Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

IMPLEMENTACIÓN DE NLP Y ANÁLISIS TEMÁTICO PARA LA DETECCIÓN DE SESGOS DE GÉNERO

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestra en Ciencia de Datos

Presenta:
Concepción Haydé Martínez Landa

Director:
Dr. Arturo Silva Galvez

Tlaquepaque, Jalisco, 20 de mayo de 2024

IMPLEMENTACIÓN DE NLP Y ANÁLISIS TEMÁTICO PARA LA DETECCIÓN DE SESGOS DE GÉNERO

Concepción Haydé Martínez Landa

Resumen

El sesgo de género en contenidos digitales puede afectar la percepción y la representación, reforzando estereotipos y promoviendo desigualdades. Este estudio se centra en analizar cómo se manifiesta este sesgo, utilizando técnicas de Procesamiento del Lenguaje Natural (NLP) y visualización de datos. A través del análisis del conjunto de datos md_gender_bias, aplicamos NLP y Análisis Latente de Dirichlet (LDA) para desentrañar patrones temáticos y representaciones de género en distintos contextos.

El análisis en los conjuntos de datos Yelp y ConvAI2 han revelado la existencia de sesgos de género. En Yelp, si bien no se observa un sesgo de género absoluto, el análisis muestra cierta inclinación hacia palabras asociadas con roles de género masculinos tradicionales, como referencias a profesiones típicamente masculinas, consumo de carnes rojas y actividades estereotípicamente masculinas. Por otro lado, también se identificaron tópicos con una perspectiva más orientada a lo femenino según roles convencionales, con palabras vinculadas a la mujer y actividades culinarias domésticas. En ConvAI2, el sesgo de género es más notorio, con tópicos claramente divididos en experiencias e intereses típicamente masculinos, como deportes y música, contrastando con tópicos femeninos relacionados con la vida familiar, relaciones personales, actividades domésticas y roles de género tradicionales.

Para el subconjunto de Wizard, se demuestra que hay una subrepresentación en el género femenino. Esto nos lleva a la conclusión de que la selección de textos para la creación del dataset, así como el etiquetado influyen directamente en la presencia de los sesgos en el mismo.

Este estudio resalta la presencia del sesgo de género en los contenidos digitales y subraya la importancia de herramientas avanzadas para su identificación y comprensión. A través de un análisis y visualizaciones, se muestra la presencia de los sesgos de género que se manifiestan el conjunto de datos analizado, finalmente buscando incentivar el desarrollo de estrategias para promover una representación más equitativa y diversa en el ámbito digital.

Tabla de Contenidos

Índice de figuras

Página

Dedicado a Luis Alejandro Delgado, mi pareja, por su apoyo incondicional, por motivarme, echarme porras, hacerme reír y escuchar mis quejas e ideas con amor y paciencia. A Ma. Andrea Valdes, mi mejor amiga, por existir y acompañarme todos los días en mi existencia en este mundo, y por recordarme constantemente que estaba procrastinando este trabajo. A mi familia, especialmente a mi mamá y a mi hermano, por su apoyo constante. A mi papá, por permitirme verlo en mí. A mis amigos y compañeros de maestría, Marco Mendoza y Ricardo Valdez, por que sin ellos este camino no hubiera sido el mismo. Y al futuro de la Inteligencia Artificial, que

promete un horizonte lleno de posibilidades

y desafíos.

1 Introducción

1.1 Contexto

Los algoritmos, especialmente aquellos utilizados en el procesamiento y análisis de datos, aprenden de los datos que se les proporciona. Si esos datos contienen sesgos, que a menudo reflejan las desigualdades y prejuicios presentes en la sociedad, el algoritmo también aprenderá esos sesgos. El sesgo de género, una forma prevalente de discriminación, se manifiesta no solo en las interacciones cotidianas, sino también en los datos y algoritmos que consumimos a diario. [1] Como ingeniera en el campo de la tecnología, experimento esta disparidad de primera mano.

Desde una perspectiva histórica, las mujeres han estado subrepresentadas en los campos STEM (Science, Technology, Engineering, and Mathematics), en los años 70's en el contexto de la programación, las muejeres tenían mayor presencia en el área, pero al detectarse una creciente importancia en el uso de la misma, y con el incremento en la oferta y uso de ordenadores de uso personal, la mercadotecnia comenzó a direccionar estos productos y actividades hacia hombres. Generando una tendencia que, se refleja en el contenido digital moderno.

Los algoritmos de Procesamiento del Lenguaje Natural (NLP) [2] , junto con el Análisis Latente de Dirichlet (LDA) [3], que exploran y estructuran grandes volúmenes de texto identificando tópicos y patrones lingüísticos sin la necesidad de intervención humana explícita, ofrecen una herramienta potente para analizar y, idealmente, rectificar sesgos de género en contenidos digitales.

Este trabajo tiene como objetivo explorar y comprender las manifestaciones del sesgo de género en la información digital, aplicando técnicas de NLP y LDA al conjunto de datos md_gender_bias. Mediante este análisis, se busca identificar tendencias y correlaciones que revelen inclinaciones hacia ciertos géneros en diversos ámbitos, incluidos los contenidos relacionados con las áreas

de STEM y justicia social. A través de este enfoque, se aspira a desvelar la estructura subyacente de los sesgos de género en los datos, proporcionando una base para su mitigación y para promover prácticas más equitativas en la ciencia de datos y la tecnología.

Este documento se estructura de la siguiente manera:

Introducción: Este capítulo proporciona un trasfondo del problema del sesgo de género, una justificación para el estudio, y una descripción general de los objetivos del proyecto.

Metodología: Este capítulo proporciona el enfoque metodológico de la investigación, incluyendo una descripción de los datos, el análisis exploratorio, y una explicación de los modelos y métricas utilizados.

Resultados y Discusión: Este capítulo presenta y discute los hallazgos del análisis, proporcionando una discusión crítica de los resultados en el contexto del sesgo de género en datos digitales.

Conclusiones: En este capítulo final, se ofrecen conclusiones basadas en los hallazgos de la investigación, así como recomendaciones para futuras investigaciones en este campo.

1.2 *Justificación*

Los sesgos pueden perpetuar y amplificar desigualdades históricas, limitar oportunidades para grupos subrepresentados, y distorsionar la toma de decisiones basada en datos. Su rectificación no es solo una cuestión ética, sino también un paso necesario para asegurar sistemas más justos y representativos.[4] Este trabajo espera identificar el problema en conjuntos de datos que diseñados para resolver el sesgo en LLM's (Modelos de Lenguaje de Gran Escala).

La desigualdad de género ha sido una constante, con mujeres siendo relegadas principalmente a las labores del hogar y siendo excluidas de la esfera pública, incluyendo la negación del derecho al voto hasta mediados del siglo XX. [5] A pesar de las contribuciones de mujeres a la ciencia y otros campos, su reconocimiento ha sido considerablemente bajo, dejando una brecha en la narrativa histórica que destaca predominantemente los logros masculinos. [6]

Actualmente, continuamos luchando contra formas más sutiles pero igualmente perniciosas de desigualdad de género, lo que se refleja en las brechas salariales, en la representación inadecuada en posiciones de liderazgo y en los estereotipos de género que aún están profundamente arraigados en la sociedad. Esto muestra la importancia de abordar y rectificar las desigualdades y discriminaciones que enfrentan las minorías.

A nivel económico, la equidad de género puede fomentar una mayor innovación y crecimiento. Hay estudios que demuestran que las empresas con equipos diversificados son más innovadoras y tienen un mejor rendimiento financiero, ilustrando el valor económico tangible de la inclusión y la equidad. [7]

Desde una perspectiva científica, el análisis de los datos mediante técnicas de Procesamiento del Lenguaje Natural (NLP) [2] y Análisis Latente de Dirichlet (LDA) [3] juega un papel fundamental en la identificación y comprensión de los sesgos de género.

Estas técnicas, al procesar y analizar extensos volúmenes de texto, tienen el potencial de revelar patrones lingüísticos y temáticos que podrían permanecer ocultos en análisis manuales o menos sofisticados. De esta manera, NLP y LDA pueden ser herramientas cruciales para descubrir y elucidar los sesgos de género presentes en los contenidos digitales, marcando un avance importante hacia el desarrollo de tecnologías más inclusivas y representativas.

Este estudio representa una contribución valiosa para la comprensión de los sesgos de género, especialmente porque se enfoca en un conjunto de datos diseñado explícitamente para minimizar sesgos en los Modelos de Lenguaje de Gran Escala (LLM's), el cual fue desarrollado predominantemente por hombres, quienes formaron dos tercios del equipo total.

Reported Gender	Percent of Total
Man	67.38
Woman	18.34
Non-binary	0.21
Prefer not to say	14.07

Figura 1.1: Distribución de genero en las personas que trabajaron en el dataset

Este hecho puede ser indicativo de un sesgo inherente dentro del propio conjunto de datos, señalando una posibilidad de representaciones desequilibradas y perspectivas limitadas, lo que en sí mismo representa un área para el análisis y la intervención.

La urgencia de abordar las desigualdades de género en la era digital radica en varias razones:

- A medida que la tecnología y los datos influyen en más aspectos de nuestras vidas, desde decisiones laborales hasta interacciones sociales, los sesgos en estos sistemas afectan a un número cada vez mayor de personas y pueden reforzar desigualdades existentes a una escala más amplia.[1]
- Los sesgos no corregidos en los algoritmos pueden reforzar y amplificar desigualdades preexistentes. Por ejemplo, si un algoritmo sesgado se utiliza en plataformas de contratación, podría perpetuar la falta de diversidad en ciertas industrias. [1]
- La rapidez con la que se desarrolla y despliega la tecnología significa que los sesgos, una vez integrados, pueden propagarse y arraigarse rápidamente en sistemas y prácticas antes de que sean identificados y corregidos. [1]
- Las desigualdades de género en la tecnología no solo tienen implicaciones éticas, sino también económicas. La falta de diversidad puede llevar a soluciones menos innovadoras, y las empresas pueden perder talento valioso debido a entornos no inclusivos. [7]
- Abordar las desigualdades de género es un imperativo ético y de derechos humanos. Cada individuo merece oportunidades equitativas y una representación justa en todas las esferas, incluida la digital. [8]

Mi decisión de abordar este tema surge de experiencias personales y de la prevalencia del sesgo de género en nuestra sociedad. Los algoritmos con sesgos pueden distorsionar la realidad, limitar oportunidades y magnificar desigualdades preexistentes. Como mujer en ingeniería, enfrento constantemente estas desigualdades, lo que intensifica mi compromiso para llevar a cabo este estudio. Al embarcarme en este estudio, busco no solo resaltar la presencia de sesgos de género en los datos digitales, sino también proporcionar una herramienta para combatir estos sesgos, promoviendo un futuro más inclusivo y equitativo.

Mediante el proceso desarrollado para el análisis y comprensón del dataset md_gender_bias [9], este estudio busca ofrecer perspectivas visuales y estádisticas sobre los sesgos de género, con el objetivo de proporcionar una perspectivas que conduzca a soluciones más equitativas e inclusivas en el ámbito digital, garantizando una representación justa para todos.

1.3 *Problema*

Aunque los conjuntos de datos como el md_gender_bias están diseñados con el propósito de mitigar los sesgos en los modelos de

lenguaje, existe una hipótesis de que aún pueden albergar ciertos niveles de sesgo, especialmente debido a las circunstancias que rodearon su creación.

Es preocupante que el equipo que trabajó en el desarrollo del dataset md_gender_bias, tenía una representación de dos tercios de individuos identificados como hombres. Esta predominancia masculina en el equipo de desarrollo plantea una pregunta: ¿Hasta qué punto la representación desbalanceada de género en el equipo de desarrollo puede haber influido en la creación de un conjunto de datos sesgado?.

La hipótesis central de esta investigación es que el conjunto de datos md_gender_bias contiene sesgos de etiquetado influenciados por la predominancia masculina en su equipo de desarrollo. El análisis mediante técnicas de Procesamiento del Lenguaje Natural (NLP) y Análisis Latente de Dirichlet (LDA) puede encontrar estas disparidades y ofrecer una mejor visión sobre la presencia de los sesgos. Para abordar esta cuestión, se realizará un análisis del conjunto de datos empleando NLP y LDA.

Las técnicas de Procesamiento del Lenguaje Natural (NLP)[2] y Análisis Latente de Dirichlet (LDA)[3] proporcionan un marco para analizar y estructurar grandes volúmenes de datos textuales sin necesidad de depender de categorías predefinidas. A través del NLP, es posible procesar y analizar texto para identificar patrones lingüísticos y semánticos, mientras que LDA permite descubrir los tópicos subyacentes en los textos, facilitando la identificación de tendencias y relaciones ocultas que pueden ser imperceptibles con métodos supervisados. Este enfoque evita replicar sesgos presentes en etiquetas predeterminadas al no depender de ellas, permitiendo un análisis más objetivo y basado en las características propias de los datos.

La aplicación de NLP y LDA contribuye significativamente a la detección de sesgos al revelar cómo ciertos tópicos del lenguaje se asocian con géneros específicos dentro de un conjunto de datos. Por ejemplo, la identificación de tópicos dominantes y su asociación con etiquetas de género puede evidenciar sesgos en la representación de géneros en los textos. Además, el análisis de tópicos detallado ofrece una base cuantitativa para evaluar la distribución y la representación de géneros, proporcionando evidencia de posibles sesgos en el contenido y en las prácticas de etiquetado.

Este estudio, más allá de abordar la cuestión central sobre sesgos de género en conjuntos de datos, tiene como objetivo contribuir al desarrollo de tecnologías más inclusivas. Mediante el uso de técnicas de Procesamiento del Lenguaje Natural (NLP) y Análisis Latente de Dirichlet (LDA), se propone un marco metodológico para identificar y analizar sesgos de género de manera más efectiva. Este enfoque

no solo facilita un análisis profundo de los sesgos presentes en los datos, sino que también establece una base para investigaciones futuras, promoviendo el desarrollo de prácticas y tecnologías que reconozcan y mitiguen los sesgos de género.

1.4 Objetivos

1.4.1 Objetivo general

Analizar la presencia de sesgos de género en el conjunto de datos md_gender_bias utilizando técnicas de Procesamiento del Lenguaje Natural (NLP) y Análisis Latente de Dirichlet (LDA) para promover una ciencia de datos orientada hacia la equidad e inclusión.

1.4.2 Objetivos específicos

1. Examinar el conjunto de datos md_gender_bias para identificar las características fundamentales, prestando especial atención a la distribución de las etiquetas de género y la naturaleza de los contenidos en las diversas fuentes de datos.
2. Seleccionar las herramientas y modelos de NLP y LDA apropiados para el análisis del conjunto de datos.
3. Analizar los tópicos revelados por las técnicas de NLP y LDA para comprender las agrupaciones y tendencias subyacentes relacionadas con los sesgos de género.
4. Evaluar la extensión y las implicaciones de los sesgos de género detectados en el conjunto de datos, basándose en los resultados obtenidos a través de NLP y LDA.

2 Metodología

2.1 Descripción de los datos

El conjunto de datos md_gender_bias consta de siete fuentes diferentes, cada una contribuyendo a un análisis del sesgo de género desde diversas perspectivas de comunicación. Estas fuentes, a través de sus características específicas, campos de aplicación y objetivos, forman una base sólida para investigar los sesgos de género en múltiples contextos. En este trabajo solo usaremos 3: ConvAI2, Wizard y Yelp. A continuación se detalla cada una de estas fuentes de datos:

	text	chosen_topic	gender
0	Hello. I hope you might enjoy or know somethin...	Krav Maga	2
1	I have no idea what or who that might be. Why ...	Krav Maga	2
2	I think Krav Maga is a martial art sport. That...	Krav Maga	2
3	Do you know how long it's been around?	Krav Maga	2
4	According to the info I have available to me i...	Krav Maga	2
5	It sounds painful.	Krav Maga	2
6	yeah I agree. It also comes from street fighti...	Krav Maga	2
7	Sounds very violent and not the smartest thing...	Krav Maga	2
8	I loved My Little Pony toys, made by Hasbro, w...	My Little Pony	0
9	My sister was really into them, and Strawberry...	My Little Pony	0
10	Yes, they were made by a group of three develo...	My Little Pony	0
11	Yeah, I would have been 8 at that time and I t...	My Little Pony	0
12	Yes, they were such cute little colorful ponie...	My Little Pony	0
13	It's so neat that all those toys from the 80s ...	My Little Pony	0
14	It is! There was a My Little Pony series that ...	My Little Pony	0

Figura 2.1: Primeros 15 renglones del conjunto de datos “Wizard”

Fuente: Wizard
Campos: text, chosen_topic, gender.
Cantidad: 8,000 instancias.
Objetivo: Observar el sesgo en diálogos sobre tópicos específicos.

En el campo **gender** existen los siguientes valores: neutral (0), female (1), male (2)

	text	binary_label	binary_score
0	do n't visit this place .	1	0.5145
1	great low cost company .	1	0.5864
2	came here saturday for the ufc fights .	1	0.7685
3	everything was hot and crispy .	1	0.5481
4	that ' s enough to check the place out .	1	0.5814
5	the managers seemed to be running a nice place .	1	0.7540
6	it ' s very nice to feel remembered .	1	0.6504
7	she choose the <UNK> grande .	0	0.8111
8	at the joint we were treated as customers .	0	0.5128
9	nice selection of bbq and sides .	1	0.7949
10	the food is excellent .	0	0.5013
11	the menu seems to change by season as well .	0	0.6339
12	really , everything here is great .	1	0.6014
13	taiwanese sausage fried rice .	0	0.5592
14	beef with chinese hot pepper .	1	0.6474

Figura 2.2: Primeros 15 renglones del conjunto de datos “Yelp”

Fuente: Yelp
Campos: text, binary_label, binary_score
Cantidad: 2 millones de instancias.
Objetivo: Identificar sesgos en reseñas y opiniones.

En el campo **text** contiene una reseña/opinion y en el campo **binary_label** contiene los siguientes valores: female (0), male (1)

	text	binary_label	binary_score	ternary_label	ternary_score
0	hi , how are you doing ? i'm getting ready to ...	1	0.6522	2	0.4496
1	i am ! for my hobby i like to do canning or so...	1	0.6505	2	0.8413
2	that's neat . when i was in high school i plac...	1	0.8615	1	0.5290
3	i do not . but i do have a favorite meat since...	0	0.6487	2	0.5111
4	i would have to say its prime rib . do you hav...	0	0.5697	2	0.5346
5	do you have anything planned for today ? i thi...	0	0.7657	2	0.5458
6	i think i will can some jam . do you also play...	1	0.8343	2	0.5547
7	hi , how are you doing today ?	1	0.5223	1	0.3620
8	wow , four sisters . just watching game of thr...	0	0.7858	0	0.6612
9	i agree . what do you do for a living ?	1	0.6433	2	0.5219
10	interesting . i'm a website designer . pretty ...	0	0.6294	2	0.4271
11	that's awesome . i have always had a love for ...	1	0.8084	2	0.5211
12	i really enjoy free diving , how about you , h...	1	0.7435	2	0.5950
13	that's nice . moms are pretty cool too .	0	0.8143	0	0.7246
14	we all live in a yellow submarine , a yellow s...	1	0.5129	2	0.6538

Figura 2.3: Primeros 15 renglones del conjunto de datos “ConvAI2”

Fuente: ConvAI2
Campos: text, binary_label, binary_score, ternary_label, ternary_score.
Cantidad: 130,000 instancias.
Objetivo: Examinar el sesgo en conversaciones simuladas.

En el campo **text** contiene una conversación simulada, en el campo **binary_label** contiene los siguientes valores: female (0), male (1) y en el campo **ternary_label** contiene los siguientes valores: female (0), male (1), neutral (2).

La selección del conjunto de datos md_gender_bias se basa en su amplia y variada representación de contextos comunicativos, desde conversaciones y reseñas hasta descripciones de imágenes y diálogos, ofreciendo un campo amplio para el análisis del sesgo de género. La diversidad de tipos de datos y el volumen de instancias posibilitan una investigación detallada de los patrones de sesgo, apoyando un análisis enriquecido.

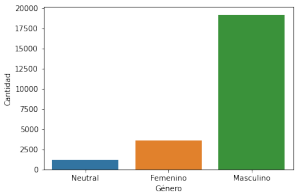
La inclusión en el conjunto de datos de campos específicos relacionados con el género en cada contexto permite una evaluación detallada del sesgo de género desde varias perspectivas, incluidas las del hablante, receptor y sujeto. Esta estructura es fundamental para identificar cómo se manifiesta el sesgo de género en el lenguaje y es crucial para desarrollar métodos más efectivos para su identificación y mitigación.

2.2 *Análisis exploratorio*

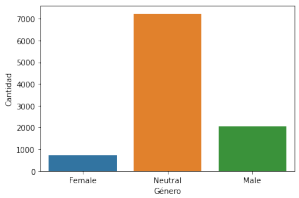
Antes de avanzar hacia visualizaciones gráficas complejas, es crucial realizar un análisis exploratorio para entender la estructura y las particularidades del conjunto de datos md_gender_bias. Esta etapa inicial se centra en examinar la distribución de género a lo largo de los diferentes segmentos del conjunto de datos, prestando atención a cualquier tendencia o patrón que pueda sugerir la presencia de sesgos de género.

Se comenzará con la creación de gráficas de barras que ilustren la distribución de género dentro de los distintos subconjuntos de datos. Para este propósito, se emplea un **Jupyter Notebook** en **Visual Studio Code** , utilizando las bibliotecas **matplotlib.pyplot** y **seaborn** para la visualización, después de organizar los datos en dataframes mediante pandas.

Este enfoque permite una inspección visual inmediata de las proporciones de género, facilitando la identificación de desequilibrios que podrían indicar sesgos. Las gráficas de barras ofrecen una representación clara de cómo se distribuye el género en cada fuente de datos, ajustando la visualización para reflejar las categorías de género relevantes, incluyendo las distinciones entre género neutral, femenino y masculino, según corresponda a la estructura de cada conjunto de datos.

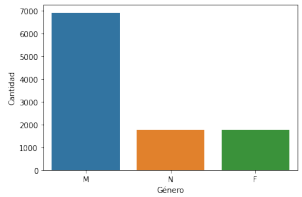


(a) Distribución de Género en conjunto de datos “Funpedia”

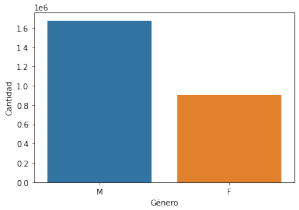


(b) Distribución de Género en “Image_Chat”

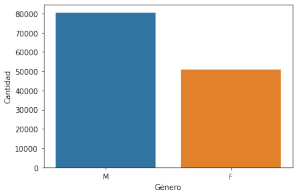
Figura 2.4: Comparación de distribuciones de Funpedia e Image_Chat



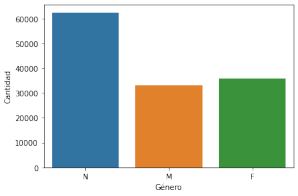
(a) Distribución de Género en conjunto de datos “Wizard”



(b) Distribución de Género en conjunto de datos “Yelp”

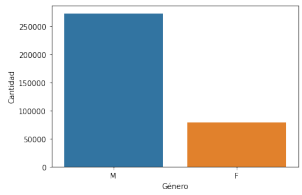


(c) Distribución de Género en conjunto de datos “Convai2 binary_label”

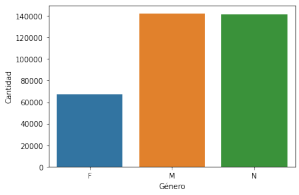


(d) Distribución de Género en conjunto de datos “Convai2 ternary_label”

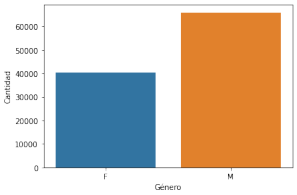
Figura 2.5: Comparación de distribuciones de Wizard, Yelp y, Convai2 en binario y ternario



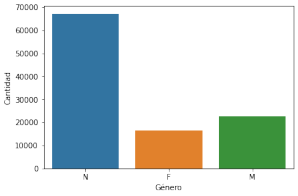
(a) Distribución de Género en conjunto de datos “Open Subtitles binary_label”



(b) Distribución de Género en conjunto de datos “Open Subtitles ternary_label”



(c) Distribución de Género en conjunto de datos “Light binary_label”



(d) Distribución de Género en conjunto de datos “Light ternary_label”

Figura 2.6: Comparación de distribuciones de Dataset Open Subtitles y Light en binario y ternario

Estas gráficas examinadas ponen de manifiesto una marcada disparidad en la distribución de las etiquetas de género, destacando particularmente la subrepresentación de las instancias etiquetadas como ’Female’. Esta tendencia se mantiene constante a través de las figuras

2.4, 2.5, 2.6 donde se observa que las categorías etiquetadas como **Male** y **Neutral** superan en número a las **Female** .

La desigualdad observada en la distribución de género sugiere un sesgo de género inherente en la anotación del conjunto de datos, donde las expresiones o contextos relacionados con las mujeres parecen subrepresentados.

Este análisis inicial, que integra métodos estadísticos con evaluación cualitativa, facilita la identificación de áreas prioritarias y anomalías, proporcionando una base sólida para las fases subsiguientes de modelado y análisis. Los resultados obtenidos preparan el terreno para optimizar las técnicas de modelado, permitiendo abordar con mayor precisión los problemas de sesgo de género en los algoritmos de aprendizaje automático.[10]

2.3 *Descripción de los modelos*

Para analizar el sesgo de género en el conjunto de datos md_gender_bias, se han seleccionado métodos específicos de Procesamiento del Lenguaje Natural (NLP) y Análisis Latente de Dirichlet (LDA), que corresponden a las particularidades de los subconjuntos de datos. Estas herramientas son adecuadas para explorar y revelar estructuras y patrones en textos extensos, lo que es crucial para los objetivos de este estudio. La elección de cada técnica se fundamenta en su eficacia para identificar temáticas y dinámicas lingüísticas en datos voluminosos, facilitando un examen profundo del sesgo de género. Los datasets seleccionados, Yelp y ConvAI2, serán analizados usando LDA para identificar tópicos predominantes y examinar cómo se representa el género dentro de estos tópicos. Para el dataset Wizard, se aplicarán técnicas de NLP para evaluar la distribución del lenguaje y sus implicaciones en la representación de género.

2.3.1 *LDA (Análisis Latente de Dirichlet) para Yelp y ConvAI2:*

LDA es un modelo de tópicos que permite identificar y agrupar conjuntos de palabras y expresiones que aparecen frecuentemente juntas en distintos documentos de un corpus[3]. Opera bajo el supuesto de que los documentos se componen de una mezcla de tópicos, donde un tópico se define como una distribución de palabras. Este método es particularmente útil en conjuntos de datos extensos, ya que puede procesar y analizar grandes volúmenes de texto, revelando los tópicos predominantes sin necesidad de etiquetado previo.

Aquí explico cómo funciona LDA de manera simplificada: **Inicialización** Se asigna aleatoriamente cada palabra de los documentos

(d) en K tópicos posibles.

Procesamiento de Textos con spaCy para Yelp y ConvAI2

El procesamiento de textos es una etapa preliminar que simplifica los textos de Yelp y ConvAI2, centrada en la tokenización para descomponer el texto en unidades básicas, la eliminación de stopwords para descartar palabras que no añaden significado, y la lematización que reduce las palabras a su forma raíz. Estos pasos ayudan a purificar y normalizar los datos, permitiendo que el análisis temático posterior se enfoque en el contenido relevante.

Existen múltiples librerías de procesamiento de lenguaje natural como NLTK, Stanford NLP, spaCy, adecuadas para manejar grandes volúmenes de datos. SpaCy, ha demostrado ser eficaz en el manejo del lenguaje natural en corpus comparables a los de Yelp y ConvAI2, destacando por su rapidez y precisión en la extracción de elementos lingüísticos esenciales [2].

Entrenamiento

LDA va a buscar la probabilidad de que una palabra "p"pertenezca a un tópico "t":

Algorithm 1: Reasignación de palabras a tópicos basada en probabilidades

Data: Conjunto de palabras $P = \{p_1, p_2, \ldots, p_n\}$ y tópicos $T = \{t_1, t_2, \ldots, t_m\}$

Result: Cada palabra p_i reasignada al tópico t_k con la máxima probabilidad condicional

for *cada palabra* $p_i \in P$ **do**

  for *cada tópico* $t_j \in T$ **do**

    Calcular $P(p_i|t_j)$ usando la fórmula:

    $P(p_i|t_j) = P(t_j|d) \cdot P(p_i|t_j)$;

  end

  Encontrar $t_k = \arg\max_{t_j} P(p_i|t_j)$;

  Reasignar p_i al tópico t_k;

end

Este algoritmo calcula la probabilidad condicional de que una palabra pertenezca a uno de los tópicos disponibles, utilizando la fórmula de probabilidad compuesta. Para cada palabra, determina el tópico con la mayor probabilidad y reasigna la palabra a este tópico, optimizando así la asignación de palabras en función de su relevancia

temática dentro de los tópicos identificados.

Este paso se repite iterativamente. En cada iteración, el modelo ajusta las asignaciones de tópicos para las palabras, con el objetivo de mejorar la coherencia de los tópicos: palabras que tienden a aparecer juntas en los documentos deberían tener una alta probabilidad de ser asignadas al mismo tópico.

Resultado

Al final, LDA ofrece dos resultados principales: una distribución de tópicos en cada documento y una distribución de palabras en cada tópico.

Justificación

Se eligió LDA para los datasets de Yelp y ConvAI2 debido a su robustez en el manejo de datos textuales complejos y su capacidad para desvelar la estructura temática subyacente. En el contexto de Yelp, con sus millones de reseñas, LDA facilita la identificación de patrones temáticos que podrían reflejar sesgos de género en las opiniones. Para ConvAI2, de la misma forma, LDA ayuda a identificar patrones temáticos que podrían reflejar sesgos de género en las conversaciones simuladas.

Objetivos

1. Identificar los tópicos predominantes dentro de los conjuntos de datos de Yelp y ConvAI2.
2. Examinar si existen diferencias significativas en la representación de géneros dentro de los tópicos identificados.
3. Utilizar LDA como herramienta para detectar y cuantificar sesgos de género indirectos, manifestados a través de la frecuencia y el contexto de palabras relacionadas con género en los tópicos identificados, ofreciendo una perspectiva más profunda sobre la dinámica de género en los datos analizados.

2.3.2 *Dataset Wizard con Nubes de Palabras:*

Nos centraremos en cuantificar la presencia de cada tópico y asociar estas frecuencias con la identidad de género correspondiente, traduciendo estos datos cuantitativos en representaciones visuales comprensibles donde el tamaño de cada término indica su frecuencia de aparición.

Justificación

La visualización a través de nubes de palabras se vuelve particularmente efectiva aquí, ya que permite una comprensión inmediata de las tendencias y desigualdades presentes en el conjunto de datos, resaltando la representación de género de manera intuitiva. Esta estrategia simplifica el proceso de análisis, centrándose en la relación entre los tópicos y el género, facilitando la identificación visual de desequilibrios o sesgos de género asociados a ciertos tópicos.

Objetivos

1. Utilizar nubes de palabras para obtener perspectivas visuales sobre la distribución de género en relación con los tópicos predefinidos en el dataset "Wizard".
2. Combinar las nubes de palabras con un análisis de frecuencia en el mismo dataset.
3. Revelar áreas donde el sesgo de género es más evidente, a través de la visualización y el análisis de frecuencias.
4. Facilitar la identificación de tópicos que pueden ser propensos a representaciones desiguales de género.

La selección de técnicas como LDA, spaCy para preprocesamiento de textos, y el análisis de frecuencia complementado con nubes de palabras, se basa en un examen cuidadoso de los subconjuntos de datos y las demandas específicas del estudio sobre el sesgo de género.

2.4 *Descripción de las métricas*

Tras aplicar técnicas como LDA para análisis temático, spaCy para el preprocesamiento de textos, y el análisis de frecuencia junto con visualizaciones

mediante nubes de palabras en los subconjuntos del conjunto de datos md_gender_bias, resulta esencial definir métricas específicas que evalúen la efectividad de estos enfoques para descubrir y comprender los sesgos de género. Estas métricas deben permitir una evaluación de cómo cada técnica contribuye a identificar y representar los sesgos de manera equitativa entre géneros. En los siguientes apartados, se explican las razones detrás de la elección de estas métricas.

Coherencia de Tópicos:

La coherencia de un tópico es una métrica que evalúa la coocurrencia de las palabras más representativas dentro de un tópico identificado por un modelo de tópicos, como LDA (Latent Dirichlet Allocation). Se refiere a qué tan semánticamente relacionadas están entre sí estas palabras clave. Una alta coherencia indica que las palabras principales de un tópico comparten fuertes conexiones semánticas, lo que sugiere que el tópico es comprensible e interpretable para los humanos. Por ejemplo, en un tópico relacionado con tecnología, se esperaría encontrar palabras estrechamente relacionadas como ”software”, ”programación”, ”hardware” y ”computadora”.

Al evaluar la coherencia de los tópicos generados por un modelo como LDA, se busca determinar la calidad y relevancia de estos tópicos dentro del corpus analizado. Una alta coherencia sugiere que el modelo ha logrado capturar tópicos significativos y bien definidos, lo que facilita su interpretación y análisis posteriores.

En el algoritmo 2 se presenta la ecuación para el calculo de coherencia de tópicos donde M es el conjunto de las M palabras más probables en el tópico t y Cv representa el calculo de la coherencia.

Algorithm 2: Cálculo de la coherencia de tópicos

$C_v(t) \leftarrow \frac{C_v(t)}{M(M-1)/2}$

Diversidad Léxica:

La diversidad léxica es una métrica que calcula la relación entre el número total de palabras únicas y el número total de palabras en el corpus, después de la limpieza y normalización de los textos utilizando una herramienta como spaCy. Se define como el cociente entre el número de palabras únicas (vocabulario) y el número total de palabras

en el corpus. El algoritmo 3 muestra el procedimiento.

Algorithm 3: Cálculo de la diversidad léxica

$LD \leftarrow \frac{vocab\_size}{total\_palabras}$ Calcular diversidad léxica

Una diversidad léxica más alta indica un amplio rango de vocabulario y tópicos representados en el corpus, sugiriendo una representación rica y variada del lenguaje. Por otro lado, una diversidad léxica baja puede indicar un vocabulario restringido o sesgado, lo que podría limitar la calidad del análisis posterior.

El objetivo de evaluar la diversidad léxica es determinar la efectividad del preprocesamiento de textos realizado con spaCy en enriquecer la calidad del corpus para análisis posteriores. Una diversidad léxica adecuada asegura que la representación de género no esté limitada por un vocabulario restringido o sesgado, permitiendo un análisis más completo y preciso de los patrones y sesgos de género presentes en los datos.

Balance de Género en Tópicos:

El balance de género en tópicos es una métrica que mide la proporción de documentos o menciones asociadas con identidades de género específicas dentro de cada tópico identificado. Se calcula a través de un análisis de frecuencia y la técnica de LDA (Latent Dirichlet Allocation), en el Algoritmo 4 vemos n de t en femenino y masculino que representan la cantidad de datos con sus respectivas etiquetas.

Algorithm 4: Cálculo del balance de género en tópicos

$GB(t) \leftarrow \frac{\text{máx}(n_{t,masculino}, n_{t,femenino})}{\text{mín}(n_{t,masculino}, n_{t,femenino})}$

Un balance equitativo de género en la representación de tópicos indica una inclusión y representación justa de ambos géneros en los tópicos identificados. Por otro lado, desbalances significativos en la proporción de géneros dentro de ciertos tópicos pueden señalar la presencia de sesgos de género en los datos.

Visibilidad de Género en Nubes de Palabras:

La visibilidad de género en nubes de palabras es una métrica que

evalúa la prominencia de palabras asociadas a diferentes géneros en visualizaciones gráficas, basándose en su frecuencia de aparición en el corpus. Se calcula mediante la técnica de nubes de palabras y el análisis de frecuencia.

Algorithm 5: Cálculo de la visibilidad de género en nubes de palabras

$$GV(p) \leftarrow \frac{freq[p]}{total\_palabras}$$

La visibilidad de términos relacionados con género en las nubes de palabras refleja su relevancia y frecuencia en el corpus, proporcionando una medida intuitiva de la representación de género. Una mayor visibilidad de palabras asociadas a un género específico sugiere una mayor presencia y representación de ese género en los datos.

Las métricas descritas, como la coherencia de tópicos, la diversidad léxica, el balance de género en tópicos y la visibilidad de género en nubes de palabras, han sido seleccionadas para medir aspectos relacionados con la representación equitativa de género y la neutralidad en los patrones identificados. El balance de género en tópicos permite identificar desigualdades en la representación de géneros dentro de los tópicos identificados, mientras que la visibilidad de género en nubes de palabras facilita la identificación visual de la prominencia relativa de términos asociados a géneros específicos. Al aplicar estas métricas en el análisis de los conjuntos de datos, se busca encontrar la existencia de sesgos de género implícitos en los datos. La identificación y cuantificación de estos sesgos mediante las métricas propuestas constituye un paso crítico para garantizar que las perspectivas y conclusiones derivadas del análisis mantengan la máxima objetividad y neutralidad posible.

2.5 *Descripción de los experimentos*

La fase de experimentación y análisis detallado es central en el estudio del sesgo de género utilizando LDA, spaCy, y visualizaciones como nubes de palabras. Esta sección explica los enfoques específicos adoptados para cada técnica y subconjunto de datos dentro del md_gender_bias, con el fin de investigar la presencia y las implicaciones del sesgo de género en los hallazgos derivados de estas metodologías.

Carga de Datos del md_gender_bias:

- **Librería**: datasets
- **Procedimiento**: A través de la función load_dataset, se cargan los

datos necesarios directamente en el entorno de Jupyter Notebook, facilitando el acceso inmediato a los conjuntos para su análisis.

Algorithm 6: Cargar y explorar el conjunto de datos de Yelp

```
[1] data ← load_dataset("md_gender_bias", yelp_inferred)
df ← data["train"].to_pandas()
print(df['binary_label'].value_counts())
```

Preprocesamiento con spaCy en los Subconjuntos Yelp y ConvAI2:

Librerías: spaCy, pandas, re.

Algorithm 7: Preprocesamiento de textos con spaCy

```
Data: Texto a procesar
Result: Lista de lemas de tokens no triviales
Function preprocess(Texto):
    nlp ← spacy.load("en_core_web_sm");
    doc ← nlp(Texto);
    tokens ← {};
    foreach token in doc do
        if not token.is_stop and not token.is_punct and not
         token.is_space then
            token_lemma ← token.lemma_.lower();
            if re.match("[a − zA − Z]3,", token_lemma) then
                tokens.append(token_lemma);
    end
    tokens ← {token for token in tokens if token};
return
```

LDA para los Subconjuntos Yelp y ConvAI2:

- **Librerías**: gensim, pandas, numpy, matplotlib y WordCloud.
- **Procedimiento**:
  - Con pandas, se selecciona aleatoriamente el 10 % de los datos de Yelp y el 50 % de ConvAI2 para facilitar el manejo de los recursos computacionales.
  - Preparación de Datos para LDA: Las listas obtenidas con el tokenizador se usan para formar el corpus para LDA utilizando corpora.Dictionary de gensim. Este corpus sirve de base para el entrenamiento del modelo.
  - Configuración y Ejecución del Modelo LDA: Se establece el número de tópicos (35 para Yelp y 20 para ConvAI2) y se entrena

el modelo LDA. La coherencia de tópicos guía la optimización de parámetros.

- Número de Tópicos: Se determina experimentando con un rango de 5 a 50 tópicos, eligiendo aquel que mejora la coherencia temática.

- Visualización de Tópicos con Nubes de Palabras: Para cada tópico identificado por el modelo LDA, se generan nubes de palabras. Se utilizan las librerías matplotlib y WordCloud para generar estas visualizaciones.

Algorithm 8: Aplicar LDA al conjunto masculino

Data: Corpus masculino y diccionario asociado
Result: Modelo LDA entrenado sobre el corpus masculino
Function `LdaModel`(*corpus, dictionary, num_topics, random_state, update_every, chunksize, passes, alpha, per_word_topics*)**:**
    lda_model_masculino ← models.LdaModel(corpus=corpus, id2word=dictionary, num_topics=num_topics, random_state=random_state, update_every=update_every, chunksize=chunksize, passes=passes, alpha=alpha, per_word_topics=per_word_topics);
return

Frecuencias y Nube de Palabras para el Subconjunto Wizard:

- **Librerías Utilizadas**: pandas, matplotlib, WordCloud

- **Procedimiento**:

- Agrupación por Tópico y Género, y Creación de Diccionario de Frecuencias: El dataframe se agrupa de acuerdo a las columnas 'chosen_topic' y 'gender', y se calcula el tamaño de cada grupo. Esto permite computar la frecuencia de cada tópico desglosado por género

  Se transforma la agrupación en un diccionario que mapea cada 'chosen_topic' a su frecuencia correspondiente.

Algorithm 9: Agrupar y contar por tópico y género, crear diccionario de frecuencias

Data: DataFrame df con las columnas 'chosen$_t$*opic*'*y*'*gender*'
Result: Diccionario de frecuencias de los tópicos elegidos
Function `GroupAndCount(`*df*`):`
    df ← df.groupby(['chosen_topic',
     'gender']).size().reset_index(name='count');
    print(df.head(15));
    frequencies ← df.set_index('chosen_topic')['count'].to_dict();
    return *frequencies*;
return

- Generación de Nubes de Palabras: Utilizando la librería WordCloud, se genera una nube de palabras a partir del diccionario de frecuencias. La visualización es ajustada para reflejar la distribución de género mediante colores específicos: rosa para femenino, gris para neutro, y azul para masculino, facilitando así la interpretación visual de cómo diferentes géneros se asocian con los tópicos del dataset Wizard.

Algorithm 10: Generar nube de palabras

Data: Diccionario de frecuencias 'frequencies'
Result: Generación de una nube de palabras visualizada
Function `GenerateWordCloud(`*frequencies*`):`
    wordcloud ← WordCloud(width=800, height=400,
     background_color='white');
    wordcloud.generate_from_frequencies(frequencies);
return

- Aplicación de Función de Color y Visualización: Se define una función de color personalizada que aplica el color correspondiente a cada género basado en la asociación de género de cada tópico. Finalmente, la nube de palabras coloreada se muestra, ofreciendo una representación gráfica inmediata de la relación entre tópicos y género en el conjunto de datos.

Algorithm 11: Algoritmo para aplicar colores en una nube de palabras basado en el género

Data: DataFrame df con las columnas 'chosen$_{t}opic'y'gender'$

Result: Muestra una nube de palabras con colores basados en el género

Function `ColorFunc(`*word*`)`**:**

    gender ← df.loc[df['chosen$_{t}opic'] ==
     word,' *gender*'].*values*[0] **if** *gender* == *0* **then**
        return *grey*;
    else if *gender* == *1* **then**
        return *pink*;
    else
        return *blue*;
    end

plt.figure(figsize=(10, 5));
plt.imshow(wordcloud.recolor(color_func=`ColorFunc()`),
 interpolation='bilinear');
plt.axis('off');
plt.show();

Esta sección ha delineado los procedimientos empleados para evaluar el sesgo de género en el conjunto de datos md_gender_bias. Utilizando una combinación de técnicas de procesamiento de texto con LDA y spaCy, junto con visualizaciones como nubes de palabras, hemos establecido un marco metodológico. Estos enfoques han permitido una inspección temática de los datos, y también han facilitado la visualización de cómo las diferencias de género se manifiestan y pueden ser analizadas.

3 Resultados y discusión

3.1 Resultados y Discusión

A continuación, se discuten los resultados obtenidos de cada uno de los subconjuntos analizados (Yelp, ConvAI2 y Wizard), enfatizando las tendencias identificadas y su implicación en la representación de género.

Yelp: La aplicación del Análisis Latente de Dirichlet (LDA) a los datos de Yelp arrojó un valor de coherencia de tópicos de .67, indicando una buena coherencia en la agrupación temática realizada.

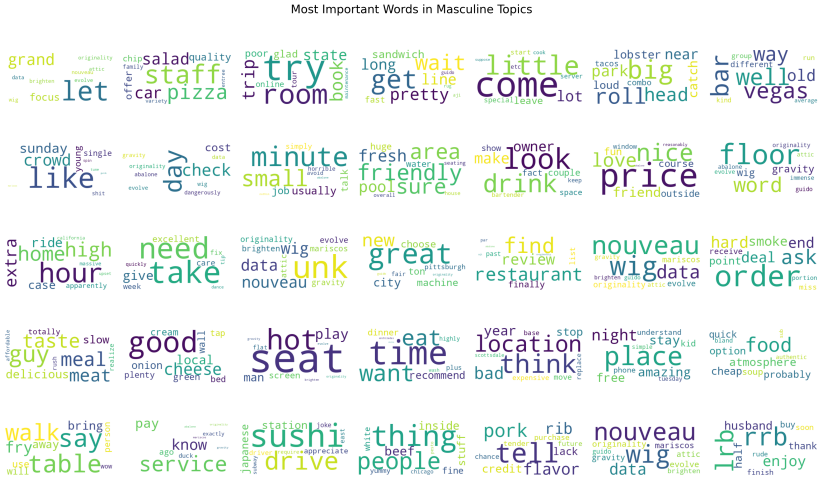


Figura 3.1: Nube de Palabras de Tópicos Masculinos en Yelp

Las palabras asociadas a cada tópico se seleccionan en función de su relevancia o contribución al tópico, algunas palabras pueden ser muy frecuentes en un corpus pero poco informativas sobre el tópico, como preposiciones o palabras comunes. LDA tiende a dar más importancia a palabras que son característicamente relevantes para un tópico, aunque no sean extremadamente frecuentes. Esto ayuda a diferenciar los tópicos

Figura 3.2: Nube de Palabras de Tópicos Femeninos en Yelp

entre sí.

Encontramos una presencia de palabras en tópicos asociados con el género masculino como ”guido”, ”guy”, ”man” que se refieren directamente a personas hombres. Además, las profesiones como ”chef”, ”cook”, ”server”, ”bartender” y ”driver” tradicionalmente han sido ocupaciones con una mayor presencia masculina.

Por otro lado, palabras como ”rib”, ”duck”, ”beef” y ”pork” suelen asociarse a platos de carnes y podrían tener una leve connotación masculina, ya que históricamente el consumo de carne roja ha sido más común entre los hombres.

Entonces, aunque no hay un sesgo absoluto, esas son algunas de las palabras en los tópicos masculinos que podrían sugerir una asociación más frecuente con el género masculino según los estereotipos.

También se identificaron palabras en tópicos asociados con el género femenino como: ”girl”, ”waitress”, ”wife”, ”lady”, ”mom”, ”daughter”. Estas hacen referencia directa a mujeres o roles típicamente asociados con lo femenino.

También, hay algunas otras palabras que podrían tener una leve connotación femenina según estereotipos de género tradicionales: ”salad”, ”cheese”, ”tomato”, ”cream”, ”butter”, ”baked”, ”sauce”, ”dessert”, ”bacon”, ”sandwich”, ”vegetable”, ”cookie”, ”almond”. Algunos alimentos como ensaladas, quesos, tomates, cremas, mantequilla, productos horneados, salsas, postres, galletas y almendras a menudo se han vinculado más con preferencias y roles culinarios femeninos.

Hay ausencia de palabras claramente asociados al género masculino en contraste con los otros tópicos que sí contenían algunas palabras con

connotaciones más masculinas.

La presencia de ciertas palabras parece apuntar a un lente o narrativa con una perspectiva más orientada a lo femenino según roles de género convencionales.

Las siguientes palabras se repiten en los grupos de tópicos de ambos géneros:

”entertained”, ”detract”, ”sala”, ”furnish”, ”wind”, ”tandoori”, ”restore”, ”preferably”, ”nostalgic”, ”mike”

Estas palabras por la naturaleza de las críticas de restaurantes en Yelp, se inclinan hacia un contexto de servicios de hospitalidad o restaurantes.

La presencia de ciertas palabras sí parece reflejar ciertas inclinaciones sesgadas según los roles de género convencionales, si bien no se muestra un sesgo de género absoluto, el análisis revela la existencia de ciertos sesgos implícitos de género en los tópicos.

ConvAI2: En ConvAI2, el valor de coherencia de tópicos fue de .61

Figura 3.3: Nube de Palabras de Tópicos Masculinos en ConvAI2

Encontramos una presencia de palabras en tópicos asociados con el género masculino como: ”dad”, ”son”, ”husband”, ”guy”, ”man” las cuales hacen referencia directa al género masculino. Otras palabras como ”driver”, ”truck”, ”car”, ”basketball”, ”hockey”, ”band”, ”guitar”, ”rock”, y las referencias a "beer"también tienden a asociarse más con intereses y actividades estereotípicamente masculinas.

Parece contener un número considerable de palabras que sugieren un sesgo o inclinación hacia experiencias, actividades e intereses típicamente vinculados a la masculinidad según los estereotipos de género tradicionales.

También se identificaron palabras en tópicos asociados con el

Figura 3.4: Nube de Palabras de Tópicos Femeninos en ConvAI2

género femenino como: ”wife”, ”mom”, ”daughter”, ”girl”, ”girlfriend” hacen referencia directa al género femenino. Otras como ”waitress”, ”married”, ”single”, ”parent”, ”child”, ”baby” y ”pregnancy” suelen asociarse más con roles e experiencias femeninas tradicionales. Además, ”color”, ”hair”, alimentos como ”vegetable”, ”cake”, ”cookie”, ”vegan”, ”veggie”, actividades como ”cook”, ”bake”, ”garden”, mascotas como ”cat”, ”allergy”, emociones como ”sad” y descripciones como ”beautiful”, también tienden a tener una asociación mayor con el género femenino según estereotipos convencionales.

Los tópicos femeninos contienen un número importante de palabras que sugieren un sesgo hacia experiencias, actividades, roles e intereses típicamente vinculados a la femineidad según los roles de género tradicionales.

Las palabras que aparecen en ambos como ”time”, ”sound”, ”movie”, ”game”, ”food”, ”eat”, ”hope”, ”nice”, ”people” son bastante genéricas y podrían aparecer en muchos contextos.

La presencia de ciertas palabras parece reflejar sesgos de género bastante marcados. Se podría afirmar que sí existe un sesgo de género considerable en este conjunto de datos.

Los tópicos se pueden encontrar en la sección de anexos. 4.2

Wizard:

El análisis de frecuencias y la visualización mediante nubes de palabras del subconjunto Wizard permiten observar cómo cada tópico

presente en el conjunto de datos, está asociado a un género específicos, ya sea masculino, femenino ó neutral. La predominancia de personas famosas como "Stephen King", "Bruno Mars", "Justin Bieber.[en] contextos masculinos, frente a la de contextos femininos como "Donna Karan", "Jane Austinïlustra cómo las normas culturales y los prejuicios pueden influir en la curación de conjuntos de datos, en especial en el balance del dataset donde pueda existir la misma representación para ambos géneros.. La nube de la figura 3.5 reveló una mayor presencia de etiquetas masculinas, lo que confirma la hipótesis inicial de sesgo en la composición del conjunto de datos.

Primeros 15 renglones de las frecuencias y género en Wizard

	chosen_topic	gender	count
0	Adam Levine	2	206
1	Alexander McQueen	2	48
2	Alfred Hitchcock	2	77
3	Alison Sudol	1	10
4	Amazon Echo	2	42
5	Amazon Kindle	0	44
6	Arnold Schwarzenegger	2	155
7	Bachelor's degree	0	86
8	Baltimore Orioles	2	143
9	Barbie Girl	0	46
10	Beastie Boys	2	164
11	Black Jack (gum)	2	16
12	Black Rock Desert	1	28
13	Bob Ross	2	130
14	Bon Iver	2	70
15	Border Collie	0	217

Estos hallazgos destacan una representación desequilibrada de género, con una tendencia hacia la sobre-representación de lo masculino y la subrepresentación de lo femenino.

Nuestro análisis se ve reforzado por los valores de coherencia de tópicos, lo que indica que los tópicos identificados son significativos y representativos de los datos analizados. La evidencia de sesgo en la visualización de los datos subraya la necesidad de abordar estas desigualdades en las etapas de curación y etiquetado de datos, así como en la selección de textos para el dataset, para desarrollar tecnologías

Figura 3.5: Nube de Palabras de frecuencias en Wizard - masculinas(azules), femeninas(rosas) y gris(neutral)

más inclusivas y equitativas.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

La exploración de 3 datasets del conjunto de datos md_gender_bias ha desvelado la presencia de sesgos de género, corroborando la hipótesis inicial de que la composición mayoritariamente masculina del equipo de etiquetado influiría en la representación de género. Desde el análisis exploratorio inicial, que reveló una visibilidad menor de las etiquetas femeninas a través de simples visualizaciones de barras, hasta el análisis cualitativo proporcionado por las técnicas de NLP y LDA, cada paso ha subrayado una tendencia hacia la sobre-representación masculina.

El empleo de herramientas como spaCy para NLP y gensim para LDA, fundamentado en una exploración completa de los datos, permitió un enfoque hacia la identificación y análisis de los sesgos presentes. Este método confirmó la validez del análisis con alta coherencia, sino que también, a través de una visualización directa, evidenció la predominancia de etiquetas masculinas, sugiriendo un sesgo implícito en la curación de los datos.

El análisis de los tópicos extraídos refleja una interpretación cualitativa de los datos que, aunque dependiente de la subjetividad del investigador, ofrece una ventana a las tendencias y sesgos subyacentes en el conjunto de datos. Este enfoque reveló una desproporción en la representación de tópicos de género, destacando cómo el sesgo de género es un reflejo de la cantidad de datos, y la calidad y profundidad temática.

La significativa presencia de sesgo en un conjunto de datos ampliamente utilizado para la evaluación y desviación de modelos de lenguaje natural es alarmante. Refleja desigualdades en la representación de género y también cómo las normas y prejuicios culturales pueden permear en la curación de conjuntos de datos. Este estudio enfatiza la urgencia de adoptar enfoques conscientes en la compilación de datos y desarrollar estrategias para contrarrestar los sesgos durante el aprendizaje automático.

Además, este trabajo insta a reconsiderar el propósito de conjuntos

de datos como md_gender_bias en la promoción de una tecnología más equitativa e inclusiva. Aunque estos datos pueden reflejar las dinámicas y sesgos del mundo actual, es crucial cuestionar si perpetuar estas condiciones es compatible con el objetivo de avanzar hacia una tecnología que desafíe las estructuras de poder existentes. En lugar de replicar el status quo, es esencial integrar principios de equidad, diversidad e inclusión en todas las etapas del ciclo de vida de los datos para fomentar el desarrollo de tecnologías que sean justas y representativas para todos.

En conclusión, este estudio contribuye al entendimiento del sesgo de género en los modelos de aprendizaje automático y también resalta la necesidad de una vigilancia continua y acción correctiva. La tecnología debe reflejar un compromiso con los principios de igualdad de género y diversidad, garantizando su funcionamiento de manera justa y equitativa para todos. Este desafío requiere un replanteamiento fundamental de cómo se recopilan, curan y utilizan los conjuntos de datos, enfatizando la creación de marcos que promuevan una representación más equitativa y la construcción de futuros tecnológicos inclusivos.

4.2 *Trabajo futuro*

La investigación presentada ha dejado en claro la presencia y las implicaciones del sesgo de género dentro de los conjuntos de datos utilizados para entrenar modelos de lenguaje natural, apuntando hacia la necesidad de desarrollar estrategias efectivas para mitigar estos sesgos. Como continuación de este trabajo, se proponen varias líneas de acción y estudio para avanzar hacia la creación de tecnologías más inclusivas y equitativas:

- Desarrollo de un Marco Integral para la Mitigación del Sesgo: La creación de una metodología que guíe la recolección, curación y utilización de conjuntos de datos para el aprendizaje automático, enfocado en minimizar los sesgos de género. Este marco debe incorporar técnicas para la identificación temprana de sesgos, estrategias de corrección y pautas para la inclusión equitativa de género en todos los aspectos del desarrollo de modelos de IA.
- Diversificación de Equipos de Curación de Datos: Aumentar la diversidad entre los equipos encargados de recopilar y etiquetar los conjuntos de datos. La inclusión de perspectivas variadas puede ayudar a identificar y mitigar sesgos implícitos durante las etapas tempranas del proceso de curación de datos.
- Desarrollo de Tecnologías de Debiasing Automatizado: Investigación y desarrollo de herramientas automatizadas que puedan detectar

y corregir sesgos de género en grandes conjuntos de datos. Estas tecnologías, basadas en algoritmos avanzados de IA, podrían ofrecer una solución para mejorar la equidad de género en los sistemas de IA.

- Evaluación Continua de Modelos de IA: Implementar sistemas de evaluación continua que monitoreen y analicen el rendimiento de los modelos de IA en términos de equidad de género. Esto incluye el desarrollo de métricas específicas para evaluar la representación y el tratamiento de diferentes géneros, garantizando que los modelos se ajusten y mejoren constantemente.
- Promoción de la Investigación Interdisciplinaria: Fomentar la colaboración entre expertos en IA, sociólogos, psicólogos y especialistas en estudios de género para desarrollar una comprensión más profunda de cómo se manifiestan los sesgos de género en los sistemas tecnológicos y cómo estos pueden ser abordados.
- Educación y Sensibilización: Educar a los desarrolladores de IA, curadores de datos y al público general sobre la importancia de la equidad de género en la tecnología. Programas de formación y campañas de sensibilización pueden ayudar a crear conciencia sobre los impactos negativos de los sesgos de género y promover prácticas más inclusivas.

Este trabajo abre la puerta a una amplia gama de proyectos futuros, desde la creación de nuevos conjuntos de datos libres de sesgo hasta el desarrollo de algoritmos que incorporen principios de equidad desde su concepción. La tarea de mitigar el sesgo de género en la tecnología es compleja y requiere un compromiso a largo plazo, pero es fundamental para asegurar que los avances en IA beneficien equitativamente a toda la sociedad.

Bibliografía

[1] C. Criado Perez, *LA MUJER INVISIBLE: DESCUBRE CÓMO LOS DATOS CONFIGURAN UN MUNDO HECHO POR Y PARA LOS HOMBRES*. Los Tres Mundos, BARCELONA: Seix Barral, 2020.

[2] S. Foci, "Corpus analysis with spacy," 2023.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, jan 2003. Submitted 2/02; Published 1/03.

[4] G. Morales Ramírez, "Problemática antropológica detrás de la discriminación generada a partir de los algoritmos de la inteligencia artificial," *Mundomys Educación*, vol. 34, no. 2, pp. 1–16, 2023.

[5] H. M. Varela Guinot, "Iguales, pero no tanto: El acceso limitado de las mujeres a la esfera pública en méxico," *CONfines relacion. internaci. ciencia política*, vol. 8, no. 16, pp. 39–67, 2012.

[6] P. Escribano López, "Mujeres en, por y para la ciencia women in, by and for science," *Dossiers Feministes*, vol. 14, pp. 151–174, 2010.

[7] O. of American States. Secretary General, *Desigualdad e inclusión social en las Américas : 14 ensayos*. No. OEA/Ser.D/XV.11 in OAS. Documentos oficiales, Organization of American States, 2014.

[8] U. de Igualdad de Género (UIG) de la Comisión Nacional de los Derechos Humanos (CNDH), "Política elaborada por la uig del a cndh," tech. rep., Comisión Nacional de los Derechos Humanos, Ciudad de México, nov 2020. Nasnia Oceransky Woolrich, Titular de la UIG. Jessica Marjane Durán Franco, Jefa de Departamento de Política Institucional de la UIG.

[9] E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams, "Multi-dimensional gender bias classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 314–331, Association for Computational Linguistics, nov 2020.

[10] J. Diaz-Ramirez, "Aprendizaje automático y aprendizaje profundo," *Ingeniare. Rev. chil. ing.*, vol. 29, no. 2, pp. 180–181, 2021. [citado 2024-04-07].

Anexos

Tópicos Masculino en Yelp

- Topic 0: let, grand, focus, wig, data, originality, evolve, nouveau, attic, brighten
- Topic 1: staff, pizza, salad, car, quality, offer, chip, family, variety, entree
- Topic 2: try, room, book, trip, state, glad, poor, online, tour, maintenance
- Topic 3: get, wait, pretty, long, line, sandwich, fast, guido, aji, rug
- Topic 4: come, little, lot, leave, special, start, server, etc, cook, suppose
- Topic 5: big, roll, head, park, near, catch, lobster, loud, combo, tacos
- Topic 6: bar, well, vegas, way, old, different, group, run, kind, average
- Topic 7: like, crowd, sunday, young, single, shit, spin, tune, guido, mariscos
- Topic 8: day, check, cost, wig, data, originality, evolve, abalone, gravity, dangerously
- Topic 9: minute, small, usually, job, talk, simply, horrible, avoid, nothin, abalone
- Topic 10: friendly, area, sure, fresh, pool, huge, water, seating, overall, house
- Topic 11: look, drink, make, owner, couple, fact, show, space, bartender, keep
- Topic 12: price, nice, love, friend, fun, course, outside, window, reasonably, abalone
- Topic 13: floor, word, wig, gravity, originality, evolve, abalone, attic, immense, guido

- Topic 14: hour, home, high, extra, ride, case, apparently, massive, california, upset
- Topic 15: take, need, give, excellent, care, week, fix, tip, quickly, dance
- Topic 16: unk, wig, nouveau, data, originality, evolve, mariscos, gravity, attic, brighten
- Topic 17: great, new, city, machine, choose, ton, pittsburgh, fair, originality, guido
- Topic 18: find, restaurant, review, finally, list, past, par, abalone, originality, aji
- Topic 19: wig, nouveau, data, originality, evolve, mariscos, gravity, attic, brighten, guido
- Topic 20: order, ask, hard, deal, end, smoke, point, receive, miss, portion
- Topic 21: guy, taste, meal, meat, delicious, slow, totally, realize, rush, affordable
- Topic 22: good, cheese, local, onion, wall, cream, green, bed, plenty, tap
- Topic 23: seat, hot, play, man, screen, flat, originality, evolve, gravity, brighten
- Topic 24: time, want, eat, recommend, dinner, plus, highly, wash, enchiladas, abalone
- Topic 25: think, location, bad, year, stop, move, expensive, base, replace, scottsdale
- Topic 26: place, night, stay, amazing, free, kid, understand, phone, tuesday, simple
- Topic 27: food, atmosphere, cheap, option, probably, quick, soup, bland, authentic, sub
- Topic 28: say, table, walk, fry, bring, away, use, will, person, wow
- Topic 29: service, know, pay, ago, duck, exactly, mariscos, abalone, originality, gravity
- Topic 30: sushi, drive, station, japanese, appreciate, driver, joke, east, require, subway
- Topic 31: thing, people, beef, inside, stuff, fine, white, yummy, chicago, pesto

- Topic 32: tell, pork, flavor, rib, credit, lack, purchase, tender, future, chance
- Topic 33: wig, nouveau, data, originality, evolve, mariscos, gravity, attic, brighten, guido
- Topic 34: rrb, l

Tópicos Femenino en Yelp

- Topic 0: food, place, fry, bring, chef, sample, apple, pack, wind, entertained
- Topic 1: flavor, pay, onion, disappointed, nicely, blend, detract, furnish, preferably, entertained
- Topic 2: little, small, family, lot, plate, fun, cool, establishment, deal, japanese
- Topic 3: suppose, interesting, describe, lighting, intimate, detract, sala, entertained, restore, furnish
- Topic 4: menu, taste, server, walk, half, offer, fish, warm, include, cocktail
- Topic 5: salad, cheese, wonderful, tomato, wind, tandoori, detract, sala, furnish, entertained
- Topic 6: look, table, sure, portion, sit, seat, area, dress, morning, wedding
- Topic 7: hand, butter, yes, product, oil, body, meatball, tree, entertained, furnish
- Topic 8: restore, tandoori, mike, preferably, nostalgic, entertained, sala, furnish, wind, detract
- Topic 9: big, different, large, pick, help, cut, chain, color, job, ring
- Topic 10: good, drink, girl, instead, okay, window, refill, baked, wash, site
- Topic 11: find, waitress, meal, night, cream, ice, location, husband, money, spend
- Topic 12: restaurant, burger, day, start, customer, perfect, appetizer, town, end, stop
- Topic 13: unk, like, upscale, cor, entertained, sala, wind, mike, preferably, nostalgic

- Topic 14: wait, dinner, excellent, long, overall, extra, make, finally, woman, course
- Topic 15: selection, kitchen, medium, dry, wrong, rare, cost, pink, slow, properly
- Topic 16: say, tell, hot, tea, egg, french, toast, mom, cold, manager
- Topic 17: price, wine, special, bit, pricey, head, dip, occasion, wise, insect
- Topic 18: order, cook, bar, worth, remember, literally, complaint, certainly, sala, entertained
- Topic 19: friendly, definitely, way, experience, lunch, old, option, patio, quickly, meet
- Topic 20: love, take, recommend, people, dog, actually, lady, absolutely, today, highly
- Topic 21: well, expect, bite, have, employee, unfortunately, nostalgic, tandoori, restore, entertained
- Topic 22: meat, add, texture, somewhat, flavorless, wind, entertained, detract, tandoori, sala
- Topic 23: try, ask, happy, hour, fan, lovely, remain, detract, tandoori, sala
- Topic 24: dish, consider, completely, forget, sala, wind, furnish, preferably, nostalgic, entertained
- Topic 25: come, think, friend, steak, breakfast, probably, return, will, pancake, happen
- Topic 26: pretty, sauce, favorite, give, bacon, hard, flavorful, daughter, beautiful, salmon
- Topic 27: rrb, lrb, chicken, bad, home, house, glass, atmosphere, close, share
- Topic 28: get, amazing, dessert, leave, review, realize, detract, wind, furnish, tandoori
- Topic 29: wife, fresh, enjoy, sandwich, drive, crispy, vegetable, trip, recently, crave
- Topic 30: know, thing, quality, bread, high, olive, basket, dipping, sala, restore
- Topic 31: service, nice, staff, greet, smile, poor, shake, balance, furniture, tandoori

- Topic 32: restore, tandoori, mike, preferably, nostalgic, entertained, sala, furnish, wind, detract
- Topic 33: delicious, need, item, free, treat, coconut, cookie, almond, gluten, enjoyable
- Topic 34: time, eat, entertained, wind, furnish, sala, restore, nostalgic, bertha, preferably

Tópicos Masculino en ConvAI2

- Topic 0: home, use, farm, computer, glad, fall, tech, game, vacation, stay
- Topic 1: people, well, swim, life, fast, lose, soon, crazy, care, think
- Topic 2: like, sport, kind, band, movie, listen, game, hey, car, eat
- Topic 3: great, kid, dad, person, truck, single, driver, fix, car, band
- Topic 4: favorite, read, thing, right, maybe, book, beach, card, coffee, check
- Topic 5: live, enjoy, travel, fish, bad, tell, cook, cold, world, concert
- Topic 6: time, lot, today, see, fly, spare, hey, help, movie, free
- Topic 7: day, okay, spend, son, real, wife, hurt, hey, hear, husband
- Topic 8: play, yeah, long, food, come, hour, guitar, basketball, hockey, mean
- Topic 9: know, living, let, language, hey, car, plane, important, different, band
- Topic 10: yes, sound, job, watch, true, major, movie, game, sport, need
- Topic 11: fun, get, ride, bike, beautiful, exciting, absolutely, online, wonderful, ready
- Topic 12: work, thank, class, ski, university, pretty, kind, finish, hear, help
- Topic 13: love, try, drive, easy, worry, seattle, listen, band, movie, car
- Topic 14: school, lol, art, high, teacher, shop, keep, believe, teach, local
- Topic 15: family, interesting, study, look, speak, english, super, tall, foot, folk

- Topic 16: cool, music, wow, awesome, rock, totally, afraid, course, hair, listen
- Topic 17: silence, good, hello, pizza, visit, have, sushi, hey, pretty, game
- Topic 18: want, haha, friend, sure, walk, place, mountain, stamp, help, late
- Topic 19: nice, new, hope, city, hike, man, trip, pool, europe, neighbor

Tópicos Femenino en ConvAI2

- Topic 0: pet, wow, year, life, wife, funny, move, open, bakery, time
- Topic 1: get, book, need, haha, purple, okay, finish, baby, high, believe
- Topic 2: want, thank, new, come, marry, interesting, currently, apartment, waitress, time
- Topic 3: home, stay, son, mom, public, bake, care, walk, watch, daughter
- Topic 4: like, kid, day, hope, play, true, vegan, future, time, sound
- Topic 5: love, nice, girl, cute, small, town, speak, time, read, sound
- Topic 6: yes, cook, italian, forest, time, office, fun, sound, mom, graduate
- Topic 7: favorite, color, hair, lol, vegetable, cake, cookie, diet, turn, hour
- Topic 8: good, cool, friend, teacher, different, artist, time, mom, sound, awesome
- Topic 9: great, live, yeah, pretty, maybe, hate, veggie, tall, cheese, time
- Topic 10: enjoy, college, garden, meet, ask, sad, die, sing, hike, bird
- Topic 11: think, thing, child, big, brown, grow, beautiful, fan, allergy, horrible
- Topic 12: eat, food, lot, kind, hobby, fast, chef, time, meat, sound
- Topic 13: old, parent, married, happy, girlfriend, miss, single, winter, study, yea
- Topic 14: dog, cat, name, time, awesome, care, long, sound, walk, person

- Topic 15: work, animal, amazing, buy, night, far, everyday, treat, time, sound
- Topic 16: petition, disease, nineteen, mex, tex, subscribe, favor, rate, juanita, organize
- Topic 17: people, sorry, look, hard, movie, close, probably, black, sleep, wear
- Topic 18: hello, family, today, try, sibling, hear, ice, city, author, lonely
- Topic 19: know, school, make, drink, red, write, eye, prefer, real, morning