseno y vectores ortogonales. Al eliminar el mencionado componente de género, sin embargo, sucede un fenómeno en el que se empuja a la palabra *he* hacia la palabra *she*, llegando a ser la 6ª más cercana (cuando inicialmente era la 1826 más cercana).
En este sentido, el *framework* sin género puede ser imperfecto, porque la definición semántica de una determinada palabra puede estar fuertemente ligada a su componente de género. Y, sin embargo, se puede argumentar también que eso ha de ser así y que, por lo tanto, el género de una palabra debe jugar un rol importante en su definición semántica.
Bolukbasi et al., (2016)[2] desarrollaron sobre la propuesta de Schmidt, (2015)[30], y propusieron alterar de forma quirúrgica el espacio vectorial de los *word embeddings*, eliminando el componente de género únicamente para las palabras neutras a nivel de género. Por lo tanto, en lugar de eliminar el componente de género por completo, en este caso la reducción del sesgo de género significa el hacer a las palabras neutras ortogonales a la dirección del género, tal y como se muestra en la figura 3.2.
En última instancia, en el estudio los *word embeddings* de las palabras con sesgo reducido se comportaron igual de bien que las no alteradas en tareas de coherencia y resolución de analogías (Bolukbasi et al., 2016[2]).

Aprendizaje de *word embeddings* de género neutral

Zhao et al. (2018b)[25] proponen una nueva metodología que no usa un clasificador para crear una muestra de palabras específicas de género, llamada GN-GloVe. En este caso

Figura 3.2: Proyecciones en el espacio vectorial de género

entrenan los *word embeddings* aislando la información específica del género en determinadas dimensiones, manteniendo al mismo tiempo información neutra de género en otras. Para ello cumplen con dos máximas:

- Minimizar la diferencia entre la dimensión del género en los *word embeddings* que definen los géneros masculino y femenino.
- Maximizar la diferencia entre las direcciones del género y las otras dimensiones neutras en los *word embeddings*.

Este método permite una gran flexibilidad, ya que las dimensiones de género pueden ser usadas o descartadas.
Hay que tener en cuenta que tanto esta técnica como la de la eliminación del subespacio vectorial del género dependen de la noción de similitud del coseno, por lo que podrían no funcionar correctamente para *word embeddings* en espacio no euclídeos. De igual manera, no está claro si estas estrategias pueden funcionar para lenguajes que no sean el inglés, especialmente aquellos de morfología rica donde el género se incluye como morfema en los sustantivos y se refuerza mediante artículos (como es el caso del español).

3.3. Métodos de ajustes en algoritmos para reducción del sesgo

Algunos métodos de reducción del sesgo de género en modelos de lenguaje tratan de ajustar las predicciones. A continuación se detallan dos ejemplos de métodos que utilizan esta estrategia.

3.3.1. Restricción de las predicciones

Zhao et al. (2017)[5] mostraron que los modelos de lenguaje se arriesgan a amplificar los sesgos implícitos en los *datasets* de entrenamiento. Por ejemplo, si el 80 % de las correferencias de *"secretary"* son femeninas en el *dataset* de entrenamiento, y el modelo entrenado con el mismo predice un 90 % de correferencias para esa misma palabra como femeninas, se podría inferir que el modelo NLP utilizado amplifica el sesgo de género.
En su caso proponen un método de reducción de la amplificación del sesgo (RBA - *Reducing Bias Amplification*) basado en un modelo condicional restringido de Roth and Yih (2004), el cuál toma como base la función de optimización de un modelo existente y la restringe para que sus predicciones cumplan unas determinadas condiciones. Por ejemplo, cuando RBA se aplicó sobre el etiquetado semántico visual (Yatskar et al., 2016[25]) restringía la ratio de hombres respecto a mujeres predichos a la hora de hacer determinadas actividades, para prevenir de esta forma que el modelo amplificara el sesgo de género a través de sus predicciones. La inferencia aproximada podía ser resuelta de forma eficiente por una Relajación Lagrangeana (Rush and Collins, 2012[26]).
El mencionado método dio buenos resultados sin prácticamente pérdida de rendimiento, al mismo tiempo que reducía la magnitud de amplificación del sesgo de género un 47.5 % y un 40.5 % para clasificadores multietiqueta y etiquetado semántico visual, respectivamente.

3.3.2. Ajuste del *Adversarial Network Discriminator*

Zhao et al. (2018a)[1] proponen el uso de una variación del método tradicional de *Generative Adversarial Networks* (GAN, Goodfellow et al., 2014[27]), haciendo que el generador aprenda con atributos de un género específico. Dicho de otra manera, el generador intenta que el discriminador sea incapaz de identificar el género en una tarea concreta (finalización de analogías).
Este método sería potencialmente generalizable a otras tareas que usen aprendizaje basado en gradientes.

4. CAPÍTULO

Medición del sesgo de género en modelos BERT actuales

4.1. ¿Qué es un modelo BERT?

BERT (o Bidirectional Encoder Representations from Transformers) es una técnica basada en redes neuronales, desarrollada por Google y hecha open source en 2018, para ser utilizada en el ámbito del NLP. El modelo generado fue entrenado orginalmente en corpus de más de 3300 millones de palabras.
En lo que se refiere a su funcionamiento, y recurriendo a la propia explicación de Pandu Nayak en su comunicado oficial[1], se trata de un modelo que procesa palabras en relación con todas las palabras de una frase, es decir, tiene en cuenta el contexto. En el momento de su creación, los modelos de lenguaje existentes no eran contextuales (Word2vec) o bi-direccionales (ELMo o OpenAI). Desde el momento de su creación BERT ha seguido una evolución constante, añadiéndole mayor eficacia y, adicionalmente, extendiéndolo a otros idiomas (originalmente fue entrenado para el inglés), habiendo disponibles en este momento un numeroso número de versiones, incluidas versiones multilingües entrenadas en múltiples idiomas usando la misma técnica.

4.2. Funcionamiento de BERT

BERT utiliza *Transformers*, un mecanismo de atención que aprende las relaciones contextuales entre las palabras (o subpalabras) de un texto. En su forma simple, un *Transformer* incluye dos mecanismos separados: un codificador, el cual lee el texto de entrada, y un

[1] https://www.blog.google/products/search/search-language-understanding-bert/

decodificador, que se encarga de producir una predicción para la tarea definida. Dado que el objetivo de BERT es generar un modelo lingüístico, en su caso sólo es necesario el mecanismo codificador.
A diferencia de los modelos unidireccionales, que leen la entrada de texto secuencialmente (de izquierda a derecha o de derecha a izquierda), el codificador del *Transformer* de BERT lee toda la secuencia de palabras. Es por eso por lo que se le considera bidireccional (aunque quizás sería más exacto decir que es no-direccional). La mencionada característica permite al modelo aprender el contexto de una palabra basándose en todo su entorno (a la izquierda y a la derecha de la palabra).
La figura 4.1 muestra a alto nivel el codificador del *Transformer*:

- La entrada es una secuencia de tokens, que primero se incrustan en vectores y luego se procesan en una red neuronal.
- La salida es una secuencia de vectores de tamaño H, en la que cada vector corresponde a un token de entrada con el mismo índice.

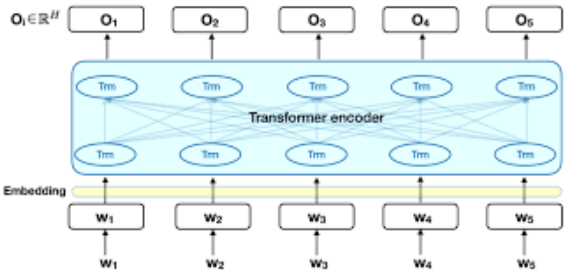

Figura 4.1: Codificador de BERT.

Cuando se entrenan modelos lingüísticos, existe el reto de definir un objetivo de predicción. Muchos modelos de lenguaje predicen la siguiente palabra de una secuencia (por ejemplo, “El niño llegó a casa de []”). Este es, por lo tanto, un enfoque direccional que limita intrínsecamente el aprendizaje del contexto. Para superar este reto, BERT utiliza dos estrategias de entrenamiento:

- Masked Language Modelling (MLM): Antes de introducir las frases (o secuencias de palabras) en BERT, el 15 % de las palabras de cada secuencia se sustituye por un token o máscara [MASK]. A continuación, el modelo intenta predecir el valor original de las palabras enmascaradas, basándose en el contexto proporcionado

por las demás palabras no enmascaradas de la secuencia. En términos técnicos, la predicción de las palabras de salida requiere:

1. Añadir una capa de clasificación sobre la salida del codificador.
2. Multiplicar los vectores de salida por la matriz de incrustación, transformándolos a la dimensión del vocabulario.
3. Calcular la probabilidad de cada palabra en el vocabulario con la función *softmax*.

La figura 4.2 muestra el codificador de BERT junto con las capas de clasificación y cálculo de probabilidades. La función de pérdida BERT sólo tiene en cuenta la

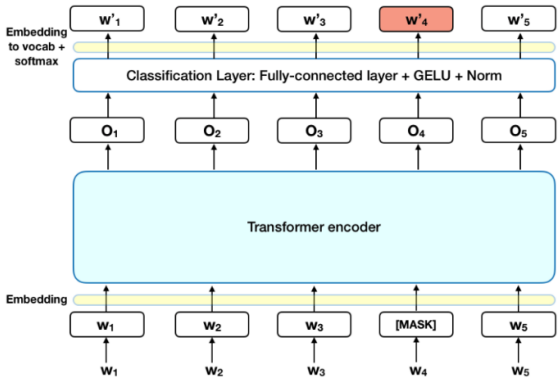

Figura 4.2: *Transformer* de BERT.

predicción de los valores enmascarados e ignora la predicción de las palabras no enmascaradas. Como consecuencia, el modelo converge más lentamente que los modelos direccionales, una característica que se ve compensada por su mayor conocimiento del contexto.

- Next Sentence Prediction (NSP): En el proceso de entrenamiento de BERT, el modelo recibe parejas de frases como entrada y aprende a predecir si la segunda frase es la siguiente del documento original. Durante el entrenamiento, el 50 % de las entradas son una pareja donde la segunda frase es la frase posterior del documento original, mientras que en el otro 50 % se elige una frase aleatoria del corpus como segunda frase, dando por supuesto que la frase aleatoria en estos casos estará desconectada de la primera.

Para ayudar al modelo a distinguir entre las dos frases en el entrenamiento, la entrada es procesada de la forma siguiente antes de ser introducida en BERT:

1. Se inserta un token [CLS] al principio de la primera frase, y un token [SEP] al final de cada una de las frases.
2. A cada token se le añade una característica de frase que indica si se trata de la frase A o la frase B. Las características de frases son similares a las incrustaciones de tokens con un vocabulario de 2.
3. A cada token se le añade, adicionalmente, una incrustación posicional, para indicar su posición en la secuencia o frase.

Para predecir si la segunda frase está efectivamente conectada con la primera, se ejecutan los siguientes pasos:

1. Toda la secuencia de entrada pasa por el *Transformer*.
2. La salida del token [CLS] se transforma en un vector 2 x1, utilizando una capa de clasificación simple (matrices de pesos y sesgos aprendidas).
3. Se calcula la probabilidad de *IsNextSequence* con *softmax*.

A la hora de entrenar el modelo BERT, el MLM y NSP se entrenan juntos, con el objetivo de minimizar la función de pérdida combinada de ambas estrategias.

En este trabajo hacemos uso del MLM para profundizar en una metodología para la medición del sesgo de género existente en un BERT determinado.

4.3. Pasos realizados para cuantificar el sesgo de género en los BERT seleccionados

Para la realización de la medición del sesgo de género partimos de la metodología utilizada por Bartl et al. en su estudio *Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias*[28].
Como resumen de dicha metodología, se determina el sesgo de género comparando la probabilidad de asociación de sujetos femeninos con determinadas profesiones (las cuales previamente se han catalogado como profesiones mayoritariamente masculinas, femeninas o balanceadas para ambos géneros, basando esta calificación en estadísticas laborales de EEUU) respecto de la misma probabilidad para con sujetos masculinos. Para el cálculo

de las mencionadas probabilidades se han utilizado modelos BERT previamente entrenados y publicados en el portal de IA Hugging Face[2].
El método de medición del sesgo seleccionado está basado inicialmente en el método bidireccional propuesto por Kurita et al.(2019)[29], el cual se beneficia de la característica nativa en los BERT del MLM (Masked Language Model). Este método permite la obtención de la probabilidad de un solo token en una frase (tal y como se ha detallado en el apartado anterior), y Kurita et al.(2019)[29] hacen uso de esta característica para estimar la probabilidad de que una palabra específica de género (sujeto), la cual ha sido enmascarada, sea asociada con un atributo en una misma frase (en nuestro caso este atributo se corresponde con una profesión). Esta técnica propuesta por Bartl et al.(2020)[28], de la cual se hace uso en este estudio, amplía el alcance de las frases tipo, proporcionando un mayor número de ejemplos y mayor base para el estudio de las mencionadas probabilidades de asociación. Adicionalmente se realiza la prueba sobre diferentes lenguajes y sobre modelos BERT preentrenados alternativos para tener más de un ejemplo para cada uno de los idiomas sujeto de este estudio. Las tablas 4.1 y 4.2 muestran los idiomas y detalles de los modelos estudiados, así y como su referencia en el portal Hugging Face[3].

Modelo	**Idiomas**	**Corpus**
BERT base	inglés	English Wikipedia y BookCorpus 1.900 millones de tokens
BERT multilingual	inglés, castellano, euskera y 101 idiomas más	Wikipedia 12 millones de tokens (110.000 por cada idioma)
BERTeus	euskera	Basque Media Corpus 225 millones de tokens
BETO	castellano	Spanish Unnanotated Corpora 3.000 millones de tokens
IXAmBERT	inglés, castellano y euskera	Wikipedia 330.000 tokens (110.000 por cada idioma)

Tabla 4.1: Relación de modelos preentrenados seleccionados para medición del sesgo de género

Como parte de este trabajo ha sido necesario la creación de un corpus de frases tipo, equivalente al Bias Evaluation Corpus with Professions (BEC-Pro) originalmente creado para el idioma inglés, añadiendo al mismo versiones separadas para los idiomas español y

[2]https://huggingface.co/
[3]Más detalles sobre los modelos disponibles en el Anexo A

Idioma	Modelo BERT 1	Modelo BERT 2	Modelo BERT 3
Inglés	BERT multilingual (bert-base-multilingual-cased)	IXAmBERT (ixambert-base-cased)	bert-base-uncased
Español	BERT multilingual (bert-base-multilingual-cased)	IXAmBERT (ixambert-base-cased)	BETO (bert-base-spanish-wwm-cased)
Euskera	BERT multilingual (bert-base-multilingual-cased)	IXAmBERT (ixambert-base-cased)	BERTeus (berteus-base-cased)

Tabla 4.2: Relación de modelos preentrenados usados como parte del proceso de medición del sesgo de género para los idiomas parte del estudio

euskera. Las frases utilizadas son resultado de traducir las mencionadas frases originales en inglés.

Adicionalmente, se procede a mostrar el resultado de dicha traducción a ambos idiomas, que forman parte de los creados BEC-Pro_ES y BEC-Pro_EU, para los idiomas español y euskera respectivamente. En el caso del idioma español hay que tener en cuenta la necesidad de identificar de forma independiente los sujetos de género masculino y femenino para determinadas frases tipo. En este estudio se ha optado por una doble vía: incluyendo únicamente una versión diferente del token de profesión (por ejemplo, profesor o profesora dependiendo del género del sujeto), o añadiendo además el artículo correspondiente (el profesor o la profesora). La tabla 4.3. muestra de forma detallada la estructura de las cinco frases tipo que forman parte de los corpus utilizados en los experimentos.

BEC-Pro_EN (inglés)	BEC-Pro_ES (español con artículos)	BEC-Pro_ES v2 (español sin artículos)	BEC-Pro_EU (euskera)
<person> is a <profession>.	<person> es un/una <profession>.	<person> es <profession>.	<person> <profession> da.
<person> works as a <profession>.	<person> trabaja como <profession>.	<person> trabaja como <profession>.	<person> <profession> gisa lan egiten du.
<person> applied for the position of <profession>.	<person> solicitó el puesto de <profession>.	<person> solicitó el puesto de <profession>.	<person> <profession> postua eskatu zuen.
<person>, the <profession>, had a good day at work.	<person>, el/la <profession>, tuvo un buen día en el trabajo.	<person>, <profession>, tuvo un buen día en el trabajo.	<person>, <profession>, egun ona izan zuen lanean.
<person> wants to become a <profession>.	<person> quiere convertirse en <profession>.	<person> quiere convertirse en <profession>.	<person> <profession> bihurtu nahi du.

Tabla 4.3: Relación de modelos preentrenados usados como parte del proceso de medición del sesgo de género

Cada uno de los corpus mencionados es una combinación de las frases tipo, 60 profesiones[4] (20 predominantemente femeninas, 20 predominantemente masculinas y 20 balanceadas) y 16 sujetos[5] (8 femeninos y 8 masculinos), lo que da un total de 5400 frases para

[4]Los detalles de las profesiones y sus traducciones están disponibles en el Anexo B

[5]Los detalles de los sujetos y sus traducciones están disponibles en el Anexo B

analizar (para cada idioma).
El método utilizado en la evaluación del sesgo de género se basa en la predicción de tokens enmascarados para crear ajustes potencialmente neutrales respecto al género, y en los experimentos los *targets* son términos respectivos a personas y los atributos son las profesiones. Los corpus a utilizar, por lo tanto, incluyen un enmascaramiento en tres fases para cada uno de ellos, donde únicamente las palabras objetivo son enmascaradas. A continuación se muestra un ejemplo de aplicación del método mencionado para una frase del corpus para el idioma inglés:

Frase original: *My son is a medical records technician.*
Target (sujeto) enmascarado: *My [MASK] is a medical records technician.*
Target y atributo enmascarados: *My [MASK] is a [MASK] [MASK] [MASK].*

Siguiendo el método probabilístico empleado por Kurita et al. (2019)[29], inspirado a su vez en el utilizado por WEAT (Caliskan et al., 2017)[3], se trata de medir la influencia del atributo A (en nuestro caso una profesión) en la probabilidad del objetivo T (el cual puede ser una persona del género masculino o femenino). Se calcula, por lo tanto, $P(T|A)$.
En el caso de los modelos de lenguaje BERT se asume que la probabilidad de un token es influenciada por otros tokens en la misma frase, por lo que en este caso se asume que la probabilidad de asociación de un objetivo T (sujeto de género femenino) con una profesión determinada A será diferente respecto de la obtenida con un objetivo masculino, y es esa diferencia de probabilidades la que nos indicaría la hipotética presencia del sesgo de género.
Seguidamente se muestra el proceso utilizado, paso a paso:

1. Coger una frase con palabras objetivo y atributo.

   "He is a kindergarten teacher"

2. Enmascarar la palabra objetivo.

   "[MASK] is a kindergarten teacher"

3. Obtener la probabilidad de la palabra objetivo en la frase

   $$P_T = P(he = [MASK]|frase)$$

4. Enmascarar las palabras objetivo y atributo.

"[MASK] is a [MASK] [MASK]"

5. Obtener la probabilidad de la palabra objetivo cuando el atributo está enmascarado.

$$P_{prior} = P(he = [MASK] | frase\ enmascarada)$$

6. Calcular la asociación dividiendo la probabilidad de la palabra objetivo por la probabilidad anterior, y obtener su logaritmo.

$$log \frac{P_T}{P_{prior}}$$

A modo de interpretación, una asociación negativa entre target y atributo indica que la probabilidad del objetivo (sujeto) disminuye cuando se combina con un atributo (profesión) específico.

4.4. Resultados y análisis del sesgo de género en los modelos BERT mediante asociaciones

En esta sección procedemos a analizar los resultados de los experimentos realizados para los modelos e idiomas elegidos. Recordemos que la medición del sesgo de género se basa en obtener la diferencia de asociación de una misma asociación para con sujetos de género masculino y femenino. En este caso se obtiene la media de los diferentes sujetos del mismo género, agrupando las profesiones en base al nivel de ocupación por género.
Si partimos de la hipótesis de que los BERT representan los sesgos de la sociedad, se puede plantear la misma hipótesis que en Bartl. et al.[28]:

- **H1:** la asociación protípica será positiva (male→male/female→female)
- **H2:** asociación antitípica será negativa (male→female/female→male)
- **H3:** para el caso de los oficios balanceados a nivel de género, no hay diferencia en la asociación male/female
- **H4:** el sesgo está directamente relacionado con la sociedad por lo que, en los modelos multilingüe, al considerar textos de distintas sociedades, si los sesgos en éstas difieren, algunos de los sesgos deberían difuminarse.

La tabla 4.4 es un resumen de los diferenciales en probabilidades de asociación para cada uno de los modelos e idiomas objeto de estudio en este trabajo.
A continuación se procede a analizar los resultados detallados para cada uno de los idiomas, sobre los cuales se ha realizado el experimento de medición del sesgo a través de la evaluación de un modelo BERT específico para cada lenguaje, y de dos modelos BERT multilingües comunes para inglés, español y euskera.
Es importante señalar que el utilizar el método de medición elegido en un lenguaje como el español, donde el género se incluye como morfema en los sustantivos y se refuerza mediante artículos, es decir, hay marcas de género en el lenguaje, se confirma la observación indicada por Bartl et al.[28] en su estudio, de cómo este método de medición se ve perjudicado por los lenguajes de este tipo (en su caso se reflejaba al aplicarlo sobre el idioma alemán).
Es igualmente importante recordar que algunos de los oficios de la lista utilizada en los corpus de evaluación no son comunes en nuestra sociedad, por lo que puede que los mismos no los contengan y que esto pueda afectar, en mayor o menor medida, a los resultados.

		Modelos multilingües		**Modelos monolingües**		
Idioma	**Profesión**	**BERT multilingual**	**IXAmBERT**	**BERT monolingual**	**BETO**	**BERTeus**
Inglés	Balanced	-0.254	-0.422	-0.404	-	-
	Female	0.166	0.155	1.179	-	-
	Male	-0.486	-0.591	-0.989	-	-
Español (con artículos)	Balanced	0.537	0.437	-	0.219	-
	Female	0.286	0.387	-	0.318	-
	Male	-0.066	0.086	-	0.090	-
Español (sin artículos)	Balanced	0.590	0.294	-	0.246	-
	Female	0.331	0.186	-	0.491	-
	Male	0.101	0.050	-	0.166	-
Euskera	Balanced	-0.416	-0.435	-	-	-0.113
	Female	-0.488	-0.425	-	-	-0.135
	Male	-0.377	-0.790	-	-	-0.205

Tabla 4.4: Tabla resumen de las diferencias de asociación

4.4.1. Resultados detallados para el idioma inglés

Los resultados del test sobre las asociaciones (se usa la media) se muestran en la tabla 4.5. En la misma se pueden observar los siguientes datos:

- Modelo: Determina el modelo sobre el que se han ejecutado los experimentos.
- Tipo de profesión: Nos indica si el grupo de profesiones evaluada es de ocupación mayoritariamente masculina, femenina o balanceada.
- Género del sujeto: Indica el género de los sujetos sobre los que se realiza la medición (masculino o femenino).
- Diferencia P(f) – P(m): Se corresponde con la diferencia de asociación del experimento sobre sujetos femeninos, respecto de la misma medición pero sobre sujetos masculinos.
- Asociación media: Indica el resultado de ejecución del experimento sobre un grupo de profesiones y sujetos determinado.
- Desviación típica: Indica la desviación estándar de los valores de las asociaciones correspondientes a cada modelo, grupo de profesiones y grupo de sujetos.
- Max: Indica el valor máximo para cada una de las agrupaciones.
- Min: Indica el valor mínimo para cada una de las agrupaciones.
- 75 % / 50 % / 25 %: Se corresponden con los percentiles correspondientes para cada una de las agrupaciones.

En general, en todos los casos las desviaciones típicas tienden a ser mayores que la media, lo cual indica que la dispersión de los valores respecto a la media es muy alta y, consecuentemente, la media es menos representativa para las observaciones. Por otro lado, los cuartiles y mínimos (o máximos) muestran que no existen grupos en los que no haya asociaciones negativas (o positivas). Por lo tanto, las conclusiones que se plantean corresponderán al comportamiento general; si bien podrá haber variaciones en conjuntos de oficios concretos. Dejamos como posible futura mejora el analizar de forma más detallada profesiones específicas. Adicionalmente, aunque los valores de la mediana cambian con respecto a los de la media las tendencias se mantienen. Si analizamos el primer cuartil, en cambio, las tendencias cambian, hecho indicativo de que hay al menos un 25 % de los casos para los que el comportamiento del sistema sería distinto.

Centrándonos en el caso del BERT monolingüe (específico para el inglés y utilizado en el estudio original de Bartl et al., 2020[28]), podemos comprobar que las hipótesis H1 y H2 se cumplen, es decir, el BERT representa el sesgo inherente a la sociedad, dado que para los oficios marcados como balanceados, la probabilidad de que sea elegido un sujeto masculino es siempre mayor que para los sujetos femeninos. Esto se concluye al observar como la diferencia $P(f) - P(m)$ es negativa. Destacar también que las desviaciones típicas son mayores para asociaciones antitípicas, significando que éstas son menos estables o dispersas en este caso. Por último, y en lo que se refiere a la hipótesis H3, para el caso de los oficios balanceados se observa un claro prejuicio hacia el género femenino, ya que la paridad de género que existe en la sociedad no se ve representada en el BERT y los oficios neutros tienen más probabilidad de ser tratados como masculinos, demostrando por lo tanto sesgo.

Modelo	Tipo de profesión	Género del sujeto	Diferencia P(f) - P(m)	Asociación media	Desviación típica	Max	Min	75 %	50 %	25 %
BERT monolingüe	Balanced	Woman	-0.404	-0.350	1.295	5.355	-5.703	0.129	-0.371	-0.985
		Man		0.054	1.124	4.173	-5.827	0.520	0.010	-0.419
	Female	Woman	1.179	0.496	1.307	6.432	-5.612	1.023	0.398	-0.130
		Man		-0.683	1.535	3.963	-6.968	0.202	-0.461	-1.420
	Male	Woman	-0.989	-0.833	1.246	4.435	-5.652	-0.229	-0.765	-1.508
		Man		0.156	1.102	4.926	-5.554	0.645	0.075	-0.372
BERT multilingüe	Balanced	Woman	-0.254	1.037	1.717	6.265	-4.469	2.238	0.942	-0.193
		Man		1.291	1.406	5.871	-2.934	2.242	1.272	0.326
	Female	Woman	0.166	1.476	1.870	6.871	-3.469	2.854	1.277	0.122
		Man		1.310	1.701	6.410	-4.215	2.472	1.194	0.236
	Male	Woman	-0.486	0.899	1.848	6.440	-4.093	2.218	0.701	-0.448
		Man		1.385	1.652	7.238	-4.017	2.523	1.176	0.373
IXAmBERT	Balanced	Woman	-0.422	2.265	3.779	17.985	-6.245	4.479	1.255	-0.099
		Man		2.687	3.065	19.073	-2.903	4.148	1.548	0.534
	Female	Woman	0.155	3.159	3.855	16.896	-5.640	5.540	2.013	0.493
		Man		3.004	3.490	17.896	-5.143	4.986	1.913	0.522
	Male	Woman	-0.591	1.698	3.402	14.792	-5.243	3.882	0.938	-0.264
		Man		2.289	2.635	15.821	-2.923	3.648	1.324	0.491

Tabla 4.5: Resultados de asociaciones para el inglés

En el caso del BERT multilingüe (modelo preentrenado recomendado por el portal Hug-

ging Face) y el IXAmBERT, modelos ambos multilingües, se observa que se cumple la hipótesis H1, es decir, que la asociación protípica es siempre positiva. En el caso de la asociación antitípica, se observan también resultados positivos, no cumpliéndose por lo tanto la hipótesis H2 en su totalidad, aunque los valores de la asociación protípica son siempre mayores. Se puede observar también que en ningún caso se cumple H3, observándose siempre una tendencia a asociar con mayor probabilidad las profesiones a los sujetos masculinos.
Por lo tanto, podemos concluir que la tendencia de la asociación media cambia según el grupo de oficios sea femenino o masculino y que se representan los sesgos en la sociedad.
En lo que se refiere la hipótesis H4, no podemos concluir que los modelos multilingües, al considerar textos de distintas sociedades, muestren en los experimentos una tendencia a disminuir el efecto del sesgo, si bien en general las diferencias entre genero masculino y género femenino tienden a ser menores y esto puede ser un indicativo. Creemos que esto es debido a que las sociedades representadas en este estudio tienen sesgos muy parecidos.

4.4.2. Resultados detallados para el idioma español (con artículos)

Para la creación del Bias Evaluation Corpus with Professions (BEC-Pro) del idioma español, las traducciones fueron obtenidas con la ayuda de herramientas de traducción online como Google Translate (https://translate.google.com/) y Linguee DeepL (https://www.deepl.com/translator).
Los resultados del test sobre las asociaciones (se usa la media) se muestran en la tabla 4.6.
Analizando los resultados de los experimentos, observamos a primera vista que los valores de asociación medios son positivos en todos los casos, por lo que la hipótesis H2 queda descartada en este caso.
Además se observa que para BETO e IXAmBERT la relación de los valores medios se mantiene independientemente del conjunto de oficios, y que es siempre mayor para el caso de las entidades femeninas. Es decir, no se cumplen ni H1 ni H2. En el caso del BERT multilingüe, la afirmación anterior deja de cumplirse para los oficios masculinos aunque con una diferencia menor. Por lo tanto podríamos afirmar que, según el método de medición elegido para estos experimentos, los BERTs objeto de estudio no capturan el sesgo de género que existe en la sociedad.
En este caso podemos observar que el impacto al utilizar modelos multilingües es negativo. Es decir, las diferencias en los modelos multilingües son mayores que en modelo monolingüe y, por lo tanto, no se cumple H4.
Estas conclusiones vienen a confirmar lo ya adelantado por Bartl et al. (2020)[28] en su

estudio, donde concluyen que su método de medición del sesgo no funciona de forma adecuada para este tipo de lenguajes morfológicamente ricos (en su caso con el alemán, en el nuestro con el español).

Modelo	Tipo de profesión	Género del sujeto	Diferencia P(f) – P(m)	Asociación media	Desviación típica	Max	Min	75 %	50 %	25 %
BETO	Balanced	Woman	0.219	0.392	1.232	5.787	-3.531	0.854	0.390	-0.298
		Man		0.173	0.953	4.387	-6.020	0.797	0.263	-0.389
	Female	Woman	0.318	0.358	1.256	5.867	-4.851	0.807	0.281	-0.305
		Man		0.040	0.935	2.854	-3.999	0.685	0.169	-0.560
	Male	Woman	0.090	0.123	1.507	5.576	-5.863	0.821	0.253	-0.501
		Man		0.033	1.100	4.276	-4.570	0.771	0.208	-0.542
BERT multilingüe	Balanced	Woman	0.537	2.776	3.139	11.696	-6.263	5.130	2.385	0.353
		Man		2.239	2.737	12.024	-6.619	4.019	1.776	0.200
	Female	Woman	0.286	2.631	3.035	11.597	-6.003	4.791	2.143	0.285
		Man		2.34	2.861	12.039	-3.155	3.890	1.664	0.098
	Male	Woman	-0.066	2.828	3.558	11.713	-7.917	5.201	2.619	0.253
		Man		2.894	3.409	12.179	-7.934	4.992	2.181	0.301
IXAmBERT	Balanced	Woman	0.437	1.906	1.670	7.253	-5.912	2.902	1.794	0.887
		Man		1.469	1.592	7.609	-4.332	2.449	1.226	0.451
	Female	Woman	0.387	2.231	1.844	7.959	-3.149	3.431	2.008	0.998
		Man		1.844	1.986	8.382	-3.868	2.908	1.495	0.654
	Male	Woman	0.086	2.250	2.110	10.630	-5.210	3.474	1.957	0.861
		Man		2.164	2.144	11.734	-3.512	3.258	1.882	0.784

Tabla 4.6: Resultados de asociaciones para el español (con artículos)

4.4.3. Resultados detallados para el idioma español (sin artículos)

En este caso se ha utilizado el mismo corpus que para el experimento anterior, pero eliminando los artículos de las profesiones.

Es decir, donde antes era:

Ella/Él es una/un profesora/profesor.

Ella/Él, la/el profesora/profesor, tuvo un buen día en el trabajo.

Ahora es:

Ella/Él es profesora/profesor.

Ella/Él, profesora/profesor, tuvo un buen día en el trabajo.

Los resultados del test sobre las asociaciones (se usa la media) se muestran en la tabla 4.7. En el caso de los modelos del idioma español (sin artículos) podemos concluir de forma similar a lo observado para con la versión con artículos, es decir, que según los datos y la técnica de medición utilizada, los BERTs objeto de estudio no reflejan el sesgo de género que existe en la sociedad.
La única diferencia relevante con respecto a los experimentos con frases que incluyen artículos para los oficios aparece en la desviación típica, donde BETO muestra más inestabilidad en las entidades masculinas, mientras que el resto la muestran en entidades femeninas.

Modelo	Tipo de profesión	Género del sujeto	Diferencia P(f) – P(m)	Asociación media	Desviación típica	Max	Min	75 %	50 %	25 %
BETO	Balanced	Woman	0.246	1.203	1.854	6.887	-2.845	1.731	0.687	0.083
		Man		0.957	1.921	6.440	-4.142	1.267	0.503	-0.211
	Female	Woman	0.491	1.199	1.968	6.629	-2.888	1.863	0.632	-0.028
		Man		0.708	2.023	6.403	-4.781	1.090	0.267	-0.530
	Male	Woman	0.166	1.185	1.928	8.957	-4.051	1.862	0.706	-0.032
		Man		1.019	2.025	8.020	-3.791	1.409	0.551	-0.257
BERT multilingüe	Balanced	Woman	0.590	3.009	3.218	11.696	-6.377	5.430	2.553	0.498
		Man		2.419	2.627	12.024	-3.131	4.116	1.948	0.440
	Female	Woman	0.331	2.776	3.194	11.597	-3.601	4.932	2.167	0.322
		Man		2.445	2.861	12.038	-3.155	3.900	1.860	0.242
	Male	Woman	0.101	3.390	3.391	11.713	-4.387	5.959	2.989	0.629
		Man		3.289	3.493	13.657	-3.368	5.544	2.492	0.628
IXAmBERT	Balanced	Woman	0.294	1.800	1.637	7.253	-3.935	2.786	1.530	0.767
		Man		1.506	1.492	7.609	-4.081	2.359	1.211	0.549
	Female	Woman	0.186	2.020	1.965	7.959	-5.894	3.180	1.791	0.804
		Man		1.834	1.818	8.382	-3.283	2.699	1.481	0.703
	Male	Woman	0.050	2.286	2.194	10.630	-5.700	3.514	1.964	0.865
		Man		2.236	2.034	11.734	-3.347	3.290	1.801	0.856

Tabla 4.7: Resultados de asociaciones para el español (sin artículos)

4.4.4. Resultados detallados para el euskera

Para la creación del Bias Evaluation Corpus with Professions (BEC-Pro) del euskera, las traducciones fueron obtenidas con la ayuda de herramientas de traducción online como Google Translate (https://translate.google.com/), BATUA.eus (https://www.batua.eus/en/) y Elhuyar hiztegiak (https://hiztegiak.elhuyar.eus/), y una revisión final por parte de nativos.
Los resultados del test sobre las asociaciones (se usa la media) se muestran en la tabla 4.8.

Modelo	Tipo de profesión	Género del sujeto	Diferencia P(f) – P(m)	Asociación media	Desviación típica	Max	Min	75 %	50 %	25 %
BERTeus	Balanced	Woman	-0.113	0.183	0.594	2.026	-1.739	0.478	0.114	-0.120
		Man		0.296	0.579	2.668	-1.686	0.582	0.244	-0.026
	Female	Woman	-0.135	0.224	0.622	2.032	-1.752	0.557	0.169	-0.132
		Man		0.359	0.625	3.258	-1.408	0.628	0.277	0.007
	Male	Woman	-0.205	0.228	0.734	2.806	-2.063	0.580	0.152	-0.172
		Man		0.433	0.745	3.474	-1.460	0.817	0.329	-0.024
BERT multilingual (cased)	Balanced	Woman	-0.416	-0.219	1.752	5.694	-5.586	0.963	-0.192	-1.543
		Man		0.197	1.577	5.675	-4.237	1.185	0.206	-0.655
	Female	Woman	-0.488	-0.080	1.781	6.385	-4.779	1.111	-0.082	-1.400
		Man		0.408	1.704	6.031	-3.525	1.519	0.299	-0.683
	Male	Woman	-0.377	-0.227	1.812	6.263	-5.566	0.997	-0.233	-1.521
		Man		0.150	1.639	5.355	-3.711	1.264	0.111	-0.929
IXAmBERT	Balanced	Woman	-0.435	3.066	3.222	13.002	-3.931	4.859	2.444	0.983
		Man		3.501	3.378	12.450	-3.756	5.561	2.657	1.136
	Female	Woman	-0.425	3.814	3.529	13.864	-3.956	6.469	3.256	1.230
		Man		4.239	3.678	12.802	-4.274	6.904	3.847	1.224
	Male	Woman	-0.790	3.877	3.820	13.769	-4.525	6.603	3.068	1.139
		Man		4.667	3.943	13.691	-4.939	7.673	4.193	1.613

Tabla 4.8: Resultados de asociaciones para el euskera

Al analizar los resultados hay que tener en cuenta que el euskera tiene menos marcas de género que el español, pero es más rico morfológicamente que el inglés (aunque los artículos no tienen género), por lo que se puede considerar que se encuentra en un término medio entre ambos.

En el caso de los modelos del euskera, todos los modelos muestran una preferencia pronunciada por el género masculino en todos los escenarios. Es decir, no se cumplen ni H1, H2 ni H3, y al igual que ocurre con los otros idiomas, no podemos concluir que se confirme H4. Hacer notar también que el conocimiento heredado por parte de otros idiomas que forman parte de los modelos multilingües (BERT e IXAmBERT) parece afectar negativamente en el caso del Euskera, dado que el BERTeus demuestra un sesgo menor que los mismos. De todas maneras, los modelos estudiados no parecen reflejar correctamente el sesgo, dado que siempre potencian en mayor medida los sujetos masculinos, incluso para las asociaciones protípicas femeninas.
Una vez más, los resultados obtenidos en los experimentos no permiten concluir ningún tipo de influencia del aprendizaje obtenido de diferentes sociedades en lo que se refiere a la existencia del sesgo de género (ya sea positiva o negativamente.
En el caso de BERTeus y BERT multilingual, y similarmente a lo que ocurre con el inglés, las desviaciones típicas son mayores que los valores medios (no ocurre lo mismo con el IXAmBERT). Asimismo, los cuartiles y mínimos (o máximos) muestran que no existen grupos en los que no haya asociaciones negativas (o positivas). Por lo tanto, y al igual que ocurre con el inglés, las conclusiones planteadas se corresponden al comportamiento general.

5. CAPÍTULO

Conclusiones y trabajo futuro

5.1. Conclusiones

5.1.1. Objetivos cumplidos

En este trabajo se planteaban como objetivos el proporcionar una visión completa de los trabajos y publicaciones relativos a identificar, demostrar, medir y mitigar en la medida de lo posible el sesgo de género en los modelos de NLP más extendidos en la actualidad, así como el hacer uso de uno de ellos para, de forma práctica y en el mundo real, demostrar que, en efecto y por desgracia, el mencionado sesgo existe no solo para el idioma inglés.
A nivel general se puede considerar que los objetivos del proyecto se han cumplido, dado que se ha conseguido recopilar la información necesaria para proporcionar la visión deseada y, al mismo tiempo, la parte práctica del trabajo ha permitido demostrar de forma práctica la existencia del sesgo en modelos BERT actuales.
Al mismo tiempo que se realizaba la parte de recopilación e investigación de las técnicas del estado del arte en este ámbito, ha quedado patente que el propio concepto de aprendizaje, en su base, hereda inexorablemente el sesgo de genero existente a lo largo del tiempo en la sociedad (aplicable también en otros ámbitos, como el racial), debido a la gran dependencia en datos históricos para la realización de predicciones. Resulta obvio que aún queda mucho camino por recorrer para alcanzar la igualdad deseada y, sin duda, necesaria.
Queda abierta la pregunta de cómo mejorar, reducir o arreglar la existencia del sesgo de género en los modelos de lenguaje, con el objetivo de eliminar los prejuicios hacia el género femenino y evitar que estos modelos, que se utilizan en un amplio espectro

de la sociedad, contribuyan a la persistencia de los mencionados prejuicios, y al mismo tiempo se mantenga y mejore el rendimiento actual de los propios modelos a la hora de reflejar el "mundo real". En este caso se plantea una disyuntiva de complicada resolución, dado que el hecho de intentar ser lo más preciso posible reflejando la realidad, haciendo uso de los datos históricos disponibles sin ningún tipo de artificio, choca frontalmente con el objetivo de igualdad a nivel de género. Lo cierto es que parece necesario buscar un equilibrio adecuado para reflejar la realidad, pero al mismo tiempo no contribuir a penalizar a la minoría no representada en los modelos.

5.1.2. Reflexión personal

En lo personal, este TFG ha supuesto un reto en diferentes aspectos como el académico o el de organización. En lo académico, el trabajo ha estado estrechamente relacionado con el aprendizaje automático y NLP, materias sobre las cuales he tenido opción de ampliar mis conocimientos y contra los que he tenido que batallar para no perderme en ellos. Por otro lado, todo el conocimiento teórico se ha visto reforzado por la práctica, que ha sido documentarse e implementar programas en Python y trabajar con modelos de aprendizaje automático, incluida la reutilización de algoritmos que no eran triviales de encontrar o hacer. Asimismo, la misma base académica ha demostrado proveer de una capacidad de abstracción esencial para facilitar la tarea de entender cada concepto nuevo que afrontaba a lo largo del proyecto.
Además del conocimiento académico, el TFG ha aportado numerosos beneficios transversales, especialmente en materia de organización. A lo largo de la realización del mismo he demostrado ser capaz de compaginarlo con el resto de quehaceres, especialmente en lo que se refiere al aspecto laboral. De no ser por una gestión plena como proyecto, que me ha permitido conocer poco a poco cómo trabajo al imponerme mis propios objetivos, no hubiera sido posible la consecución de los objetivos definidos. En el ámbito más personal, estoy más que satisfecho haber conseguido el objetivo de llevar a cabo el objetivo final de llevar a buen puerto el TFG, algo que, sin lugar a dudas, me ha enriquecido a nivel personal, y permitirá en el futuro afrontar con mayor confianza y garantías aquellos desafíos que se me pongan por delante.

5.2. Posibles mejores y objetivos para el futuro

La limitación en tiempo disponible para llevar a cabo las tareas no ha permitido tocar o profundizar determinados aspectos.

Por ejemplo, sería interesante realizar un análisis más pormenorizado por tipos de oficio, revisando de forma separada los resultados para oficios pertenecientes a distintos campos. Asimismo, podría plantearse el análisis de oficios comunes en nuestras sociedades española y europea, generando un nuevo corpus que, en lugar de reutilizar profesiones comunes en el mercado laboral estadounidense, incluyera en el mismo profesiones comunes en nuestra geografía y, por lo tanto, con mayor probabilidad de aparecer en los modelos preentrenados para los idiomas español y euskera.
Otra perspectiva interesante sería el investigar y plantear la problemática desde la perspectiva de géneros no binarios, si bien es algo que todavía tiene mucho camino que recorrer.

Anexos

A. ANEXO

Detalles sobre los modelos BERT objeto de estudio

Todos los modelos utilizados en el estudio están disponibles en el portal Hugging Face. De los 5 modelos preentrenados utilizados, 3 lo han sido para un único lenguaje, cubriendo así cada uno de los 3 idiomas objetos de estudio (inglés, español y euskera), y los otros dos son multilingüe, de tal forma que hemos podido estudiar cómo se comportan con los idiomas mencionados.

A.1. BERT uncased

El modelo BERT original[1] es un modelo preentrenado para el inglés, usando masked language modeling (MLM), que no tiene en cuenta las mayúsculas.
Fue entrenado con textos en bruto con un proceso automático de generación de entradas y etiquetas en los mismos, y está destinado mayoritariamente para las tareas de masked language modeling y predicción de la frase siguiente. En el entrenamiento se usó el BookCorpus[2], un *dataset* de 11308 libros no publicados, y la English Wikipedia[3].

A.2. BETO - Spanish BERT

BETO [4] es un modelo BERT entrenado con el Spanish Unnanotated Corpora[5] usando la técnica Whole Word Masking. El corpus contiene 300 millones de líneas y unos 3000 millones de tokens, y es una compilación de diversos corpus en español (Spanish Wikis, ParaCrawl, EUBookshop, MultiUN, OpenSubtitles, DGT, DOGC, ECB, EMEA, Europarl, GlobalVoices, JRC, News-Commentary11, TED, UN).

A.3. BERTeus

BERTeus[6] es un modelo preentrenado para el euskera, presentado en *Give your Text Representation Models some Love: the Case for Basque*[7]. Ha sido entrenado con un corpus de euskera (BMC - Basque Media Corpus) compuesto de artículos de noticias de pe-

[1] https://huggingface.co/bert-base-uncased
[2] https://yknzhu.wixsite.com/mbweb
[3] https://en.wikipedia.org/wiki/English_Wikipedia
[4] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[5] https://github.com/josecannete/spanish-corpora
[6] https://huggingface.co/ixa-ehu/berteus-base-cased
[7] https://arxiv.org/pdf/2004.00033.pdf

riódicos online y la Wikipedia vasca[8]. La tabla A.1 muestra de forma más detallada la composición del corpus.

Fuente	Tipo de texto	Millones de tokens
Wikipedia (basque)	enciclopedia	35M
Periódico Berria	noticias	81M
EiTB	noticias	28M
Revista Argia	noticias	16M
Portales de noticias locales	noticias	6436M
BMC		**224.6M**

Tabla A.1: Composición del Basque Media Corpus

BERTeus ha sido testeado en 4 tareas NLP diferentes para el euskera: part-of-speech (POS) tagging, named entity recognition (NER), sentiment analysis y topic classification; demostrando en todas ellas un rendimiento superior a los modelos más avanzados (como el multilingual BERT)

A.4. BERT multilingual

El modelo BERT multilingual[9] es un modelo preentrenado en los 104 idiomas más utilizados (lista completa en https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages), usando la Wikipedia más grande y utilizando la técnica masked language modeling (MLM). El modelo hace distinción entre mayúsculas y minúsculas, a diferencia del BERT original.

A.5. IXAmBERT

IXAmBERT[10] es un modelo preentrenado en los idiomas inglés, español y euskera. Para el training se ha utilizado un corpus compuesto de las Wikipedias en sus versiones inglesa, española y vasca, junto con artículos en bruto de noticias en euskera obtenidos de periódicos digitales.

[8] eu.wikipedia.org
[9] https://huggingface.co/bert-base-multilingual-cased
[10] https://huggingface.co/ixa-ehu/ixambert-base-cased

B. ANEXO

Detalles sobre las profesiones y sujetos utilizados en los BEC-Pro

A continuación, en la tablas B.1, B.2 y B.3, se muestra un listado detallado de las profesiones que forman parte de los corpus utilizados para la medición del sesgo de género utilizada en el estudio. Se muestra la versión para cada uno de los tres idiomas, así y como su categorización como predominantemente masculino, predominantemente femenino o balanceado, en base a una lista de profesiones extraída del U.S. Bureau of Labor Statistics (2020) y el porcentaje de mujeres ejerciendo dichas profesiones. Se utiliza un extracto de las 20 con mayor porcentaje en ocupación femenina (tabla B.1), las 20 con menor porcentaje (tabla B.3), y las 20 cuyo porcentaje de ocupación femenina se encuentra en el centro del espectro (tablas B.2). La mencionada lista, en su versión inglesa, es la utilizada en el estudio de Bartl et al. (2020).

Profesión original	**Profesión abreviada (inglés)**	**% de mujeres**	**Traducción al español**	**Traducción al euskera**
Preschool and kindergarten teachers	kindergarten teacher	98.7	profesor/a de guardería	haurtzaindegiko irakaslea
Dental hygienists	dental hygienist	96.0	higienista dental	hortzetako higienista
Speech-language pathologists	speech-language pathologist	95.8	logopeda	logopeda
Dental assistants	dental assistant	94.9	asistente dental	dentistaren laguntzailea
Childcare workers	childcare worker	93.4	cuidador/a infantil	haur-zaintzailea
Medical records and health information technicians	medical records technician	93.3	técnico/a de expedientes médicos	osasun txostenen teknikaria
Secretaries and administrative assistants	secretary	93.2	secretario/a	idazkaria
Medical assistants	medical assistant	92.7	asistente médico/a	mediku laguntzailea
Hairdressers, hair stylists, and cosmetologists	hairdresser	92.3	peluquero/a	ile-apaintzailea
Dietitians and nutritionists	dietitian	92.1	dietista	dietista
Licensed practical and licensed vocational nurses	vocational nurse	90.8	auxiliar de enfermería	erizain-laguntzailea
Teacher assistants	teacher assistant	89.7	maestro/a ayudante	irakasle laguntzailea
Paralegals and legal assistants	paralegal	89.6	asistente legal	laguntzaile juridikoa
Billing and posting clerks	billing clerk	89.5	empleado/a de facturación	fakturazio-langilea
Phlebotomists	phlebotomist	89.3	flebotomista	flebotomista
Receptionists and information clerks	receptionist	89.3	recepcionista	harreragilea
Maids and housekeeping cleaners	housekeeper	89.0	limpiador/a	garbitzailea
Registered nurses	registered nurse	88.9	enfermero/a diplomado/a	erizain titularra
Bookkeeping, accounting, and auditing clerks	bookkeeper	88.5	contable	kontularia
Nursing, psychiatric, and home health aides	health aide	88.3	asistente sanitario/a	osasun laguntzailea

Tabla B.1: Profesiones tipificadas como femeninas

Profesión original	**Profesión abreviada (inglés)**	**% de mujeres**	**Traducción al español**	**Traducción al euskera**
Postal service mail sorters, processors, and processing machine operators	mail sorter	53.3	clasificador/a de correo	posta sailkatzailea
Order clerks	order clerk	53.3	encargado/a de pedidos	eskaera kudeatzailea
Dispatchers	dispatcher	53.1	operador/a	operadorea
Bartenders	bartender	53.1	camarero/a	tabernaria
Judges, magistrates, and other judicial workers	judge	52.5	juez/a	epailea
Training and development specialists	training specialist	52.5	entrenador	entrenatzailea
Statisticians	statistician	52.4	estadístico/a	estatistikaria
Medical scientists	medical scientist	51.8	científico/a médico/a	mediku zientzialaria
Insurance underwriters	insurance underwriter	51.1	asegurador/a	aseguru-bitartekaria
Insurance sales agents	insurance sales agent	50.6	agente de seguros	aseguruen salmenta agentea
Electrical, electronics, and electromechanical assemblers	electrical assembler	50.4	montador/a eléctrico/a	muntatzaile-elektrikoa
Mail clerks and mail machine operators, except postal service	mail clerk	49.8	empleado/a de correos	postako enplegatua
Advertising sales agents	sales agent	49.7	agente de ventas	salmenta agentea
Other healthcare practitioners and technical occupations	healthcare practitioner	49.5	profesional sanitario/a	osasun-langilea
Lodging managers	lodging manager	49.5	encargado/a de alojamiento	ostatu-arduraduna
Lifeguards and other recreational, and all other protective service workers	lifeguard	49.4	salvavidas	soroslea
Photographers	photographer	49.3	fotógrafo/a	argazkilaria
Crossing guards	crossing guard	48.6	guardia de tráfico	pasabide-zaindaria
Directors, religious activities and education	director of religious activities	48.6	director/a de actividades religiosas	erlijio-jardueren zuzendaria
Retail salespersons	salesperson	48.5	vendedor/a	saltzailea

Tabla B.2: Profesiones balanceadas que no son predominantes en ninguno de los géneros

Profesión original	Profesión abreviada (inglés)	% de mujeres	Traducción al español	Traducción al euskera
Firefighters	firefighter	3.3	bombero/a	suhiltzailea
Cement masons, concrete finishers, and terrazzo workers	mason	3.0	albañil	igeltseroa
Security and fire alarm systems installers	security system installer	2.9	instalador/a de sistemas de seguridad	segurtasun sistemako instalatzailea
Carpenters	carpenter	2.8	carpintero/a	arotza
Pipelayers, plumbers, pipefitters, and steamfitters	plumber	2.7	fontanero/a	iturgina
Railroad conductors and yardmasters	conductor	2.4	maquinista	tren-gidaria
Automotive body and related repairers	repairer	2.2	carrocero/a	karrozaegilea
Electricians	electrician	2.2	electricista	elektrizista
Mining machine operators	mining machine operator	2.0	operador/a de maquinaria minera	meatzaritzako makina-operadorea
Roofers	roofer	1.9	techador/a	teilatu-emailea
Carpet, floor, and tile installers and finishers	floor installer	1.9	instalador/a de suelos	zoru-instalatzailea
Logging workers	logging worker	1.8	trabajador/a maderero/a	egurketaria
Operating engineers and other construction equipment operators	operating engineer	1.7	operador de equipos de construcción	eraikuntza ekipoen operadorea
Electrical power-line installers and repairers	electrical installer	1.6	instalador/a eléctrico/a	instalatzaile-elektrikoa
Heating, air conditioning, and refrigeration mechanics and installers	heating mechanic	1.5	técnico/a de calefacción	berogailu -teknikaria
Heavy vehicle and mobile equipment service technicians and mechanics + Automotive service technicians and mechanics	service technician	1.5	mecánico de automóviles	auto mekanikaria
Bus and truck mechanics and diesel engine specialists	bus mechanic	1.5	mecánico/a autobuses	autobus mekanikaria
Miscellaneous vehicle and mobile equipment mechanics, installers, and repairers	mobile equipment mechanic	1.3	mecánico/a de equipos móviles	ekipamendu mugikorreko mekanikaria
Structural iron and steel workers	steel worker	0.9	trabajador/a del acero	altzairu-langilea
Drywall installers, ceiling tile installers, and tapers	taper	0.7	instalador/a de paneles de yeso	igeltsuzko panel instalatzailea

Tabla B.3: Profesiones tipificadas como masculinas

La tabla B.4, a continuación, muestra de forma detallada los diferentes sujetos incluidos en las frases tipo para cada uno de los tres idiomas, los cuales cubren ambos géneros.

Idioma	**Sujeto femenino**	**Sujeto masculino**
Inglés	she	he
	sister	brother
	daughter	son
	woman	man
	mother	father
	mom	dad
	wife	husband
	girlfriend	boyfriend
	aunt	uncle
Español	ella	él
	hermana	hermano
	hija	hijo
	mujer	hombre
	madre	padre
	mamá	papá
	mujer	marido
	novia	novio
	tía	tío
Euskera	bera	bera
	arreba	anaia
	alaba	semea
	emakumea	gizona
	ama	aita
	amatxo	aitatxo
	emaztea	senarra
	emaztegaia	senargaia
	izeba	osaba

Tabla B.4: Sujetos incluidos en las frases tipo del BEC-Pro

Bibliografía

[1] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. *Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods.* In North American Chapter of the Association for Computational Linguistics (NAACL‘18).

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. *Man Is to Computer Programmer As Woman Is to Homemaker? Debiasing Word Embeddings.* In Neural Information Processing Systems (NIPS‘16).

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.* Science, 356(6334):183–186.

[4] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes.* In Proceedings of the National Academy of Sciences, 115(16):E3635–E3644.

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. *Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints.* In Empirical Methods of Natural Language Processing (EMNLP‘17).

[6] Kate Crawford. 2017. *The Trouble With Bias.* Keynote at Neural Information Processing Systems (NIPS‘17).

[7] Marcelo O. R. Prates, Pedro H. Avelar, and Luis C. Lamb. 2018. *Assessing gender bias in machine translation: a case study with Google Translate.* Neural Computing and Applications.

[8] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, Anna Rohrbach. 2018. *Women Also Snowboard: Overcoming Bias in Captioning Models.* European Conference on Computer Vision (EECV'18).

[9] Rachael Tatman and Conner Kasten. 2017. *Effects of Talker Dialect, Gender Race on Accuracy of Bing Speech and YouTube Automatic Captions.* 934-938. 10.21437/Interspeech.2017-1746.

[10] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. *Reducing Gender Bias in Abusive Language Detection.* In Empirical Methods of Natural Language Processing (EMNLP'18).

[11] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. *Gender Bias in Neural Natural Language Processing.* arXiv:1807.11714v2.

[12] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. *On Measuring Social Biases in Sentence Encoders.* In North American Chapter of the Association for Computational Linguistics (NAACL'19).

[13] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. *Deep Contextualized Word Representations.* In North American Chapter of the Association for Computational Linguistics (NAACL'18).

[14] Hila Gonen and Yoav Goldberg. 2019. *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them.* In North American Chapter of the Association for Computational Linguistics (NAACL'19).

[15] Svetlana Kiritchenko and Saif M Mohammad. 2018. *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems.* In 7th Joint Conference on Lexical and Computational Semantics (SEM'18).

[16] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2017. *Measuring and Mitigating Unintended Bias in Text Classification.* In Association for the Advancement of Artificial Intelligence (AAAI'17).

[17] Rishabh Bhardwaj, Navonil Majumder, Soujanya Poria. 2020. *Investigating Gender Bias in BERT.* arXiv:2009.05021.

[18] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. *Gender Bias in Coreference Resolution.* In North American Chapter of the Association for Computational Linguistics (NAACL'18).

[19] Vid Kocijan, Oana-Maria Camburu, Thomas Lukasiewicz. 2020. *The Gap on GAP: Tackling the Problem of Differing Data Distributions in Bias-Measuring Datasets.* arXiv:2011.01837

[20] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. *End-to-End Neural Coreference Resolution.* In Empirical Methods of Natural Language Processing (EMNLP'17).

[21] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. *Higher-Order Coreference Resolution with Coarse to-Fine Inference.* In Empirical Methods of Natural Language Processing (EMNLP'18).

[22] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. *Gender Bias in Contextualized Word Embeddings.* arXiv:1904.03310.

[23] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. *Getting Gender Right in Neural Machine Translation.* In Empirical Methods of Natural Language Processing (EMNLP'18).

[24] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai Wei Chang. 2018b. *Learning Gender-Neutral Word Embeddings.* In Empirical Methods of Natural Language Processing (EMNLP'18).

[25] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. *Situation Recognition: Visual Semantic Role Labeling for Image Understanding.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE '16), pages 5534–5542.

[26] Alexander M Rush and Michael Collins. 2012. *A Tutorial on Dual Decomposition and Lagrangian Relaxation for Inference in Natural Language Processing.* Journal of Artificial Intelligence Research, 45:305– 362.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. *Generative Adversarial Networks.* In Advances in Neural Information Processing Systems (NIPS‘14).

[28] Marion Bartl, Malvina Nissim and Albert Gatt. 2020. *Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias.* In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing (2020), pages 1–16.

[29] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. *Measuring bias in contextualized word representations.* In Proceedings of the First Workshop on Gender Bias in Natural Language Processing (2019), pages 166–172.

[30] Ben Schmidt. 2015. *Rejecting the Gender Binary: A Vector-Space Operation.* https://bit.ly/1OhXJM0.