Una herramienta para superar las barreras técnicas para la evaluación de sesgos en las tecnologías del lenguaje humano

Laura Alonso Alemany[1,2], Luciana Benotti[1,2], Lucía González[1,2], Jorge Sánchez[1], Beatriz Busaniche[2], Alexia Halvorsen[2] y Matías Bordone[1,2]

[1]Sección de Computación, FAMAF, Universidad Nacional de Córdoba
[2]Fundación Vía Libre
Argentina

Resumen

El procesamiento automático del lenguaje se está volviendo omnipresente en nuestras vidas, a menudo tomando roles centrales en nuestra toma de decisiones, como lo son el elegir la redacción de nuestros mensajes y correos electrónicos, traducir nuestras lecturas o incluso sostener con nosotros conversaciones completas. *Las incrustaciones de palabras* son un componente clave de los sistemas modernos de procesamiento del lenguaje natural. Proporcionan una representación de las palabras que ha potenciado el rendimiento de muchas aplicaciones, al funcionar como una semblanza del significado.

Las incrustaciones de palabras parecen capturar una semblanza del significado de las palabras del texto sin procesar, pero, al mismo tiempo, también destilan estereotipos y prejuicios sociales que luego se transmiten a las aplicaciones finales. Estos sesgos pueden ser discriminatorios. Es muy importante detectar y mitigar esos sesgos, para evitar comportamientos discriminatorios de los procesos automatizados, que pueden ser mucho más dañinos, por su escala, que en el caso de los humanos.

Actualmente existen muchas herramientas y técnicas para detectar y mitigar los sesgos en las incrustaciones de palabras, pero éstas presentan muchas barreras para el involucramiento de personas sin conocimientos técnicos. Casualmente, la mayoría de los expertos en sesgo, ya sean científicos sociales o personas con un conocimiento profundo del contexto donde el sesgo es dañino, no tienen tales habilidades y no pueden participar en los procesos de detección de sesgos debido a las barreras técnicas.

Hemos estudiado las barreras en las herramientas existentes y hemos explorado sus posibilidades y limitaciones con diferentes tipos de usuarios. Con esta exploración, nos proponemos desarrollar una herramienta que esté especialmente dirigida a reducir las barreras técnicas y aportar el poder de exploración para atender los requerimientos de los expertos, científicos y personas en general que estén dispuestas a auditar estas tecnologías.

1 Introducción

Los modelos de aprendizaje automático y los sistemas basados en datos se utilizan cada vez más para respaldar los procesos de toma de decisiones. Dichos procesos podrían afectar los derechos fundamentales, como son el derecho a recibir una educación o el derecho a la no discriminación. Es importante que los modelos puedan ser evaluados y auditados para garantizar que tales derechos no se vean comprometidos. Lo ideal sería que un abanico más amplio de actores pudiese llevar a cabo esas

auditorías, especialmente aquellos que conocen el contexto en el que se despliegan los sistemas o aquellos que se verían afectados.

Aunque los sistemas basados en datos se pueden auditar, tales auditorías a menudo requieren habilidades técnicas que están más allá de las capacidades de la mayoría de los actores relevantes. La barrera técnica se ha convertido en un gran obstáculo para que los expertos y comunidades se involucren en la evaluación del comportamiento de los sistemas automatizados. Los detalles técnicos no sólo son una barrera para la auditoría, sino que también funcionan como una especie de oscurantismo que hace muy difícil que las personas de otros ámbitos entiendan las capacidades y limitaciones de las herramientas. Esto hace que sea muy difícil planificar políticas públicas que tengan en cuenta el impacto de estas tecnologías. Por ello, este trabajo intenta facilitar la reducción de las barreras técnicas con el fin de comprender, inspeccionar y modificar algunos procesos basados en datos. En particular, nos estamos centrando en un componente clave en el tratamiento automático del lenguaje, a saber, las *incrustaciones de palabras*.

En los últimos años, la comunidad del procesamiento del lenguaje natural (PLN) se ha preocupado cada vez más por los sesgos y estereotipos contenidos en los modelos y cómo estos sesgos pueden afectar las aplicaciones prácticas, tales como son los anuncios de trabajo personalizados. En particular, varios estudios encontraron que las representaciones de palabras aprendidas a partir de los corpus contienen asociaciones que producen efectos nocivos cuando se ponen en práctica, como son la invisibilización y la autocensura o simplemente se tornan en elementos disuasorios. Para consultar un estudio fundamental de 146 artículos que analizan el sesgo en los modelos de PLN, refiérase a Blodgett et al., (2020). Para abordar estas preocupaciones, se han propuesto muchas técnicas para medir y mitigar el sesgo codificado en tales representaciones de palabras, a saber, las incrustaciones de palabras y los modelos de lenguaje (Bolukbasi et al., 2016; Caliskan et al., 2017).

Estos enfoques para lidiar con los sesgos tienden a centrarse en cuestiones técnicas complejas. Pero, las cuestiones técnicas no suelen ser la clave para instaurar la equidad dentro de la Inteligencia artificial. La equidad requiere una comprensión adecuada de constructos sociológicos complejos que, a menudo, implican fenómenos que no se comprenden bien y mucho menos se sistematizan o formalizan. La formalización de dichos constructos para el tratamiento computacional es un desafío y requiere la participación de expertos: sociólogos, lingüistas, médicos, psicólogos, entre otros, según el dominio de aplicación de la tecnología. Sin embargo, en este documento argumentamos que las métricas y los marcos de equidad se basan en instrumentos técnicos con matices que dificultan la comprensión y la participación de las personas sin una amplia formación técnica, lo cual incluye las habilidades de programación.

En este proyecto, proponemos una metodología que facilita la exploración de los sesgos en las incrustaciones de palabras, a la vez que se toma en cuenta las necesidades específicas de la región latinoamericana. En América Latina, necesitamos expertos en el dominio para poder realizar estos análisis con autonomía, sin depender de un equipo interdisciplinario o de una capacitación, ya que éstos no suelen estar disponibles.

Este artículo está organizado de la siguiente forma. La próxima sección presenta las incrustaciones de palabras y argumenta que son la representación más simple de los significados de las palabras que se

usan ampliamente y que incorporan los sesgos presentes en los datos sobre los que se desarrollan otras tecnologías de PLN. La Sección 3 presenta trabajos relevantes en el área de diagnóstico y mitigación de sesgos en las incrustaciones de palabras. En la sección 4 se describen dos conjuntos de casos de estudio en los que dos grupos de usuarios con diferentes perfiles aplicaron esta metodología para realizar una exploración de sesgos y nuestras observaciones en el dominio de utilizabilidad y requisitos. El primer grupo está constituido por científicos de datos con diferentes trayectorias de experiencia, pero con al menos 350 horas de capacitación y educación en el dominio de la ciencia de los datos (incluidos el desarrollo y la evaluación de los modelos de aprendizaje automático). El segundo grupo está formado por científicos sociales sin capacitación previa en programación o en aspectos técnicos del aprendizaje automático. La Sección 5 explica nuestra metodología en un caso práctico que ilustra cómo se pone el poder de diagnosticar sesgos en manos de personas con el conocimiento y los roles sociales para tener un impacto en las tecnologías actuales. Finalmente, la Sección 6 incluye un resumen y un esquema de los pasos a seguir en el desarrollo del prototipo y el piloto.

2 Incrustaciones de palabras y sesgo léxico

En esta sección presentamos los conceptos básicos de cómo se representa el significado léxico en los sistemas de PLN a través de incrustaciones de palabras y luego analizamos cómo surgen los sesgos en las incrustaciones de palabras.

2.1 Conceptos básicos

Las incrustaciones de palabras son artefactos de procesamiento del lenguaje natural ampliamente utilizados que representan el significado de las palabras de forma totalmente automática, en función de su uso en grandes cantidades de texto. Por eso es necesario disponer de grandes volúmenes de texto para entrenar las incrustaciones de palabras.

La esencia de las incrustaciones de palabras consiste en representar el significado de las palabras como una similitud entre las palabras. Las palabras se consideran similares si ocurren a menudo en contextos lingüísticos similares, más concretamente, si comparten una alta proporción de contextos de co-ocurrencia. Los contextos de co-ocurrencia generalmente se representan como las n palabras que ocurren antes (y después) de la palabra objetivo que está siendo representada. En algunas estructuras más sofisticadas, los contextos pueden incluir cierta medida de orden en las palabras o estructuras sintácticas. Sin embargo, la mayoría de las mejoras en las representaciones de palabras actuales no se han obtenido mediante la adición de información sintáctica explícita, sino mediante la optimización de n para la tarea de PLN (de unas pocas palabras a unas pocas docenas de palabras) (Lison y Kutuzov, 2017).

Una vez que las palabras están representadas por sus contextos de ocurrencia (en una estructura de datos matemática llamada *vector*), la similitud entre las palabras puede capturarse y manipularse como una distancia matemática, de modo que las palabras que comparten más contextos estén más cercanas y las palabras que comparten menos contextos estén más lejanas, como se ve en la Figura 1. Para medir la distancia se utiliza la similitud del coseno (Singhal, 2001).

Las incrustaciones de palabras son un componente clave de aplicaciones como son el autocompletado de texto o la traducción automática y se ha demostrado que mejoran el rendimiento de prácticamente cualquier tarea de procesamiento de lenguaje natural a la que se han aplicado. El problema es que,

incluso si su impacto en el desempeño es en general positivo, estuvieron sesgadas sistemáticamente. Por lo tanto, aún si las incrustaciones de palabras mejoran el rendimiento general, pueden perjudicar a las comunidades que son objeto de esos sesgos.

Figura 1 - Una representación de una incrustación de palabras en dos dimensiones, que muestra cómo las palabras están más cercanas en el espacio según la proporción de co-ocurrencias que comparten.

Las incrustaciones de palabras están sesgadas porque se obtienen de grandes volúmenes de textos que tienen sesgos y prejuicios sociales subyacentes. Estos sesgos se trasladan a la representación y, por lo tanto, se transfieren a las aplicaciones. Pero dado que estas incrustaciones son artefactos complejos y opacos, que funcionan a un nivel subsimbólico, es muy difícil para una persona inspeccionarlos y detectar posibles sesgos. Esta dificultad se agudiza aún más para las personas sin amplios conocimientos en este tipo de tecnologías. A pesar de esa opacidad, las incrustaciones pre-entrenadas fácilmente disponibles son ampliamente utilizadas en soluciones socioproductivas, como son las rápidas soluciones tecnológicas para escalar y acelerar la respuesta a la pandemia de COVID 19 (Aigbe y Eick, 2021).

2.2 **La evaluación del sesgo en las incrustaciones de palabras**

Afortunadamente, dada la importancia que tienen las incrustaciones de palabras en las tecnologías del lenguaje y el impacto que pueden tener los sesgos, hemos visto surgir en los últimos años una amplia gama de herramientas y técnicas para evaluar los sesgos en las incrustaciones de palabras y en los modelos lingüísticos.

La metodología central para evaluar los sesgos en las incrustaciones de palabras consta de tres partes principales, ilustradas en la Figura 2:

1. Definir un **espacio de sesgo**, generalmente binario, delimitado por dos extremos opuestos, como en *male-female*, *young-old*, o *high-low* [*masculino-femenino*, *joven-viejo* o *alto-bajo*]. Cada uno de los extremos del espacio de sesgo se caracteriza por una lista de palabras, que se muestra en la parte superior de los diagramas de la Figura 2.

2. Evaluar el comportamiento de las **palabras de interés en este espacio de sesgo**, encontrando qué tan cerca están de cada uno de los extremos del espacio de sesgo. Esta evaluación muestra si una palabra dada está más fuertemente asociada a cualquiera de los dos extremos del sesgo y qué tan fuerte es esa asociación. En la Figura 2 se puede observar que la palabra "*nurse*" ["*enfermera/enfermero*"] está más fuertemente asociada al extremo "*femenino*" del espacio de sesgo, mientras que la palabra "*leader*" ["*líder*"] está más fuertemente asociada al extremo "*masculino*".

3. Actuar sobre las conclusiones de la evaluación. Las acciones por tomar varían enormemente según los enfoques, como se verá en la siguiente sección, pero todas ellas están dirigidas a **mitigar** la fuerza del sesgo detectado en la incrustación de palabras.

Por variados que sean los enfoques para evaluar el sesgo, cada uno de ellos se basa en **listas de palabras** para definir el espacio del sesgo que se explorará. Estas palabras tienen un impacto crucial sobre cómo y qué sesgos se detectan y mitigan, pero no son centrales en los esfuerzos dedicados a esta tarea, como se argumenta en Antoniak y Mimno, 2021. Las metodologías para elegir las palabras que componen estas listas son variadas: a veces las listas son de origen colectivo, a veces son seleccionadas a mano por los investigadores y, a veces, extraídas de trabajos anteriores en las ciencias sociales. La mayoría de ellas se desarrollan en un contexto específico y luego se utilizan en otros sin reflexionar sobre el cambio de dominio o contexto. La mayoría de los trabajos anteriores utilizan listas de palabras desarrolladas para el idioma inglés o traducciones directas del idioma inglés que no tienen en cuenta las diferencias estructurales entre los idiomas (Garg et al., 2018). Por ejemplo, en español casi todos los sustantivos y adjetivos están marcados morfológicamente con un género, pero esto no es así en inglés.

La Figura 2 ilustra las diferencias en las mediciones de sesgos léxicos entre las traducciones de las listas de palabras del inglés al español sobre dos incrustaciones de palabras diferentes en cada idioma: la incrustación en inglés se describe en Bolukbasi et al., 2016 y la incrustación en español en Cañete et al., 2020. De las 16 palabras analizadas, en inglés, ocho están más asociadas al extremo "*femenino*" del sesgo espacial, mientras que en español solo cinco de ellas lo están. Las tres palabras con diferentes posiciones son "*nurse, care and wash*" ["*enfermera/enfermero*, *cuidar* y *lavar"]*. En particular, "*nurse*" en español está marcado morfológicamente con el género masculino en la palabra "*enfermero*", por lo que no es neutral en cuanto al género. Esta figura ilustra dos cosas. Primero, el hecho de que las metodologías para la detección de sesgos desarrolladas para el inglés no son directamente aplicables a otros idiomas. Segundo, que la lista de palabras seleccionadas para analizar el sesgo tiene un fuerte impacto en el sesgo que muestra el análisis.

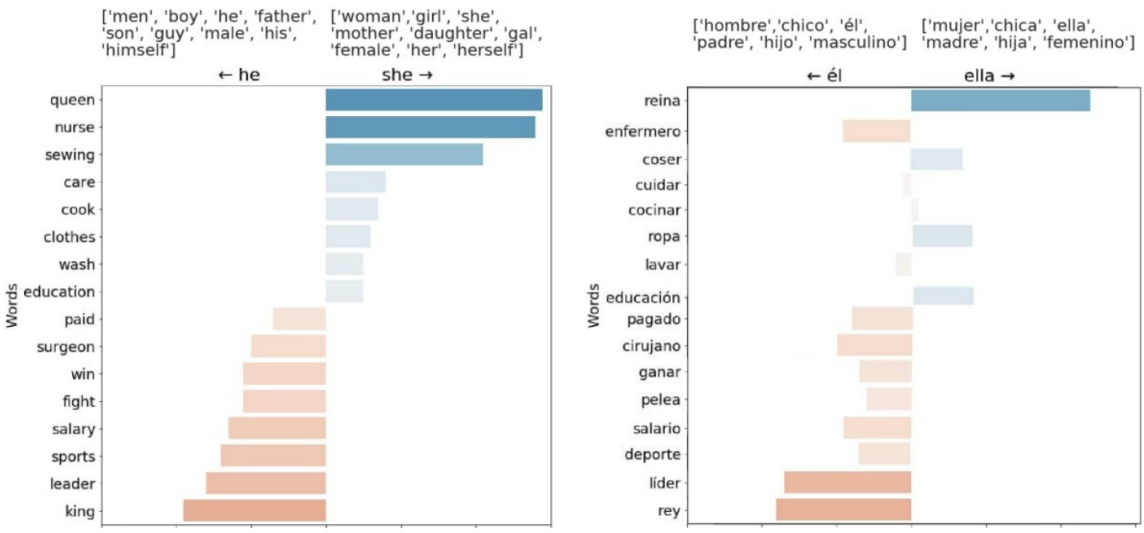

Figura 2 - Una lista de 16 palabras en inglés (izquierda) y una traducción al español (derecha) y la similitud de sus incrustaciones de palabras con respecto a la lista de palabras "woman, girl, she, mother, daughter, feminine" ["mujer, chica, ella, madre, hija, femenino"] que representan el concepto "femenino", la lista "man, boy, he, father, son, masculine" ["hombre, chico, él, padre, hijo, masculino"] que representa "masculino" y las traducciones de ambos al español. Los datos de la incrustación de palabras y el entrenamiento en inglés los describen Bolukbasi et al. (2016) y en inglés Cañete et al. (2020). De las 16 palabras de interés, en inglés, ocho están más asociadas al concepto de "femenino", mientras que en español solo cinco de ellas lo están. En particular, "nurse" en español está marcado morfológicamente con el género masculino en la palabra "enfermero", por lo que existe cierto grado de sesgo de género que debe tenerse en cuenta para explicar completamente el comportamiento de la palabra. Esta figura ilustra que las metodologías para la detección de sesgos desarrolladas para el inglés no son directamente aplicables a otros idiomas. Además, la figura ilustra que los sesgos observados dependen completamente de la lista de palabras elegidas.

2.3 Diferentes idiomas, ¿el mismo enfoque?

Como ilustra el ejemplo anterior, las diferencias lingüísticas tienen un gran impacto en los resultados obtenidos por la metodología para evaluar el sesgo. Representar las idiosincrasias del lenguaje es un objetivo crucial en nuestro proyecto, en primer lugar, porque queremos facilitar estas tecnologías a una gama más amplia de actores. En segundo lugar, porque para modelar el sesgo en un contexto determinado o en una cultura determinada es necesario hacerlo en el idioma de esa cultura.

Se han propuesto diferentes enfoques para capturar fenómenos lingüísticos específicos. Los ejemplos paradigmáticos de la variación lingüística son las lenguas con género morfológicamente marcado, que en cierta medida pueden confundirse con el género semántico. La mayoría de las propuestas para modelar el sesgo de género en idiomas con género morfológicamente marcado añaden algún constructo técnico que captura los fenómenos específicos. Ese es el caso de Zhou et al. (2019), quienes agregan un espacio adicional para representar el género morfológico, independiente de la dimensión donde modelan el género semántico. Esta complejidad adicional supone una dificultad añadida para las personas sin conocimientos técnicos.

Sin embargo, no es estrictamente necesario agregar complejidad técnica para capturar estas complejidades lingüísticas. Una explotación bien informada de las listas de palabras también puede modelar adecuadamente las particularidades lingüísticas. En el trabajo presentado aquí, adaptamos el enfoque para la evaluación del sesgo propuesto por Bolukbasi et al. (2016) y las implementamos en la caja de herramientas Responsibly. Este enfoque está destinado al inglés y no contempla el género marcado morfológicamente ni el uso diferente de los pronombres. Sin embargo, podríamos aplicar este enfoque al español, considerando lo siguiente en las listas de palabras:

- los extremos del sesgo no pueden ser definidos únicamente por los pronombres, porque los pronombres no ocurren con tanta frecuencia o en las mismas funciones en español que en inglés. Por lo tanto, las listas de palabras que definen los extremos del espacio de sesgo deben diseñarse para las particularidades del español, no traducirse tal cual.

- con respecto a las listas de palabras de interés a colocarse en el espacio de sesgo, el enfoque de Bolukbasi et al. (2016) se basa fuertemente en palabras de género neutral. Sin embargo, en español, la mayoría de los sustantivos y adjetivos están marcados morfológicamente por su género semántico (como en "enfermera", "*female nurse*", frente a "enfermero", "*male nurse*"). Para abordar esta diferencia, construimos palabras con un género neutro recurriendo a patrones como verbos, adverbios, sustantivos abstractos, sustantivos colectivos y sufijos adjetivos que tienen un género neutro, como lo es la terminación "-ble".

- no se puede hacer una evaluación adecuada del sesgo para el español solo con palabras con un género neutro, porque la mayoría de los sustantivos y adjetivos están morfológicamente marcados por su género semántico, o tienen un género morfológico incluso si no tienen una semántica de género (como lo es "mesa", "*table*", que es morfológicamente femenino). Para evaluar el sesgo también en esa amplia gama de palabras, construimos listas de palabras que contenían ambas palabras y comparamos qué tan lejos estaban posicionadas con respecto al extremo correspondiente del sesgo. En la Figura 3 se puede ver que "*female nurse*", "enfermera", está mucho más fuertemente asociada al extremo femenino del sesgo que "*male nurse*", "enfermero", el cual está asociado al extremo masculino.

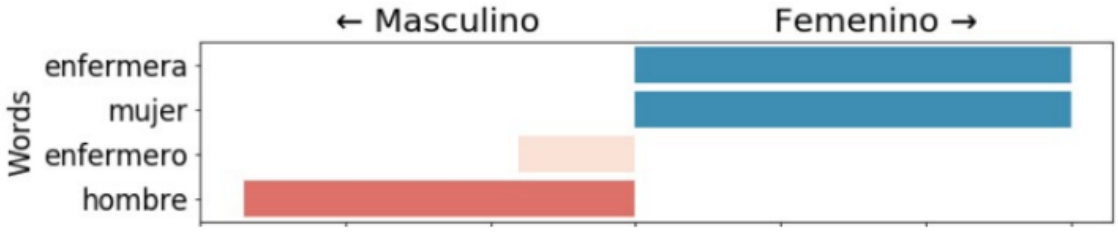

Figura 3 - Una evaluación de la manera en que el sesgo afecta las palabras con género en español. Se puede observar que "female nurse", "enfermera", está mucho más fuertemente asociada al extremo femenino del sesgo que "male nurse", "enfermero", el cual está asociado al extremo masculino.

De hecho, centrarse en una cuidadosa y consciente construcción de listas de palabras tiene el efecto secundario de poner en el punto de mira no sólo las diferencias lingüísticas, sino también otros factores culturales como los estereotipos, los prejuicios culturales o la interacción entre diferentes factores. Por lo tanto, al construir listas de palabras, se debe tener en cuenta diferentes factores, no solo el objeto principal de la investigación que es el sesgo.

3 Trabajos relevantes relacionados

En los últimos años, el estudio de los sesgos en la tecnología del lenguaje ha ido ganando relevancia, con una variedad de enfoques acompañados de críticas perspicaces (Nissim et al., 2020) y elaboraciones sucesivas que se basan en la experiencia de las propuestas iniciales (Bolukbasi et al., 2016; Gonen y Goldberg, 2019).

En esta sección revisamos en cuatro partes el trabajo relevante relacionados con el tema. En primer lugar, describimos que la mayoría de los trabajos anteriores se han centrado en un conjunto bastante reducido de sesgos y lenguajes. Luego, discutimos las limitaciones de los trabajos anteriores que se centraban en el desarrollo de algoritmos para medir y mitigar los sesgos automáticamente. Más adelante, analizamos el papel que juegan los datos de entrenamiento en el proceso y revisamos el trabajo que se enfoca en los datos en lugar de centrarse en los algoritmos. Finalmente, presentamos las herramientas existentes para la exploración de sesgos y situamos nuestra propuesta con respecto a ellas.

3.1 Raza y género, los sesgos más estudiados

La mayor parte del trabajo publicado sobre exploración y mitigación de sesgos ha sido producida por científicos informáticos basados en el hemisferio norte, en grandes laboratorios que tienen acceso a grandes cantidades de fondos, poder de cómputo y datos. Como era de esperarse, la mayor parte del trabajo se ha realizado en inglés y para los sesgos de género y raza (Garg et al., 2018; Blodgett et al., 2020; Field et al., 2021). Mientras tanto, existen otros sesgos que afectan profundamente al hemisferio sur, como son la nacionalidad, el poder y la condición social. En consonancia con el resto del área de la PLN, el trabajo se ha centrado también en los matices técnicos en lugar de en los aspectos cualitativos más impactantes, como lo es quién desarrolla la lista de palabras utilizada para las técnicas de medición y evaluación de sesgos (Antoniak y Mimno, 2021). Dado que los sesgos relacionados con el género son unos de los más estudiados, los trabajos anteriores han demostrado que las diferentes métricas de sesgo que existen para las incrustaciones de palabras, sean contextualizadas o independientes del contexto, sólo se correlacionan entre sí en los parámetros de comparación construidos para evaluar los sesgos relacionados con el género en inglés (Badilla et al., 2020).

El inglés es un idioma donde la marcación morfológica del género gramatical es residual, observable en la forma de muy pocas palabras, en su mayoría pronombres personales y algunas lexicalizaciones como “actress - actor” [“actriz - actor”]. Algunas de las suposiciones que subyacen a este enfoque parecían inadecuadas para modelar lenguajes donde un gran número de palabras tienen una marca morfológica de género gramatical, como son el español o el alemán, en los cuales la mayoría de los sustantivos o adjetivos deben expresar una marca morfológica para diferentes géneros gramaticales. La propuesta de la comunidad informática que trabajaba en materia de sesgos era un enfoque geométrico más complejo, con una dimensión que modelaba el género semántico y otra dimensión que modelaba el género

morfológico (Zhou et al., 2019). Este enfoque es más difícil de entender para las personas sin formación en informática, que suele ser el caso de los expertos en dominios sociales, quienes podrían brindar sus percepciones sobre las causas subyacentes de los fenómenos observables.

En este trabajo, hemos explorado los efectos de plasmar la complejidad de la tarea en constructos que resulten intuitivos para que los expertos del dominio viertan sus conocimientos, formulen sus hipótesis y entiendan los datos empíricos. Por otra parte, intentamos mantener la complejidad técnica de la metodología al mínimo.

Lauscher y Glavaš (2019) hacen una comparación de los sesgos en diferentes idiomas, técnicas de incrustación y textos. Zhou et al. (2019) y Gonen et al., (2019) desarrollan dos técnicas diferentes para la detección y la mitigación para los idiomas con género gramatical que se aplican luego del procesamiento. Estos enfoques agregan muchas barreras técnicas que requieren un amplio conocimiento sobre el aprendizaje automático por parte de la persona que aplica estas técnicas. Por lo tanto, no logran involucrarse de forma interactiva con expertos relevantes que no trabajan en el campo de la informática y con expertos en el dominio de aplicaciones particulares para el PLN.

3.2 **Medición y mitigación automática**

Existe un consenso (Field et al., 2021) de que lo que llamamos sesgo son los fenómenos observables (si bien sutiles) de causas subyacentes profundamente arraigadas en la dinámica social, cultural y económica. Tal complejidad va mucho más allá de las capacidades de las ciencias sociales de la mayoría de los científicos informáticos que actualmente trabajan en materia de sesgo en los artefactos de inteligencia artificial. La mayor parte del esfuerzo de investigación e innovación en curso con respecto a los sesgos se relaciona con cuestiones técnicas. En verdad, estas líneas técnicas de trabajo están dirigidas a desarrollar y consolidar herramientas y técnicas más adecuadas para tratar las cuestiones complejas que a construir una base sólida y confiable para ellas. Sin embargo, tales desarrollos han resultado, generalmente, en una complejidad técnica cada vez mayor, lo que dificulta la participación de los expertos en el dominio. Estos expertos podrían aportar precisamente el entendimiento de las causas subyacentes de la que carecen los informáticos y que podría ayudar a establecer un modelo más adecuado de las cuestiones relevantes.

Antoniak y Mimno (2021) argumentan que la variable más importante al explorar los sesgos en las incrustaciones de palabras no son las partes automatizables del problema sino la parte manual, es decir, las listas de palabras utilizadas para modelar el tipo de sesgo a explorar y la lista de palabras que deben ser neutrales. Concluyen que *las listas de palabras son probablemente inevitables, pero que ninguna herramienta técnica puede absolver a los investigadores del deber de elegir las semillas con cuidado y de manera intencional.*

Hay muchos enfoques para “el problema del sesgo” que tienen como objetivo automatizar cada paso, desde el diagnóstico del sesgo hasta su mitigación. Algunos de estos enfoques argumentan que cuando se trata de decisiones subjetivas y difíciles sobre cómo modelar ciertos sesgos, la solución es automatizar el proceso a través de un enfoque algorítmico. Guo y Caliskan (2021); Guo et al. (2022); An et al. (2022); y Kaneko y Bollegala, (2021) introducen diversos métodos para identificar automáticamente (e incluso mitigar) los sesgos en las incrustaciones de palabras y en los modelos de lenguaje.

Por el contrario, la metodología que proponemos en este trabajo esconde la complejidad técnica. Nos basamos en las ideas de Antoniak y Mimno (2021) al facilitar a expertos en dominios sin experiencia técnica el acceso a estas tecnologías, para que puedan aportar listas de palabras bien fundamentadas, al verter su conocimiento en esas listas. Sostenemos que la evaluación debe ser llevada a cabo por personas conscientes del impacto que el sesgo puede tener en las aplicaciones posteriores. Nuestra metodología se centra en ofrecer una técnica que estas personas puedan utilizar para evaluar el sesgo presente en los datos de texto, como explicamos en la siguiente sección.

3.3 **Una mirada más cercana a los datos de entrenamiento**

Brunet et al. (2019) rastrean el origen del sesgo en la incrustación de palabras hasta los datos de entrenamiento y muestran que perturbar el corpus de entrenamiento afectaría el sesgo de la incrustación resultante. Desafortunadamente, como argumentan Bender et al. (2021), la mayoría de las incrustaciones de palabras pre-entrenadas que se usan ampliamente en los productos de PLN no describen esos textos. Curiosamente, Dinan et al. (2020) muestran que los datos de entrenamiento se pueden seleccionar para mitigar los sesgos causados por datos desequilibrados. Además, Kaneko y Bollegala (2021) muestran que los datos mejor curados producen modelos menos sesgados.

Brunet et al. (2019) muestran que las técnicas de eliminación de sesgos son más eficaces cuando se aplican a los textos a partir de los cuales se inducen las incrustaciones, en lugar de aplicarlas directamente en las incrustaciones de palabras ya inducidas. Prost et al. (2019) muestran que las estrategias de mitigación demasiado simplistas en realidad empeoran las métricas de equidad en las tareas posteriores. Se requieren estrategias de mitigación más perspicaces para eliminar el sesgo de toda incrustación y no solo de las palabras utilizadas para diagnosticar el sesgo. Sin embargo, eliminar el sesgo de los textos de entrada funciona mejor. La curaduría de textos se puede hacer automáticamente (Gonen et al., 2019), pero aún no se ha demostrado que no empeore las cosas. Es mejor que los expertos del dominio diseñen estrategias de curaduría para cada caso en particular. Nuestra propuesta es ofrecer una forma de comparar las incrustaciones de palabras creadas sobre diferentes corpus.

3.4 **Herramientas existentes para la exploración de sesgos**

Existen herramientas como WordBias (Ghai et al., 2021) o Language Interpretability Toolkit [*Caja de herramientas para la interpretabilidad lingüística*] (Tenney et al., 2020) que tienen como objetivo reducir la barrera técnica que debe superarse para utilizar técnicas de detección y mitigación de sesgos. Este tipo de herramientas ofrecen implementaciones de muchas técnicas de PLN e interfaces gráficas para evitar la escritura en código.

WEFE es una biblioteca de código abierto en Python similar a WordBias en tanto ambas permiten la exploración de sesgos diferentes a la raza y el género y diferentes idiomas (Badilla et al., 2020) . Uno de los enfoques de WEFE es la comparación de diferentes métricas automáticas para la medición y mitigación de sesgos; sin embargo, para usar esta biblioteca se necesitan habilidades de programación en Python, ya que ésta no cuenta con una interfaz gráfica.

Ninguno de estos marcos fue diseñado con el objetivo de ser utilizado, en general, por científicos sociales o personas sin conocimientos técnicos y de programación. WordBias está diseñado para explorar

muchos tipos diferentes de sesgos, sin embargo, tiene una compleja interfaz gráfica que no es adecuada para la definición iterativa de listas de palabras. La Herramienta para la interpretabilidad lingüística (*Language Interpretability Tool*) incluye (entre muchas otras capacidades) visualizaciones interactivas, métricas integradas de equidad y capacidad de explicación y análisis contrafáctico, entre otros. Estas herramientas requieren una amplia comprensión del procesamiento del lenguaje natural (PLN) por parte del usuario. En general, la mayoría de las cajas de herramientas, marcos y bibliotecas que brindan funcionalidades para evaluar el comportamiento del modelo con respecto a los sesgos son confusas y opacas incluso para los desarrolladores con amplios conocimientos técnicos (Richardson et al., 2021).

En este proyecto, nos basaremos en la caja de herramientas **Responsibly** (Hod, 2018) como la biblioteca que proporciona funcionalidades básicas para nuestra herramienta. Seleccionamos Responsibly porque se enfoca en la exploración usando la métrica de sesgo más simple (la proyección de la dirección) en lugar de la comparación de diferentes mediciones automáticas de sesgo como lo hace WEFE. Responsibly se basa en trabajos previos de Bolukbasi et al. (2016) que establecen una aproximación intuitiva y eficaz al problema, pero al mismo tiempo establece algunas de las simplificaciones excesivas que se han perpetuado en la mayoría de los trabajos posteriores. Esta técnica se presenta para el idioma inglés y los sesgos de género y raza, pero se puede aplicar a cualquier sesgo que se pueda modelar como binario en idiomas que no tienen género gramatical.

Sin embargo, para poder usar la caja de herramientas Responsibly, una persona debe tener habilidades de programación en Python y un entendimiento sobre las incrustaciones de palabras, el procesamiento de lenguaje natural y sesgos en las incrustaciones de palabras. Además, esta persona, a menos que esté trabajando en un equipo interdisciplinario, también debe tener experiencia en cuáles de los sesgos presentes en las incrustaciones de palabras son relevantes para el problema y podrían afectar las aplicaciones posteriores. En nuestro proyecto, proponemos integrar las funcionalidades proporcionadas por Responsibly en una herramienta que facilita la participación de personas sin habilidades técnicas en el proceso de evaluación de sesgos.

4 Dos casos de estudio

El objetivo principal de nuestro proyecto es facilitar a personas sin conocimientos técnicos el acceso a las herramientas para la exploración de sesgos en las incrustaciones de palabras. Para ello, exploramos la utilizabilidad de nuestra adaptación de la caja de herramientas Responsibly[1] al español. En la Sección 3.4 explicamos las razones para utilizar Responsibly como base de nuestro trabajo.

Para evaluar las carencias de las herramientas disponibles y las barreras para su uso, realizamos dos estudios de utilizabilidad con diferentes perfiles de usuarios: científicos de datos novatos, la mayoría de los cuales proceden de una trayectoria no técnica, pero con una instrucción de 350 horas en materia de aprendizaje automático y científicos sociales sin conocimientos técnicos, pero con una introducción de dos horas sobre incrustación de palabras y sesgo en la tecnología del lenguaje. Nuestro objetivo era enseñar a estos dos grupos de personas cómo explorar los sesgos en las incrustaciones de palabras, al mismo tiempo que recopilamos información sobre cómo entendieron y usaron la técnica propuesta por Bolukbasi et al. (2016) para modelar espacios de sesgo. Nos centramos en las dificultades para entender

[1] https://docs.responsibly.ai/

cómo modelar el espacio de sesgo, las deficiencias para captar los fenómenos de interés y las posibilidades que ofrecen las herramientas.

Nuestros objetivos de diseño son los siguientes. En primer lugar, reducir la barrera técnica al mínimo indispensable. En segundo lugar, centrarse en la exploración y en la caracterización del sesgo (en lugar de centrarse en un diagnóstico compacto, opaco y basado en métricas). En tercer lugar, lograr una interfaz que muestre listas de palabras de una manera dinámica e interactiva que suscite, dé forma y exprese conocimiento del dominio de forma contextualizada (en lugar de tomar las listas tal y como las dan otros documentos, incluso si estos son documentos de científicos sociales). En cuarto lugar, orientar sobre los aspectos lingüísticos y culturales que pueden sesgar las listas de palabras (en lugar de limitarse a traducir las listas de palabras de otro idioma o tomar las profesiones como neutrales).

El primer grupo que estudiamos eran estudiantes al final de un nano-grado de 1 año sobre la ciencia de los datos, para un total de 180 horas y un proyecto. Todos ellos tenían algún grado de conocimientos técnicos, ya que el nanogrado les proporcionaba una amplia práctica en materia de aprendizaje automático, pero la mayoría procedía de un entorno no informático y no había recibido capacitación sobre el procesamiento del lenguaje natural como parte del curso. El segundo grupo eran periodistas, lingüistas y científicos sociales sin habilidades técnicas.

Ambos grupos recibieron un tutorial de dos horas, basado en la aplicación Jupyter Notebook que creamos para español[2], como se explica en la Sección 2, al adaptar la implementación de la caja de herramientas Responsibly (Bolukbasi et al., 2016)[3]. Los grupos realizaron el análisis sobre vectores FastText en español, entrenados con base en el *Spanish Billion Word Corpus* [*el Corpus de mil millones de palabras en español*] (Cardellino, 2019) utilizando un vocabulario de 100 mil palabras.

4.1 **Estudio de caso 1: Científicos de datos novatos**

El grupo compuesto por científicos de datos novatos recibió una explicación de una hora sobre cómo se diseñaron las herramientas y un ejemplo de cómo podrían aplicarse para explorar y mitigar el sesgo de género, el ejemplo prototípico de aplicación. Esto fue parte de un curso de ocho horas sobre Enfoques prácticos de la ética en la ciencia de datos. Como parte de la explicación, también se proporcionaron estrategias de mitigación basadas en la misma metodología, además de la apreciación de que las métricas de rendimiento no disminuyeron en un par de aplicaciones posteriores con incrustaciones mitigadas. La presentación de las herramientas se realizó para inglés y español y se explicó a los estudiantes, quienes en su mayoría eran bilingües, las analogías y diferencias entre ambos idiomas. Luego, como actividad para llevar a casa, tuvieron que trabajar en equipos para explorar un espacio de sesgo bidimensional de su elección, diferente del género.

Durante la presentación de las herramientas, los estudiantes no solicitaron aclaraciones o explicaciones extensas sobre los matices de las incrustaciones de palabras, los sesgos representados como listas de palabras o las diferencias lingüísticas entre el inglés y el español y la adaptación de la herramienta. Sospechamos que esto se debió a que estaban condicionados por la metodología del nanogrado, que se basaba en clases en las que se explicaba una metodología y se mostraba cómo se aplicaba, seguidas de

[2] Nuestro código está disponible en https://tinyurl.com/ycxz8d9e

[3] El código original para inglés está disponible en https://learn.responsably.ai/word-embedding/

sesiones prácticas en las que realmente aplicaban la metodología a otros casos. Por lo tanto, no estaban tratando de ser críticos, sino de reproducir la metodología.

Los equipos aplicaron con éxito la metodología para caracterizar sesgos distintos al género, a saber, el sesgo económico (ricos frente a pobres), el sesgo geográfico (latinoamericano frente a norteamericano), el sesgo etario (viejo frente a joven). La Figura 4 ilustra uno de sus análisis del espacio de sesgo definido por *rich* [*rico*] frente a *poor* [*pobre*], mediante la exploración de cómo se posicionaron las palabras negativas y positivas en ese espacio. La figura muestra la lista de palabras que usaron para definir los dos extremos del espacio de sesgo, los conceptos de *pobre* y *rico.* También muestra qué tan cercanas están las palabras *gorgeous* [*hermosa*] y *violence* [*violencia*] de los conceptos rico o pobre. Los estudiantes concluyeron que el concepto de pobre está más asociado con palabras con un sentimiento negativo y rico con uno positivo.

No informaron de mayores dificultades o frustraciones y, en general, manifestaron estar satisfechos con sus hallazgos al aplicar la herramienta. También emplearon las estrategias de mitigación disponibles con la caja de herramientas Responsibly, pero no hicieron ningún análisis sobre su impacto.

Como no les fue requerido por la evaluación, no informaron sobre su proceso de exploración de forma sistemática. Para obtener más información sobre el proceso de exploración, incluimos en la experiencia con los científicos sociales una observación sobre el proceso.

En general, aplicaron la herramienta de manera satisfactoria pero poco crítica. Esto no sugiere que los participantes no fueran críticos per se, sino que la forma en que se presentó la metodología, alineada con un enfoque consistente para aplicar las metodologías aprendidas a lo largo del curso, inhibió una explotación más crítica y matizada de la herramienta.

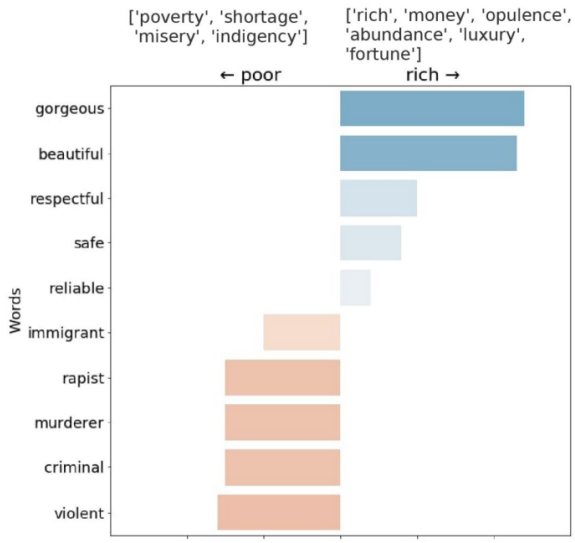

Figura 4 - Exploración del espacio de sesgo "rich" [rico] frente a "poor" [pobre] realizada por los científicos de datos, que muestra las palabras utilizadas para

definir los dos extremos del espacio de sesgo y cómo las palabras de interés, como "gorgeous" [hermosa] o "violence" [violencia], se posicionan con respecto a cada uno de los extremos. La exploración original se llevó a cabo en español y se ha traducido al inglés para facilitar la lectura.

4.2 **Estudio de caso 2: Científicos Sociales**

A los científicos sociales se les presentó la herramienta como parte de un taller de doce horas sobre herramientas para inspeccionar tecnologías de PLN. Se fomentó una visión crítica y se pidió explícitamente a los participantes su opinión para mejorar la herramienta.

Se dedicó mucho tiempo durante el taller para llevar a cabo la exploración del sesgo. Observamos y, en ocasiones, suscitamos sus procesos, desde la selección guiada de los sesgos a explorar, en función de los antecedentes y las experiencias personales, hasta la manipulación de las herramientas disponibles. Además, se establecieron conexiones explícitas entre las herramientas para explorar la incrustación de palabras disponibles a través de Responsibly y una plataforma interactiva para la comprensión del modelo de PLN, la Herramienta para la interpretabilidad lingüística (*Language Interpretability Too*l (LIT, por sus siglas en inglés)) (Tenney et al., 2020), lo que también sirvió de inspiración a los participantes para saber qué otra información podían obtener para enriquecer su análisis durante la exploración.

Dado que no se requería un informe formal, la exploración del sesgo no se describió de forma sistemática. Se exploraron diferentes sesgos, en diferentes profundidades y longitudes. Además del género y la edad, también se exploraron los detalles de origen (cubano - norteamericano) y la intersección entre la edad y la tecnología. Los participantes fueron creativos y trabajaron de forma colaborativa para encontrar palabras que representan satisfactoriamente los fenómenos que estaban tratando de evaluar. También fueron perspicaces en su análisis de los resultados: pudieron discutir diferentes hipótesis sobre por qué una palabra determinada podría estar más lejos en uno de los extremos del espacio que en otro.

La Figura 5 ilustra uno de sus análisis del espacio de sesgo definido por los conceptos de "*young*" [*joven*] y "*old*" [*viejo*], utilizando la posición de los verbos en este espacio para explorar el sesgo. En este análisis no llegaron a ninguna conclusión definitiva, sino que descubrieron que necesitaban más información sobre los datos textuales de los que se había inferido la incrustación. Por ejemplo, querían ver los contextos reales de ocurrencia de "*sleep"* [*sueño*] o "*argument*" [*argumento*] con "*old*" [*viejo*] y las palabras relacionadas, para explicar el hecho de que están más cerca del extremo del sesgo que representa el concepto "*old*". Al analizar estos resultados, el grupo también se dio cuenta de que había varios sentidos asociados a las palabras que representan el concepto "*old*", algunos de ellos positivos y otros negativos. También se dieron cuenta de que el concepto en sí mismo puede transmitir diferentes sesgos, por ejemplo, el respeto en algunas culturas o el desprecio en otras. Dichos hallazgos estaban más allá del alcance del simple análisis de la incrustación, lo que requería más datos contextuales para analizarlos adecuadamente y poder abordados posteriormente.

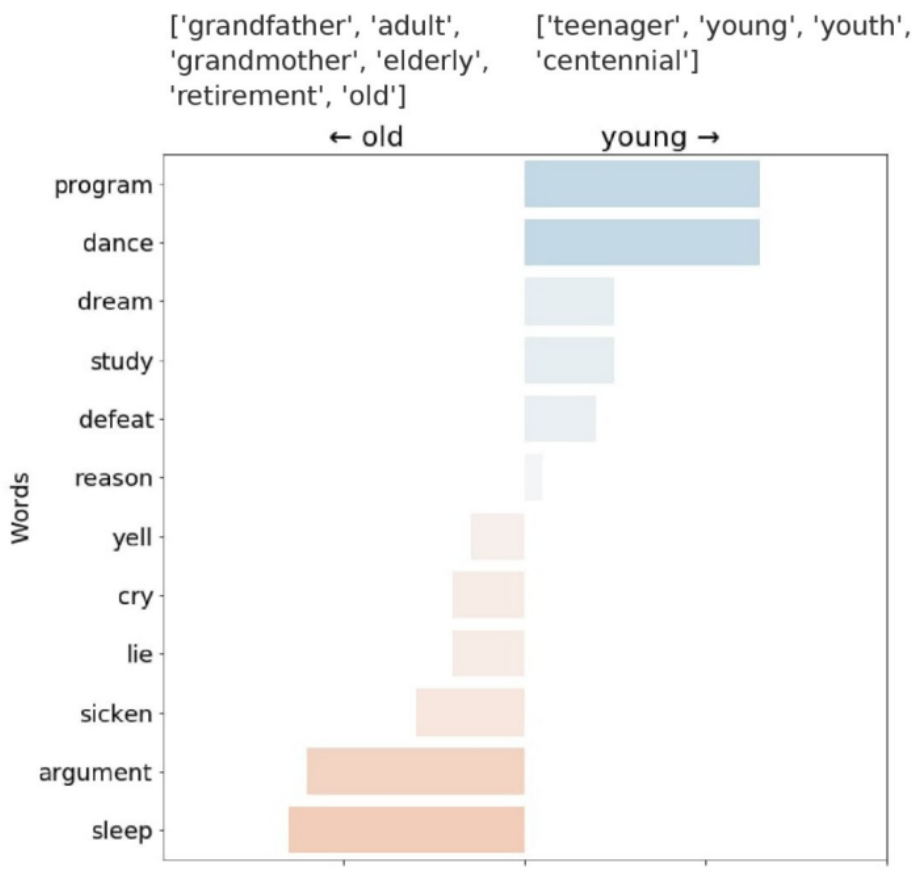

Figura 5 - Exploración del espacio de sesgo entre "old" [viejo] y "young" [joven] realizada por los científicos sociales, que muestra las palabras utilizadas para definir los dos extremos del espacio de sesgo y cómo las palabras de interés, como "dance" [bailar] o "sleep" [dormir], se posicionan con respecto a cada uno de los extremos. La exploración original se llevó a cabo en español y se ha traducido al inglés para facilitar la lectura.

Los participantes se mostraron activos y creativos al solicitar información complementaria sobre los textos de los que se habían inferido las incrustaciones de palabras. Muchos de estos requisitos estarán incluidos en el prototipo de la herramienta que estamos diseñando, como son los siguientes: las frecuencias de las palabras que se exploran, las concordancias de las palabras que se exploran, es decir, sería útil el poder examinar las producciones textuales reales al construir las listas de palabras, especialmente para los extremos que definen el espacio de sesgo, las sugerencias de palabras similares o palabras que están cercanas en el espacio incrustado, ya que a menudo es difícil encontrarlas y el espacio se representa mejor si se usan listas más largas para describir las funcionalidades de los extremos y una interfaz con el usuario que facilite la comparación entre diferentes delimitaciones (listas de palabras) del espacio de sesgo, o las mismas delimitaciones en diferentes incrustaciones.

También manifestaron que sería valioso que la herramienta les permitiera explorar diferentes incrustaciones, de diferentes lapsos, origen geográfico, publicaciones, géneros o dominios. Nuestro

prototipo incluye, con ese fin, la posibilidad de cargar un corpus y tener incrustaciones inferidas a partir de ese corpus, que luego pueden ser exploradas.

En general, notamos que los científicos sociales pidieron más contexto para sacar conclusiones durante la exploración. Fueron críticos con todo el proceso de análisis, con declaraciones como: “Siento que estoy torturando los datos”. Los científicos sociales estudian los datos en contexto, no los datos en sí mismos, como es más común en la práctica de los científicos de datos. Asimismo, el contexto en el que se presentó la herramienta también fomentó un enfoque más crítico. Pedimos a los participantes que formularan las pautas de sus exploraciones en términos de hipótesis. Tal requisito dejó en claro que se necesitaba más información sobre los datos de entrenamiento para poder formular hipótesis con mayor claridad.

4.3 **Lista de deseos para el prototipo**

En estos dos casos de estudio pudimos probar la idoneidad de algunos aspectos de nuestro enfoque, describir las deficiencias de las herramientas que estamos utilizando actualmente en esta etapa de desarrollo y también revaluar algunas decisiones de diseño con respecto a la herramienta propuesta.

A continuación, destacamos en resumen aquellas limitaciones que hemos identificado durante estos casos de estudio y que abordaremos en el prototipo. Luego discutimos las limitaciones intrínsecas de las capacidades actuales de las herramientas para la exploración de sesgos.

- Trabajar con **expresiones de múltiples palabras** como una unidad lingüística.
- Ser capaces de recuperar **los contextos reales de la ocurrencia de palabras** en el corpus de donde se ha obtenido la incrustación.
- Ser capaces de comparar cómo un mismo conjunto de palabras se relaciona con **extremos de sesgo definidos con diferentes listas de palabras.**
- Ser capaces de explorar y **comparar diferentes incrustaciones** y cómo se representa allí el mismo sesgo.
- Integrar **diferentes espacios de sesgo en una sola visualización**, de modo que podamos ver cómo una palabra se asocia a diferentes significados al mismo tiempo.

4.3.1 **Limitaciones del enfoque**

La limitación más evidente y frustrante es el hecho de que únicamente se pueden representar sesgos binarios con este enfoque, debido a la forma en que se define matemáticamente el espacio de sesgo, con sus dos extremos. Sin embargo, esta limitación se superó un poco con la interseccionalidad expresada en las propias listas de palabras, es decir, la construcción de listas de palabras que eran intersecciones, por ejemplo, una lista de palabras que es una intersección entre género y edad contendría las palabras “*grandmother*”, “*granny*”, “*old lady*” [“abuela”, “abuelita”, “anciana”] en un extremo y “*girl*”, “*lass*”, “*young lady*” [“niña”, “muchacha”, “jovencita”] en el otro.

Una de las deficiencias más interesantes que surgieron fue que no se podían tener en cuenta las ambigüedades de las palabras: ¿tiene una palabra diferentes sentidos, diferentes categorías morfosintácticas? Todas ellas se fusionan en una única representación de la palabra, lo que puede

generar ruido en la representación de un extremo del sesgo. Por ejemplo, la palabra "*gusano*" que se usa para "*Cuban*" ["*cubano*"] tiene el significado principal de "*worm*" ["*gusano"*], por lo que cuando se usa para definir un sesgo extremo para los cubanos, toma muchos contextos relacionados con los animales. Probaremos diferentes formas para abordar este problema en los estudios de utilizabilidad.

Asimismo, encontramos que, en algunos casos, uno de los extremos de la dirección del sesgo estaba lexicalizado, pero no el otro, como es el caso de "*football*" ["*fútbol americano*"]. La metodología es fuertemente binaria y no puede dar cuenta de tales casos, por frecuentes que sean. También se queda corta para dar cuenta del sesgo interseccional, aunque se pueden hacer algunas aproximaciones utilizando palabras en los extremos que incluyen una combinación de significados, representando así las intersecciones.

4.4 **Conclusiones de los estudios de caso**

Con estos casos de estudio, mostramos que reducir la complejidad técnica de la herramienta y las explicaciones al mínimo fomenta la participación de los expertos en el dominio. El proporcionar herramientas intuitivas como son las listas de palabras, en lugar de barreras como son los vectores, les permite formalizar su comprensión del problema, plasmando los conceptos y patrones relevantes en esas herramientas, a la vez que formulan hipótesis sobre esos términos e interpretan los datos. Dicho involucramiento es útil en diferentes momentos del ciclo de vida del software: el análisis de errores, el encuadre del problema, la curación del conjunto de datos y los artefactos obtenidos.

Nuestra conclusión es que se puede entender la inspección de los sesgos en las incrustaciones de palabras sin la mayor parte de los detalles técnicos subyacentes. Sin embargo, la caja de herramientas Responsibly adaptada al español necesita las mejoras que discutimos en esta sección y que desarrollamos como un plan de investigación aplicada en la Sección 6.

5 **Historia de los usuarios para nuestro prototipo**

Hasta aquí hemos motivado la necesidad de evaluar el sesgo en las tecnologías del lenguaje y, en particular, en las incrustaciones de palabras; hemos explicado nuestras diferencias en el encuadre de la solución con respecto a las herramientas existentes; hemos discutido las barreras técnicas artificiales de los enfoques existentes que dificultan la participación de expertos reales en el proceso de exploración y hemos elaborado una lista de deseos de científicos sociales que describen su herramienta ideal para la exploración del sesgo en las incrustaciones de palabras. En esta sección estamos describiendo una historia de usuarios que presenta las funcionalidades previstas de la herramienta propuesta y terminamos en la sección siguiente con un detalle de los pasos para desarrollarla, si se otorga la siguiente fase del proyecto.

Nos gustaría señalar que la historia de los usuarios se pensó originalmente como situada en Argentina, el contexto local de este proyecto. Sin embargo, para facilitar la comprensión de los lectores que no hablan español, adaptamos el caso para trabajar con el idioma inglés y, en consecuencia, situamos este caso en los Estados Unidos. El caso de estudio original en español se publicará más adelante como pieza en medios de comunicación.

Los usuarios. Marilina es una científica de datos que trabaja en un proyecto para desarrollar un software que ayude a la administración pública a clasificar las solicitudes de los ciudadanos y enviarlas al departamento más adecuado. Tomás es un trabajador social dentro de la oficina de no discriminación y quiere evaluar los posibles comportamientos discriminatorios de dicho software.

El contexto. Marilina aborda el proyecto como un problema de clasificación supervisada de texto. Para clasificar los nuevos textos de los ciudadanos, se comparan con documentos que han sido clasificados manualmente en el pasado. A los nuevos textos se les asigna la misma etiqueta que el documento que más se les parece. Calcular la similitud es un punto clave en este proceso y se puede hacer de muchas maneras: reglas de programación explícitas, a través del aprendizaje automático con ingeniería de características manual o mediante el aprendizaje profundo, donde las incrustaciones de palabras son un componente clave. Marilina observa que el último enfoque tiene la menor cantidad de errores de clasificación en los datos anteriores que ella separó para la evaluación (el llamado conjunto de prueba). Además, el aprendizaje profundo parece ser la solución preferida en estos días; a menudo se presenta como un gran avance para muchas tareas de procesamiento del lenguaje natural. Así que Marilina decide seguir esa opción.

Un componente importante del enfoque de aprendizaje profundo son las incrustaciones de palabras. Marilina decide probar una conocida incrustación de palabras pre-entrenada en el contenido de Wikipedia. Cuando lo integra en la línea de producción, el rendimiento del sistema aumenta: se clasifican correctamente más textos en su conjunto de pruebas.

La búsqueda del sesgo. Marilina decide revisar los resultados de la clasificación más allá de las cifras. Siendo descendiente de inmigrantes latinoamericanos, ella revisa documentos relacionados con este grupo social. Encuentra que las solicitudes de subvenciones para pequeñas empresas presentadas por los inmigrantes latinoamericanos o los ciudadanos de ascendencia latinoamericana a veces se clasifican erróneamente como problemas de inmigración y se envían al departamento equivocado. Estos errores conllevan un proceso más largo, en promedio, para la atención de estas solicitudes y a veces se pierden las solicitudes mal clasificadas. En algunos casos, esto hace que el solicitante abandone el proceso.

El hallazgo de sistematicidades en los errores. Intrigada por este comportamiento de la automatización del proceso, investiga más a fondo cómo se clasifican las solicitudes de los inmigrantes, en comparación con las de los no inmigrantes. Al igual que hizo con las solicitudes latinoamericanas, encuentra que los documentos presentados por otros inmigrantes tienen una tasa de clasificación errónea más alta que las solicitudes de quienes no son inmigrantes. Sospecha que otros grupos sociales pueden sufrir tasas de clasificación errónea más altas, pero se enfoca en los inmigrantes latinoamericanos porque tiene una mejor comprensión de la idiosincrasia de ese grupo y esto puede ayudarla a establecer una base para seguir investigando. Ella encuentra algunos patrones en las clasificaciones erróneas. En particular, encuentra que algunos negocios específicos, como son las peluquerías o las panaderías, acumulan más errores que otros.

El hallazgo del componente responsable del sesgo. Ella rastrea los detalles de cómo se procesan dichos documentos y descubre que se consideran más similares a otros documentos que no están relacionados con las actividades profesionales, sino con la inmigración. La incrustación de palabras es el componente del procesamiento que determina las similitudes, por lo que analiza la incrustación. Ella descubre que la

biblioteca resposibly.ai sirve para inspeccionar las incrustaciones de palabras de sesgo y utiliza algunas de sus funcionalidades para evaluar el sesgo: la proyección de palabras neutras en la dirección del sesgo o la métrica para medir el sesgo. Define un espacio de sesgo con "*Latin American*" ["*Latinoamericano*"] en un extremo y "*North American*" ["*Norteamericano*"] en el otro y comprueba la posición relativa de algunas profesiones con respecto a esos dos extremos, como se puede ver en la Figura 6, izquierda. Encuentra que, como sospechaba, algunas de las palabras relacionadas con el ámbito profesional están más fuertemente vinculadas a palabras relacionadas con latinoamericano que a palabras relacionadas con norteamericano; es decir, palabras como "*hairdresser*" ["*peluquera*"] están más cercanas al extremo latinoamericano. Sin embargo, las palabras más fuertemente asociadas a "norteamericano" no corresponden a sus intuiciones. No sabe cómo proceder con esta inspección más allá de los hallazgos anecdóticos y cómo tomar medidas con respecto a los hallazgos. Es entonces cuando pide ayuda a la oficina de no discriminación.

La evaluación de los daños. La oficina de no discriminación designa a Tomás para la tarea de evaluar el comportamiento discriminatorio del software. Informado por Marilina sobre sus hallazgos, él descubre que las clasificaciones erróneas implican algún daño para las personas afectadas que se tipifica entre las prácticas discriminatorias que la oficina trata de prevenir. La clasificación errónea implica que los procesos se hacen más largos que para otras personas, porque necesitan ser reclasificados manualmente antes de que realmente puedan ser atendidos. A veces, simplemente son desestimados por el funcionario público equivocado, lo que da lugar a una desigual prestación de los servicios. En muchos casos, el propio error tiene un efecto negativo en la autopercepción del emisor, haciéndole sentir menos merecedor y desincentivando la búsqueda de la asistencia o incluso la iniciativa empresarial. Tomás puede ver el resultado del sistema, pero no puede ver la razón de las clasificaciones (erróneas) del sistema, ya que desconoce los detalles técnicos del procesamiento.

La comprensión de las incrustaciones de palabras sin tecnicismos innecesarios. Tomás entiende que hay un componente subyacente del software que está impactando en el comportamiento de la clasificación. Marilina le explica qué es una incrustación de palabras pre-entrenada, y que una incrustación de palabras es una proyección de palabras desde un espacio disperso donde cada contexto de ocurrencia es una dimensión en un espacio denso donde hay menos dimensiones, obtenido con una red neuronal y que cada palabra es un vector con valores en cada una de esas dimensiones. Tomás siente que entender la incrustación está más allá de sus capacidades. Entonces Marilina se da cuenta y le explica que las palabras se representan como un resumen de sus contextos de ocurrencia en un corpus de textos, pero esto no se puede ver directamente, sino que se explora usando la similitud entre palabras, para que las palabras más similares estén más cercanas.

La búsqueda de una herramienta intuitiva para la exploración de sesgos. Ella le muestra algunas de las herramientas disponibles para evaluar el sesgo en la caja de herramientas Responsibly, pero Tomás no puede programar y se siente abrumado por el código. Luego busca una herramienta que no requiera este tipo de experiencia y encuentra el prototipo desarrollado por nuestro proyecto. Esta es una herramienta accesible para Tomás, que presenta los conceptos clave de forma intuitiva y que puede manipular de forma autónoma. Con ello Tomás se siente empoderado y continúa con la exploración.

La familiarización con el corpus subyacente a la incrustación. Para empezar, Tomás quiere explorar las palabras que se consideran similares a "*Latin American*" ["*latinoamericano*"], porque quiere ver qué

palabras pueden estar fuertemente asociadas al concepto, aparte de lo que ya había observado Marilina. Se da cuenta que la incrustación ha sido entrenada con textos de periódicos. La mayoría de las noticias que contienen la palabra latinoamericano trata sobre catástrofes, problemas y otras noticias negativas de los países latinoamericanos o retratan a los latinoamericanos estereotipados, refiriéndose a las costumbres típicas de sus países de origen más que a sus facetas como ciudadanos en los Estados Unidos. Con respecto a los negocios y las profesiones, los latinoamericanos tienden a ser representados de acuerdo con los estereotipos y ocupaciones históricas predominantes de ese grupo social en los Estados Unidos, como son los trabajadores de la construcción, meseros, y peones agrícolas, entre otros.

Concluye que este corpus y, en consecuencia, la incrustación de palabras que de él se deriva, contiene muchos estereotipos sobre los latinoamericanos que luego se trasladan al comportamiento del software de clasificación, al asociar ciertas actividades profesionales y grupos demográficos más fuertemente con la inmigración que con los negocios. Posiblemente tendrán que encontrar una mejor incrustación de palabras, que no tenga esos sesgos ni que éstos sean tan marcados, pero él quiere caracterizar los sesgos primero para poder hacer una evaluación más rápida en otras incrustaciones de palabras.

La comprensión de la exploración de sesgos en incrustaciones de palabras sin tecnicismos innecesarios. Tomás debe centrar su exploración de la incrustación de palabras en el sesgo de interés, en este caso, en el latinoamericano frente al norteamericano. Marilina recurre a los materiales disponibles de nuestra herramienta para explicarle fácilmente a Tomás la definición y exploración de sesgos. Él capta rápidamente los conceptos del espacio del sesgo, la definición del espacio mediante las listas de palabras, la evaluación mediante la observación de cómo se colocan las palabras dentro de ese espacio y la exploración mediante la modificación de las listas de palabras, tanto en la definición del espacio como en su colocación en el espacio. Él obtiene más información sobre las posibilidades de las técnicas y sobre posibles malentendidos leyendo ejemplos y viendo breves tutoriales que se pueden encontrar en la herramienta. Entonces comprende que la ambigüedad de las palabras puede oscurecer los fenómenos que uno quiere estudiar, que la frecuencia de las palabras tiene un gran impacto y que los fenómenos específicos del idioma, como son el género gramatical o los niveles de formalidad, deben tenerse en cuenta cuidadosamente.

La formalización de un punto de partida para la exploración de los sesgos. Tomás es ahora capaz de sistematizar la exploración de sesgos, con el objetivo final de hacer un informe y tomar medidas informadas y basadas en principios para prevenir y reparar cualquier comportamiento discriminatorio que el proceso automatizado pudiese haber desplegado. En primer lugar, construye los conjuntos de palabras que representarán cada uno de los extremos del espacio de sesgo. Se da cuenta de que el enfoque de Marilina con una sola palabra en cada extremo no es muy sólido, porque puede estar fuertemente influenciado por las propiedades de esa sola palabra. Por ello define cada uno de los extremos del espacio de sesgo con listas de palabras más largas y experimenta con diferentes listas y con cómo éstas determinan la posición relativa de sus palabras de interés. Las palabras de interés son las palabras que se están posicionando en el espacio de sesgo, palabras que Tomás quiere caracterizar con respecto a este sesgo porque sospecha que su caracterización es una de las causas del comportamiento discriminatorio del software de clasificación.

Para encontrar palabras a ser incluidas en las listas de palabras de los extremos, Tomás recurre a la funcionalidad que permite encontrar las palabras más cercanas en la incrustación. Usando “*Latin*

American" *[*"*latinoamericano*"] como punto de partida, encuentra otras palabras similares como "*Latino*" ["*Latin*"] y también nacionalidades de América Latina.

También explora los contextos de sus palabras de interés. Al hacer esto, descubre que "*shop*" ["*comprar*"] ocurre en muchos más contextos de los que había imaginado originalmente, muchos con diferentes significados, por ejemplo, la abreviatura de Photoshop. Esto le hace pensar que esta palabra probablemente no sea un buen indicador del tipo de comportamiento en las palabras que está tratando de caracterizar. También encuentra que algunas profesiones que inicialmente le interesaban, como "*capoeira trainer*" ["*entrenador de capoeira*"], son muy poco frecuentes y su caracterización no tiene correspondencia con su intuición sobre el significado y el uso de la palabra, por lo que las descarta.

Finalmente, está satisfecho con la definición que le brindan las listas de palabras que se pueden ver en la Figura 6, a la derecha. Con esa lista de palabras, la caracterización de las palabras de interés muestra tendencias que tienen correspondencia con las malas clasificaciones del sistema final: las solicitudes de estilistas, panaderos, modistas de origen o ascendencia latina son mal clasificadas con más frecuencia que las solicitudes para otro tipo de negocios.

El informe sobre los sesgos y la propuesta de una estrategia de mitigación. Tomás puede, con esta caracterización, realizar un informe detallado del comportamiento discriminatorio del sistema de clasificación. Desde el principio, él pudo describir las sistematicidades que se pueden encontrar en los errores, las cuales afectan con mayor frecuencia a las personas de ascendencia latinoamericana que solicitan subsidios para cierto tipo de negocios. Sin embargo, su comprensión de la incrustación de palabras subyacente le permite describir un patrón de comportamiento que va más allá de los casos que realmente ha podido ver como mal clasificados por el sistema y predecir otros casos.

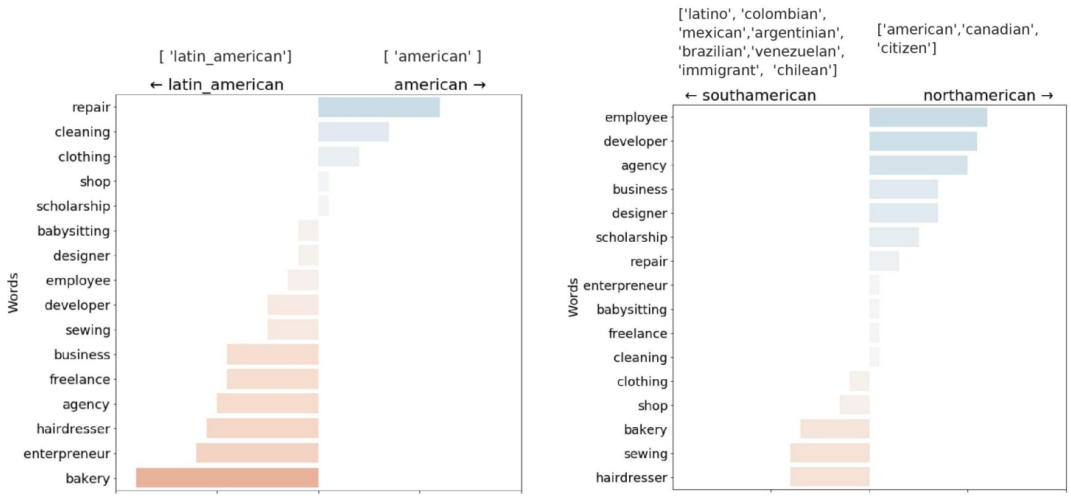

Figura 6 - Diferentes caracterizaciones del espacio de sesgo "Latin American" vs "North American" ["latinoamericano" vs " norteamericano"], con diferentes listas de palabras creadas por un científico de datos (izquierda) y un científico social (derecha) y el efecto diferente para definir el espacio de sesgo como se refleja en la posición de las palabras de interés (columna de la izquierda).

Además, comprender el patrón de comportamiento le permite describir las propiedades del corpus subyacente que contribuirá a una mejor incrustación de palabras, con menos sesgos. Puede proponer estrategias como la edición de las oraciones que contienen estilistas, diseñadores y panaderos para mostrar una mezcla más equilibrada de nacionalidades y etnias en ellas.

Mayor amplitud y profundidad. Si el tiempo lo permite, también podrá explorar la interseccionalidad y compararla con otras incrustaciones, todo en una interfaz visual con conceptos intuitivos.

6 Propuesta de investigación aplicada

En esta sección describimos primero las funcionalidades que incluirá el prototipo. A continuación, enumeramos las actividades planeadas para implementar dichas funcionalidades. En la última subsección, explicamos el papel crucial que tendrán las experiencias piloto en el desarrollo iterativo de nuestro prototipo. También presentamos a nuestros socios regionales para las experiencias piloto.

6.1 Funcionalidades del prototipo

Los principios de diseño y las funcionalidades del prototipo que priorizaremos son los siguientes:

- Centrarse en la **exploración** del sesgo, no en el diagnóstico basado en métricas o en la mitigación dentro de las incrustaciones de palabras, como es el caso de los enfoques existentes.

- Facilitar **la comparación de las listas de palabras** utilizadas para definir el espacio de sesgo, como un medio para evaluar los efectos de las diferentes palabras y sus combinaciones para definir el espacio de sesgo.

- Permitir **la comparación entre diferentes colecciones de textos**: diferentes épocas, diferentes regiones, diferentes autores, diferentes medios, etc.

- Facilitar la **inspección** de los contextos de ocurrencia de las palabras en las listas, su frecuencia y cualquier otra información que permita identificar las razones de los comportamientos o sesgos insospechados, como son las palabras poco frecuentes que están fuertemente asociadas a otras palabras simplemente por ocurrencias fortuitas.

- Caracterizar **expresiones con múltiples palabras** como una unidad léxica.

- Reducir la **complejidad técnica** al mínimo necesario y sustituir conceptos altamente técnicos por conceptos más intuitivos siempre que sea posible.

6.2 Actividades planificadas

En este artículo, hemos presentado una metodología para el tipo de participación que puede enriquecer los enfoques para la exploración de sesgos de los artefactos de PLN con el conocimiento de dominio

necesario para modelar adecuadamente los problemas de interés. Hasta este punto de nuestra investigación, nos hemos basado en las herramientas proporcionadas por la caja de herramientas Responsibly y nuestra adaptación al español. Dado nuestro trabajo de campo con diferentes tipos de usuarios, ahora estamos en el punto donde podemos comenzar el desarrollo de una herramienta independiente que integre tanto las capacidades técnicas de los enfoques existentes como los requisitos de diseño valorados por nuestro enfoque y los usuarios que estudiamos.

En esta sección detallamos las actividades que llevaremos a cabo para desarrollar dicho prototipo.

El desarrollo del prototipo será iterativo, siguiendo una metodología ágil que será validada en varios puntos del proceso de desarrollo con estudios de utilizabilidad que describimos en el siguiente subapartado, seguido de un piloto más completo. El cuadro 1 organiza las siguientes actividades en siete meses desde julio de 2022 hasta enero de 2023.

1. Proporcionar una interfaz gráfica inspirada en algunas de las ideas del área de Interacción Humano-Computadora (HCI, por sus *siglas en inglés*) en la herramienta *WordBias* que presentamos en la Sección 3.

2. Proporcionar visualizaciones comparativas que registren el historial de las interacciones con el prototipo, lo cual permite comparar:

    - variaciones en los extremos del espacio de sesgos
    - variaciones a través de diferentes incrustaciones, diacrónicamente (p. ej., periódicos)
    - combinaciones de diferentes espacios (es decir, interseccionalidad)

3. Alojar nuestro prototipo en *huggingface*[4], para importar incrustaciones pre-entrenadas y ofrecer nuestra herramienta a la comunidad de practicantes de PLN que tiene *huggingface*.

4. Tener la posibilidad de entrenar incrustaciones de palabras, dado un corpus y proporcionar métricas de confiabilidad de la incrustación palabra por palabra.

5. Mostrar la siguiente información adicional sobre las palabras:

    - la frecuencia con respecto al tamaño del corpus
    - las concordancias, el contexto de las ocurrencias
    - la mayoría de las palabras similares
    - la similitud promedio con las palabras más similares

[4] https://huggingface.co/

6. Definir métricas que evalúen la calidad de las listas de palabras, en función de sus propiedades estadísticas.

7. Ampliar las incrustaciones con n-gramas como se ha explicado anteriormente con expresiones de múltiples palabras.

8. Sugerir estrategias de mitigación que impliquen la comparación de diferentes corpus o la modificación de los corpus originales. Por ejemplo, un corpus en español podría neutralizar el género antes de entrenar las incrustaciones de palabras usando el género neutral 'e'.

9. Evaluar si es posible aplicar nuestra metodología de exploración de sesgos a los métodos de incrustación contextual utilizados en grandes modelos lingüísticos (Zhao et al., 2019; Sedoc y Ungar, 2019).

10. Estudios de utilizabilidad para el desarrollo ágil que describimos a continuación.

11. Integración con la práctica de las políticas públicas.

Meses	J	A	S	O	N	D	E
1. interfaz gráfica	X		X				
2. visualizaciones	X	X					
3. *huggingface*			X				
4. entrenamiento de incrustaciones			X				
5. información sobre las palabras				X			
6. calidad de la lista de palabras				X			
7. incrustaciones de n-gramas				X	X		
8. mitigación del corpus			X			X	
9. estudios de utilizabilidad			X		X		
10. política pública					X	X	X
11. entrega de prototipos							X

Cuadro 1: Planificación de las actividades mensuales desde julio de 2022 hasta enero de 2023.

6.3 **Estudios de utilizabilidad y piloto**

El objetivo general de nuestro proyecto es desarrollar herramientas utilizables por cualquier persona sin perfil técnico o especialización en lenguaje natural de IA [s*iglas para Inteligencia Artificial*] y procesamiento de datos, pero, al mismo tiempo, buscamos promover el uso de esta herramienta dentro de las comunidades de las ciencias sociales, para integrarla con prácticas que estas comunidades han establecido. Por eso, estamos trabajando con FLACSO (Universidad Latinoamericana de Ciencias Sociales) como socio estratégico para realizar estudios de utilizabilidad y un piloto.

FLACSO es una institución con una larga trayectoria en la región, con sedes en Argentina, Costa Rica, Ecuador y México y estudiantes e investigadores de toda Latinoamérica, que trabajan en áreas de las ciencias sociales vinculadas a los estudios de género, migraciones, pueblos originarios, cultura y comunicación, economía, cambio climático y cuestiones emergentes desde la perspectiva de las políticas públicas. En la sede argentina, FLACSO cuenta con un área de bioética, propiedad intelectual y bienes públicos. Algunos miembros de nuestro equipo y del equipo ampliado de Fundación Vía Libre son egresados, profesores e investigadores de FLACSO, por lo que contamos con alianzas bien establecidas y una larga trayectoria de trabajo conjunto entre Fundación Vía Libre y FLACSO, Argentina.

Visualizamos dos tipos de interacción con la comunidad FLACSO. Primero, estudios de utilizabilidad con pequeños grupos focales para obtener realimentación cualitativa inmediata sobre las características que estaremos implementando durante el desarrollo del prototipo y que nos permitan realizar ajustes rápidos en el desarrollo. Estos están planificados como “estudios de utilizabilidad” dentro del plan de investigación aplicada y tienen como objetivo elegir las opciones más adecuadas para la interfaz gráfica del usuario, las visualizaciones, la utilizabilidad de las funcionalidades para cargar y entrenar las incrustaciones y para inspeccionar los corpus subyacentes y los contextos de ocurrencia de palabras.

Luego, si se otorga la siguiente fase del proyecto, planeamos implementar un estudio en grupos más grandes, para evaluar las dificultades y potencialidades de la herramienta al usarla con una población más amplia de usuarios. Luego, usaremos estos hallazgos para desarrollar materiales complementarios para facilitar y promover el uso de esta herramienta en diferentes contextos: tutoriales, videos, piezas informativas, documentación y páginas de ayuda en línea.

Nos proponemos hacer un esfuerzo especial para facilitar la integración de esta herramienta y enfoques similares como parte integrante de los procesos de seguimiento, evaluación y mitigación de los comportamientos discriminatorios de las tecnologías lingüísticas. Centraremos nuestros esfuerzos en integrar la herramienta dentro del flujo de trabajo de los agentes de políticas públicas, ya sean entidades reguladoras o gubernamentales que necesitan tomar decisiones basadas en datos. Estamos pensando en casos de uso dentro del área de protección al consumidor, como son la inspección de sesgos en tecnologías del lenguaje de proveedores estatales o los productos con grandes audiencias o la evaluación de sesgos cuando se reciben quejas de los usuarios.

6.4 **Integración con la política pública**

Como se indicó en la introducción, uno de nuestros objetivos al reducir las barreras técnicas para acceder a las herramientas de evaluación de sesgos es precisamente hacer posible que los tomadores de

decisiones comprendan las capacidades y limitaciones de las aplicaciones de procesamiento del lenguaje actuales y que puedan planificar políticas públicas que tengan en cuenta el impacto de estas tecnologías.

Dentro del alcance de nuestro trabajo, investigaremos cómo se puede integrar esta herramienta con los esfuerzos para evaluar si las tecnologías del lenguaje cumplen con las normas y estándares legales. Parece necesaria una herramienta de evaluación de sesgos para llevar a cabo las evaluaciones de cumplimiento, porque el objetivo declarado de transparencia de las tecnologías no parece ser suficiente para resolver los problemas derivados de los potenciales problemas asociados.

También investigaremos cómo se puede integrar esta herramienta de forma más proactiva para prevenir comportamientos discriminatorios antes de que se realicen las automatizaciones. Estamos pensando en referentes colaborativos para detectar diferentes tipos de sesgos por los que pueden pasar las herramientas antes de ser implementadas, como una norma de calidad para el software con impacto social.

Varias organizaciones internacionales, incluida la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], han iniciado procesos para la creación de marcos éticos para el desarrollo, adopción e implementación de tecnologías de inteligencia artificial en la vida pública. Una de las pautas de dichos marcos establece que toda tecnología debe tener varias métricas y herramientas para una evaluación de impacto adecuada, antes de que ésta se ponga en funcionamiento. Estaremos trabajando en esta herramienta con este marco de referencia.

6.4.1 **Una nota sobre el marco legal: Derechos Humanos en América Latina**

La Declaración Universal de los Derechos Humanos aprobada en 1948 establece el marco normativo general diciendo que “Todos los seres humanos nacen libres e iguales en dignidad y derechos y, dotados de razón y conciencia, deben comportarse fraternalmente los unos con los otros” y continúa diciendo que “Toda persona tiene los derechos y libertades enunciados en esta Declaración, sin distinción alguna por motivos de raza, color, sexo, idioma, religión, opinión política o de otra índole, origen nacional o social, posición económica, nacimiento o cualquier otra condición”.

Sin embargo, la sociedad humana está lejos de cumplir esta visión de quienes promovieron y promueven la igualdad y la justicia como derechos fundamentales. La discriminación por motivos de etnia, color, género, idioma, religión, opinión, posición económica y diversidad de capacidades sigue estando a la orden del día a pesar de los marcos legales establecidos para su erradicación. En los últimos años, las tecnologías del lenguaje han sido un importante agente de discriminación, sobre todo por su escala masiva, donde un solo programa puede afectar a millones de personas a través de millones de dispositivos que funcionan las 24 horas del día. Los comportamientos discriminatorios de los sistemas automatizados pueden estar sujetos a leyes contra la discriminación.

En Argentina, desde 1994, la Declaración Americana de los Derechos y Deberes del Hombre; la Declaración Universal de Derechos Humanos; la Convención Americana sobre Derechos Humanos; el Pacto Internacional de Derechos Económicos, Sociales y Culturales; y el Pacto Internacional de Derechos Humanos tienen rango constitucional. Los Derechos Civiles y Políticos y su Protocolo Facultativo, la Convención para la Prevención y la Sanción del Delito de Genocidio, la Convención Internacional sobre la

Eliminación de Todas las Formas de Discriminación Racial, la Convención sobre la Eliminación de Todas las Formas de Discriminación contra la Mujer; la Convención contra la Tortura y Otros Tratos o Penas Crueles, Inhumanas o Degradantes; la Convención sobre los Derechos del Niño; en las condiciones de su vigencia, tienen jerarquía constitucional, no derogan artículo alguno de la primera parte de esta Constitución y deben entenderse como complementarios de los derechos y garantías por ella reconocidos. Sólo podrán ser denunciados, en su caso, por el Poder Ejecutivo Nacional, previa aprobación de las dos terceras partes de todos los miembros de cada Cámara.

Ese mismo marco constitucional es compartido por los países de toda América Latina, región desde la cual realizamos esta investigación. Es este marco de Derechos Humanos el que establece la base normativa que luego deriva en regulaciones específicas sobre cada uno de los ámbitos abordados en este campo. Sin embargo, entendemos que no es necesaria una legislación específica para los usos y aplicaciones de los sistemas automatizados de toma de decisiones, sino la posibilidad de aplicar los marcos normativos y legales vigentes, con el consiguiente cumplimiento de las garantías y derechos establecidos con el más alto rigor legal.

Referencias

Steve Aibuedefe Aigbe y Christoph Eick. 2021. Learning domain-specific word embeddings from covid-19 tweets. En *2021 IEEE International Conference on Big Data (Big Data)*, páginas 4307–4312.

Haozhe An, Xiaojiang Liu y Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. En *Findings of the Association for Computational Linguistics: ACL 2022*, páginas 1139–1152, Dublin, Irlanda. Association for Computational Linguistics.

Maria Antoniak y David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Páginas 1889–1904, en línea. Association for Computational Linguistics.

Pablo Badilla, Felipe Bravo-Márquez y Jorge Pérez. 2020. WEFE: the word embeddings fairness evaluation framework. En *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, páginas 430–436.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major y Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Páginas 610–623, New York, NY, EUA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III y Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 5454–5476, en línea. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama y Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. En *Proceedings of the 30th*

International Conference on Neural Information Processing Systems, NIPS'16, páginas 4356–4364, Red Hook, NY, EUA. Curran Associates Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson y Richard Zemel. 2019. Understanding the origins of bias in word embeddings. En *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, páginas 803–811. PMLR.

Aylin Caliskan, Joanna J. Bryson y Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cristián Cardellino. 2019. Spanish Billion Words Corpus and Embeddings.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang y Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. En *Workshop Practical Machine Learning for Developing Countries: learning under limited resource scenarios at International Conference on Learning Representations (ICLR)*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela y Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 8173–8188, en línea. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem y Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 1905–1925, en línea. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky y James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Bhavya Ghai, Md Naimul Hoque y Klaus Mueller. 2021. Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. En *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, EUA. Association for Computing Machinery.

Hila Gonen y Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. En *NAACL-HLT*.

Hila Gonen, Yova Kementchedjhieva y Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gendermarking languages? En *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, páginas 463–471, Hong Kong, China. Association for Computational Linguistics.

Wei Guo y Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. En *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, páginas 122–133, New York, NY, EUA. Association for Computing Machinery.

Yue Guo, Yi Yang y Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. En *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1012–1023, Dublin, Irlanda. Association for Computational Linguistics.

Shlomi Hod. 2018. Responsibly: Toolkit for auditing and mitigating bias and fairness of machine learning systems. [en línea; accesado <hoy>].

Masahiro Kaneko y Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. En *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, páginas 212–223, en línea. Association for Computational Linguistics.

Anne Lauscher y Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. En *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, páginas 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Pierre Lison y Andrey Kutuzov. 2017. Redefining context windows for word embedding models: An experimental study. En *Proceedings of the 21st Nordic Conference on Computational Linguistics*, páginas 284–288, Gothenburg, Suiza. Association for Computational Linguistics.

Malvina Nissim, Rik van Noord y Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Flavien Prost, Nithum Thain y Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. En *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, páginas 69–75, Florencia, Italia. Association for Computational Linguistics.

Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom y Henriette Cramer. 2021. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. En *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, EUA. Association for Computing Machinery.

João Sedoc y Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. En *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, páginas 55–61, Florencia, Italia. Association for Computational Linguistics.

Amit Singhal. 2001. Modern information retrieval: A brief overview. . *IEEE Data Engineering Bulletin*, 24(4):35–43.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif y Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, páginas 107–118, en línea. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez y Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell y KaiWei Chang. 2019. Examining gender bias in languages with grammatical gender. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 5276–5284, Hong Kong, China. Association for Computational Linguistics.