Sesgos fantásticos:

Qué son y dónde encontrarlos

VALENTIN BARRIERE

PhD por Télécom Paris, Francia. Profesor Asistente del Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile e investigador joven en el Centro Nacional de Inteligencia Artificial (CENIA), Chile. Su área de investigación es la inteligencia artificial.

vbarriere@dcc.uchile.cl

RESUMEN. Los modelos de *deep learning* tienden a aprender correlaciones de patrones en grandes conjuntos de datos. Cuanto más grandes son estos sistemas, más complejos son los fenómenos que pueden detectar y más datos necesitan para esto. El uso de la inteligencia artificial se está volviendo cada vez más ubicuo en nuestra sociedad, y su impacto crece cada día. Las promesas que conlleva dependen, en gran medida, de su uso justo y universal, como el acceso a la información o la educación para todos. En un mundo de desigualdades, pueden ayudar a alcanzar las áreas más desfavorecidas. Sin embargo, tales sistemas universales deben ser capaces de representar a la sociedad, sin beneficiar a algunos a expensas de otros. No debemos reproducir las desigualdades observadas en todo el mundo, sino educar a estas inteligencias artificiales para que las superen. Hemos visto casos en los que estos sistemas usan información de género, raza o incluso clase de maneras que no son apropiadas para resolver sus tareas. En lugar de un razonamiento causal real, se basan en correlaciones espurias, lo que generalmente llamamos sesgo. En este artículo, primero definimos qué es un sesgo en términos generales. Esto nos ayuda a desmitificar el concepto de sesgo, a entender por qué podemos encontrarlos en todas partes y por qué a veces son útiles. En segundo lugar, nos enfocamos en la noción de lo que generalmente se considera un sesgo negativo, el que queremos evitar en *machine learning*, antes de presentar una taxonomía general que contiene los sesgos más comunes. Finalmente, concluimos analizando métodos clásicos para detectarlos, mediante conjuntos de datos especialmente diseñados de plantillas y algoritmos específicos, y también métodos clásicos para mitigarlos.

¿Estamos hablando de justicia?

En la sociedad...

La justicia (*fairness*) en los modelos de inteligencia artificial (IA) es una gran preocupación para la sociedad actual. Los algoritmos sesgados (*biased*) están sacudiendo a la sociedad de formas que nunca imaginamos, a menudo reforzando las desigualdades y la discriminación existentes. Imagine un algoritmo decidiendo quién obtiene un préstamo, quién es contratado o incluso quién obtiene la libertad bajo fianza. Si estos algoritmos están sesgados, pueden apuntar injustamente a ciertos grupos, como las comunidades minoritarias o las mujeres, perpetuando la injusticia y la discriminación.

En la justicia penal, los algoritmos sesgados pueden llevar a tasas más altas de encarcelamiento para las minorías, mientras que en la contratación, pueden favorecer a los hombres sobre mujeres igualmente calificadas. La atención médica tampoco es inmune a este problema: los algoritmos sesgados pueden resultar en recomendaciones de tratamiento de menor calidad para grupos subrepresentados, con consecuencias graves. Las implicaciones éticas de estos sesgos van más allá de la discriminación directa, alcanzando también cuestiones de responsabilidad (*accountability*) y transparencia. Estos algoritmos a menudo operan como cajas negras, haciendo que sus procesos de toma de decisiones sean opacos e inimpugnables. Esta falta de transparencia erosiona la confianza y plantea serios problemas de responsabilidad. ¿Quién es responsable cuando un algoritmo discrimina?

La respuesta no es simple, pero la solución comienza con diversificar los datos, implementar técnicas de detección de sesgos y fomentar la colaboración entre tecnólogos, éticos y legisladores. Concienciar al público sobre estos sesgos digitales es crucial. Si bien los algoritmos prometen eficiencia e innovación, su despliegue ético debe priorizar la justicia y la equidad para garantizar que beneficien a todos por igual, sin profundizar las divisiones sociales existentes.

A medida que la IA se vuelve cada vez más omnipresente, la búsqueda de un alto rendimiento impulsa a los modelos a volverse más complejos, a menudo basándose en asociaciones correlacionales. Sin embargo, también exigimos que estos modelos presenten un comportamiento imparcial, respetando la diversidad y basándose en la causalidad en lugar de la correlación. Sin embargo, no olvidemos que los sesgos están en todas partes, ya que casi nada en el mundo es pura aleatoriedad, las estructuras están en todas partes.

Las matemáticas pueden definirse como el estudio de las estructuras, pero las estructuras similares comparten un sesgo estructural común, lo que puede complicar el esfuerzo de usarlas sin perpetuar correlaciones injustas. Esto crea un dilema significativo: ¿cómo podemos aprovechar los antecedentes (*priors*) útiles del mundo para tomar decisiones sin caer en la trampa de las correlaciones sesgadas no causales que pueden ser dañinas en algunos casos sensibles? Por ejemplo, la tasa de criminalidad es más alta en ciertos subgrupos de la población, una elección probabilística no es una razón causal para saber que un individuo cometerá un delito aunque pueda ser estadísticamente probable según las hipótesis. Observatoire des Inégalités [1] informó que los jóvenes negros y árabes, que son subgrupos de la población con una tasa de criminalidad más alta, son más propensos a ser controlados por la policía. Es una práctica común llamada perfil racial, que sanciona basándose en prejuicios sociales y étnicos, y conduce

¿Quién es responsable cuando un algoritmo discrimina?

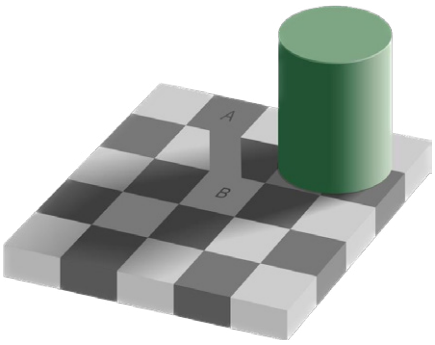


Figura 1. La ilusión de sombra en el tablero de ajedrez.

a un tratamiento injusto y un sesgo sistémico en las prácticas de aplicación de la ley [2]. El desafío radica en eliminar dichos sesgos perjudiciales de los modelos de IA sin sacrificar su capacidad para tomar decisiones informadas y probabilísticas, basadas en las estructuras inherentes del mundo. En otras palabras: ¿cómo eliminar los sesgos perjudiciales en los modelos de IA?

...y en los modelos de IA

Asegurar la justicia en machine learning (ML) es una tarea compleja y desafiante. En primer lugar, los modelos de ML se entrenan con datos del mundo real, que inherentemente contienen sesgos relacionados con la raza, género, religión, clase social o cualquier otro factor. Como resultado, estos modelos pueden no sólo aprender, sino también amplificar estos sesgos preexistentes [3], lo que lleva a resultados problemáticos.

En segundo lugar, a pesar de un entrenamiento y pruebas rigurosas y exhaustivas, crear sistemas que se comporten de manera justa en todas las situaciones y culturas sigue siendo un desafío significativo [4,5]. Por ejemplo, considere un chatbot entrenado en español castellano. Puede funcionar bien al interactuar con usuarios de España, pero puede tener dificultades para entender y responder adecuadamente a la jerga evolutiva utilizada por los adolescentes. Cuando es utilizado por hablantes de Chile o Argentina, el chatbot podría tener dificultades para adaptarse a los modismos únicos y la nueva polisemia de esas regiones, lo que resulta en mala comunicación y frustración (o risas, pero con un rendimiento deficiente). Incluso dentro de su contexto inicial, como entre los castellanoparlantes, el chatbot podría encontrar problemas con dialectos específicos o eventos locales, revelando resultados inesperados e injustos después de su implementación.

En tercer lugar, definir lo que entendemos por "justicia" es inherentemente complejo, ya que no existe un estándar universalmente aceptado, ya sea para decisiones humanas o de máquinas. Determinar los criterios de justicia adecuados para un sistema requiere equilibrar la experiencia del usuario, consideraciones culturales, sociales, históricas, políticas, legales y éticas, cada una con posibles compensaciones. Incluso en situaciones aparentemente sencillas, puede haber desacuerdo sobre qué constituye justicia, ya que puede depender de los valores de los individuos [6,7], complicando el establecimiento de políticas de IA, especialmente en un contexto global.

Finalmente, la IA generativa se está convirtiendo en una parte cada vez más importante de nuestras vidas y genera cada vez más contenido en Internet, que es la principal fuente de datos para entrenar nuevos modelos. El sesgo inherente de los modelos generativos contemporáneos es un "veneno" para los futuros modelos generativos, ya que reducirá la calidad de los datos de entrenamiento futuros [8].

No obstante, es esencial esforzarse por mejorar continuamente hacia sistemas "más justos", y aunque la justicia absoluta sigue siendo elusiva, hay formas de tender hacia ella eliminando los sesgos de los modelos existentes. Pero antes de hablar de desviar la IA, empecemos por lo simple: ¿qué es un sesgo? Mostraremos en lo siguiente que el sesgo tiene muchos sentidos, comenzando por sus diversas definiciones y usos en el lenguaje.

¿Qué son los sesgos?

Definición general de los sesgos

Primero, pensemos en el sesgo matemático de un modelo lineal. Este es el valor que un modelo emitirá cuando las entradas sean cero. Un modelo está sesgado cuando emite un valor diferente de cero al transformar el vector nulo. Este sesgo ayuda al modelo a ajustar los datos, porque todos los datos están sesgados.

En segundo lugar, pensemos en los sesgos cognitivos. Estos son sesgos que todos los humanos compartimos. Pueden verse como una diferencia entre la realidad y nuestra percepción del mundo. Pueden ser muy básicos, como una simple ilusión óptica,[1] o de mayor nivel de complejidad, construidos sobre conceptos sociales. A un nivel de complejidad más bajo, se puede ver un ejemplo en la Figura 1, donde se aprecia un tablero de ajedrez con un cilindro sobre él: los cuadrados oscuros fuera de la sombra parecen más oscuros que los cuadrados blancos en la sombra, aunque en

1 Que actúa muy parecido a un sesgo cognitivo [9].

realidad tienen el mismo color gris oscuro. Otro caso en el que podemos separar grupos de personas según su sesgo es el famoso vestido dorado y blanco, versus negro y azul.[2]

A un nivel de complejidad más alto, podemos mencionar, entre otros, el efecto Dunning-Kruger, el sesgo de disponibilidad o el sesgo de confirmación. El efecto Dunning-Kruger, ilustrado en la Figura 2, es la tendencia de un individuo con conocimientos o competencias limitadas en un campo dado a sobreestimar sus propias habilidades en ese campo.[3] Uno de los mayores libros de divulgación sobre sesgos cognitivos humanos es "Pensar rápido, pensar despacio" de Kahneman [10], quien ganó un Premio Nobel en Economía por su trabajo de vida sobre psicología del juicio y la toma de decisiones. Estos sesgos pueden verse como una desviación entre nuestra percepción y la realidad.

En tercer lugar, podemos nombrar los sesgos sociales, como las normas sociales que son sesgos culturales. Puede ser tan simple como la forma adecuada de vestirse, pero también implica aspectos de comportamiento. Las expectativas de las personas de diferentes grupos sociales variarán con respecto al grupo, dependiendo de lo que se espera de uno de sus miembros. Cuando dos individuos de diferentes culturas no conocen las normas sociales de los otros, pueden experimentar lo que se conoce como un *choque cultural*. Jonathan H. Turner[4] define el choque cultural como las "diferencias en los valores y creencias culturales que ponen a las personas en conflicto entre sí". En algunas culturas, decir "no" puede verse como una señal de debilidad, mientras que en otras como una señal de fortaleza. Tenga en cuenta que las normas sociales dependen de los grupos sociales, por lo que incluso puede encontrar diferencias dentro de una cultura, las normas pueden cambiar con respecto al grupo social: no puede tener el mismo tipo de comportamiento como hombre o como mujer en muchos países.

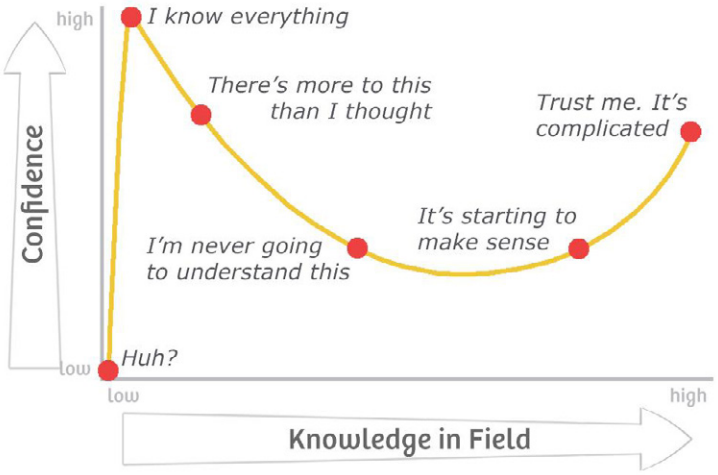


Figura 2. Efecto Dunning-Kruger.

Los sesgos no son fundamentalmente malos, simplemente son una desviación de una norma o valor (subjetiva y definida). Por ejemplo, los sesgos cognitivos provienen de la estructura de nuestro cerebro, que no es aleatoria. Todos compartimos una parte significativa de nuestro ADN, que podría verse como no sesgado si estuviera centrado en cero: una secuencia aleatoria de nucleótidos,[5] como una distribución uniforme centrada en cero. Las diferencias en el color del cabello, de los ojos o de la piel, en la forma de la nariz o de la boca pueden ser representativas de un fenotipo, que incluye patrones similares en el ADN, ¡sesgos de nuevo al comparar un grupo con otro!

Dependemos de los sesgos cada vez que tomamos una decisión, son sumamente útiles para seleccionar las opciones más probables. En su artículo ACL, Meister et al. [11] muestran que un sesgo intuitivo para un modelo de lenguaje[6] es la frecuencia de uso actual de las diferentes palabras. Refleja la forma de hablar, que es un sesgo: sabes que un argentino y un chileno no usarán las mismas palabras. Para esto, Meister et al. inicializan el término de sesgo de la capa lineal final del modelo de lenguaje con la distribución de log-unigramas de las palabras. Esto ayuda a usar conocimiento (*prior knowledge*) previo para minimizar la pérdida (*loss*) de una manera muy *naive* y vincula las dos nociones de sesgo que introdujimos antes.

2 https://upload.wikimedia.org/wikipedia/en/2/21/The_dress_blueblackwhitegold.jpg.

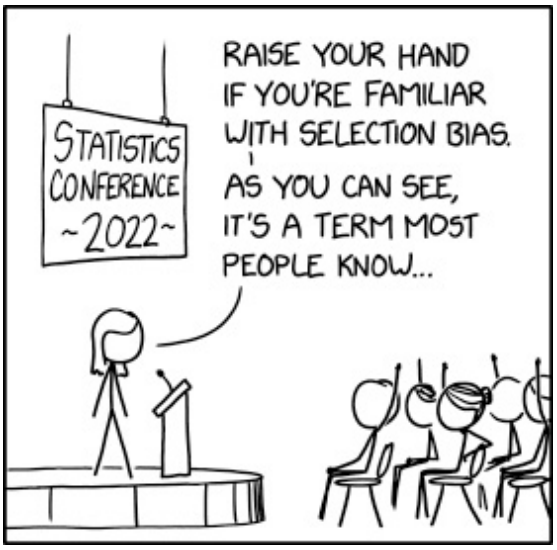
3 El sesgo de confirmación es la tendencia del cerebro a valorar nueva información que apoya ideas existentes, y el sesgo de disponibilidad es la tendencia del cerebro a concluir que una instancia conocida es más representativa del todo de lo que realmente es.

4 Profesor de sociología en la Universidad de California, Riverside.

5 A, T, G y C.

6 Un modelo destinado a predecir la siguiente palabra conociendo las anteriores, como el utilizado al escribir un texto en un smartphone.

El desafío radica en eliminar dichos sesgos perjudiciales de los modelos de IA sin sacrificar su capacidad para tomar decisiones informadas.



Fuente: https://xkcd.com/2618/

Figura 3. Un ejemplo de sesgo de cobertura de los cómics de xkcd #2618.

Taxonomía de los sesgos comunes en ciencia

Existe una taxonomía completa de sesgos, los que provienen de una brecha entre la percepción y la verdad. Pueden en última instancia impactar la validez y la fiabilidad de los hallazgos, lo que lleva a una mala interpretación de los datos. Más sesgos se describen en el artículo "Types of Bias in Research: Definition & Examples".[7]

Sesgo de reporte. Este sesgo ocurre cuando la frecuencia de eventos, propiedades y resultados registrados en un conjunto de datos no refleja con precisión su prevalencia en el mundo real. A menudo surge porque las personas tienden a documentar circunstancias inusuales o memorables, asumiendo que los eventos ordinarios no valen la pena mencionarse: el fenómeno existe pero la gente no lo reporta. A otro nivel, puede resultar de la tendencia a publicar sólo experimentos exitosos o resultados positivos, lo que crea una percepción sesgada de la efectividad de un modelo y da lugar a malentendidos sobre sus verdaderas capacidades y limitaciones.

Sesgo de automatización. Es la tendencia a preferir los resultados producidos por sistemas automatizados sobre aquellos producidos por sistemas no automatizados, independientemente de las tasas de error respectivas. Así es como se llega a un chatbot deficiente que no entiende la solicitud especial de un usuario, que debería ser manejada por un humano.

Sesgo de selección. Ocurre si los ejemplos de un conjunto de datos se eligen de una manera que no refleja su distribución en el mundo real.[8] Este sesgo puede tomar varias formas. Por ejemplo, el sesgo de cobertura surge cuando algunos grupos están insuficientemente representados en los datos de entrenamiento, como encuestar en un aula de Ciencias de la Computación para sondear cuánto sabe la gente sobre programación (ver Figura 3). El sesgo de participación ocurre cuando sólo las personas interesadas responderán a un estudio y el sesgo de muestreo se debe a una no aleatorización de las respuestas (¡o de los datos durante el entrenamiento!).

Sesgo de representación. Este sesgo es algo parecido al sesgo de selección pero con una sutileza. Ocurre cuando los datos recopilados sólo representan un subgrupo de la población, aunque representen la realidad. El hecho de que en su mayoría los hombres sean CEOs no significa que el género sea una característica del éxito como CEO.

Sesgo de atribución de grupo. Implica generalizar en exceso las características, basándose en observaciones limitadas de individuos, al grupo completo al que pertenecen. Puede ser un sesgo de grupo interno, que tiende a favorecer a los individuos del mismo grupo que el experimentador, o sesgo de homogeneidad de grupo externo que tenderá a percibir a los miembros de un grupo externo como más similares entre sí de lo que realmente son, como un científico de datos que habla castellano creando una categoría para el español castellano y otra para el resto de las variaciones.

7 https://www.scribbr.com/category/research-bias/.

8 Es diferente del sesgo de reporte.

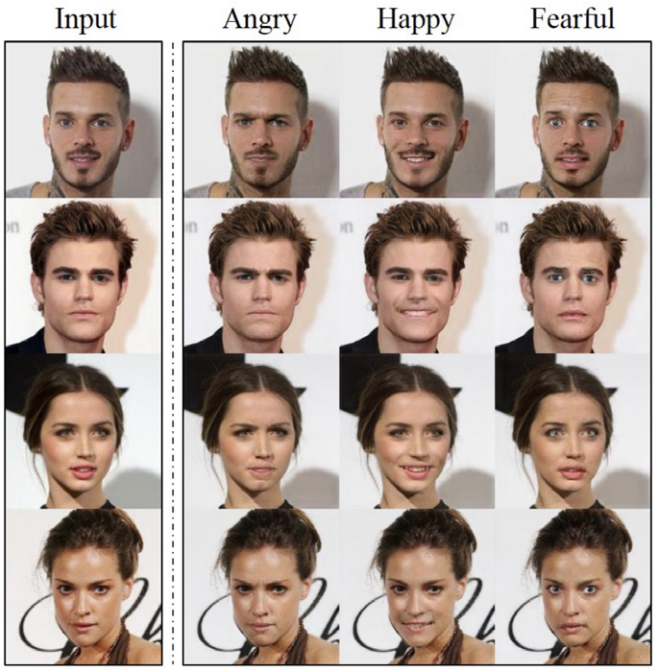


Figura 4. El modelo Stargan, entrenado sobre conjuntos de datos que contienen principalmente caras blancas jóvenes como CelebA.

Los modelos de machine learning se entrenan con datos del mundo real, que inherentemente contienen sesgos [...] Estos modelos pueden no sólo aprender, sino también amplificar estos sesgos.

Sesgo implícito. Este sesgo ocurre cuando se hacen suposiciones basadas en los propios modelos mentales y experiencias personales que no necesariamente se aplican de manera más general. Se llama sesgo de confirmación cuando se procesan los datos de manera que afirmen creencias e hipótesis preexistentes, como descartar fuera de distribución sin causa, o sesgo del experimentador cuando se condiciona el experimento para alcanzar la conclusión esperada.[9]

Ejemplo de sesgos en ML, NLP y LLMs

Hoy en día, los modelos necesitan muchos más datos para entrenarse, pero los datos disponibles son los que hay en Internet, lo que no necesariamente representa el mundo real. Y aunque reflejen con precisión la distribución de la vida real, no hay garantía de que no estén sesgados, ya que el mundo real en sí contiene sesgos inherentes. De hecho, con distribuciones sesgadas, el modelo tenderá a sobreajustarse a características espurias específicas: aquí vienen los sesgos.

En el procesamiento del lenguaje natural (NLP), los sesgos están en todas partes, comenzando por los datos [12], las anotaciones [5,13] e incluso las instrucciones de la campaña de anotación [14]. Entre otras cosas, los modelos de NLP pueden arrastrar sesgos morales [15], sociales [16] o políticos [17]. La cuantificación del sesgo social es un tema destacado en la investigación reciente. Puede estar en datos multimodales como el etiquetado de imágenes [18] o simplemente en texto general [19].

En general, los principales sesgos provienen de sesgos de selección. La mayor parte de los datos de Internet están centrados en Occidente. Los modelos generativos como StarGAN ([20]; ver Figura 4) principalmente creaban caras de personas blancas debido a su conjunto de datos de entrenamiento (CelebA de Liu et al. [21]). Esto sigue ocurriendo con los nuevos modelos como Stable Diffusion o Dall-E2. Aunque ahora estos modelos son más diversos, crean estereotipos generando imágenes de hombres musulmanes con la cabeza cubierta, empleadas de limpieza mujeres, pero las personas productivas las identifica como hombres blancos y las personas que acceden a servicios sociales como de piel oscura [22]. Esto sigue siendo la principal fuente de sesgo para los grandes modelos de lenguaje (LLMs) hoy en día.

Sesgo de selección de características. En los algoritmos de machine learning tradicionales, el sesgo de selección a nivel de características puede ser un problema. Por ejemplo, priorizar características como el nivel de ingresos y el vecindario en un modelo de aprobación de préstamos puede inducir sesgos

9 Un ejemplo de sesgo del experimentador se describe en https://xkcd.com/882/.

Los sesgos no son fundamentalmente malos, simplemente son una desviación de una norma o valor (subjetiva y definida).

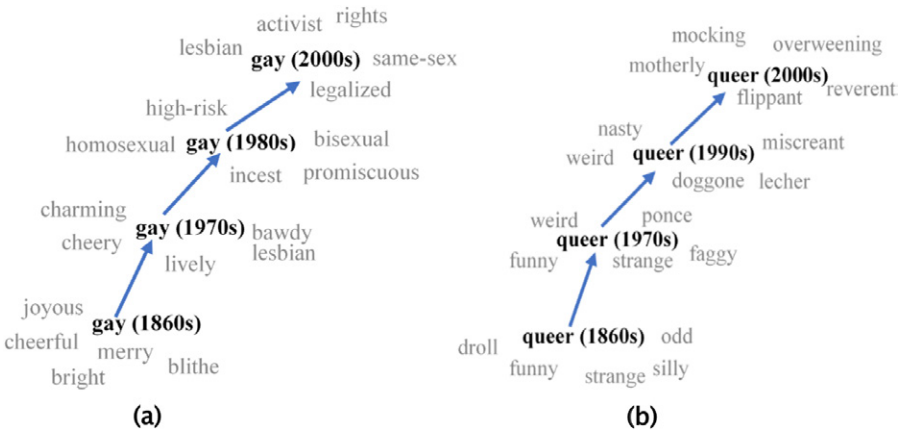


Figura 5. Evolución del significado de las palabras *gay* y *queer* a lo largo del tiempo [32].

socioeconómicos, perjudicando a los solicitantes de áreas de bajos ingresos. Esto se debe a que son variables confusas (*confounding variables*) [23,24], creando correlaciones espurias entre las variables objetivo y de entrada. Estas características sesgadas pueden llevar a predicciones sesgadas del modelo, perpetuando la desigualdad.

Sesgo de anotación. Al igual que con los datos, las anotaciones pueden verse afectadas por un claro sesgo de selección. De hecho, debido a las preferencias culturales, las personas reaccionarán de manera diferente a varios fenómenos subjetivos como el discurso de odio o la aceptabilidad social. Santy et al. [5] muestran que para estas dos tareas las anotaciones varían con respecto a la demografía de los anotadores.

Sesgo cultural. Los datos de una cultura pueden estar sobrerrepresentados en comparación con otra, lo que resulta en un comportamiento no igualitario del modelo. Primero, Naous et al. [25] mostraron que los LLMs tienen sesgos negativos hacia la cultura árabe: el modelo tendrá menos referencias culturales y estará menos alineado con las creencias, normas y costumbres humanas de un grupo cultural subordinado. Este es un sesgo de selección. En segundo lugar, un LLM tenderá a asimilar los estereotipos culturales encontrados en los datos de entrenamiento, tendiendo a una amplificación de los prejuicios culturales existentes dentro de las salidas del modelo. Estos son sesgos de atribución de grupo.

Sesgos lingüísticos. Algunos idiomas son prominentes en el conjunto de entrenamiento, por ejemplo, GPT-3 se ha entrenado con 50 veces más inglés que francés, que es el segundo idioma en términos de datos de entrenamiento. Un LLM confundirá el español chileno y argentino, aunque no confundirá el inglés irlandés y escocés. Esto se debe a un sesgo de selección, ya que los datos de entrenamiento no representan el mundo real. Pero también a un sesgo de representación, ya que los datos de la Web contienen muchas más referencias irlandesas/escocesas que chilenas/argentinas.

Sesgos ideológicos y políticos. En los datos de entrenamiento, algunos sesgos políticos e ideológicos están más representados que otros. Argyle et al. [26] muestran que es posible replicar los puntos de vista de subpoblaciones demográficamente diversas de Estados Unidos al solicitar a los LLMs que actúen como una persona de un lado político específico. Sin embargo, tenderán a favorecer más ciertas perspectivas políticas o ideologías y serán más propensos a representar estereotipos de grupos subdominantes [27].

Sesgos demográficos. Los datos de entrenamiento muestran una representación desigual (o falta) de ciertos grupos demográficos. Esto puede tomar la forma de sesgos geográficos: varios trabajos [28–30] señalaron que los LLMs muestran un conocimiento geográfico deficiente sobre algunas partes del mundo. Los sesgos también pueden centrarse más en los individuos, como un sesgo de clase social; Curry et al. [31] muestran que los LLMs desfavorecen a los grupos socioeconómicos menos privilegiados.

Sesgos temporales. Dado que los datos se seleccionan durante un período de tiempo específico, limita los contextos históricos al informar sobre eventos actuales, pero también las tendencias u opiniones. Primero, los significados semánticos de las palabras cambian con el tiempo [33] (ver Figura 5), pero afecta los resultados incluso en un corto período [34]. ¿Quién querría un LLM con la visión dominante de las mujeres de principios de los años sesenta?, ¿o uno que tenga la opinión dominante de la esclavitud que tenían las personas en el siglo XVIII?

Sesgos de confirmación. Los LLMs están entrenados para alinearse con las creencias de los usuarios y tenderán a ser más asertivos con respecto a los datos de entrenamiento asertivos, como las opiniones firmes. Además, como desean satisfacer al usuario, los LLMs pueden tender a la retención selectiva de información para crear contenido cognitivamente atractivo en lugar de informativo.

¿Dónde encontrarlos?

Aunque se "cuelan" en todas partes, existen técnicas más o menos exitosas para detectar sesgos en los datos o en los modelos.

En los datos

Valor de características faltantes. Para datos tabulares, un valor faltante puede ser el lugar de un sesgo. Si algunos valores faltan de un grupo objetivo específico, entonces puede indicar que está subrepresentado. Por ejemplo, si los datos de ingresos de un grupo demográfico en particular a menudo faltan, el modelo podría malinterpretar el estado económico de ese grupo.

Valores de características inesperados. Los valores de características inesperados podrían indicar errores de entrada de datos u otras inexactitudes que conducen a un sesgo. Una edad negativa para algunas personas podría sesgar el modelo contra ciertos grupos de edad. Además, identificar valores inesperados ayuda a detectar valores atípicos que pueden representar desproporcionadamente a un grupo minoritario. Si el modelo se entrena con estos valores atípicos sin corrección, podría generalizarse mal en ellos.

Asimetría de datos. Si ciertos grupos o características pueden estar sub o sobrerrepresentados en relación con su prevalencia en el mundo real, puede introducir un sesgo en el modelo. Y aunque refleje con precisión la prevalencia del mundo real, puede que no sea la mejor idea entrenar con ellos y repetir las desigualdades del mundo real. Observar la distribución de datos con respecto a los diferentes grupos objetivo es una buena opción. Es una manera fácil de verificar si hay una asimetría importante en los datos representados entre los diferentes grupos objetivo. De manera similar, el número de muestras por grupo objetivo también es importante para no favorecer al grupo sobrerrepresentado.

Figura 6. Sesgo de "latinas" de la versión 1.5 de Stable Diffusion comparada con la versión 2.1.

Anotación no representativa. Pedir la demografía de los anotadores al recopilarlos ahora se considera esencial. Santy et al. [5] mostraron que la noción de discurso de odio y aceptabilidad social varía mucho entre diferentes culturas, es decir, lo que puede ser socialmente aceptado para algunos individuos puede estar totalmente prohibido para otros. Por esta razón, las anotaciones que se utilizarán para el aprendizaje supervisado no deben estar sesgadas hacia un subgrupo de la población global que se verá afectada por este modelo.

Datos de entrenamiento (pre)sucios. Cuando se busca recopilar un conjunto de datos masivo para preentrenar modelos fundamentales, es probable que los datos no sean perfectos. Limpiar el conjunto de datos para eliminar texto duplicado ayuda a reducir los sesgos, pero también a mejorar el rendimiento de los modelos: Hernández et al. [35] señalan que "para un modelo de 1B (mil millones) de parámetros, cien duplicados son dañinos; en 175B, incluso unos pocos duplicados podrían tener un efecto desproporcionado". Limpiar eliminando contenido de discurso de odio y pornografía (que forma la mayor parte de la Web) también es una buena idea. Por ejemplo, Tiku et al. [22] descubrieron que al usar la versión 1.5 de Stable Diffusion, "latina" produce imágenes altamente sexuales, ya que el 20% de las leyendas que contienen esta palabra fueron juzgadas como inseguras por un clasificador NSFW (Figura 6).

En los modelos

Positividad sobre grupos objetivo. Analizar la predicción de salida de un sistema con respecto a diferentes grupos objetivo puede decir mucho sobre su funcionamiento, especialmente cuando las salidas pueden verse como positivas o negativas. Por ejemplo, si un sentimiento es malo para personas árabes, la empleabilidad es menor para mujeres o si la predicción de reincidencia es mayor para personas negras, puede que no sea una buena señal.

Robustez y estabilidad. En machine learning, un sesgo puede verse como un cambio de decisión influenciado por una variable no causal. Ribeiro et al. [37] utilizan contrafactuales para verificar la robustez, como cambiar algunas palabras o atributos en una oración y ver si afecta el comportamiento del modelo.

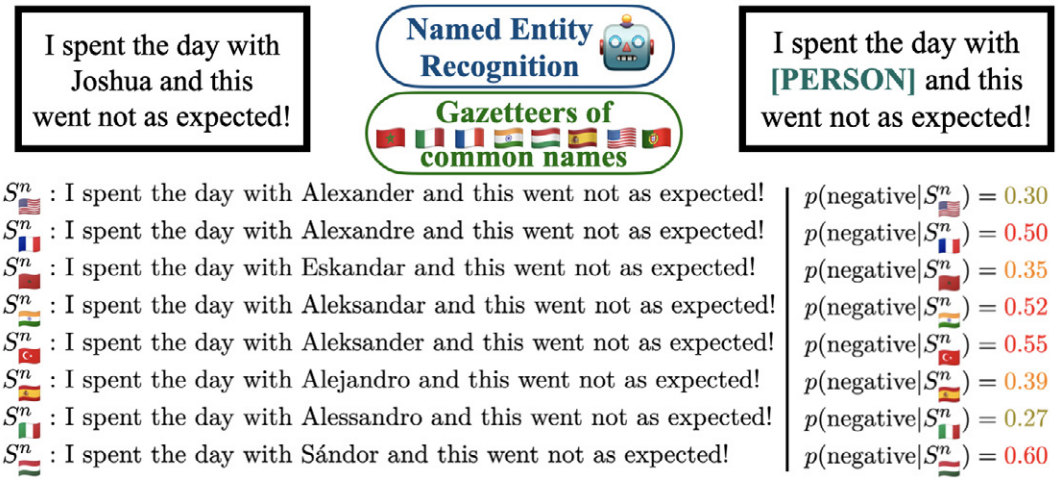


Figura 7. Detección de sesgos con respecto a los países, utilizando ejemplos contrafactuales sobre un sistema de análisis de sentimientos, que debería suponer emitir las mismas predicciones.

Si la salida del modelo etiqueta específicamente como positiva o negativa, se puede inferir un sesgo con respecto a los cambios de atributos. De hecho, al usar nombres como proxy, es posible estimar el sesgo de un modelo con respecto a la procedencia de un nombre ([23,24]; ver Figura 7).

Rendimientos heterogéneos sobre grupos objetivo. Una definición simple de un sistema sesgado es que se desempeña mal en datos de un grupo objetivo. Si un sistema de reconocimiento facial no puede funcionar con personas negras, entonces está sesgado. La tasa de falsos positivos y falsos negativos por subgrupo puede ayudar a entender qué grupos experimentan un rendimiento desproporcionadamente peor o mejor. Este sesgo oculto puede ser perjudicial, especialmente en aplicaciones como la puntuación de crédito o los diagnósticos médicos.

Prueba de choque en la vida real. Una de las mejores pruebas es poner el sistema en contacto con los usuarios. Rara vez opera perfectamente cuando se aplica a datos reales y en vivo. Cuando ocurre un problema, se debe evaluar si refleja desventajas sociales existentes y determinar su influencia en las personas afectadas.

Dependemos de los sesgos cada vez que tomamos una decisión, son sumamente útiles para seleccionar las opciones más probables.

¿Cómo mitigarlos?

Existen muchas formas de mitigar los sesgos negativos de los modelos. Aquí presentamos algunas de ellas.

Sobremuestreo/Submuestreo. Si un grupo objetivo está subrepresentado o sobrerrepresentado, podría afectar el rendimiento del modelo. Una solución simple es proceder a un muestreo.

Muestras ponderadas. Otra solución sería ponderar la pérdida de la función, simplemente por el inverso de la proporción de cada muestra de grupo objetivo (si tienes un 90% del grupo A y un 10% del grupo B, entonces puedes ponderar las muestras del grupo A por $\frac{1}{0.9}$ y las del grupo B por $\frac{1}{0.1}$).

Función objetivo que refleja justicia. También se puede crear una función para ayudar a reducir el impacto de las muestras sesgadas como una pérdida focal que reduce el impacto de las muestras fáciles en la actualización del peso, como una pérdida focal desesgada [38] o su versión no supervisada de Orgad and Belinkov [39]. Esta última se basa en un detector de éxito que se supone predice si el modelo principal, sin conocer la tarea, tendrá éxito en la predicción: si puede predecir el éxito del modelo principal en una muestra, entonces podría contener características sesgadas y debería tener un peso reducido en la pérdida.

Aumento de datos. Sharma et al. [40] proponen crear, para cada muestra que contenga un atributo del grupo objetivo, una nueva muestra que tenga las mismas características (excepto el(los) atributo(s) protegido(s)) pero con el valor opuesto del atributo protegido y la misma etiqueta. Por ejemplo, la oración "John

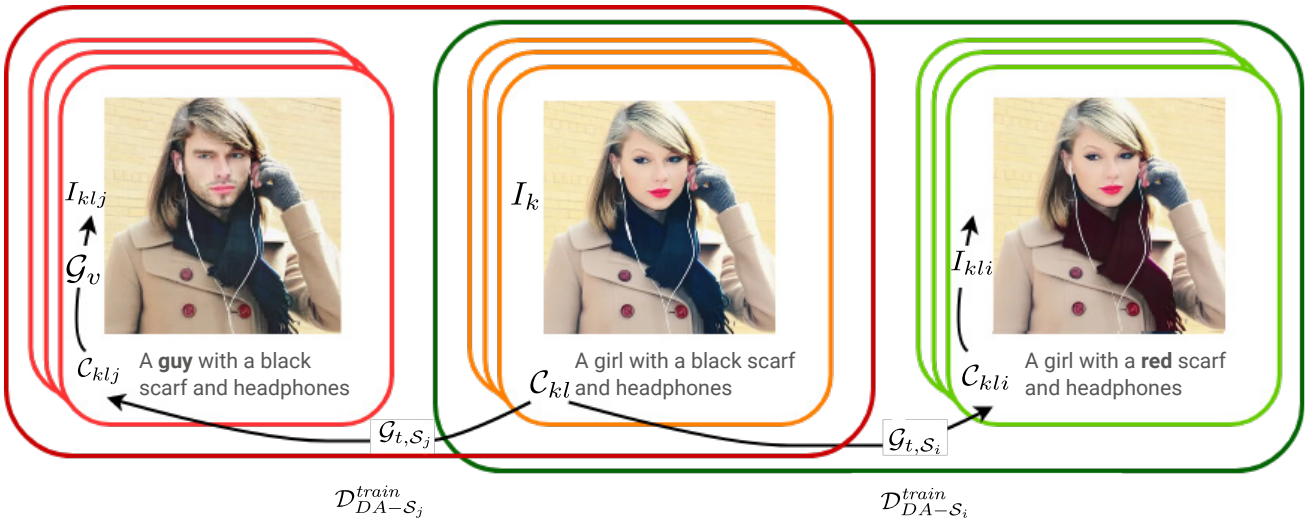


Figura 8. Ejemplo de aumento de datos multimodales que fomenta la diversidad [36].

es ingeniero y le encanta el snowboard" se convertiría en "Jane es ingeniera y le encanta el snowboard". Esta técnica también se puede usar para eliminar sesgos (no)sociales y forzar a los modelos multimodales a adaptarse a asociaciones menos comunes que las que están en el conjunto de datos inicial, como un limón azul (ver Figura 8 de [36]).

Pérdida adversarial. La *fairness through blindness* es una técnica utilizada para maximizar la capacidad del clasificador para predecir la clase, minimizando al mismo tiempo la capacidad de una red adversarial para predecir una variable protegida [41,42]. Por ejemplo, si una representación de modelo de alta dimensión de un currículum vitae puede usarse para predecir si una persona debe ser empleada y al mismo tiempo no puede usarse para predecir el género de una persona, entonces debería ser independiente del género del solicitante.

Perspectivas humanas. ¡A través del uso de diferentes perspectivas! Discutir con expertos del dominio, científicos sociales, legisladores y psicólogos para tener un punto de vista diferente sobre el impacto de un trabajo. Algunos artículos recientes [43,44] proponen redefinir las raíces del problema utilizando argumentos y enfoques alejados de los planteados clásicamente en la informática, para fenómenos como la diversidad o la empatía [5,45].

Perspectivas de máquinas. También se puede integrar el hecho de que a veces no hay una única respuesta para una pregunta, y se necesitan o deben representarse muchas perspectivas en los modelos. Este es en realidad un nuevo campo de investigación en NLP [46,47]. Por ejemplo, un LLM puede solicitarse con demografía para representar diversas perspectivas [48]. Esto permite que el modelo se vea obligado a adaptarse a diferentes subgrupos de la población en los que se utiliza.

Integración de todas las etiquetas. En el mismo sentido, es posible no entrenar en una agregación de anotaciones que representan la verdad objetiva sino en la distribución de anotaciones. Esto permite representar mejor a los anotadores de tareas subjetivas como el reconocimiento de discurso de odio o emociones [49].

Alineación del usuario. Las técnicas utilizadas por los LLMs como el *reinforcement* en preferencias humanas pueden ayudar a reducir algunos sesgos, como la generación de discurso de odio o contenido ofensivo. Sin embargo, también se ha demostrado que reducen la universalidad del modelo [50].

Conclusión

Los sesgos están en todas partes. Pueden ser útiles en el proceso de razonamiento, ya que estructuran el mundo, y aunque puedan representar la realidad, pueden ser perjudiciales para los subgrupos de la población. Es importante primero detectar sus impactos, lo cual es posible utilizando algunos de los métodos que presentamos en este documento y, segundo, minimizarlos para tender hacia modelos de IA más justos. Esto es posible mediante la optimización de objetivos específicos con funciones de pérdida (*loss functions*), el aprendizaje adversarial o el uso de aumento de datos, etc. Este paso es esencial si queremos cumplir las promesas de la IA para un mundo más justo. ▪

REFERENCIAS

[1] Observatoire des inégalités. Des contrôles de police très inégaux selon la couleur de la peau, 2021.

[2] N. Jounin, F. Ahmadouchi, A. Bachiri, B. Bakhayokho, J. Bihet, R. Bouali, N. Cognasse, S. El Mellah, C. Gicquel, M. Josse, Y. Kettal, N. Krumnow, A. Mimoun, L. Mokrani, J. Mongongnon, P. Orsini, C. Otto, L. Rondou, A. Tamega, L. Tilbourg, E. H. Touré, and U. Tubeuf. Control features: Identity checks, appearance and ways of life of students in the Île-de-France. Deviance et Societe, 39(1):3–29, 2015. ISSN 03787931.

[3] M. Hall, L. van der Maaten, L. Gustafson, M. Jones, and A. Adcock. A Systematic Study of Bias Amplification. Trustworthy and Socially Responsible Machine Learning (TSRML) at Neurips, 1(1), 2022.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning, 2021. ISSN 15577341.

[5] S. Santy, J. T. Liang, R. L. Bras, K. Reinecke, and M. Sap. NLPositionality: Characterizing Design Biases of Datasets and Models. 1:9080–9102, 2023.

[6] T. Sorensen, L. Jiang, J. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, and Y. Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. (Archive 2011), 2023.

[7] N. Mirzakhmedova, J. Kiesel, M. Alshomary, M. Heinrich, N. Handke, X. Cai, V. Barriere, D. Dastgheib, O. Ghahroodi, M. A. Sadraei, E. Asgari, L. Kawaletz, H. Wachsmuth, and B. Stein. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. In LREC-COLING, 2024.

[8] R. Taori and T. B. Hashimoto. Data Feedback Loops: Model-driven Amplification of Dataset Biases. In Proceedings of Machine Learning Research, volume 202, pp. 33883–33920, 2023.

[9] D. Baer. Kahneman: Your Cognitive Biases Act Like Optical Illusions, 2017.

[10] D. Kahneman. Thinking, Fast and Slow. 2011.

[11] C. Meister, W. Stokowiec, T. Pimentel, L. Yu, L. Rimell, and A. Kuncoro. A Natural Bias for Language Generation Models. In ACL, volume 2, pp. 243–255, 2022.

[12] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer. Detection of abusive language: The problem of biased datasets. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1:602–608, 2019.

[13] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 5884–5906, 2022.

[14] M. Parmar, S. Mishra, M. Geva, and C. Baral. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, pp. 1771–1781, 2023.

[15] K. Hämmerl, B. Deiseroth, P. Schramowski, J. Libovický, C. A. Rothkopf, A. Fraser, and K. Kersting. Speaking Multiple Languages Affects the Moral Bias of Language Models. In Findings of ACL: ACL 2023, pp. 2137–2156, 2022.

[16] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5477–5490, 2020.

[17] S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In ACL, volume 1, pp. 11737–11762, 2023.

[18] Y. Hirota, Y. Nakashima, and N. García. Quantifying Societal Bias Amplification in Image Captioning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022-June:13440–13449, 2022. ISSN 10636919.

[19] P. Czarnowska, Y. Vyas, and K. Shah. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. Transactions of the Association for Computational Linguistics, 9:1249–1267, 2021. ISSN 2307387X.

[20] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8789–8797, 2018.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, volume 2015 Inter, pp. 3730–3738, 2015.

[22] N. Tiku, K. Schaul, and S. Y. Chen. This is how AI image generators see the world, 2023.

[23] V. Barriere and S. Cifuentes. Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method with Least Confounding Variables. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 1511–1518, Torino, Italia, 5 2024. ELRA and ICCL.

[24] V. Barriere and S. Cifuentes. A Study of Nationality Bias in Names and Perplexity using Off-the-Shelf Affect-related Tweet Classifiers. Submitted to EMNLP, 2024.

[25] T. Naous, M. J. Ryan, A. Ritter, and W. Xu. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. ACL, 2024.

[26] L. P. Argyle, C. A. Bail, E. C. Busby, J. R. Gubler, T. Howe, C. Rytting, T. Sorensen, and D. Wingate. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. Proceedings of the National Academy of Sciences of the United States of America, 120(41):1–8, 2023. ISSN 10916490.

[27] A. Liu, M. Diab, and D. Fried. Evaluating Large Language Model Biases in Persona-Steered Generation. 2024.

[28] R. Manvi, S. Khanna, M. Burke, D. Lobell, and S. Ermon. Large language models are geographically biased. arXiv preprint arXiv:2402.02680, 2024.

[29] J. Dunn, B. Adams, and H. T. Madabushi. Pre-Trained Language Models Represent Some Geographic Populations Better than Others. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 12966–12976, 2024.

[30] N. Godey, E. de la Clergerie, and B. Sagot. On the Scaling Laws of Geographical Representation in Language Models. In LREC-COLING, 2024.

[31] A. C. Curry, G. Attanasio, Z. Talat, M. B. Zayed, and D. Hovy. Classist Tools: Social Class Correlates with Performance in NLP. (1964), 2024.

[32] Y. Shi and L. Lei. The evolution of LGBT labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. English Today, 36(4):33–39, 2020.

[33] D. Schlechtweg, A. Hätty, M. del Tredici, and S. S. i. Walde. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 732–746, 2020.

[34] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados. TimeLMs: Diachronic Language Models from Twitter. 2022.

[35] D. Hernández, T. Brown, T. Conerly, N. DasSarma, D. Drain, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume, S. Johnston, B. Mann, C. Olah, C. Olsson, D. Amodei, N. Joseph, J. Kaplan, and S. McCandlish. Scaling Laws and Interpretability of Learning from Repeated Data. pp. 1–23, 2022.

[36] V. Barriere, F. Del Río, A. Carvallo, C. Aspillaga, E. Herrera-Berg, and C. Buc. Targeted Image Data Augmentation Increases Basic Skills Captioning Robustness. In S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, editors, Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pp. 243–257, Singapore, 12 2023. Association for Computational Linguistics.

[37] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond Accuracy: Behavioral Testing of NLP Models. ACL, 2020.

[38] R. K. Mahabadi, Y. Belinkov, and J. Henderson. End-to-end bias mitigation by modelling biases in corpora. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 8706–8716, 2020.

[39] H. Orgad and Y. Belinkov. BLIND: Bias Removal With No Demographics. In ACL, volume 1, pp. 8801–8821, 2022.

[40] S. Sharma, Y. Zhang, J. M. Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney. Data augmentation for discrimination prevention and bias disambiguation. In AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 358–364, 2020.

[41] Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 11–21, 2018.

[42] L. Wang, Y. Yan, K. He, Y. Wu, and W. Xu. Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3740–3750, 2021.

[43] A. Curry and A. C. Curry. Computer says "No": The Case Against Empathetic Conversational AI. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 8123–8130, 2023.

[44] M. Valette. What Does Perspectivism Mean? An Ethical and Methodological Countercriticism. In 3rd Workshop on Perspectivist Approaches to NLP, NLPerspectives 2024 at LREC-COLING 2024 - Workshop Proceedings, pp. 111–115, 2024.

[45] S. Tafreshi, V. Barriere, and S. Buechel. WASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, EACL 2021, pp. 92–104, 2021.

[46] G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, and A. Uma. Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022. In Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, 2022.

[47] G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli. Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024. In Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024, 2024.

[48] S. A. Hayati, M. Lee, D. Rajagopal, and D. Kang. How Far Can We Extract Diverse Perspectives from Large Language Models? 2023.

[49] Y. Kim and J. Kim. Human-Like Emotion Recognition: Multi-Label Learning From Noisy Labeled Audio-Visual Expressive Speech. In ICASSP, pp. 5104–5108, 2018.

[50] T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi. Position: A Roadmap to Pluralistic Alignment. In Proceedings of the 41 st International Conference on Machine Learning, 2024.