Tarea 1: Ética, sesgo y equidad

Instituto Tecnológico de Costa Rica
Maestría en ciencias de la computación
Prof. María Auxiliadora Mora Cross
II Semestre 2024
Aprendizaje Automático
Estudiante: Danny Xie-Li, Miguel G. Abreu
Fecha de entrega: Viernes 02 de Agosto

Curso: Procesamiento de lenguaje natural Trabajo 1

1 Opinión respecto al tema del sesgo en modelos de Procesamiento del Lenguaje Natural (PLN)

Ambos videos se centran en el problema del sesgo en los modelos de **PLN**. Los temas clave son el sesgo de género en los modelos de lenguaje y cómo se filtra este sesgo a las tareas de lenguaje subsecuentes, en particular la resolución de correferencias. Se revisan en detalle varios enfoques para la detección, medición y mitigación del sesgo en los modelos de lenguaje, así como los problemas técnicos y éticos relacionados con el tema.

Los modelos de **PLN** tienen un impacto significativo en diversas aplicaciones críticas, como la búsqueda de empleo y el mantenimiento de la ley. La transparencia en estos modelos es esencial para garantizar que se sigan las mejores prácticas y mitigar cualquier sesgo inherente(Sun et al., 2019). El sesgo presente en estos modelos puede perpetuar y aumentar las desigualdades existentes, afectando negativamente a las poblaciones minoritarias(Daneshjou et al., 2021).

A pesar de las técnicas propuestas para reducir el sesgo en los modelos de IA(Sun et al., 2019), como la eliminación de componentes sesgados de los *embeddings*, estas técnicas no son completamente efectivas. El sesgo puede persistir de maneras sutiles y difíciles de eliminar por completo. Este es un desafío continuo en el campo del **PLN** con el que los investigadores deben lidiar constantemente, subrayando la necesidad de técnicas más avanzadas y flexibles (Talat et al., 2022), (Sengupta and Srivastava, 2022).

2 Estrategia para evitar problemas de sesgo y equidad en proyectos de Ciencia de Datos

Para abordar los problemas de sesgo y equidad en proyectos de Ciencia de Datos, es necesario definir varias estrategias respaldadas por varias investigaciones recientes. A continuación, se detallarán algunas de estas estrategias:

1. **Implementación de auditorías y marcos de mitigación de sesgo**: Es esencial que los sistemas de Ciencia de Datos incluyan en su estrategia de crecimiento y desarrollo, auditorías periódicas con el objetivo de poder identificar y corregir en el menor tiempo posible, problemas derivados de los sesgos. Una de las herramientas mas importantes para la implementación de estas auditorías, podría ser ***AEQUITAS*** (Saleiro et al., 2020), su esencia es la de abordar y contener las múltiples manifestaciones de sesgo e injusticia en la IA, además busca promover como principio general la noción de equidad de la IA.

2. **Desarrollo y uso de metodologías de aprendizaje interactivo y explicable**: Es necesario la incorporación de enfoques de aprendizaje interactivo, que logren brindar al usuario una retroalimentación continua para la mejora de la equidad en los modelos de lenguaje. Por ejemplo en el enfoque ***FairCaipi*** (Heidrich et al., 2023), se permite a los usuarios proporcionar comentarios sobre las diferentes predicciones y explicaciones realizadas por el modelo. Posteriormente, estos modelos, pueden ser ajustados dinamicamente en función de la retroalimentación continua por parte de los usuarios. Este enfoque facilita la detección de diferentes tipos de sesgos que no se previeron inicialmente.

3. **Uso de estrategias de procesamiento y selección de características**: El uso de estrategias de procesamiento y selección de características es fundamental para evitar la perpetuación y amplificación de los sesgos en los modelos de aprendizaje automático (Salazar et al., 2021). A continuación se detallarán algunas de las estrategias que pueden ser implementadas:

    - **Eliminación de atributos sensibles**: Implica la eliminación de atributos sensibles como género, raza o grupo etnico cel conjunto de datosantes de entrenar el modelo. La idea detrás de esta estrategia es evitar que el modelo utilice directamente estas características para tomar decisiones (Valentim et al., 2019). Este enfoque puede no ser suficiente ya que podrían existir tributos que se correlacionan estrechamente con los que fueron eliminados, introduciendo sesgos.

    - **Corrección de ruido en las etiquetas**: Se centra en las correcciones a realizar en las etiquetas de los datos que pueden habr sido asignada de manera

sesgada. Este enfoque ayuda a asegurar que el modelo no aprenda patrones discriminatorios presentes en los datos (Silva et al., 2023).

- **Enfoques causales y redes bayesianas**: Las redes bayesianas causales permiten la modelación y el análisis de las relaciones de causa-efecto que existe entre variables. Este tipo de técnicas ayudan, en gran medida, en la identificación y mitigación de los diferentes tipos de sesgos, ya que garantiza que las decisiones sean basadas en relaciones causales legítimas (Oneto and Chiappa, 2020).

3 Problemas de sesgo en proyectos de NLP

Los problemas de sesgo en proyectos de NLP, se encuentran bien documentados y afectan significativamente la equidad y eficacia de los sistemas. A continuación se describirán algunos de los casos de problemas de sesgos en sistemas de NLP:

1. **Sesgo de género en la resolución de correferencias**: La resolución de correferencias es una tarea de vital importancia en el campo del NLP, consiste en identificar cuando diferentes expresiones en un texto se refieren a la misma entidad. Pongamos un ejemplo para entrar en contexto:

   - En la frase **"María es Ingeniera. Ella trabaja en una empresa tecnológica"**, el modelo debería interpretar que **"Ella"** hace referencia directa a **María**, sin embargo esto no siempre es así y es producto a los sesgos implícitos en los datos de entrenamiento lo que pudiera llevar a asociaciones erróneas.

   En un estudio realizado por (Lu et al., 2018) destacó que los sistemas de resolución de correferencias muestran un sesgo de género significativo. Los modelos entrenados con conjuntos de datos que reflejan los diferentes sesgos sociales existentes tienden a asociar profesiones tradicionalmente realizadas por hombres, como **"Ingeniero"**, pronombres masculinos. Por otro lado profesiones que comunmente son practicadas por mujeres como **"Enfermera"**, suele asigmar pronombres femeninos. Un ejemplo de lo explicado anteriormente, se puede ver en mas detalle en la sección 4.3.2

   Como se mencionaba anteriormente, los conjuntos de datos utilizados para el entrenamiento de los modelos de NLP pueden contener sesgo histórico y sociales. Si estos datos incluyen muchas más referencias a hombres como **ingenieros** y a mujeres como **enfermeras**, este simplemente aprenderá estas asociaciones para las predicciones realizadas con posterioridad. Otro problema es la amplificaciones de estos sesgos, un modelo puede contribuir a esto debido a que se encuentra ímplicito en los datos, lo que significa que si no se introdujo explícitamente el modelo puede llegar a inferirlo por diversas asociaciones.

2. **Sesgo en los modelos de inferencia de lenguaje natural (NLI)**: El sesgo de género en los modelos de inferencia de lenguaje natural es un problema crítico que se manifiesta de diversas maneras, afectando la precisión y la equidad de estos sistemas. Los modelos de NLI están diseñados para determinar la relación entre dos oraciones, como si una premisa implica o contradice una hipótesis, o si son neutrales entre sí. Sin embargo, cuando estos modelos se entrenan en conjuntos de datos que contienen sesgos de género, pueden reflejar y perpetuar estos sesgos en sus predicciones (MacCartney, 2009).

   Los modelos de NLI, como **BERT**, **RoBERTa** y **BART**, han demostrado ser propensos a errores inducidos por el género. Esto significa que estos modelos, cuando se enfrentan a oraciones que contienen pronombres o términos relacionados con el género, pueden hacer inferencias basadas en estereotipos de género. Por ejemplo, podrían asociar automáticamente profesiones o roles específicos con un género particular, como asumir que un "doctor" es siempre masculino y una "enfermera" es femenina.

   Un estudio realizado por (Sharma et al., 2021) evaluó el sesgo de género en estos modelos al construir tareas de desafío que emparejan premisas neutrales en cuanto a género con hipótesis específicas de género. Los resultados mostraron que estos modelos eran propensos a errores de predicción inducidos por el género, reflejando un sesgo significativo hacia ciertos estereotipos de género.

   Este estudio destacó la necesidad de metodologías más efectivas para evaluar y mitigar el sesgo de género en los modelos de NLI.

3. **Sesgo en los Conjuntos de Datos y Representaciones de Palabras**: Los modelos de procesamiento del lenguaje natural (NLP) a menudo se entrenan en grandes conjuntos de datos no curados, lo que significa que estos datos no han sido cuidadosamente seleccionados o limpiados para eliminar sesgos inherentes. Estos conjuntos de datos suelen reflejar los prejuicios y estereotipos presentes en la sociedad, ya que se obtienen de diferentes fuentes en Internet, donde los sesgos humanos están presentes en abundancia.

   Uno de los componentes clave de muchos modelos de NLP son las representaciones de palabras, también conocidas como **word embeddings**, que mapean palabras a vectores en un espacio de alta dimensión. Ejemplos comunes de **word embeddings** incluyen **Word2Vec** y **GloVe**. Estos *embeddings* capturan relaciones semánticas entre palabras basándose en su contexto de uso en los datos de entrenamiento. Sin embargo, debido a que los datos de entrenamiento pueden contener sesgos, estos sesgos también se capturan en las representaciones de palabras.

	SNLI (I)				MNLI (I)			
	$(\Delta P)(\downarrow)$		B $(\downarrow)$		$(\Delta P)(\downarrow)$		B $(\downarrow)$	
	Male	Female	Male	Female	Male	Female	Male	Female
BERT	51.43	**33.6**	98.19	**37.22**	27.16	**20.51**	95.23	**40.11**
RoBERTa	28.75	**25.44**	83.33	**71.33**	27.4	**15.38**	94.85	**30.44**
BART	35.22	**34.54**	89.14	**56.33**	16.99	**15.46**	90.47	**39.22**
	SNLI (O)				**MNLI (O)**			
	$(\Delta P)(\downarrow)$		B $(\downarrow)$		$(\Delta P)(\downarrow)$		B $(\downarrow)$	
	Male	Female	Male	Female	Male	Female	Male	Female
BERT	49.16	**32.72**	96.19	**32.83**	34.36	**27.02**	92.52	**40.5**
RoBERTa	28.46	**24.35**	78.76	**67.94**	30.58	**18.18**	93.8	**30.38**
BART	34.54	**33.19**	86.04	**51.88**	22	**19.65**	87.9	**37.61**

Figure 1: Análisis detallado de cómo varía el sesgo con respecto a las ocupaciones dominadas por hombres y mujeres. Los números en negrita indican el mejor valor para cada métrica en los dos géneros. El sesgo de los empleos con predominio de hombres es comparativamente mayor que el de los empleos con predominio de mujeres. ΔP denota la diferencia absoluta media en las probabilidades de vinculación de las hipótesis masculinas y femeninas y B denota el número de veces que la probabilidad de vinculación de la hipótesis que coincide con el estereotipo fue mayor que la de su contraparte. **Tomado de (Sharma et al., 2021)**

Un estudio realizado por (Liu et al., 2023) demostró que las representaciones de palabras preentrenadas como Word2Vec y GloVe contienen sesgos significativos. Por ejemplo, se observó que la palabra “mujer" estaba asociada con términos relacionados con la familia y roles domésticos, mientras que la palabra “hombre" estaba asociada con términos relacionados con la carrera profesional y roles públicos. Este tipo de asociaciones reflejan y perpetúan estereotipos de género.

4 Análisis del conjunto de datos BiasBios (De-Arteaga et al., 2019)

El conjunto de datos *Bias in Bios* (De-Arteaga et al., 2019) es una colección a gran escala diseñada para estudiar el sesgo de género en la clasificación de ocupaciones. Esta tarea, que emplea aprendizaje automático, puede producir resultados perjudiciales para las personas al analizar biografías textuales con el atributo sensible del género binario. El conjunto de datos incluye 28 ocupaciones diferentes y mantiene una proporción de género de 53.9% masculino y 46.1% femenino en los tres conjuntos de datos: entrenamiento,

validación y prueba. Específicamente, el conjunto de datos se divide en 257,000 observaciones para entrenamiento, 99,000 observaciones para prueba y 40,000 observaciones para validación. La cantidad de datos por profesión sigue una distribución sesgada, con una desviación estándar de 5.726 y un promedio de 3.57. Esto indica una representación desigual de las profesiones, donde algunas están sobrerrepresentadas en comparación con otras debido al sesgo inherente, lo que introduce un desbalance en la representación de cada profesión.

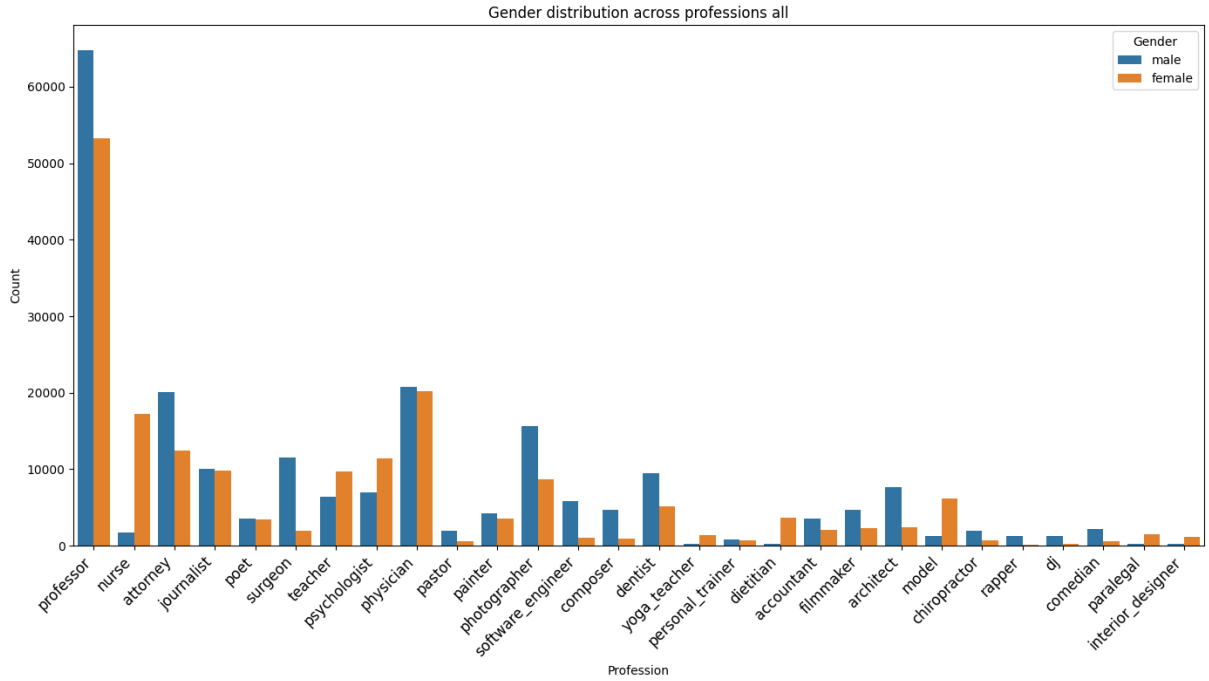

Figure 2: Gráfico de barras de la distribución de género en las diferentes profesiones. Se destaca que la profession con mayor representación es professor.

A partir de los datos, se observa que 20 de las 28 profesiones, lo que equivale al 71.42%, tienen una mayor representación de hombres que de mujeres, mostrando un gran sesgo y desbalance en términos de género. En el top 10 de profesiones con mayor proporción de hombres se encuentran ocupaciones como rapero (90.30%), DJ (85.80%), cirujano (85.18%), ingeniero en software (84.20%), compositor (83.62%), arquitecto (76.32%), pastor (84.20%), quiropráctico (73.67%) y cineasta (67.05%). En contraste, el top 10 de profesiones con mayor proporción de mujeres incluye dietista (92.28%), enfermera (90.84%), asistente legal (84.87%), profesora de yoga (84.50%), modelo (82.72%), diseñadora de interiores (80.76%), psicóloga (62.07%) y profesora (60.22%) en este conjunto de datos.

4.1 Palabras más frecuentes por género

4.1.1 Metodología

Para identificar las palabras más frecuentes por género en el conjunto de datos, se utilizó la biblioteca *Natural Language Toolkit* (nltk) de Python, una colección de herramientas y programas para el procesamiento del lenguaje natural. Se removió los datos que tenían la profesión de profesor, ya que representaban un 30% del total de datos. Se empleó el operador de conteo para determinar la frecuencia de las palabras en el texto. Como método de preprocesamiento, se eliminaron las *stopwords*, excluyendo los pronombres como "he", "him", "his", "himself", "she", "her", "hers", "herself", los cuales son indicadores de género. El objetivo fue identificar la frecuencia de estos pronombres en el texto en relación con las profesiones.

Sea T el texto que consiste en una secuencia de tokens, denotada como $T = \{w_1, w_2, w_3, \ldots, w_n\}$, donde w_i representa cada token en el texto y n es el total de tokens en el texto, utilizando como operador de separación el espacio para definir el token. Se define V como el conjunto de vocabulario, que es el conjunto de palabras únicas en T, denotado $V = \{v_1, v_2, v_3, \ldots, v_m\}$, donde v_j representa la j-ésima palabra única en el texto y m es el número total de palabras únicas.

Definimos la función de frecuencia como $f : V \to \mathbb{N}$, que asigna a cada palabra en el vocabulario la frecuencia de su aparición en el texto. Para cada palabra $v_j \in V$, la función de frecuencia $f(v_j)$ se define como:

$$f(v_j) = |\{w_i \in T \mid w_i = v_j\}|$$

En otras palabras, $f(v_j)$ es la cardinalidad del conjunto de palabras en T que son iguales a v_j. Esto representa el número de veces que v_j aparece en el texto T.

4.1.2 Resultados

- **Femenino**: En la figura 4 se puede observar la nube de palabras de los tokens más comunes en el conjunto de datos generadas con el algoritmo TF-IDF. Por otro lado, usando la función de conteo se pudo identificar los tokens más frecuentes en el texto fueron *she* (305,725), *her* (209,676), *university* (90,100), *research* (53,929), *medical* (43,527), *years* (35,998), *health* (34,564), hospital (33,019), *school* (32,186), *medicine* (29,994), y *nurse* (12,353). Estos tokens podrían estar relacionados con profesiones como dietista y enfermería, donde hay una mayor representación del género femenino.

- **Masculino**: En la figura 4 se puede observar la nube de palabras de los tokens más comunes en el conjunto de datos generadas con el algoritmo TF-IDF. Por otro lado, usando la función de conteo se pudo identificar los tokens más frecuentes en el texto

fueron *he* (253,737), *his* (150,928), *medical* (37,746), *hospital* (28,495), *blue* (27,286), y *medicine* (24,426). Estos tokens podrían estar asociados con profesiones como quiropráctico y cirujano, donde hay una mayor representación del género masculino en los datos.

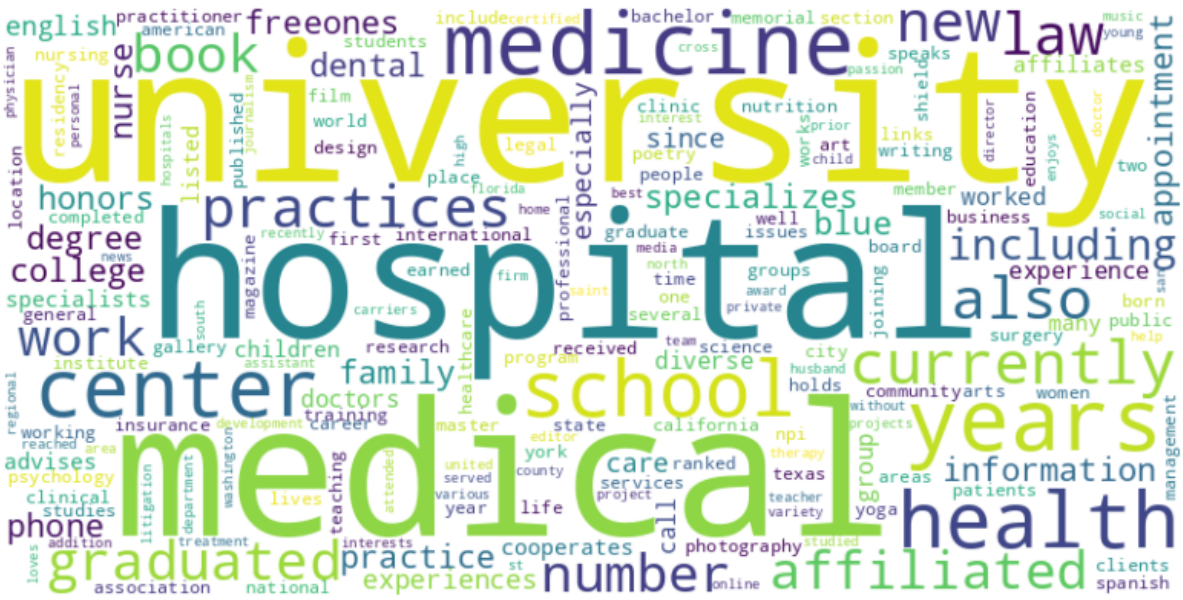

Figure 3: Nube de palabras generada con el algoritmo TF-IDF, mostrando las palabras más comunes en el conjunto de datos, excluyendo la profesión de *professor* en el género femenino.

Figure 4: Nube de palabras generada con el algoritmo TF-IDF, mostrando las palabras más comunes en el conjunto de datos, excluyendo la profesión de *professor* en el género masculino.

4.2 Técnicas de correción del sesgo

Según la revisión literaria de los autores (Sun et al., 2019) destacan differentes métodos para mitigar el bias usando manipulación de datos. Proponen trabajar en el corpora del texto y su representación, y algoritmos de predicción, pero discutiremos técnicas para mitigar en el cuerpo del texto para bias en el género dentro del texto.

- **Augmentación de datos**: Los autores (Zhao et al., 2018) propuso crear un conjunto de datos aumentado que sea idéntico al conjunto de datos original pero sesgado hacia el género opuesto. El proceso de augmentación implica intercambiar el género en cada oración del conjunto original para generar su equivalente con el género modificado. Además, se realiza una anonimización de nombres, reemplazando todas las entidades con nombre por entidades anonimizadas, eliminando así las asociaciones de género. El modelo se entrena utilizando la combinación del conjunto de datos original con la versión anonimizada y el conjunto de datos aumentado.

- **Etiquetado de género**: En las traduciones automáticas confundir el género puede llevar a prediciones inexactas, debido a sesos en el conjunto de entrenamiento. Mitigar esto al agregar una etiqueta que indica el género de la fuente del punto de datos al comienzo de cada punto de datos. Por ejemplo, "I'm happy" cambiaría a "MALE I'm happy". En teoría, codificar la información de género en las oraciones podría mejorar las traducciones en las que el género del hablante afecta la traducción (Sun et al., 2019).

- **Fine-tuning de sesgo**: Los conjuntos de datos no sesgados pueden ser escasos, pero se pueden usar conjuntos de datos no sesgados para tareas relacionadas mediante el aprendizaje por transferencia (*fine-tuning*). Esto minimiza el sesgo en los modelos antes de afinarlos con conjuntos de datos más sesgados para la tarea objetivo. (Sun et al., 2019)

- **Máscara de género**: Para enmascarar el género en el texto, utilizamos una etiqueta que identifica artículos y pronombres que pueden revelar el género. En este enfoque inicial, se propuso el siguiente mapeo:

  - `"he"` se etiqueta como `[PRONOUN]`
  - `"she"` también se etiqueta como `[PRONOUN]`
  - `"his"` se clasifica como `[POSSESSIVE]`
  - `"her"` se marca como `[POSSESSIVE]`
  - `"him"` se etiqueta como `[OBJECT]`

  Por ejemplo, considera la siguiente oración: "She joined IBM in 2007 and has been part of the Java Technology Center. She has worked on JAXB and JAXWS on

JDK6, and currently she is part of JDK7 development." Esta oración se transforma en: `[PRONOUN] joined IBM in 2007 and has been part of the Java Technology Center. [PRONOUN] has worked on JAXB and JAXWS on JDK6, and currently [PRONOUN] is part of JDK7 development.`

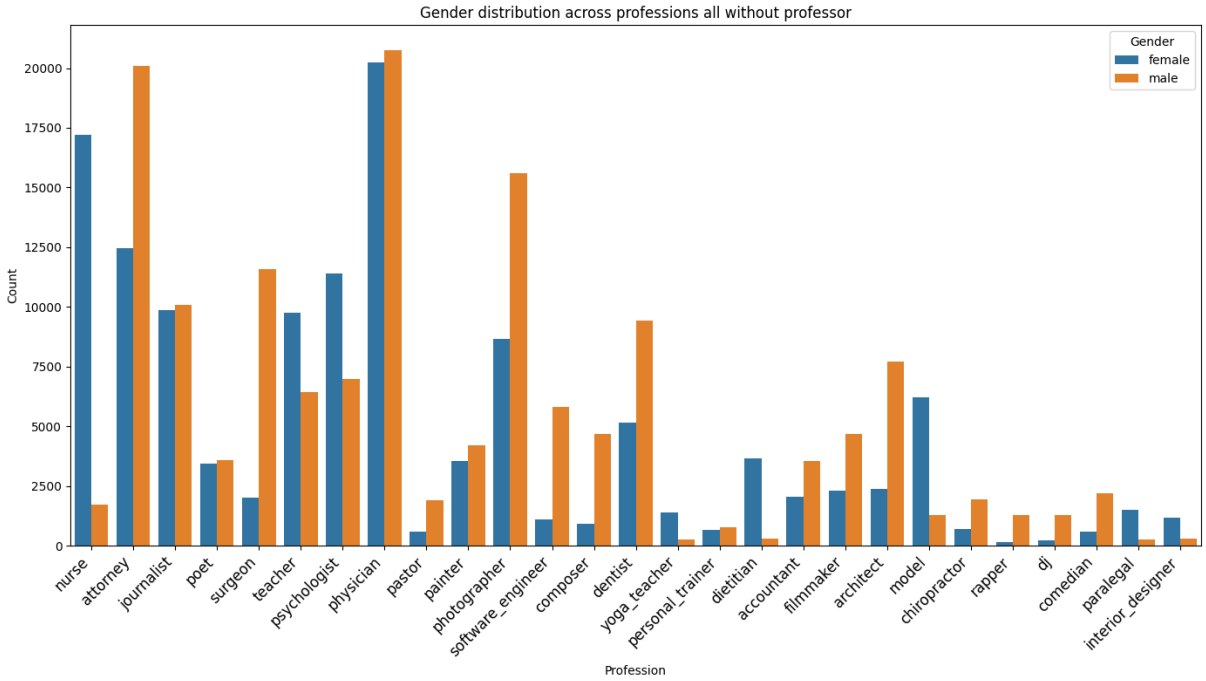

Figure 5: Gráfico de barras de la distribución de profesiones sin la profesión *professor*

4.3 Experimentos con modelos contextuales *transformers* utilizando clasificación *zero-shot*

4.3.1 Metodologia

Ahora nos planteamos la siguiente pregunta: ¿Los modelos preentrenados contextuales de transformers presentan sesgos? Para investigar esto, evaluamos dos modelos diferentes:

- **Cross-encoder/NLI-RoBERTa-Base** (Reimers and Gurevych, 2019): Entrenado con un corpus de 570,000 pares de oraciones en inglés, etiquetados manualmente para clasificación equilibrada con las etiquetas de "entailment" (consecuencia), "contradiction" (contradicción) y "neutral" (neutral).
- **facebook/bart-large-mnli** (Lewis et al., 2019): Entrenado con el corpus Multi-Genre Natural Language Inference (MultiNLI), una colección de 433,000 pares de oraciones anotados con información de inferencia textual.

Para nuestra evaluación, hemos seleccionado dos profesiones representativas: ingeniería en software y enfermería. La ingeniería en software presenta una notable predominancia

masculina en el conjunto de datos, con un 84.20% de representaciones masculinas frente a las femeninas. En contraste, la profesión de enfermería muestra una representación femenina mucho mayor, alcanzando el 90.84%. Empleamos una metodología de clasificación en cero disparos (zero-shot classification), en la que el modelo debe asignar una de las dos etiquetas a un texto dado. Posteriormente, evaluamos el rendimiento del modelo utilizando el *F1 Score*, *recall* y la precisión para cada profesión en función del género. Para esto, utilizamos un conjunto de prueba compuesto por 200 muestras para cada género, totalizando 800 muestras (2 profesiones x 2 géneros x 200 muestras cada uno). Utilizamos el conjunto de datos de prueba para obtener los resultados y, a continuación, aplicamos el enmascaramiento de género como se detalla en la sección 4.2.

4.3.2 Resultados

En la tabla 1 para el modelo NLI-RoBERTa-Base se observa que el F1 score para la profesión de enfermera (nurse) es significativamente alto para el género femenino (female), con un valor de 0.997494, y para el género masculine (male), con un valor de 0.915989, lo que resulta en una diferencia de 0.081505. En contraste, para la profesión de ingeniero de software, la diferencia en el F1 score es menor, de 0.023, lo que indica un sesgo menor en comparación con la profesión de enfermera. Por otro lado, el modelo bart-large-mnli presenta un F1 score de 0.969 para enfermera en el género femenino y de 0.892 para el género masculino, con una diferencia menor para la profesión de ingeniero de software. Esto sugiere que el sesgo de género en el rendimiento de estos modelos varía dependiendo de la profesión.

Modelo	**Profesión**	**Género**	**Precisión**	**Recall**	**F1 Score**
NLI-RoBERTa-Base	*nurse*	*male*	1.0	0.845	0.915989
NLI-RoBERTa-Base	*nurse*	*female*	1.0	0.995	0.997494
NLI-RoBERTa-Base	*software engineer*	*male*	1.0	0.875	0.933333
NLI-RoBERTa-Base	*software engineer*	*female*	1.0	0.835	0.910082
bart-large-mnli	*nurse*	*male*	1.0	0.805	0.891967
bart-large-mnli	*nurse*	*female*	1.0	0.940	0.969072
bart-large-mnli	*software engineer*	*male*	1.0	0.995	0.997494
bart-large-mnli	*software engineer*	*female*	1.0	0.985	0.992443

Table 1: Métricas de rendimiento para diferentes profesiones y géneros utilizando varios modelos con el conjunto de datos BiasBios.

Después de aplicar la técnica de enmascaramiento de género, se observa una mejora en la clasificación de la profesión de enfermería para el género masculino, como se detalla en la tabla 2, en comparación con los datos sin procesar en la tabla 1. Sin embargo, para la profesión de ingeniero de software, la diferencia es mínima y no muestra una mejora significativa respecto a los datos sin procesar. Además, al utilizar el modelo

Modelo	Profesión	Género	Precisión	Recall	F1 Score
NLI-RoBERTa-Base	*nurse*	*male*	1.0	0.920	0.958333
NLI-RoBERTa-Base	*nurse*	*female*	1.0	0.985	0.992443
NLI-RoBERTa-Base	*software engineer*	*male*	1.0	0.800	0.888889
NLI-RoBERTa-Base	*software engineer*	*female*	1.0	0.820	0.901099
bart-large-mnli	*nurse*	*male*	1.0	0.800	0.888889
bart-large-mnli	*nurse*	*female*	1.0	0.910	0.952880
bart-large-mnli	*software engineer*	*male*	1.0	0.980	0.989899
bart-large-mnli	*software engineer*	*female*	1.0	0.950	0.974359

Table 2: Métricas de rendimiento para diferentes profesiones y géneros utilizando varios modelos con el conjunto de datos BiasBios utilizando enmascaramiento de género.

`bart-large-mnli`, no se identifican diferencias notables al aplicar ambos métodos de procesamiento de datos.

4.3.3 Discusión

En este estudio, se observa que los modelos contextuales *transformers NLI-RoBERTa-Base* y *bart-large-MNLI*, preentrenados en grandes conjuntos de datos, presentan sesgos de género en relación con profesiones. Estos sesgos se manifiestan cuando los modelos procesan indicadores como artículos, pronombres u otros elementos que sugieren género. Dado que los datos actuales reflejan las desigualdades de la sociedad, estos sesgos pueden influir en las decisiones basadas en dichos modelos que han sidos preentrenados con estos datos. Por lo tanto, es fundamental desarrollar modelos justos, especialmente en contextos de toma de decisiones. Además, la aplicación de técnicas para mitigar el sesgo puede contribuir a construir modelos más equitativos.

4.3.4 Limitaciones

La técnica utilizada para enmascarar el género es relativamente sencilla y fácil de implementar, ya que se basa en un diccionario y requiere ajustes para manejar las diversas variaciones. Por ello, se limita a los términos existentes en el diccionario, y el costo de manterlo.

4.3.5 Trabajo futuro

Dado que este análisis se basa en una única ejecución del experimento, se recomienda realizar pruebas adicionales para evaluar el sesgo de los modelos en distintas profesiones y llevar a cabo análisis estadísticos para determinar su significancia. Además, es fundamental evaluar múltiples profesiones en el contexto de clasificación *zero-shot*. Para trabajos futuros, se sugiere emplear modelos de lenguaje grandes como agentes para preprocesar los

datos y eliminar cualquier indicio de género. Una mejora en el enmascaramiento podría incluir la adición de artículos neutros en inglés, como *they* y *theirs*.

References

Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA dermatology*.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Heidrich, L., Slany, E., Scheele, S., & Schmid, U. (2023). Faircaipi: A combination of explanatory interactive and fair machine learning for human and machine bias reduction. *Machine Learning and Knowledge Extraction*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://arxiv.org/abs/1910.13461

Liu, R., Wu, Y., & Xing, Y. (2023). The presence of bias based on pre-trained language models. *Applied and Computational Engineering*.

Lu, K., Mardziel, P. (, Wu, F., Amancharla, P., & Datta, A. (2018). Gender bias in neural natural language processing. *ArXiv*, *abs/1807.11714*.

MacCartney, B. (2009). *Natural language inference* [Doctoral dissertation] [Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Última actualización - 2023-02-23]. https://www.proquest.com/dissertations-theses/natural-language-inference/docview/305018371/se-2

Oneto, L., & Chiappa, S. (2020). Fairness in machine learning. *ArXiv*, *abs/2012.15816*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

Salazar, R., Neutatz, F., & Abedjan, Z. (2021). Automated feature engineering for algorithmic fairness. *Proc. VLDB Endow.*, *14*, 1694–1702.

Saleiro, P., Rodolfa, K. T., & Ghani, R. (2020). Dealing with bias and fairness in data science systems: A practical hands-on tutorial. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*.

Sengupta, K., & Srivastava, P. R. (2022). Causal effect of racial bias in data and machine learning algorithms on user persuasiveness discriminatory decision making: An empirical study. *ArXiv*, *abs/2202.00471*.

Sharma, S., Dey, M., & Sinha, K. (2021). Evaluating gender bias in natural language inference. *ArXiv*, *abs/2105.05541*.

Silva, I. O. E., Soares, C., Sousa, I., & Ghani, R. (2023). Systematic analysis of the impact of label noise correction on ml fairness. *ArXiv*, *abs/2306.15994*.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. https://arxiv.org/abs/1906.08976

Talat, Z., Blix, H., Valvoda, J., Ganesh, M. I., Cotterell, R., & Williams, A. (2022). On the machine learning of ethical judgments from natural language. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 769–779.

Valentim, I., Lourenço, N., & Antunes, N. (2019). The impact of data preparation on the fairness of software systems. *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, 391–401.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. https://arxiv.org/abs/1804.06876