Discovering Bias Indicators in LLM Behavior

Sofía Desimone[1,2] and Laura Alonso Alemany[1]

[1] Facultad de Matemática, Astronomía, Física y Computación
Universidad Nacional de Córdoba, Argentina
[2] CONICET

Abstract. Bias detection in texts generated by large language models (LLMs) often relies on predefined lexical lists or distributional differences, but is limited by lexical and morphological variation, which disperses statistical evidence. In this work in progress, we propose constructing equivalence classes of n-grams from LLM-generated texts using lemmatization to group variants with similar semantic content. This approach reduces dispersion and concentrates statistical evidence. We then apply an adaptation of Pointwise Mutual Information (PMI) to estimate differential associations between groups in a contrastive generation setting, enabling the identification of bias indicators that emerge directly from the data while improving robustness to low-frequency events. We present a preliminary exploration with promising results.

Keywords: Bias in LLMs, Subtle Bias, Unsupervised bias discovery

Descubriendo indicadores de sesgo en el comportamiento de grandes modelos de lenguaje

Abstract. La detección de sesgos en textos generados por modelos de lenguaje de gran escala (LLMs) suele basarse en listas léxicas predefinidas o en diferencias distribucionales, pero se ve limitada por la variación léxica y morfológica, que dispersa la evidencia estadística. En este trabajo en curso proponemos construir clases de equivalencia de n-gramas a partir de textos generados por LLMs, utilizando lematización para agrupar variantes con contenido semántico similar. Este enfoque reduce la dispersión y concentra la evidencia estadística. Sobre estas unidades, aplicamos una adaptación de Pointwise Mutual Information (PMI) para estimar asociaciones diferenciales entre grupos en un contexto de generación contrastiva, permitiendo identificar indicadores de sesgo que emergen directamente de los datos y mejorar la robustez frente a eventos de baja frecuencia. Presentamos una primera exploración prospectiva de esta propuesta que muestra resultados prometedores.

Keywords: Sesgo en modelos de lenguaje, Sesgos sutiles, Descubrimiento no supervisado de sesgos

1 Introducción y Motivación

La detección de sesgos en textos generados por grandes modelos de lenguaje (LLMs) suele basarse en la identificación de asociaciones estadísticas entre expresiones lingüísticas y grupos sociales (Gallegos et al., 2024). Estas asociaciones suelen analizarse mediante listas de palabras u oraciones predefinidas, lo que limita la capacidad de capturar la diversidad de realizaciones lingüísticas presentes en los textos (Bai et al., 2025; Nadeem et al., 2021; Nangia et al., 2020; Panda et al., 2025).

Como alternativa, los enfoques empíricos permiten descubrir indicadores de sesgo a partir de diferencias en la distribución de expresiones en contextos comparables, identificando patrones que emergen directamente de los datos. Sin embargo, la variación léxica y morfológica introduce ruido y dispersión en las observaciones, ya que expresiones con significados similares pueden aparecer en múltiples formas, dificultando identificar distribuciones diferenciales robustas.

Para mitigar este problema, proponemos construir clases de equivalencia de n-gramas que agrupen expresiones con contenido semántico similar. Entre los diferentes métodos para encontrar expresiones con semántica comparable, encontramos que los *word embeddings* ofrecieron un nivel de abstracción demasiado grueso para modelar indicadores de sesgo. Por ello, utilizamos lematización para reducir la dispersión de ocurrencias y concentrar la evidencia estadística asociada a un mismo contenido conceptual. Sobre estas abstracciones aplicamos una adaptación de Pointwise Mutual Information (PMI) para estimar asociaciones diferenciales entre grupos en un esquema de generación contrastiva. La PMI ha sido reconocida como de utilidad para identificar asociaciones entre grupos sociales y expresiones de sesgo (Aka et al., 2021; Valentini et al., 2023). Mediante PMI es posible identificar indicadores de sesgo que emergen directamente de los datos, al tiempo que con las abstracciones que proponemos en este artículo se reduce la sensibilidad a eventos de baja frecuencia, obteniendo estimaciones más estables y representativas.

La propuesta consta de tres componentes: (1) generación contrastiva de textos por grupo social, (2) construcción de clases de equivalencia y (3) cálculo de asociación entre grupos sociales y clases de equivalencia mediante PMI.

2 Metodología

2.1 Generación contrastiva de textos

El primer paso de nuestra metodología consiste en generar datos sintéticos en contextos controlados donde pueda emerger comportamiento diferencial entre grupos sociales. Para ello, utilizamos prompts que fijan un contexto y varían únicamente el grupo social referido. Este enfoque contrastivo permite analizar cómo los LLMs generan texto para distintos grupos sociales bajo condiciones equivalentes e identificar patrones diferenciales. En caso de observarse, estos patrones pueden destacarse para que personas expertas evalúen sus posibles efectos perjudiciales.

2.2 Construcción de clases de equivalencia

Para mejorar la representatividad estadística de las asociaciones entre grupos y expresiones lingüísticas necesitamos reducir la variabilidad de estas expresiones. En exploraciones iniciales observamos que trabajar sobre las formas de las palabras sin ninguna abstracción efectivamente dificultaba encontrar patrones estadísticamente significativos.

Para evitarlo, nuestro objetivo es trabajar con grupos de expresiones textuales que se corresponden a significados equivalentes. Sobre el vocabulario de los textos generados, consideramos que cada palabra pertenece a la clase de equivalencia que agrupa todas las variaciones morfológicas de un mismo lema. Algunas exploraciones prospectivas con técnicas de clustering y medidas de semejanza basadas en *word embeddings* resultaron en abstracciones demasiado gruesas (*coarse-grained*) que terminan agrupando expresiones que no consideramos equivalentes a efectos de nuestros objetivos de representación, por lo cual descartamos esa aproximación.

Para n-gramas de mayor longitud (bigramas, trigramas y tetragramas), se agrupan en una misma clase de equivalencia todos aquellos n-gramas con el mismo conjunto de lemas, independiente de orden y excluyendo palabras vacías. Se descartan los n-gramas que empiezan o terminan con palabras vacías.

En esta etapa, excluimos los n-gramas que contienen vocabulario explícitamente asociado a los grupos en comparación, ya que su distribución diferencial está determinada por su significado inherente y no por patrones de sesgos sutiles.

2.3 PMI sobre clases de equivalencia

Para estimar la asociación entre una clase de equivalencia y los grupos en comparación, adaptamos la métrica PMI a nuestras unidades de análisis. Utilizamos un puntaje de asociación diferencial $BiasScore(BS)$, definido como la diferencia entre los valores de PMI de una unidad respecto de cada grupo. Para un n-grama x y dos grupos G_1 y G_2:

$$BS(x) = PMI(x, G_1) - PMI(x, G_2) = \log\left(\frac{p(x \mid G_1)}{p(x \mid G_2)}\right) \tag{1}$$

En nuestro caso, si GI denota el grupo de interés, GC el grupo de control y A una clase de equivalencia, el puntaje queda definido como:

$$BS(A) = \log\frac{p(A \mid GI)}{p(A \mid GC)} \approx \log\frac{M_{A,GI}}{M_{\cdot GI}} - \log\frac{M_{A,GC}}{M_{\cdot GC}} \tag{2}$$

donde $M_{A,GI}$ y $M_{A,GC}$ son las frecuencias de A en cada grupo y $M_{\cdot GI}$ y $M_{\cdot GC}$ los totales de unidades observadas. La frecuencia de una clase se obtiene sumando las frecuencias de todos los n-gramas que la componen.

La proporción $\frac{M_{\cdot GC}}{M_{\cdot GI}}$ depende únicamente del tamaño total de cada conjunto y puede considerarse constante al comparar n-gramas de distinta longitud. En la práctica, este factor puede estimarse a partir de las frecuencias totales de unigramas en cada grupo, que proporcionan una aproximación del tamaño relativo de cada conjunto.

3 Prueba de la Metodología con Discapacidad

Para analizar la metodología propuesta, realizamos una exploración prospectiva sobre un grupo de interés inicial, el de las personas con discapacidad, usando `llama-3.3-70B`. Para ello, utilizamos pares de prompts que fijan un mismo escenario y varían únicamente en el grupo social, como "*Escribe una historia sobre Ana, una persona ciega que tiene una reunión de trabajo*" para el grupo de interés y "*Escribe una historia sobre Ana, una persona con anteojos que tiene una reunión de trabajo*" para el grupo de control.

La Tabla 1 muestra ejemplos de clases de equivalencia y sus puntajes BS, junto con la frecuencia de cada n-grama (entre paréntesis). Valores positivos indican mayor asociación con textos que mencionan discapacidad, mientras que valores negativos corresponden al grupo de control. El valor absoluto refleja la intensidad de la asociación.

El agrupamiento de variaciones morfosintácticas permite acumular evidencia sobre el comportamiento de n-gramas poco frecuentes. Por ejemplo, expresiones como "podido superar el obstáculo" o "pudo superar los obstáculos", que aparecen de forma aislada, pueden ser consideradas como instancias de una misma clase de equivalencia cuya frecuencia agregada alcanza 37 ocurrencias. Esto evita que la evidencia quede fragmentada entre variantes superficiales y permite tratar estas expresiones de manera conjunta dentro de una misma unidad de análisis, en lugar de considerarlas como n-gramas aislados de baja frecuencia.

Podemos observar también que las clases de equivalencia tienden a agrupar expresiones con significados muy cercanos, que, a efectos de lo que queremos representar, funcionan como expresiones equivalentes.

En particular, las expresiones con puntajes altos de BS se vinculan con narrativas de superación y limitación, mientras que las expresiones con puntajes negativos se asocian a descripciones más neutrales o a atributos profesionales mejor valorados socialmente. Este comportamiento sugiere que el método permite identificar configuraciones lingüísticas consistentes con estereotipos asociados a la discapacidad presentes en los textos generados.

4 Conclusiones y Trabajo Futuro

En este trabajo presentamos un método para reducir el impacto de la variación léxica y morfológica en la detección de sesgos en el comportamiento de modelos de lenguaje. Construimos clases de equivalencia que agrupan expresiones lingüísticas que se corresponden con un mismo significado a efectos de análisis de sesgo. Estas clases de equivalencia permiten agregar las ocurrencias de diferentes expresiones, y de esta manera reducir el número de observaciones infrecuentes. Así se pueden aplicar medidas de asocación que sólo modelan adecuadamente las observaciones con una cierta frecuencia mínima, como PMI.

En una exploración preliminar prospectiva mostramos que este método efectivamente agrupa expresiones equivalentes, produce clases con mayores frecuencias de ocurrencia y la medida de sesgo final obtiene resultados muy prometedores.

Clase de equivalencia	BS
"pesar de las limitaciones" (33), "pesar de esta limitación" (10), "pesar de sus limitaciones" (10)	3.15
"puede superar los obstáculos" (15), "pueden superar los obstáculos" (4), "podido superar los obstáculos" (6), "podía superar los obstáculos" (2), "podido superar el obstáculo" (1), "puedo superar los obstáculos" (1), "pueden superar obstáculos" (4), "puede superar obstáculos" (4)	2.72
"trabajaba en una tienda" (11), "trabajar en una tienda" (1), "trabajando en la tienda" (9), "trabajar en la tienda" (1), "trabajaba en la tienda" (1), "trabajando en una tienda" (2), "tienda donde trabajaba" (3),	0.50
"trabajaba en una empresa" (9), "trabajando en la empresa" (3), "trabajar con la empresa" (1), "trabajar con su empresa" (1), "trabajando en una empresa" (3), "empresa donde trabajaba" (9)	-1.08
"puede ser una oportunidad" (4), "podría ser una oportunidad" (1), "pueden ser una oportunidad" (1), "pueden ser oportunidades" (13)	-2.06
"tomó notas detalladas" (2), "tomando notas detalladas" (7), "tomaba notas detalladas" (3), "tomar notas detalladas" (5)	-2.13

Table 1. Ejemplos de clases de equivalencia de n-gramas y sus puntajes de asociación (BS). Entre paréntesis se indica la frecuencia de cada n-grama en el corpus.

Como trabajo futuro, aplicaremos los puntajes de estas clases de equivalencia para identificar fragmentos textuales con expresiones asociadas a cada grupo. Estos fragmentos facilitan que personas expertas puedan hacer un análisis contextualizado determinar potenciales impactos perjudiciales.

References

Aka, O., Burke, K., Bauerle, A., Greer, C., & Mitchell, M. (2021). Measuring model biases in the absence of ground truth. *AIES*.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *PNAS*, *122*(8).

Gallegos, I. O., Rossi, R., Rottger, P., Cabrera, A., & Schiebinger, L. (2024). Bias and fairness in large language models. *ACM Computing Surveys*.

Nadeem, M., Bethke, A., & Reddy, S. (2021, August). StereoSet: Measuring stereotypical bias in pretrained language models. In *ACL*.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020, November). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP*.

Panda, S., Agarwal, A., & Patel, H. L. (2025). Accesseval: Benchmarking disability bias in large language models. https://arxiv.org/abs/2509.22703

Valentini, F., Rosati, G., Blasi, D., Fernandez Slezak, D., & Altszyler, E. (2023, July). On the interpretability and significance of bias metrics in texts: A PMI-based approach. In *ACL*.