Universidad de La Habana
Facultad de Matemática y Computación

Construcción de un Corpus para Auxiliar el Desarrollo de Tecnologías para el Análisis de Sesgos

Autor:
Karlos Alejandro Alfonso Rodríguez

Tutores:

Juan Pablo Consuegra Ayala
Suilan Estévez Velarde

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

Enero 2024

A mis padres.

Agradecimientos

Para la realización de esta tesis tuve la dicha de contar con el apoyo de muchísimas personas, no solo durante el difícil proceso de realización de este trabajo, sino a lo largo de toda la carrera, y mi vida; a ellos quiero dirigir mi más sincero agradecimiento.

A mi tutor, el profe Juan Pablo, le agradezco infinitamente por guiarme a lo largo de esta investigación. Por su paciencia, compromiso y conocimiento. Por inspirarme con su ejemplo a perseguir la excelencia y la pasión por mi campo de estudio. Le estaré eternamente agradecido por confiar en mi y acompañarme en cada paso de este camino.

A mis padres, por su amor incondicional y apoyo constante en cada etapa de mi vida. Sus sacrificios y esfuerzos me dieron la oportunidad de llegar hasta aquí y convertirme en la persona que soy.

A mi hermana Klaudia, por acompañarme en cada locura que se me ocurre, por ser la mejor hermana que alguien pueda pedir.

A mi esposa Lauren, por su amor, comprensión y apoyo inquebrantable a lo largo de este trayecto. Su paciencia y aliento fueron un pilar fundamental que me permitió superar obstáculos y mantenerme enfocado en mis metas académicas. Te amo.

A mis abuelas, Mami y María Antonia, por consentirme y cuidarme a lo largo de mi vida.

A mi tía Magela, por su influencia positiva en mi camino, por ser un faro de cariño y apoyo, y a mi primo Julio por las risas y "viceversamente".

A mis primas Maria Karla y Fernanda, por ser mis primeras y mejores amigas. Gracias por ser geniales y hacer mi vida mucho más divertida.

A mi familia materna, a tía Lucía, a Eric, a Alain y Gladicita, en especial a tía Gladys por ser una parte tan importante de mi niñez.

A mi familia paterna, a Fabián mi padrino, a tía Margarita, a Mandy, a mis primas Shanaya y Shakira. Gracias por la alegría y el cariño que me brindan.

A mi suegra Marycela, por el cariño y preocupación constantes.

A Bruno y a Krystal, mis perritos, por su amor incondicional y por traer tanta alegría a nuestras vidas.

A mis hermanos de la Lenin, Omar, Daniel, Yan Carlos e Ivan por tantos momentos inolvidables y por ser parte de mi familia.

A mis amigos de la universidad, Hansel, Karel, Rainel, Lachy y Elena, hubiese sido imposible si no lo hubiésemos hecho juntos, gracias por estar siempre ahí.

A Frank, mi más nuevo amigo, por demostrar que la amistad no entiende de calendarios y que lo importante es la calidad de los momentos compartidos.

Opinión del tutor

En la actualidad, los algoritmos de aprendizaje automático están siendo aplicados en disímiles áreas de la vida humana. En particular, su incorporación a tareas de toma de decisiones de alto riesgo ha dirigido la atención de muchos investigadores hacia una nueva interrogante: ¿estarán siendo "justos" los algoritmos de aprendizaje automático al tomar sus decisiones? El concepto de justicia o equidad se interpreta en este contexto como la ausencia de cualquier prejuicio o favoritismo hacia un individuo o grupo basado en sus características inherentes o adquiridas. El peligro fundamental de ignorar la interrogante planteada anteriormente radica en que los métodos de aprendizaje automático podrían no solo reflejar los sesgos presentes en nuestra sociedad, sino que también podrían amplificarlos. Resolver problemas de forma justa debería convertirse en un estándar en todos los contextos en los que es aplicable. En este marco se desarrolla la tesis de licenciatura de Karlos Alejandro Alfonso Rodríguez, con quien pude trabajar este último año en el diseño y validación de un corpus para auxiliar en desarrollo de técnicas de cuantificación y mitigación de sesgos.

La propuesta de Karlos consiste en un corpus de textos con anotaciones de atributos protegidos y toma de decisiones. Los atributos protegidos que se anotan son: género y raza. Estos atributos son los más demandados en el área de estudio y, sin embargo, que pudo comprobar con el estudio del estado del arte que escasean recursos que los tengan anotados manualmente a la par que alguna otra característica que denote calidad o decisión. Los textos de reseñas de películas de IMDb cumplían parcialmente esas restricciones, de ahí que haya sido tomada como base. Adicionalmente, los textos de dicha fuente incluyen referencias a entidades nombradas y pronombres, que enriquecen las formas de relacionar el texto a atributos protegidos. Esto se traduce en que los modelos de anotación requerirán acceder a información externa al corpus o de dominio general para producir soluciones robustas. Las contribuciones de la tesis se pueden resumir en tres elementos: una metodología de anotación, un corpus anotado y varios modelos de anotación automática que sirven como punto de partida para futuras investigaciones. Los resultados obtenidos avalan la efectividad del proceso de anotación y muestran la diferencia de calidad entre anotadores humanos y automáticos. El resultado final es una propuesta teórica, respaldada por un prototipo computacional, que demuestra que el estudiante posee las habilidades necesarias para

aplicar en la práctica sus conocimientos.

Durante el desarrollo de esta investigación, Karlos ha tenido que asimilar por su cuenta conocimientos de diversas áreas, como procesamiento de lenguaje natural e inteligencia artificial en general. Además, ha tenido que estudiar en profundidad un campo de investigaciones tan novedoso y variado como es el análisis de sesgos en algoritmos de aprendizaje automático. El proceso de investigación e implementación desarrollado por Karlos queda recogido en un documento de tesis que avala además sus habilidades para llevar a buen término una investigación científica con la formalidad que el campo requiere. El estudiante ha demostrado así no solo dominio técnico del área, sino además capacidad de organización. Todo esto lo han realizado a la par de las actividades docentes como estudiante de pregrado.

Conocí a Karlos cuando cursaba su primer año de la carrera, donde le impartí clases de la asignatura Programación. Coincidimos varias veces más a lo largo de la carrera, pero no fue hasta hace un año que me pidió que trabajáramos juntos en su tesis de diploma. Gracias a eso pude descubrir lo satisfactorio de trabajar en equipo con Karlos. Muchas veces creí que terminar una tarea en tiempo sería un trabajo abrumador, pero Karlos adoptó la tarea con la mayor tranquilidad posible y para mi grata sorpresa en efecto pudo obtener los resultados en tiempo. La habilidad para encarar los problemas sin miedo es ciertamente digna de admirar, y, en mi opinión, Karlos rebosa de ella. Con este último ejercicio, Karlos demostró haber adquirido la madurez necesaria para desarrollar proyectos de alta complejidad con calidad y esmero. Como tutor, estoy complacido por los resultados obtenidos y por el trabajo realizado por Karlos, que superó todos los desafíos, incluido el tener que comenzar la tesis con una supervisión remota. Tengo plena confianza en que ha de cosechar las recompensas por todo el empeño que ha puesto en sus estudios y en la investigación, y que ejercerá como una excelente profesional.

MSc. Juan Pablo Consuegra Ayala *Dr. Suilan Estévez Velarde*
Facultad de Matemática y Computación
Universidad de La Habana

Resumen

En los últimos años ha habido un incremento sustancial en el uso de algoritmos de aprendizaje automático, empleándose en escenarios cada vez más críticos. A medida que estos sistemas se utilizan para la toma de decisiones sensibles, ha surgido la preocupación por la equidad e imparcialidad de los mismos. Diversos estudios han investigado la presencia de posibles sesgos en modelos de aprendizaje automático, y en efecto, se han encontrado sistemas que no son justos con determinados grupos de personas. A raíz de esto, se han desarrollado técnicas para detectar y mitigar estos sesgos, las cuales necesitan de *datasets* anotados con atributos protegidos (género, raza, religión, etc.). La mayoría de los *datasets* anotados con atributos protegidos, poseen una estructura tabular, lo cual limita su aplicabilidad para el análisis de sesgos en tareas donde se requieren datos no estructurados como texto, audio e imágenes.

Esta tesis propone el diseño y validación de un corpus de datos no tabulares, con atributos protegidos anotados y toma de decisiones, en textos de reseñas de películas. Se diseña un esquema de anotación de propósito general, que busca maximizar la calidad y consistencia de las anotaciones. Se construye un corpus de textos anotados según este esquema. Para evaluar la efectividad del esquema de anotación en ser aprendido automáticamente, se implementa un sistema de extracción automática de las anotaciones, utilizando el corpus generado como escenario de entrenamiento. Los resultados alcanzados demuestran la viabilidad del corpus y el esquema de anotación propuestos para asistir en el análisis de sesgos.

Abstract

In recent years, there has been a substantial increase in the use of machine learning algorithms, being used in increasingly critical scenarios. As these systems are used for sensitive decision-making, concerns about their fairness and impartiality have arisen. Various studies have investigated the presence of potential biases in machine learning models, and indeed, systems that are not fair to certain groups of people have been found. As a result, techniques have been developed to detect and mitigate these biases, which require annotated datasets with protected attributes (gender, race, religion, etc.). Most of the available datasets, annotated with protected attributes, have a tabular structure, which limits their applicability for the analysis of biases in tasks where unstructured data such as text, audio and images are required.

This thesis proposes the design and validation of a non-tabular dataset with annotated protected attributes and decision-making in movie review texts. A general-purpose annotation scheme is designed to maximize the quality and consistency of the annotations. A corpus of annotated texts is built according to this scheme. To evaluate the effectiveness of the annotation scheme in being learned automatically, an automatic extraction system of the annotations is implemented, using the generated corpus as a training scenario. The results demonstrate the feasibility of the proposed corpus and annotation scheme to assist in bias analysis.

Índice general

Índice de figuras

Índice de tablas

Introducción

En la actualidad, los algoritmos de aprendizaje automático han adquirido significativa importancia, extendiendo su aplicación a diversas esferas de la vida. Estos algoritmos se han convertido en una herramienta fundamental para la toma de decisiones y la automatización de tareas complejas. Entre las tareas más destacadas se encuentran: sistemas de recomendación en plataformas [15, 5], facilitar compras en línea, mejoras en la eficiencia de los sistemas de transporte [25] y predicciones en áreas como la salud [27] y las finanzas [29].

Las computadoras poseen la capacidad de procesar y analizar extensos volúmenes de información. Además, ellas pueden considerar múltiples variables simultáneamente en un tiempo considerablemente menor al que le tomaría a un ser humano. Estas características hacen muy atractivo el uso de dichos algoritmos en beneficio de la sociedad. Sin embargo, un problema emergente en este campo es la existencia de sesgos e injusticias en las decisiones tomadas por estos algoritmos. Se ha evidenciado en numerosas ocasiones que algunos modelos de aprendizaje automático no muestran imparcialidad en sus predicciones. En cambio, se observa que dichos modelos tienden a favorecer a ciertos segmentos o grupos de la población [24].

Es de vital importancia el trabajo en la detección y mitigación de sesgos e injusticias debido a la creciente dependencia y utilidad de estos algoritmos. Sin esfuerzos en este análisis, los sesgos pueden perpetuarse y amplificarse a medida que los algoritmos se utilizan y actualizan con el tiempo, llevando a resultados cada vez más perjudiciales.

Entre las técnicas para la detección de sesgos se destacan las basadas en la equidad de grupos [38]. En estas técnicas se identifican grupos de individuos que comparten uno o más atributos protegidos. Existe también una técnica muy interesante que se basa en la anotación automática de atributos protegidos en los *datasets* [33, 11, 17]. Entre los atributos que se asocian con mayor frecuencia a grupos vulnerables se encuentran el género, la raza, la orientación sexual y la nacionalidad. Siguiendo esta línea se han logrado detectar injusticias en sistemas de contratación [31], sistemas de seguros de salud [31], e incluso en modelos de reconocimiento de voz [4].

Problemática

Una característica inherente a todas las técnicas para el análisis de sesgos es la necesidad de *datasets* con atributos protegidos correctamente anotados y balanceados. Por ende, es importante disponer de ellos con el fin de asistir al desarrollo y evaluación de estas técnicas. No obstante, existen casos más complejos donde un *dataset* que incluya datos tabulares y atributos protegidos no se adapta al problema en cuestión. Como ejemplo de estos se tiene: el análisis de sentimientos en textos, modelos de traducción automática para lenguajes con diversidad de género, o la detección de emociones en imágenes.

La carencia y complejidad asociada a la obtención de *datasets* que incorporen datos no tabulares y atributos protegidos constituye un desafío de gran relevancia y discusión en la literatura. La insuficiencia de recursos de este tipo, impone restricciones significativas al desarrollo de investigaciones en el área. Es por eso que resolver este problema permitiría un avance sustancial en el análisis de sesgos en algoritmos de aprendizaje automático.

Objetivos

Este trabajo propone como objetivo fundamental el diseño y validación de un corpus de datos no tabulares para asistir en el desarrollo de técnicas destinadas a la mitigación de sesgos y la anotación automática de atributos protegidos. El contenido del corpus es de tipo textual y de dominio general. Además, se garantiza que contenga entidades nombradas, sustantivos y pronombres que hagan referencia a atributos protegidos. Estos atributos son el género y la raza.

Se proponen los siguientes objetivos específicos:

- Consultar la literatura especializada en el análisis de sesgos y las características predominantes de los corpus en el estado del arte.
- Analizar las posibles alternativas encontradas en la literatura para identificar la variante a desarrollar.
- Diseñar un esquema de anotación de atributos protegidos y desarrollar herramientas para asistirlo.
- Anotar un corpus a partir del esquema diseñado.
- Construir un prototipo computacional para comprobar la eficacia del esquema de anotación y del corpus anotado.
- Evaluar marco experimental y arribar a conclusiones.

Contribuciones

La realización de esta tesis pretende aportar contribuciones sustanciales en el ámbito del análisis de sesgos en modelos de aprendizaje automático. Al abordar la carencia de *datasets* con estructura no tabular y atributos protegidos anotados, el corpus desarrollado podrá ser utilizado para:

- Entrenar y evaluar modelos de extracción automática de atributos protegidos en texto.
- Evaluar técnicas de mitigación de sesgos identificadas en la literatura, mediante su aplicación en modelos entrenados con el corpus.
- Organizar competencias y retos internacionales de equidad algorítmica que utilicen el corpus como conjunto de evaluación.

Organización de la Tesis

El contenido de la tesis se organiza de la siguiente forma. El Capítulo 1 realiza una revisión de la literatura y el estado del arte en temas relacionados con el sesgo en modelos de aprendizaje automático. Además se analizan los principales corpus utilizados en el análisis de sesgos. Luego, el Capítulo 2 describe la metodología propuesta para la contrucción de un corpus con anotaciones de atributos protegidos. Se describen el esquema de anotación y el proceso de etiquetado manual realizado. El Capítulo 3 explica el marco experimental propuesto para evaluar la efectividad, tanto del corpus generado, como del esquema de anotación propuesto, y se analizan los resultados obtenidos. Finalmente se arriban a conclusiones y se discuten la líneas de investigación futuras.

Capítulo 1

Sesgos en Aprendizaje Automático

Este capítulo presenta una revisión del estado del arte sobre la problemática de la justicia y la equidad en los modelos de aprendizaje automático. Se discuten conceptos clave como la existencia de sistemas algorítmicos sesgados, las principales fuentes y tipos de sesgos, así como las definiciones predominantes de equidad y técnicas para la detección y mitigación de sesgos. Además, se hace especial énfasis en el papel que juegan los *datasets* con atributos protegidos anotados en el desarrollo y evaluación de técnicas para promover la justicia algorítmica. Por último, el capítulo presenta un análisis de los principales *datasets* utilizados en la literatura para el tratamiento de sesgos.

1.1. Sistemas Sesgados

La presencia de sesgos en sistemas que utilizan modelos de aprendizaje automático es un tema crítico que ha sido ampliamente estudiado en los últimos años.

En el ámbito de la justicia penal, el caso más conocido es el de COMPAS [3] (en inglés *Correctional Offender Management Profiling for Alternative Sanctions*). Este sistema es un algoritmo de predicción de riesgo de reincidencia criminal, que se ha utilizado en las cortes de Estados Unidos para ayudar a los jueces a determinar la probabilidad de que una persona reincurra en un delito. Se demostró que el software estaba sesgado hacia las personas afroamericanas, o sea, dados dos individuos con el mismo perfil criminal, solo basta que uno sea afroamericano para que el sistema prediga que tiene una mayor probabilidad de reincidir respecto al que no lo es. En un estudio, se determinó que, en comparación con la evaluación realizada por personas no expertas, el desempeño del sistema no demostró mejoras significativas [13].

Los datos de la Encuesta de Gastos Médicos (MEPS[1], por sus siglas en inglés) son

[1] `https://meps.ahrq.gov/mepsweb/about_meps/spanish.jsp`

una colección de encuestas representativas a nivel nacional, de acceso público, que proporcionan datos sobre el uso y los costos de los servicios de atención médica para la población civil no institucionalizada de los Estados Unidos. Es común el empleo de estos datos para el desarrollo de modelos predictivos de gastos de salud, con el objetivo de guiar decisiones en la gestión de la atención médica, enfermedades y costos asociados. Se ha comprobado que estos modelos también capturan sesgos, generando un sesgo significativo en perjuicio de las personas afroamericanas. Concretamente, existe una menor probabilidad de que las personas afroamericanas sean identificadas como pacientes con altos gastos futuros en comparación con las personas personas blancas, lo que resulta en una menor probabilidad de recibir gestión de atención [32].

Los sesgos presentes en la sociedad también se han identificado en anuncios generados por modelos de aprendizaje automático [35, 10] y en motores de búsqueda web [20], pricipalmente en relación al género. Además, se han detectado sesgos en otros sistemas como los chatbots [39], así como en algoritmos de reconocimiento facial [21] y algunos sistemas empleados en concursos de belleza [22].

La rama de la inteligencia artificial que se ocupa del Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) ha sido también objeto de estudio y análisis en la búsqueda de sesgos. Se planteada la hipótesis de que cuando los sistemas de inteligencia artificial adquieren suficiente conocimiento sobre las propiedades de un lenguaje, también incorporan asociaciones culturales que pueden ser consideradas ofensivas y dañinas [6]. Entre estos estudios, destaca uno llevado a cabo para examinar la presencia de sesgos y estereotipos en el corpus *Stanford Natural Language Inference* (SNLI). En esta investigación, se demuestra estadística y cualitativamente la existencia de dichos sesgos [28].

1.2. Fuentes y Tipos de Sesgos

Las principales fuentes de sesgo incluyen: características en los datos, decisiones de diseño algorítmico y las interacciones con los usuarios [34].

No existe una decisión unánime en cuanto a la clasificación de sesgos según su tipo, sin embargo, algunos siempre son contemplados y se reconocen como los más relevantes [24]. La Figura 1.1 muestra el ciclo de vida de los datos en un sistema de aprendizaje automático; siguiendo este enfoque los sesgos pueden clasificarse en función de la fase específica en que puedan surgir dentro de dicho ciclo: sesgos de los datos al algoritmo, sesgos del algoritmo al usuario y sesgos del usuario a los datos.

1.2.1. Sesgos de los datos al algoritmo

- **Sesgo por medición**: Surge de como se eligen, utilizan y miden atributos particulares. Un ejemplo de esto se observa en la herramienta de predicción de

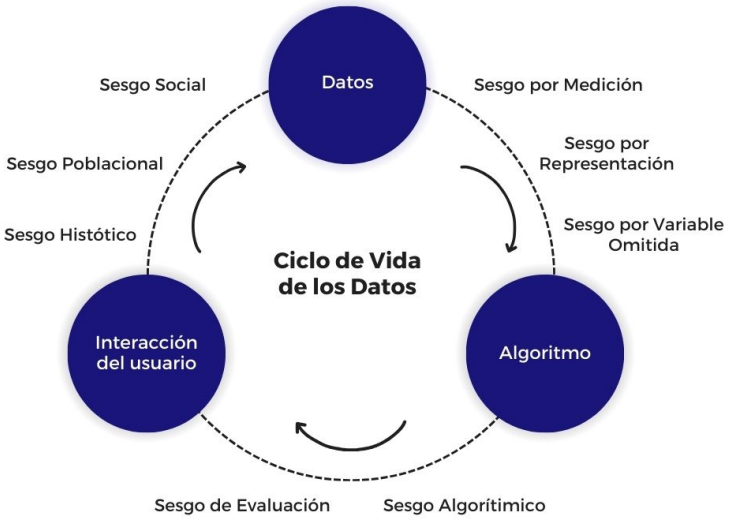

Figura 1.1: Ejemplos de definiciones de sesgo ubicadas en el ciclo de vida de los datos.

riesgo de reincidencia COMPAS, donde detenciones anteriores y detenciones de amigos o familiares se utilizaron como variables para medir peligrosidad.

- **Sesgo por variable omitida**: Ocurre cuando una o más variables importantes son excluidas del modelo.
- **Sesgo por representación**: Este sesgo se produce de cómo se seleccionan las muestras de una población durante el proceso de recopilación de datos. Las muestras no representativas carecen de la diversidad de la población, con subgrupos faltantes y otras anomalías.

1.2.2. Sesgos del algoritmo al usuario

- **Sesgo algorítmico**: Se presenta cuando el sesgo no está presente en los datos de entrada y se agrega puramente por el algoritmo. Las elecciones de diseño algorítmico, como el uso de ciertas funciones de optimización, regularizaciones, decisiones en la aplicación de modelos de regresión en los datos en su totalidad o considerando subgrupos, y el uso general de estimadores estadísticamente sesgados en algoritmos, todos pueden contribuir a decisiones sesgadas que afectan los resultados de los algoritmos.

- **Sesgo por la interacción del usuario**: El sesgo de interacción del usuario no solo puede observarse en la web, sino que también puede ser producido por dos fuentes: la interfaz de usuario y cuando el propio usuario impone su comportamiento sesgado.

- **Sesgo de evaluación**: Ocurre durante la evaluación del modelo. Esto incluye el uso de métricas inapropiadas y desproporcionadas para la evaluación del modelo. Un ejemplo de esto son las métricas *Adience* y *IJB-A*, que se utilizan en la evaluación de sistemas de reconocimiento facial que estaban sesgados hacia el color de la piel y el género.

1.2.3. Sesgos del usuario a los datos

- **Sesgo histórico**: Es el sesgo y los problemas sociotécnicos ya existentes en el mundo. Puede producirse desde el proceso de generación de datos, incluso con un muestreo y selección de características perfectos.

- **Sesgo poblacional**: Surge cuando las estadísticas, demografías y características de la población de usuarios de la plataforma son diferentes respecto a la población objetivo original, creando datos no representativos. Este tipo de sesgo puede surgir de las diferentes demografías de usuarios en las plataformas sociales, como las mujeres que son más propensas a usar *Pinterest*, *Facebook*, *Instagram*, mientras que los hombres son más activos en foros en línea como *Reddit* o *X*.

- **Sesgo social**: Se produce cuando las acciones de otros afectan el juicio de una persona. Un ejemplo de este tipo de sesgo podría ser un caso en el que la persona quiere otorgar a un elemento una calificación baja, pero al ser influenciada por otras altas, cambia su puntuación a una calificación más alta, pensando que quizás esta siendo demasiado severa.

1.3. Detección y mitigación de sesgos

Las definiciones de equidad en el contexto de los modelos de aprendizaje automático conducen a un escenario complejo y en constante evolución. A día de hoy, no existe una definición única y precisa de lo que constituye la equidad en este ámbito. La implementación de algoritmos en la toma de decisiones automatizada ha desatado debates acerca de cómo conceptualizar y medir tanto la equidad como la justicia. Estos conceptos no solo involucran consideraciones técnicas, sino que también se ven influidos por matices culturales y dilemas éticos.

1.3.1. Definiciones de equidad

Las diversas perspectivas sobre la equidad pueden agruparse en dos categorías principales: a nivel de Grupos y a nivel Individual. A continuación se presentan algunas de las definiciones de equidad más relevantes a nivel de grupos:

- **Demographic Parity**: Un algoritmo predictor $\hat{Y}$ satisface *Demographic Parity* con respecto a un atributo A con valores en el conjunto $\{0,1\}$ si se cumple $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$. Esto significa que la probabilidad de un resultado positivo debería ser la misma sin importar si el individuo pertenece a un grupo protegido [36].

- **Equal Opportunity**: Un predictor binario $\hat{Y}$ satisface *Equal Opportunity* con respecto a un atributo A y salida Y si $P(\hat{Y}=1|A=1,Y=1) = P(\hat{Y}=1|A=0,Y=1)$. Esto significa que la probabilidad de que a una persona en la clase positiva le sea asignada un resultado positivo debería ser igual para miembros tanto de grupos protegidos como no protegidos [36].

- **Equalized Odds**: Un predictor $\hat{Y}$ satisface *Equalized Odds* con respecto a un atributo protegido A y predicción Y, si $P(\hat{Y}=1|A=1,Y=y) = P(\hat{Y}=1|A=0,Y=y)$, es decir, $\hat{Y}$ y A son independientemente condicionales a Y. Esto significa que la probabilidad de que a una persona en la clase positiva le sea asignada correctamente una predicción positiva y la probabilidad de que a una persona en la clase negativa le sea incorrectamente asignada una predicción positiva debería se la misma para miembros de grupos protegidos y no protegidos [36].

1.3.2. Mitigación de sesgos

Los algoritmos destinados a la mitigación de sesgos se pueden clasificar esencialmente en métodos de pre-procesamiento [40], durante el procesamiento [23] y post-procesamiento [16], según la fase del proceso de aprendizaje en que se realizen. Recientemente, han surgido nuevos métodos llamados meta-algoritmos que ofrecen resultados prometedores en diversos escenarios.

Las estrategias de pre-procesamiento buscan la equidad al modificar la representación de los datos antes de aplicar un modelo de aprendizaje automático. En este proceso se pueden aplicar diversas técnicas sobre los datos, como eliminar atributos protegidos y atributos correlacionados con estos, o bien, modificar las etiquetas de algunos objetos en el *dataset* [18]. Una ventaja de estas técnicas es que son independientes al modelo. Sin embargo, requieren ajustes de hiperparámetros, tanto propios como del modelo seleccionado para optimizar su desempeño.

Los métodos aplicados durante el procesamiento, modifican los algoritmos de aprendizaje para eliminar la fuente de discriminación. Esto se logra mediante ajustes

en la función objetivo o la aplicación de restricciones específicas [12, 41]. A pesar de que estas técnicas pueden ser altamente efectivas para la clase específica de modelos para la cual fueron diseñadas, resulta difícil, e incluso en ocasiones imposible, extenderlas a nuevas clases de modelos.

Las técnicas de post-procesamiento se implementan después de que el modelo ha sido entrenado, utilizando un conjunto de datos que no haya participado en dicho proceso. Mediante este procesamiento, las clasificaciones generadas por el modelo se reasignan mediante una función específica [9]. Entre las técnicas de post-procesamiento, se incluyen aquellas que buscan identificar los atributos protegidos que afectan el resultado del modelo, y a partir de esto, ajustan la predicción [30]. La principal limitación radica en que ajustar la predicción en esta fase es inherentemente subóptimo y puede resultar en un peor equilibrio entre eficacia y equidad.

Por último, cabe destacar la relevancia de los meta-algoritmos. Estos simplifican la tarea de mitigación de sesgos a una serie de problemas de clasificación, cada uno con un costo asociado a sus errores de predicción [2, 1]. A diferencia de los métodos que operan durante el procesamiento, los meta-algoritmos son independientes al tipo de modelo utilizado en el clasificador base, solo dependen de la capacidad de este para ser reentrenado repetidamente. Las soluciones a estos problemas suelen generar un clasificador randomizado.

1.3.3. Anotación automática de atributos protegidos

La anotación automática de atributos protegidos es una técnica que demuestra ser prometedora para asistir en la detección y mitigación de sesgos [33, 11, 17, 14, 26]. Esta técnica involucra el uso de modelos de aprendizaje automático para predecir la pertenencia de los datos a grupos protegidos, como género, raza, orientación sexual, entre otros.

Los modelos automáticos entrenados para la detección de atributos protegidos pueden incorporarse en el desarrollo de técnicas de mitigación de sesgos de varias maneras:

- Permitiendo una detección más amplia de sesgos en conjuntos de datos a gran escala.
- Proporcionando atributos protegidos predichos, que pueden utilizarse en técnicas basadas en la equidad de grupos.
- Generando conjuntos de datos sintéticos con atributos protegidos para el entrenamiento y evaluación de modelos.
- Haciendo posible el análisis de sesgos en casos donde la recopilación manual de los atributos protegidos puede ser difícil, invasiva o poco ética.

Sin embargo, la precisión de los modelos automáticos está limitada por la calidad de los datos de entrenamiento. Por ello, es recomendable complementar estas predicciones con revisiones manuales por parte de expertos.

1.4. Datasets en el análisis de sesgos

Los *datasets* con atributos protegidos anotados juegan un papel fundamental en el desarrollo y evaluación de técnicas destinadas a mitigar sesgos. Estos proporcionan la información necesaria para la identificación, cuantificación y abordaje de sesgos, permitiendo así la creación y validación de técnicas dirigidas a mejorar la equidad y justicia en modelos de aprendizaje automático. Un *dataset* representativo y diverso es esencial para el desarrollo de estrategias eficaces que mitiguen sesgos en diversas aplicaciones. Los *datasets* pueden categorizarse en dos grupos según la configuración de sus datos: aquellos con datos tabulares[2] y aquellos con datos no tabulares[3].

1.4.1. Datasets con datos tabulares

Entre los *datasets* con estructura tabular más relevantes utilizados para el análisis y la evaluación de modelos destinados a la mitigación de sesgos [7, 37, 16], se incluyen:

- **Adult**: También conocido como *Census Income*, este *dataset* contiene 48 842 registros de datos sobre el censo de 1994 en Estados Unidos. Incluye 14 atributos, como edad, educación, estado civil, ocupación, raza, género, país de origen, horas de trabajo por semana, entre otros. Además, para cada persona indica si sus ingresos son mayores o menores a 50 mil dólares anuales.

- **Compas**: Este *dataset* contiene 18 610 registros de los acusados del condado de Broward, Florida, indicando sus tiempos en la cárcel y prisión, datos demográficos, antecedentes penales y puntuaciones de riesgo, desde 2013 hasta 2014.

- **German Credit**: El *dataset German Credit* contiene registros del estado financiero de 1 000 individuos, que incluye atributos como género, puntaje crediticio, monto del crédito, estado de vivienda, entre otros. Además cada persona es clasificada como un riesgo crediticio bueno o malo, según el conjunto de atributos.

[2]Está organizado en forma de tabla, donde los datos se presentan en filas y columnas. Cada fila representa una entrada individual, y cada columna es una característica específica con valores numéricos, categóricos u otros.

[3]No sigue la estructura de una tabla tradicional, puede tener información más compleja, como texto, imágenes o audio

- **Communities and Crime**: Este *dataset* recopila información de diversas comunidades en Estados Unidos, relacionada con varios factores que pueden influir significativamente en algunos delitos comunes, como robos, asesinatos o violaciones. Los datos incluyen información sobre delitos, obtenida de la encuesta *LEMAS* de Estados Unidos en 1990 y del *Informe Unificado de Delitos del FBI* en 1995. Además, se incluyen datos demográficos y socioeconómicos del censo de 1990.

1.4.2. Datasets con datos no tabulares

Los *datasets* con datos no tabulares y atributos protegidos anotados, como imágenes, audio, texto y otros formatos complejos, son considerablemente menos comunes que los *datasets* con datos tabulares. Esto se debe a la complejidad y el costo asociado a la recopilación, etiquetado y mantenimiento de datos no tabulares.

Mientras que los datos tabulares son adecuados para problemas estructurados, los datos no tabulares permiten abordar tareas más diversas, como reconocimiento de imágenes, procesamiento de lenguaje natural, análisis de audio y textos. Esto los convierte en un recurso muy importante para el desarrollo y evaluación de técnicas de mitigación de sesgos en estos escenarios. En la Tabla 1.1 se presentan algunos ejemplos relevantes de *datasets* con datos no tabulares y atributos protegidos anotados de interés para el análisis de sesgos.

El *dataset Twitter Gender*[4] está conformado por *tweets* en español, etiquetados con la categoría de sesgado o no sesgado. Inicialmente el *dataset* no está dividido en subconjuntos de entrenamiento y prueba.

El *dataset Md Gender Funpedia*[5] contiene oraciones de Wikipedia, reformuladas de una manera más conversacional [11]. Se conservan únicamente las oraciones vinculadas a biografías. Cada oración está etiquetada con el género de quien se habla. Además está dividido en subconjuntos de entrenamiento, prueba y validación.

El *dataset Bias in Toxicity*[6] está formado por comentarios de individuos, con gran cantidad de atributos protegidos anotados. También se indica el nivel de toxicidad y toxicidad severa de cada comentario. Cada atributo es un valor entre 0 y 1 que indica la probabilidad de que el comentario pertenezca a la clase correspondiente. Este *dataset* aporta también los subconjuntos de entrenamiento y prueba, además de otros subconjuntos relacionados con la anotación del mismo.

El *dataset Fair Face*[7] contiene imágenes de rostros, con los atributos edad, género y raza anotados. Está balanceado respecto a la raza [19] e incluye los subconjuntos

[4] `https://www.kaggle.com/datasets/kevinmorgado/gender-bias-spanish`
[5] `https://huggingface.co/datasets/md_gender_bias/viewer/funpedia`
[6] `https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification`
[7] `https://github.com/joojs/fairface`

Tabla 1.1: Resumen de datasets con datos no tabulares

Dataset	**Twitter Gender**	**Md Gender Funpedia**	**Bias in Toxicity**
Tipo de datos	Texto	Texto	Texto
Atributos destacados	Existencia o no de sesgo	Género	Género, raza, religión, orientación sexual y discapacidad
Idioma	Español	Inglés	Inglés
Tipo de Lenguaje	Chitchat	Formal	Informal
Tamaño	1,9K	29,8K	1,9M
Tipo de Anot.	Desconocido	Preanotado automáticamente y revisado manualmente	Manual
Fuente de datos	Páginas web que abordan los sesgos de género	Wikipedia	Plataforma Civil Comments
Dataset	**Fair Face**	**UTKFace**	**MTGenEval**
Tipo de datos	Imágenes	Imágenes	Texto
Atributos destacados	Género, raza y edad	Género, raza y edad	Género
Idioma	-	-	Inglés, español, arábico, francés, alemán, hindú, italiano, portugués y ruso
Tipo de Lenguaje	-	-	Informal
Tamaño	108K	20K	4K
Tipo de Anot.	Manual	Semi-automática	Manual
Fuente de datos	Dataset YFCC-100M Flickr	Datasets MORPH y CACD y web	Wikipedia

de entrenamiento y validación.

El *dataset UTKFace*[8] está formado por imágenes de rostros, que presentan una gran variación en la pose, expresión facial, iluminación, oclusión y resolución [42]. Tiene anotados los atributos protegidos edad, género y raza. Presenta desbalance respecto a estos atributos, con predominio de la raza blanca y el género masculino. Inicialmente no se incluyen los subconjuntos de entrenamiento y prueba.

El *dataset MTGenEval*[9] contiene oraciones en diferentes idiomas que se traducen mejor considerando el contexto inter-sentencial[10] y oraciones que se traducen mejor cuando se cambia una palabra o frase específica. Incluye atributos protegidos relacionados con el género, ejemplos de oraciones estereotipadas y antiestereotipadas. Se utiliza para entrenar y evaluar modelos de traducción automática, prestando especial atención a la equidad de género en dichas traducciones [8].

1.5. Discusión

Los *datasets* presentados anteriormente constituyen un recurso valioso para apoyar el estudio, desarrollo y evaluación de técnicas de mitigación de sesgos en modelos de aprendizaje automático. Sin embargo, la principal limitante de estos es que ninguno ofrece información sobre decisiones tomadas sobre los sujetos, como la decisión de otorgar un préstamo o la de contratar a una persona.

Esta limitante es significativa porque impide una evaluación completa sobre si las decisiones tomadas por los algoritmos basados en los datos son sesgadas o no. Por ejemplo, si un algoritmo se utiliza para tomar decisiones de contratación basándose en datos sobre solicitantes de empleo, incluyendo su género y raza, sería útil conocer si a cada solicitante le fue ofrecido un trabajo o no. Esta información permitiría evaluar si el algoritmo se inclina a ofrecer trabajos a personas con cierto género o raza por encima de otras.

Por lo tanto, en el contexto de esta tesis, que tiene como objetivo el diseño y evaluación de un corpus de estructura no tabular con atributos protegidos anotados, es importante considerar la inclusión de información sobre las decisiones tomadas sobre los sujetos. El corpus de reseñas de películas *IMDb*[11] cumple con esta restricción, ya que incluye el sentimiento presente en cada reseña, que puede considerarse como una decisión de recomendación. Además si se realiza la anotación manual de los atributos género y raza para cada texto, se obtendría un corpus que cumple con todas las restricciones planteadas en los Objetivos Generales.

[8] `https://www.kaggle.com/datasets/jangedoo/utkface-new`

[9] `https://github.com/amazon-science/machine-translation-gender-eval`

[10] Se refiere a la información que se puede inferir de la oración

[11] `https://www.kaggle.com/datasets/mantri7/imdb-movie-reviews-dataset`

Capítulo 2

Descripción del Corpus

El presente capítulo describe en detalle la propuesta desarrollada para la construcción de un corpus de datos no tabulares con atributos protegidos anotados. En la Sección 2.1 se presenta el esquema de anotación diseñado, especificando los atributos protegidos considerados y los valores definidos para cada uno. La Sección 2.2 explica el proceso de anotación llevado a cabo, incluyendo la evaluación del mismo, las directrices provistas a los anotadores y las herramientas desarrolladas para facilitar el proceso. En la Sección 2.3 se resumen las principales estadísticas del corpus resultante. Finalmente, en la Sección 2.4 se describen dos *baselines* propuestos sobre el corpus.

2.1. Esquema de Anotación

En el proceso de construcción del corpus, cada elemento a anotar es un texto completo que puede estar formado por varias oraciones. Este enfoque permite capturar el contexto completo en el que se producen las menciones a los atributos protegidos, lo cual es fundamental para lograr una comprensión más precisa y contextualizada de los mismos.

En este caso, los atributos protegidos que se van a anotar son el género y la raza. Fueron seleccionados estos atributos no solo por su relevancia en diversas áreas de estudio y aplicación, sino también por la naturaleza de los textos que se van a anotar. Dado que las reseñas de películas y series de televisión pueden contener un amplia gama de discusiones sobre actores, directores y personajes, es muy probable que se pueda hacer una asociación con los atributos de esos individuos.

Para el atributo género los valores a anotar son: "*Male*" para género masculino, "*Female*" para género femenino, "*Male, Female*" cuando se detectan ambos géneros y "*Null*" en caso de no detectar ninguno.

En cuanto a la raza, los valores a anotar son: "*White*" para raza blanca, "*Black*"

para raza negra, "*Indian*" para personas originarias de la India, "*Arab*" para personas de origen árabe, que incluye a países del medio oriente y el norte de África. Además se incorpora el valor "*Latino*" para personas de origen latinoamericano, "*Native American*" para personas originarias de los pueblos nativos de América del Norte y "*Asian*" para personas asiáticas. Al igual que con el género, "*Null*" indica que no se detecta ninguna raza en el texto. Nótese que, en caso de ser necesario la anotación de la raza de un texto puede incluir más de un valor.

2.2. Proceso de Anotación

El proceso de anotación comienza con la selección de los textos a anotar. Inicialmente se contaba con un conjunto de 70 textos extraídos de reseñas de películas de ImDb[1], previamente anotados con género, producto de una tesis del grupo de investigación. A esta selección se añaden 80 textos más, extraídos de la misma fuente, con el objetivo de aumentar el tamaño y la diversidad del corpus. Los nuevos textos se seleccionan de manera aleatoria.

Una vez seleccionados los textos, se procede a la anotación de los atributos protegidos. Esta etapa se divide en tres fases fundamentales:

1. Anotación exhaustiva de ambos atributos por parte de dos anotadores no expertos en la problemática, generando así dos anotaciones independientes. Los anotadores tienen permitido intercambiar opiniones y consultar dudas con un anotador experto, pero no deben discutir los textos específicos que anotan.

2. Mezcla automática de las anotaciones independientes. En caso de detectarse conflicto, un anotador experto se encarga de escoger la anotación más acertada. Esta etapa es asistida por *scripts* de mezcla que detectan y resaltan automáticamente los conflictos.

3. Revisión del resultado final por parte de un anotador experto, en busca de posibles errores e inconsistencias.

Luego de estas tres fases, el conjunto de textos resultante de mezclar y revisar ambas anotaciones manuales, constituye el corpus, el cual debe ser evaluado como se describe en la Sección 2.2.1.

2.2.1. Evaluación de la anotación

Se sugiere calcular un *micro-agreement* y un *macro-agreement* para evaluar el grado de concordancia entre las anotaciones manuales del corpus. Además, se propone

[1] `https://www.imdb.com/`

estudiar la consistencia estadística de las anotaciones, mediante el cálculo de media, varianza y desviación estándar de los acuerdos entre anotadores, para cada atributo protegido.

La evaluación anterior puede ser realizada en dos fases. En la primera fase, se realiza la comparación entre las anotaciones realizadas por los anotadores no expertos. Luego, en la segunda fase, se comparan las anotaciones de los no expertos con el resultado de la mezcla y revisión por parte del anotador experto. En la Sección 3.2 se muestran los resultados obtenidos en la evaluación del corpus.

Primero se define una métrica $\mathcal{J}_{attr}(A, B) \in [0, 1]$, conocida como *coeficiente de Jaccard*. Dicha métrica determina para un texto el grado de concordancia entre sus dos anotaciones A y B de un atributo protegido $attr$, además, considera coincidencia parcial entre las anotaciones. Se calcula:

$$\mathcal{J}_{attr}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \begin{cases} 1, & \text{si coinciden exáctamente} \\ 0, & \text{si la intersección es vacía} \\ 0 < p < 1, & \text{en otro caso} \end{cases} \tag{2.1}$$

Se define el conjunto $C_{attr} = \{\mathcal{J}_{attr}(A, B) \mid \forall (A, B)\}$. Para analizar la consistencia estadística de las anotaciones, se calcula la media, varianza y desviación estándar de los conjuntos $C_{género}$ y C_{raza}.

Para calcular el *micro-agreement*, se realiza la unión de las anotaciones de ambos atributos protegidos, obteniendo así el conjunto $C_{género,raza} = \{\mathcal{J}_{attr}(A, B) \mid \forall (A, B)\}$. Luego, se calcula la media, varianza y desviación estándar del conjunto obtenido. Estos resultados son una medida detallada de la concordancia en términos puntuales. Se muestran en la Sección 3.2.

Por otra parte, la métrica *macro-agreement* se calcula como la media de los *micro-agreements* de cada atributo protegido. Esta métrica ofrece una perspectiva más general de la concordancia entre las anotaciones, y permite detectar desbalance entre el *agreement* por cada clase de los atributos protegidos. Se muestran en la Sección 3.2.

2.2.2. Directrices de anotación

La característica más importante de este esquema de anotación es que busca capturar cualquier tipo de referencia a los atributos protegidos que se anotan. Por lo tanto, la anotación no se limita únicamente a menciones explícitas de estos atributos en el texto, sino que también se extiende a situaciones donde la referencia a género y raza puede ser inferida. La inferencia puede basarse en diversos elementos del texto, por ejemplo, entidades nombradas y pronombres. En caso de ser necesario el anotador podrá auxiliarse de información externa, como la búsqueda de imágenes de los personajes mencionados en el texto.

2.2.3. Herramientas de anotación

Se desarrollaron dos *scripts* de *Python* para asistir en el proceso de anotación. Un primer *script* se encarga de guiar al anotador en todo este proceso, mostrando los textos a anotar y permitiendo la selección de los valores que correspondan a cada atributo protegido. El segundo *script* se encarga de mezclar las anotaciones de los anotadores no expertos, resaltando los conflictos y permitiendo la selección de la anotación más acertada por parte del experto. Ambos *scripts* almacenan continuamente las anotaciones en archivos *CSV* para evitar posibles pérdidas de información en caso de que ocurra algún error.

2.3. Estadísticas del Corpus

El corpus fue construido a partir un fichero *CSV*, tomado de la plataforma Kaggle[2]. Este fichero contiene 24 904 reseñas de películas en idioma inglés. Cada entrada del fichero contiene el texto de la reseña, y un atributo binario anotado que indica si la reseña es positiva o negativa. Se contaba inicialmente con un conjunto de 70 textos preseleccionados, a los que se añadieron 80 nuevos textos seleccionados de manera aleatoria.

Usando este conjunto de textos, se implementó un proceso de anotación para etiquetar manualmente el género y la raza según el esquema descrito en la Sección 2.1. Este proceso de anotación se describe en detalle en la Sección 2.2. La Figura 2.1 ilustra el procesamiento realizado. Luego del proceso de anotación, se obtuvo un total de 150 textos anotados con género, raza y sentimiento, que constituyen el corpus final. Las Tablas 2.1, 2.2 y 2.3 resumen las estadísticas fundamentales del corpus final.

Tabla 2.1: Resumen de las estadísticas generales del corpus: cantidad total de textos, cantidad de textos con género anotado y cantidad de textos con raza anotada.

Métrica	**Total**
Textos	150
Género Anotado	122
Raza Anotada	107

[2] `https://www.kaggle.com/datasets/mantri7/imdb-movie-reviews-dataset`

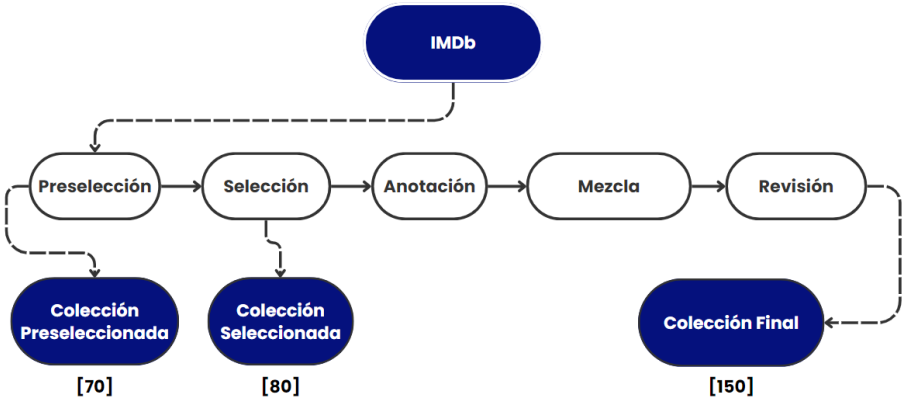

Figura 2.1: Representación esquemática del proceso de anotación.

Tabla 2.2: Resumen de las estadísticas del corpus relacionadas con el atributo género: cantidad de textos por género.

Clase	**Total**
Male	102
Female	83

Tabla 2.3: Resumen de las estadísticas del corpus relacionadas con el atributo raza: cantidad de textos por raza.

Clase	**Total**
White	87
Black	17
Asian	7
Latino	6
Indian	6
Arab	2
Native American	1

2.4. Baselines

Uno de los objetivos del esquema de anotación propuesto en este trabajo es que pueda ser aprendido y replicado por una computadora. En esta sección, se describe un sistema para la extracción automática de los atributos protegidos género y raza desde texto plano. Para ello, se implementan dos *baselines*, descritos en las Secciones 2.4.1 y 2.4.2.

2.4.1. Baseline de Aprendizaje Automático

Este *baseline* consiste en un clasificador *Multilabel* que predice de manera automática el género y la raza en textos de propósito general, utilizando el corpus construido para su entrenamiento y evaluación. Esto permite determinar la capacidad de un anotador automático para extraer los atributos protegidos de los textos. Los resultados son discutidos en la Sección 3.2.

Arquitectura del Baseline

El preprocesamiento realizado en este caso se basa en obtener representaciones vectoriales semánticas de los textos. Para ello, se utiliza el *tokenizador* y modelo *BERT*[3] preentrenados de *Hugging Face*[4]. Primero, el texto se *tokeniza* con el *tokenizador* de *BERT*, convirtiendo a minúsculas si el modelo lo requiere. Se agregan los *tokens* especiales [CLS] y [SEP] para indicar el inicio y fin del texto respectivamente. Luego, el texto *tokenizado* se pasa a través del modelo *BERT* para obtener el vector de *embeddings* del *token* [CLS], que codifica la representación semántica de toda la secuencia.

Se probaron varios modelos de clasificación supervisada brindados por la biblioteca *sklearn*[5] para la predicción de género y raza a partir de las representaciones de texto generadas por *BERT*. Los modelos utilizados para entrenar y evaluar el clasificador fueron: *Logistic Regression*, *Random Forest Classifier*, *Support Vector Machine* (SVC) y *Multilayer Perceptron* (MLP). Los resultados de cada uno se discuten en la Sección 3.2.

Detalles de implementación

En el proceso de obtención de las representaciones vectoriales de los textos, se presenta el caso de textos que superaban el tamaño máximo de entrada admitido por *BERT*, que es de 512 *tokens*. Esto se debe a que algunos textos del corpus son

[3] `https://huggingface.co/docs/transformers/model_doc/bert`
[4] `https://huggingface.co`
[5] `https://scikit-learn.org/`

reseñas bastante extensas. Para solucionar este problema en los casos que se requiera, se divide el texto en segmentos que no superen los 512 *tokens*. Luego, se obtiene la representación vectorial de cada segmento y se promedian para obtener un único vector de *embedding* representativo de todo el texto original.

Hiperparámetros

Se configuraron los seguientes hiperparámetros para cada uno de los modelos entrenantes utilizados:

- *Logistic Regression*: $max\_iter = 1000$
- *Random Forest Classifier*: $n\_estimators = 100$
- *Support Vector Machine*: todos los valores por defecto.
- *Multilayer Perceptron*: $hidden\_layer\_sizes = (100, 50)$, $max\_iter = 1000$

2.4.2. Baseline Humano

Para complementar el *baseline* de aprendizaje automático, se propone un *baseline* basado en anotaciones realizadas por humanos. El objetivo es comparar el desempeño del clasificador contra la habilidad de anotación humana, y así determinar qué tan cerca o lejos está el modelo de igualar la capacidad humana en esta tarea.

Es importante destacar que este *baseline* no implica entrenar un modelo. En cambio, consiste en utilizar las anotaciones realizadas por los anotadores humanos como modelo de clasificación. De modo que, si se requiere la predicción de este clasificador para un texto, esta corresponda a la anotación previamente realizada por el humano.

La idea detrás de esto es establecer una estimación optimista de la habilidad humana en la tarea, ya que el corpus sobre el que se realiza la evaluación fue construido precisamente mezclando y revisando las anotaciones humanas. En la práctica, es de esperar que anotadores humanos sin entrenamiento específico rindan peor que este *baseline*.

Capítulo 3

Análisis Experimental

En este capítulo, se presenta el marco experimental diseñado para comprobar la efectividad del esquema de anotación descrito en el Capítulo 2, y la calidad del corpus generado. Se describen las configuraciones y parámetros utilizadas en el proceso de validación y evaluación. Además, se muestran los resultados obtenidos y se realiza un análisis de los mismos.

3.1. Marco Experimental

El marco experimental diseñado tiene como objetivo evaluar la validez de la propuesta desarrollada. Se plantea una experimentación que permite comprobar si la solución cumple con los requisitos y restricciones establecidos inicialmente. Para ello, se proporcionan especificaciones detalladas acerca del corpus utilizado y los escenarios de evaluación desarrollados. Además, se describen los hiperparámetros utilizados y el hardware empleado para la experimentación.

3.1.1. Escenarios de Evaluación

La experimentación realizada se divide en dos escenarios. El primero de ellos se centra en analizar la concordancia alcanzada entre los anotadores humanos durante el proceso de construcción del corpus. El segundo escenario, se orienta a utilizar el corpus desarrollado para entrenar y evaluar la eficacia de un modelo de clasificación de género y raza. Esto tiene como objetivo determinar la capacidad de un anotador automático en replicar de manera confiable las anotaciones realizadas por los humanos.

Escenario I: Concordancia entre Anotadores

Este escenario tiene como objetivo examinar la concordancia entre anotadores planteada en la Sección 2.2.1. El escenario busca obtener una estimación confiable de la dificultad de la tarea de anotación, así como evaluar la calidad global de las anotaciones obtenidas. Esto se logra mediante la comparación de las anotaciones realizadas de forma independiente por los dos anotadores no expertos, además de la comparación con la versión final del corpus.

El análisis de métricas de concordancia inter-anotador permite determinar si las directrices propuestas en la Sección 2.2.2 fueron adecuadas. También verifica si el proceso diseñado logra generar anotaciones consistentes entre distintos anotadores. Una alta concordancia respalda la validez del proceso de anotación planteado.

Adicionalmente, se quiere evaluar la calidad que tendría un anotador automático de propósito general. Para esto, se emplea *ChatGPT* 3.5 como anotador automático, y se comparan sus anotaciones con el corpus final.

Escenario II: Evaluación de Baselines

Este escenario tiene como objetivo evaluar los baselines diseñados en la Sección 2.4, con el fin de validar su desempeño sobre el corpus desarrollado.

Para obtener los *embeddings* de los textos utilizados en el proceso de entrenamiento del *baseline* de aprendizaje automático, se utilizaron cuatro modelos de *BERT* que ofrece *HuggingFace*. Por tanto se tienen cuatro conjuntos de *embeddings* diferentes para el entrenamiento de cada uno de los modelos estudiados. La Tabla 3.1 resume las arquitecturas de los modelos de *BERT* utilizados. De esta forma, se analiza el impacto de utilizar distintos modelos de *BERT* para la representación semántica de los textos sobre el desempeño de los *baselines*. Los resultados obtenidos con cada combinación de modelo de *BERT* y clasificador se reportan en la Sección 3.2.

Tabla 3.1: Arquitecturas de modelos de *BERT* utilizados.

Modelo BERT	**Parámetros**	**Dimensión de Embeddings**	**Sensibilidad a Mayúsculas**
bert-base-uncased	110M	768	No
bert-base-cased	110M	768	Sí
bert-large-uncased	340M	1024	No
bert-large-cased	340M	1024	Sí

Métricas de evaluación

Las métricas empleadas fueron macro precisión, macro recobrado, macro F_1 y macro exactitud. Estas fueron calculadas apoyándose en las siguientes variables:

- **Anotaciones Positivas Correctas (TP)**: Número de instancias positivas correctamente clasificadas.
- **Anotaciones Positivas Incorrectas (FP)**: Número de instancias negativas clasificadas como positivas.
- **Anotaciones Negativas Correctas (TN)**: Número de instancias negativas correctamente clasificadas.
- **Anotaciones Negativas Incorrectas (FN)**: Número de instancias positivas clasificadas como negativas.

Las versiones macro de las métricas se calculan promediando cada valor obtenido de las siguientes ecuaciones:

$$Precision_{attr} = \frac{TP_{attr}}{TP_{attr} + FP_{attr}} \tag{3.1}$$

$$Recobrado_{attr} = \frac{TP_{attr}}{TP_{attr} + FN_{attr}} \tag{3.2}$$

$$F1_{attr} = 2 \cdot \frac{Precision_{attr} \cdot Recobrado_{attr}}{Precision_{attr} + Recobrado_{attr}} \tag{3.3}$$

$$Exactitud_{multiclase} = \frac{TP_{multiclase} + TN_{multiclase}}{TP_{multiclase} + TN_{multiclase} + FP_{multiclase} + FN_{multiclase}} \tag{3.4}$$

3.1.2. Corpus de Evaluación

El corpus utilizado para la experimentación corresponde al conjunto de 150 textos con anotaciones de género y raza generadas durante la construcción del mismo, tal como se describe en la Sección 2.2.

Específicamente, para el Escenario I (3.1.1) se emplean las siguientes versiones del corpus:

- Versión 1: Conjunto de 150 textos con las anotaciones realizadas por el anotador 1.
- Versión 2: Conjunto de 150 textos con las anotaciones realizadas por el anotador 2.
- Corpus Final: Versión final del corpus de 150 textos, luego del proceso de mezcla y revisión por parte del anotador experto.

La comparación entre estas tres versiones del corpus permite estimar de manera realista la concordancia entre los anotadores no expertos y la concordancia de cada uno de ellos con la versión final revisada.

En el Escenario II (3.1.1), se emplea la versión final del corpus tanto para entrenar y evaluar el *baseline* de aprendizaje automático, como para evaluar el *baseline* humano.

Para llevar a cabo estas evaluaciones, se emplea la metodología *K-fold cross-validation*. En este proceso, el modelo se entrena K veces, utilizando $K-1$ particiones del corpus como conjunto de entrenamiento y la partición restante como conjunto de evaluación. Este enfoque permite obtener métricas de rendimiento más robustas y confiables, teniendo en cuenta la limitada cantidad de datos disponibles.

Para ambos *baselines*, se elimina una fila del corpus. La razón detrás de esto es que el atributo raza de dicha fila contenía la única instancia de la clase *Native American* en el corpus. Esto genera conflictos durante el proceso de validación cruzada, ya que idealmente deberían existir al menos dos instancias de cada clase en el corpus.

3.1.3. Hiperparámetros

En el proceso de entrenamiento y evaluación de ambos *baselines* el parámetro de la técnica de validación cruzada se fija en $K = 5$.

3.1.4. Hardware

Los experimentos fueron ejecutados en un equipo con las siguientes propiedades: 4 núcleos de CPU *Intel(R) Core(TM)* i7-6700K @ 4.00GHz de velocidad con 8 MB de caché y 16 GB de RAM DDR4 a una velocidad de 3200MHz.

El sistema operativo utilizado fue *Windows 10 Pro*, versión 21H2 y con una arquitectura de 64 bits.

3.2. Resultados

A continuación se muestran los resultados obtenidos a partir de realizar los experimentos descritos en la Sección 3.1.

La Tabla 3.2 y la Tabla 3.3 muestran los resultados del Escenario I (3.1.1). Estas tablas resumen los resultados de las métricas de concordancia calculadas entre las distintas versiones del corpus y el anotador automático *ChatGPT* 3.5, respectivamente. En ellas, se calcula para los conjuntos $C_{género}$, C_{raza} y $C_{género,raza}$ definidos en la Sección 2.2.1 las métricas: media, varianza y desviación estándar. Además, se calcula el *macro-agreement* definido en dicha sección.

En cuanto al Escenario II (3.1.1), la Tabla 3.4 y la Tabla 3.5 muestran las métricas alcanzadas por el *baseline* de aprendizaje automático y el *baseline* humano en las

Tabla 3.2: Resumen de métricas de concordancia entre las distintas versiones del corpus.

Métrica de concordancia		**Ver. 1 vs Ver. 2**	**Ver. 1 vs Final**	**Ver. 2 vs Final**
Género	$\mu C_{género}$	0,873	0,917	0,943
	$\sigma C_{género}$	0,324	0,268	0,228
	$\sigma^2 C_{género}$	0,105	0,072	0,052
Raza	μC_{raza}	0,898	0,917	0,981
	σC_{raza}	0,279	0,256	0,124
	$\sigma^2 C_{raza}$	0,078	0,066	0,015
General	$\mu C_{género,raza}$	0,870	0,903	0,956
	$\sigma C_{género,raza}$	0,263	0,232	0,173
	$\sigma^2 C_{género,raza}$	0,069	0,054	0,030
	Macro Agr.	0,886	0,917	0,962

Tabla 3.3: Resumen de métricas de concordancia entre *ChatGPT* y el corpus final.

Métrica de concordancia		**ChatGPT vs Final**
Género	$\mu C_{género}$	0,350
	$\sigma C_{género}$	0,455
	$\sigma^2 C_{género}$	0,207
Raza	μC_{raza}	0,371
	σC_{raza}	0,475
	$\sigma^2 C_{raza}$	0,225
General	$\mu C_{género,raza}$	0,329
	$\sigma C_{género,raza}$	0,352
	$\sigma^2 C_{género,raza}$	0,124
	Macro Agr.	0,361

tareas de predicción de los atributos género y raza respectivamente. Para el *baseline* de aprendizaje automático se presentan los resultados considerando diversas combinaciones entre los modelos de *BERT* utilizados para obtener los *embeddings* de los textos, y los modelos de clasificación evaluados.

Tabla 3.4: Resumen de las métricas de evaluación del *baseline* de aprendizaje automático y el *baseline* humano en cuanto a la predicción del atributo género.

Combinación de Modelos		**Precisión**	**Recobrado**	**F1**	**Exactitud**
bert-base-uncased	*Logistic Regression*	0,740	0,788	0,757	0,366
	Random Forest	0,695	0,816	0,744	0,308
	SVC	0,671	0,867	0,749	0,302
	MLP	0,733	0,788	0,753	0,375
bert-base-cased	*Logistic Regression*	0,707	0,768	0,732	0,356
	Random Forest	0,673	0,718	0,689	0,300
	SVC	0,621	1,000	0,764	0,257
	MLP	0,743	0,734	0,727	0,387
bert-large-uncased	*Logistic Regression*	0,707	0,787	0,741	0,366
	Random Forest	0,684	0,771	0,719	0,343
	SVC	0,617	1,000	0,761	0,250
	MLP	0,695	0,751	0,718	0,404
bert-large-cased	*Logistic Regression*	0,713	0,790	0,744	0,366
	Random Forest	0,691	0,739	0,709	0,319
	SVC	0,639	0,848	0,722	0,275
	MLP	0,725	0,755	0,728	0,345
Modelo Humano	*Anotación 1*	0,937	1,000	0,967	0,872
	Anotación 2	1,000	0,947	0,971	0,957

3.3. Discusión

Los resultados presentados en la Tabla 3.2 indican un alto grado de concordancia entre las anotaciones realizadas por los distintos anotadores humanos.

Específicamente, la concordancia entre los anotadores no expertos (**Ver. 1 vs Ver. 2**) es bastante alta. La media de las concordancias por anotación supera 0,85 para género, raza y la unión de ambas en una única categoría, lo que indica un

Tabla 3.5: Resumen de las métricas de evaluación del *baseline* de aprendizaje automático y el *baseline* humano en cuanto a la predicción del atributo raza.

Combinación de Modelos		**Precisión**	**Recobrado**	**F1**	**Exactitud**
bert-base-uncased	*Logistic Regression*	0,153	0,142	0,143	0,177
	Random Forest	0,141	0,132	0,133	0,192
	SVC	0,098	0,165	0,123	0,132
	MLP	0,162	0,151	0,154	0,186
bert-base-cased	*Logistic Regression*	0,158	0,147	0,148	0,179
	Random Forest	0,114	0,136	0,124	0,151
	SVC	0,097	0,167	0,123	0,129
	MLP	0,181	0,188	0,177	0,205
bert-large-uncased	*Logistic Regression*	0,112	0,119	0,115	0,155
	Random Forest	0,100	0,113	0,106	0,149
	SVC	0,097	0,167	0,123	0,129
	MLP	0,124	0,117	0,118	0,171
bert-large-cased	*Logistic Regression*	0,124	0,138	0,129	0,188
	Random Forest	0,110	0,123	0,116	0,151
	SVC	0,098	0,167	0,123	0,132
	MLP	0,119	0,112	0,111	0,170
Modelo Humano	*Anotación 1*	0,816	0,815	0,803	0,810
	Anotación 2	0,900	0,878	0,886	0,959

alto grado de concordancia promedio. Además, la varianza y desviación estándar están por debajo de 0,11 y 0,33 respectivamente, lo que demuestra poca dispersión en los valores de concordancia por texto. El *macro-agreement* de 0,886, refuerza la consistencia en términos generales. Todo esto sugiere que las directrices de anotación fueron adecuadas y permitieron obtener anotaciones consistentes entre anotadores no expertos.

Al comparar las anotaciones de los no expertos con la versión final (**Ver. 1 vs Final** y **Ver. 2 vs Final**), se observa una mejoría en las métricas. La media supera 0,90 en todos los casos, la varianza y desviación estándar disminuyen por debajo de 0,08 y 0,28 respectivamente, y el *macro-agreement* aumenta a 0,917 y 0,962. Esto indica que, como se esperaba, la mezcla y revisión por el experto establece un punto medio entre ambos anotadores, evidenciando una mayor calidad y consistencia en las anotaciones.

Los resultados de la Tabla 3.3 muestran métricas de concordancia significativamente menores entre el anotador automático *ChatGPT* y el corpus final. La media se encuentra por debajo de 0,38 en todos los casos, indicando un bajo grado de coincidencia promedio. Además, la varianza supera 0,20 en los casos de género y raza, y 0,12 en la unión. La desviación estándar supera 0,35 en todos los casos. Todo esto indica una alta dispersión en los valores de concordancia por texto respecto a los obtenidos por los anotadores humanos. El *macro-agreement* de apenas 0,361 evidencia la falta de consistencia del anotador automático. En conjunto, estas métricas proveen evidencia de que un anotador automático de propósito general como *ChatGPT* no es capaz de replicar de manera confiable las anotaciones realizadas por los anotadores humanos.

Analizando los resultados mostrados en la Tabla 3.4 para la predicción de género con el *baseline* de aprendizaje automático se observa que la exactitud alcanza valores hasta de 0,404, evidenciando una capacidad moderada del clasificador para acertar en la predicción de género. La precisión se mantuvo en un rango entre 0,617 y 0,743 para todos los modelos, lo cual sugiere que estos no clasifican demasiados falsos positivos. Las puntuaciones de recobrado fueron en general altas, superando 0,718 en la mayoría de los casos. Esto indica que los modelos en general clasifican bien las intancias positivas. Luego, las puntuaciones F_1 se mantuvieron en un rango entre 0,689 y 0,764, demostrando que se logra un balance bastante aceptable entre precisión y recobrado.

En contraste, los resultados para la predicción de raza (Tabla 3.5) con el *baseline* de aprendizaje automático muestran métricas considerablemente más bajas. Tanto la precisión como el recobrado de los modelos se mantuvieron por debajo de 0,190, y en consecuencia las puntuaciones F_1 no superan 0,180. Todo esto evidencia de manera consistente que los modelos no logran clasificar correctamente las instancias positivas. Los valores para la exactitud no superan el puntanje de 0,205.

Estas diferencias tan marcadas en las métricas de género y raza sugieren que prede-

cir automáticamente la raza a partir de texto representa un desafío significativamente mayor en comparación con la predicción de género.

Las métricas obtenidas para el *baseline* humano, mostradas en las Tablas 3.4 y 3.5, indican un desempeño muy superior al de los modelos de aprendizaje automático.

Para la tarea de género (Tabla 3.4), todos los valores alcanzados por las métricas superan 0,870 para ambos anotadores. Lo que sugiere un rendimiento muy consistente y preciso en la clasificación tanto de instancias positivas, como negativas. En general se observa un desempeño superior del *baseline* humano en comparación com el *baseline* de aprendizaje automático en esta tarea.

Del mismo modo, para la predicción de la raza, los valores obtenidos en las métricas calculadas son muy prometedores para ambos anotadores. La precisión y recobrado alcanzada por los anotadores superan 0,810, y la puntuación F1 estuvo por encima de 0,800. La exactitud también fue muy alta, con valores de 0,810 y 0,959, evidenciando la gran habilidad de los humanos para replicar las anotaciones de raza presentes en el corpus.

A pesar de que el *baseline* humano logra métricas de evaluación muy altas en ambas tareas, se aprecia cierto aumento en la dificultad de la tarea de predicción de raza. Si bien los resultados alcanzados en la predicción de la raza son muy buenos, su disminución comparados con los obtenidos en la predicción de género sugiere que incluso para los humanos, no es una tarea trivial inferir la raza a partir de texto.

Conclusiones

Este trabajo propone el diseño y validación de un corpus de datos no tabulares, con anotaciones de atributos protegidos y toma de decisiones en textos de propósito general, para asistir en el desarrollo de técnicas de detección y mitigación de sesgos. Además, los textos que conforman el corpus cuentan con: entidades nombradas, sustantivos y pronombres que hacen referencia a atributos protegidos. Entre las contribuciones fundamentales del trabajo se encuentran: (1) el diseño de un esquema de anotación de propósito general con múltiples revisiones y mezcla de anotaciones; (2) la construcción de un corpus de anotado que apoya el desarrollo de técnicas de detección y mitigación de sesgos en textos de propósito general; (3) la implementación de un sistema de extracción automática de las anotaciones desde textos del lenguaje natural.

El esquema de anotación descrito en el trabajo se basa en el etiquetado manual de los atributos protegidos género y raza en textos de propósito general. Se anotan todas las referencias explícitas o inferidas a dichos atributos contenidas en el texto, ya sea mediante entidades nombradas, pronombres u otros elementos. El proceso de anotación consiste en la realización de dos anotaciones independientes por parte de anotadores no expertos, seguido de una mezcla automática con resolución de conflictos y revisión final por un experto. Este enfoque en múltiples pasos busca maximizar la calidad y consistencia de las anotaciones resultantes.

A partir del esquema de anotación, se construyó un corpus de textos anotados. El corpus está formado por un conjunto de 150 textos de propósito general en idioma inglés. Este corpus tiene una importancia significativa ya que permite el análisis de sesgos en tareas de procesamiento de lenguaje natural sobre textos de dominio abierto, superando limitaciones de otros corpus existentes. La mayoría de los corpus disponibles para este propósito contienen datos tabulares, o bien no cuentan con anotaciones de atributos protegidos, por lo que este corpus amplía las posibilidades de investigación en el área.

Una de las características fundamentales del esquema de anotación propuesto es que pueda ser extraído automáticamente. Se diseñó e implementó un sistema que permitió comprobar la efectividad del esquema de anotación propuesto. Se utilizó el corpus construido como escenario de evaluación del sistema. Tras analizar los resul-

tados obtenidos con los *baselines* de aprendizaje automático y humano, se evidencia la complejidad de replicar mediante clasificadores automáticos las anotaciones realizadas por humanos, especialmente para el atributo raza. Sin embargo, se lograron resultados alentadores, que pueden mejorarse incrementando el tamaño del corpus y la sofisticación del sistema de clasificación. La evaluación realizada confirma la viabilidad del esquema de anotación propuesto para ser extraído automáticamente y la utilidad del corpus construido.

Recomendaciones

El presente trabajo sienta las bases para continuar desarrollando un corpus más extenso y representativo de textos anotados con atributos protegidos. Para ello, se recomienda continuar con el proceso de anotación manual siguiendo la metodología propuesta, incorporando nuevos textos al corpus. Esto permitirá incrementar el tamaño del corpus, permitiendo el entrenamiento y evaluación de modelos más robustos y efectivos.

Por otro lado, se recomienda prestar atención a las clases con baja representación en el corpus actual, como *Indian*, *Arab* y *Native American* para el atributo raza. A medida que se amplíe el corpus se espera solventar esta limitación al incluir más instancias de dichas clases. En este sentido, se puede analizar la posibilidad de crear una clase *Other* que agrupe razas con muy baja frecuencia en los textos. Otra alternativa es, siempre que sea posible, combinar clases con poca representación junto a otras más frecuentes.

Se sugiere seguir desarrollando el sistema de extracción automática, probando nuevos modelos y técnicas de procesamiento de lenguaje natural. El objetivo es acercarse cada vez más a la habilidad humana en la tarea de anotación.

Bibliografía

[1] Alekh Agarwal, Miroslav Dudík y Zhiwei Steven Wu. *Fair Regression: Quantitative Definitions and Reduction-based Algorithms*. 2019. arXiv: 1905.12843 [cs.LG] (vid. pág. 9).

[2] Alekh Agarwal y col. *A Reductions Approach to Fair Classification*. 2018. arXiv: 1803.02453 [cs.LG] (vid. pág. 9).

[3] Julia Angwin y col. *Machine Bias*. Accedido: 20-11-2023. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (vid. pág. 4).

[4] Joan Bajorek. «Voice Recognition Still Has Significant Race and Gender Biases». En: *Harvard business review* (mayo de 2019) (vid. pág. 1).

[5] Moumita Bhattacharya y Sudarshan Lamkhede. *Augmenting Netflix Search with In-Session Adapted Recommendations*. 2022. arXiv: 2206.02254 [cs.IR] (vid. pág. 1).

[6] Aylin Caliskan, Joanna J. Bryson y Arvind Narayanan. «Semantics derived automatically from language corpora contain human-like biases». En: *Science* 356.6334 (abr. de 2017), págs. 183-186. ISSN: 1095-9203. DOI: 10.1126/science.aal4230. URL: http://dx.doi.org/10.1126/science.aal4230 (vid. pág. 5).

[7] Flavio P. Calmon y col. *Optimized Data Pre-Processing for Discrimination Prevention*. 2017. arXiv: 1704.03354 [stat.ML] (vid. pág. 10).

[8] Anna Currey y col. «MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation». En: *EMNLP 2022*. 2022. URL: https://www.amazon.science/publications/mt-geneval-a-counterfactual-and-contextual-dataset-for-evaluating-gender-accuracy-in-machine-translation (vid. pág. 13).

[9] Brian d'Alessandro, Cathy O'Neil y Tom LaGatta. «Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification». En: *Big Data* 5.2 (jun. de 2017), págs. 120-134. ISSN: 2167-647X. DOI: 10.1089/big.2016.0048. URL: http://dx.doi.org/10.1089/big.2016.0048 (vid. pág. 9).

[10] Amit Datta, Michael Carl Tschantz y Anupam Datta. *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*. 2015. arXiv: 1408.6491 [cs.CR] (vid. pág. 5).

[11] Emily Dinan y col. *Multi-Dimensional Gender Bias Classification*. 2020. arXiv: 2005.00614 [cs.CL] (vid. págs. 1, 9, 11).

[12] Michele Donini y col. *Empirical Risk Minimization under Fairness Constraints*. 2020. arXiv: 1802.08626 [stat.ML] (vid. pág. 9).

[13] Julia Dressel y Hany Farid. «The accuracy, fairness, and limits of predicting recidivism». En: *Science Advances* 4.1 (2018), eaao5580. DOI: 10.1126/sciadv.aao5580. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aao5580. URL: https://www.science.org/doi/abs/10.1126/sciadv.aao5580 (vid. pág. 4).

[14] Eran Eidinger, Roee Enbar y Tal Hassner. «Age and Gender Estimation of Unfiltered Faces». En: *IEEE Transactions on Information Forensics and Security* 9.12 (2014), págs. 2170-2179. DOI: 10.1109/TIFS.2014.2359646 (vid. pág. 9).

[15] Soheil Esmaeilzadeh y col. *Abuse and Fraud Detection in Streaming Services Using Heuristic-Aware Machine Learning*. 2022. arXiv: 2203.02124 [cs.LG] (vid. pág. 1).

[16] Xavier Gitiaux y Huzefa Rangwala. «mdfa: Multi-Differential Fairness Auditor for Black Box Classifiers». En: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, jul. de 2019, págs. 5871-5879. DOI: 10.24963/ijcai.2019/814. URL: https://doi.org/10.24963/ijcai.2019/814 (vid. págs. 8, 10).

[17] Haruka Hirota, Natthawut Kertkeidkachorn y Kiyoaki Shirai. «Weakly-Supervised Multimodal Learning for Predicting the Gender of Twitter Users». En: *Natural Language Processing and Information Systems*. Ed. por Elisabeth Métais y col. Cham: Springer Nature Switzerland, 2023, págs. 522-532. ISBN: 978-3-031-35320-8 (vid. págs. 1, 9).

[18] Faisal Kamiran y Toon Calders. «Data Pre-Processing Techniques for Classification without Discrimination». En: *Knowledge and Information Systems* 33 (oct. de 2011). DOI: 10.1007/s10115-011-0463-8 (vid. pág. 8).

[19] Kimmo Karkkainen y Jungseock Joo. «FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation». En: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, págs. 1548-1558 (vid. pág. 11).

[20] Matthew Kay, Cynthia Matuszek y Sean A. Munson. «Unequal Representation and Gender Stereotypes in Image Search Results for Occupations». En: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, págs. 3819-3828. ISBN: 9781450331456. DOI: 10.1145/2702123.2702520. URL: https://doi.org/10.1145/2702123.2702520 (vid. pág. 5).

[21] David Leslie. *Understanding bias in facial recognition technologies*. Inf. téc. 2020. DOI: 10.5281/ZENODO.4050457. URL: https://zenodo.org/record/4050457 (vid. pág. 5).

[22] Sam Levin. *A beauty contest was judged by AI and the robots didn't like dark skin*. Accedido: 18-11-2023. 2016. URL: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people (vid. pág. 5).

[23] Barbara Martinez Neda, Yue Zeng y Sergio Gago-Masague. «Using Machine Learning in Admissions: Reducing Human and Algorithmic Bias in the Selection Process». En: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. SIGCSE '21. Virtual Event, USA: Association for Computing Machinery, 2021, pág. 1323. ISBN: 9781450380621. DOI: 10.1145/3408877.3439664. URL: https://doi.org/10.1145/3408877.3439664 (vid. pág. 8).

[24] Ninareh Mehrabi y col. *A Survey on Bias and Fairness in Machine Learning*. 2022. arXiv: 1908.09635 [cs.LG] (vid. págs. 1, 5).

[25] Takafumi Okuyama, Tad Gonsalves y Jaychand Upadhay. «Autonomous Driving System based on Deep Q Learnig». En: *2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*. 2018, págs. 201-205. DOI: 10.1109/ICoIAS.2018.8494053 (vid. pág. 1).

[26] G Rajendra y col. «Gender Prediction using Deep Learning Algorithms and Model on Images of an Individual». En: *Journal of Physics: Conference Series* 1998.1 (ago. de 2021), pág. 012014. DOI: 10.1088/1742-6596/1998/1/012014. URL: https://dx.doi.org/10.1088/1742-6596/1998/1/012014 (vid. pág. 9).

[27] Mrinmoy Roy y col. *Machine Learning Applications In Healthcare: The State Of Knowledge and Future Directions*. 2023. arXiv: 2307.14067 [cs.LG] (vid. pág. 1).

[28] Rachel Rudinger, Chandler May y Benjamin Van Durme. «Social Bias in Elicited Natural Language Inferences». En: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Ed. por Dirk Hovy y col. Valencia, Spain: Association for Computational Linguistics, abr. de 2017, págs. 74-79. DOI: 10.18653/v1/W17-1609. URL: https://aclanthology.org/W17-1609 (vid. pág. 5).

[29] Jaydip Sen, Rajdeep Sen y Abhishek Dutta. *Machine Learning in Finance-Emerging Trends and Challenges*. 2021. arXiv: 2110.11999 [q-fin.ST] (vid. pág. 1).

[30] William Seymour. «Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?» En: *BIAS 2018* (2018). URL: http://ceur-ws.org/Vol-2103/ (vid. pág. 9).

[31] Terence Shin. *Real-Life Examples of Discriminating Artificial Intelligence*. Accedido: 1-12-2023. 2020. URL: https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070 (vid. pág. 1).

[32] Moninder Singh y Karthikeyan Natesan Ramamurthy. *Understanding racial bias in health using the Medical Expenditure Panel Survey data*. 2019. arXiv: 1911.01509 [cs.LG] (vid. pág. 5).

[33] Valentin-Gabriel Soumah y col. *Radar de Parité: An NLP system to measure gender representation in French news stories*. 2023. arXiv: 2304.09982 [cs.CL] (vid. págs. 1, 9).

[34] Julia Stoyanovich, Bill Howe y H. V. Jagadish. «Responsible Data Management». En: *Proc. VLDB Endow.* 13.12 (ago. de 2020), págs. 3474-3488. ISSN: 2150-8097. DOI: 10.14778/3415478.3415570. URL: https://doi.org/10.14778/3415478.3415570 (vid. pág. 5).

[35] Latanya Sweeney. *Discrimination in Online Ad Delivery*. 2013. arXiv: 1301.6822 [cs.IR] (vid. pág. 5).

[36] Sahil Verma y Julia Rubin. «Fairness Definitions Explained». En: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. 2018, págs. 1-7. DOI: 10.1145/3194770.3194776 (vid. pág. 8).

[37] Guanchu Wang y col. *Mitigating Algorithmic Bias with Limited Annotations*. 2023. arXiv: 2207.10018 [cs.LG] (vid. pág. 10).

[38] Jakub Wiśniewski y Przemysław Biecek. *fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation*. 2022. arXiv: 2104.00507 [stat.ML] (vid. pág. 1).

[39] Jintang Xue y col. *Bias and Fairness in Chatbots: An Overview*. 2023. arXiv: 2309.08836 [cs.CL] (vid. pág. 5).

[40] Ke Yang y col. «Fairness-Aware Instrumentation of Preprocessing Pipelines for Machine Learning». En: *Workshop on Human-In-the-Loop Data Analytics (HILDA'20)* (). DOI: 10.1145/3398730.3399194. URL: https://par.nsf.gov/biblio/10182459 (vid. pág. 8).

[41] Muhammad Bilal Zafar y col. «Fairness Constraints: Mechanisms for Fair Classification». En: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. por Aarti Singh y Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, abr. de 2017, págs. 962-970. URL: https://proceedings.mlr.press/v54/zafar17a.html (vid. pág. 9).

[42] Zhifei Zhang, Yang Song y Hairong Qi. *Age Progression/Regression by Conditional Adversarial Autoencoder*. 2017. arXiv: 1702.08423 [cs.CV] (vid. pág. 13).