UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERÍA

Estudio de sesgos en representaciones vectoriales de palabras

Informe de Proyecto de Grado presentado por

María Fernanda Cánepa Romero
Sebastián Lagomarsino Etchandy

en cumplimiento parcial de los requerimientos para la graduación de la carrera de Ingeniería en Computación de Facultad de Ingeniería de la Universidad de la República

Supervisores

Aiala Rosá
Lorena Etcheverry
Álvaro Cabana

Montevideo, 21 de noviembre de 2023

Agradecimientos

A todas las personas que nos acompañaron en la carrera, a las que están y a las que ya no.

Fernanda Cánepa, Sebastián Lagomarsino

Resumen

Este proyecto se centró en el análisis de sesgos regionales en representaciones vectoriales de palabras (*word embeddings*) en el contexto del Río de la Plata. El objetivo principal fue explorar si los modelos de *word embeddings* entrenados en español reflejan sesgos específicos de esta área geográfica y cultural. El proyecto se llevó a cabo en dos etapas, la creación y ajuste de modelos de *word embeddings* y la evaluación de estos utilizando diversas pruebas.

Para abordar este objetivo, se utilizó la biblioteca `gensim` de procesamiento de lenguaje natural y se crearon modelos de *word embeddings* con Word2Vec y Fasttext. También se ajustaron modelos existentes de *word embeddings* al español rioplatense, con la intención de capturar de manera más precisa las particularidades léxicas y semánticas de esta región. El corpus de entrenamiento y ajuste fue formado por textos de noticias de Uruguay y Argentina.

Además, se diseñaron y adaptaron al español pruebas de evaluación de *word embeddings*. Estas pruebas se utilizaron para evaluar el rendimiento de los modelos, para identificar su capacidad de reflejar el léxico y los matices del Río de la Plata tratando de identificar el uso de palabras típicas de la región, y por último, para determinar o no la presencia de sesgos en los modelos. Las pruebas de sesgo se realizaron bajo los subespacios de estudio del género binario (femenino-masculino), la raza (blanca-negra) y el concepto de colonización (colonizado-colonizador).

En el análisis no se llegó a una conclusión definitiva sobre la existencia de sesgos específicos del Río de la Plata en los modelos de *word embeddings*. Sin embargo, uno de los logros significativos de este proyecto fue la creación de un conjunto de pruebas adaptadas al español para evaluar sesgos. Este recurso puede ser de utilidad para investigaciones futuras que busquen abordar cuestiones de sesgo en modelos de *word embeddings* en idioma español.

Palabras clave: *word embeddings*, sesgo, procesamiento del lenguaje natural, PLN.

Índice general

Capítulo 1

Introducción

En los últimos años, las representaciones vectoriales de palabras, llamadas *word embeddings*, han cobrado gran importancia como mecanismo para transmitirle a los modelos de aprendizaje automático el significado de las palabras que las personas utilizamos en lenguaje natural, convirtiéndolos en una de las herramientas más utilizadas en esta área. La calidad de las aplicaciones que utilizan *word embeddings* está muy acoplada a la calidad de estos. Esto supone un problema, dado que las estructuras de los *word embeddings* son difícilmente interpretables por las personas. El análisis de calidad de los vectores sugiere toda un área de estudio dado que estamos tratando de embeber el significado de algo que presenta una gran flexibilidad léxica y semántica, como lo es la lengua. Sumado a esta dificultad, nos encontramos con que la mayoría de las representaciones existentes, así como los estudios en el área de Procesamiento del Lenguaje Natural (PLN), se encuentran en inglés (Khurana, Koli, Khatter, y Singh, 2023).

Una problemática que se ha popularizado en las últimas décadas sobre *word embeddings* es que estas estructuras son replicadoras de sesgos (Bolukbasi, Chang, Zou, Saligrama, y Kalai, 2016a). Cuando comenzaron a surgir los modelos de inteligencia artificial (IA), sus aplicaciones eran principalmente técnicas, ocupando ámbitos de bajo impacto. Luego, se empezaron a utilizar estos modelos en ámbitos de alto impacto, donde se encontraron en mayor contacto con la sociedad. Es cuando se empieza a dar este movimiento que se evidencia lo que supone que estas estructuras porten y repliquen el sesgo de la sociedad, generando resultados con impacto directo y muchas veces negativo hacia las personas, propagando la discriminación y prejuicios a diferentes grupos sociales, étnicos o culturales (O'Neil, 2017). A partir de estos resultados comienza a surgir el área de *fairness* o equidad en aprendizaje automático, que estudia a los modelos de IA para que no generen resultados injustos basados en atributos como género, raza, entre otros (Barocas, Hardt, y Narayanan, 2019).

Es importante remarcar que la presencia de sesgo en un modelo no implica que este funcione mal, muchas veces incluso su presencia contribuye al correcto funcionamiento del modelo. Lo importante es que la aplicación final del modelo

no se vea perjudicada por la existencia del sesgo. Los buenos modelos son caros de construir, dado que requieren acceso a grandes volúmenes de datos y una gran capacidad de procesamiento, es por esto que existen muchos estudios como (Bolukbasi, Chang, Zou, Saligrama, y Kalai, 2016b) y (Zhao, Wang, Yatskar, Ordonez, y Chang, 2017) que se enfocan en el análisis y la mitigación del sesgo en los modelos existentes. Pero también es importante entender que resulta utópico pensar que podemos tener modelos que no tienen ningún tipo de sesgo, incluso la falta de sesgo en algunos casos puede generar modelos malos o poco representativos de la realidad que se intenta transmitir. Lo necesario es que los nuevos modelos no solamente se midan utilizando métricas clásicas, sino que también se midan con un enfoque ético, estudiando los sesgos que puedan portar y que estos resultados queden bien documentados, indicando y contraindicando escenarios de aplicación de estos (Bender y Friedman, 2018).

Este proyecto de investigación está centrado en el uso de *word embeddings* para representar el lenguaje natural, así como en la identificación y análisis de sesgos en estas representaciones.

Nuestro primer objetivo consistió en obtener una comprensión sólida de los *word embeddings*, investigando los métodos conocidos para su generación, utilizando diferentes fuentes de datos. Otro objetivo fue realizar un estudio del sesgo en el área de la IA, abarcando un espectro amplio desde los sesgos presentes en los datos en general y, en específico, en los recursos de lenguaje (textos y representaciones de palabras). Como siguiente objetivo, nos planteamos la aplicación de lo estudiado para llevar a cabo la generación de conjuntos de *word embeddings*. En esta etapa decidimos desarrollar conjuntos específicos para el dialecto rioplatense, partiendo de datos provenientes de Uruguay y Argentina. Como último objetivo nos propusimos analizar los sesgos en los diferentes conjuntos de *word embeddings* que generamos, estudiando diferentes dimensiones, como género, raza y profesiones entre otras.

Durante la realización del proyecto decidimos enfocarnos en el análisis de sesgos regionales, para evidenciar la importancia de la sub-representación de los dialectos en modelos de aprendizaje automático. Si bien no pudimos cumplir en su totalidad lo propuesto, logramos hacer una recopilación de pruebas de evaluación de *word embeddings* adaptándolas al idioma español, en particular pruebas enfocadas al análisis del sesgo.

En este documento presentamos en primer lugar los conceptos previos necesarios para el desarrollo del proyecto, en particular definimos los conceptos de *word embeddings* y de sesgo algorítmico, introduciendo la problemática del sesgo en los modelos de PLN. Luego, describimos la etapa de desarrollo de *word embeddings* específicos para el dialecto rioplatense y las pruebas de evaluación realizadas, con los resultados obtenidos. Por último, presentamos diferentes pruebas de análisis de sesgo en conjunto con los resultados obtenidos para cada una.

Dejamos disponibles los *notebooks* utilizados para la creación de *embeddings*, *fine-tuning* y pruebas realizadas en el repositorio del proyecto[1].

[1] Repositorio del proyecto https://gitlab.fing.edu.uy/sesgo-pln-2022/estudio-de-sesgos-en-representaciones-vectoriales-de-palabras

Además, como entregable desarrollamos un *data statement* del conjunto de *word embeddings* del proyecto *Small World Of Words*. Esta documentación busca recopilar diferentes características de los conjuntos de datos y resulta de utilidad cuando queremos hacer uso de un conjunto de datos para una determinada aplicación. Para su creación nos basamos en gran parte en el paper oficial del modelo (Cabana, Zugarramurdi, Valle-Lisboa, y De Deyne, 2023), en particular la sección de limitantes la escribimos utilizando los resultados más relevantes encontrados en este proyecto, mencionando aplicaciones que consideramos podrían generar resultados perjudiciales si utilizaran el modelo.

1.1. Cronograma

Desarrollamos este proyecto en un total de 15 meses, comenzando en agosto del 2022 y finalizando en noviembre de 2023. El cronograma que seguimos durante estos meses fue el siguiente:

- **Mes 1-3**: Estudio del estado del arte y recolección de recursos.
- **Mes 4-7**: Creación de corpus y *word embeddings*.
- **Mes 8-10**: Creación de pruebas de análisis de calidad y de sesgos de los modelos creados.
- **Mes 11-13**: Evaluación de resultados.
- **Mes 15**: Defensa.
- **Mes 1-15**: Documentación del informe y pruebas realizadas.

Capítulo 2

Conceptos previos

Dentro del procesamiento de lenguaje natural (PLN), surge la necesidad de hacer que las palabras y oraciones sean interpretables para las máquinas, de forma que no solamente puedan entender el lenguaje humano sino también interpretarlo y generarlo. Dos elementos fundamentales en el contexto actual para lograr esto son los *word embeddings* y las redes neuronales.

Los *word embeddings* también llamados vectores de palabras o incrustaciones de palabras, son representaciones vectoriales multidimensionales de las palabras. Su uso como modelos de representación de palabras surge ante la necesidad de transmitir el significado de las palabras a una computadora, permitiendo su utilización en aplicaciones de PLN (Jurafsky y Martin, 2009).

Los *word embeddings* toman su nombre de la idea de que la semántica de las palabras se encuentra embebida en un espacio vectorial. En este documento utilizaremos también la palabra *embeddings* para referirnos a estas representaciones.

La idea de representar palabras como vectores aparece en (Osgood, Suci, y Tannenbaum, 1957) y, en conjunto con la hipótesis distribucional de que el significado de las palabras está representado por las palabras que las rodean (Harris, 1954), surge lo que hoy en día es el formato estándar para representar el significado de las palabras en el PLN.

Las primeras representaciones vectoriales utilizaban formatos como *one-hot* (un vector con una entrada con valor 1 y el resto 0) y conteos de ocurrencias de palabras en documentos. Luego llegan los *word embeddings*, cuyo nombre deriva del concepto de embeber el significado de un espacio en otro. Estas estructuras suponen un gran cambio, dado que logran capturar las relaciones semánticas entre las palabras. Por ejemplo, si consideramos las palabras "hombre", "mujer", "rey" y "reina", utilizando el formato *one-hot*, trataríamos a estas palabras como partes de espacios totalmente separados unos de otros. Sin embargo, los *word embeddings* logran capturar, a través de su representación geométrica, la idea de que `rey - hombre + mujer = reina` como muestra la figura 2.1 (Mikolov, Yih, y Zweig, 2013).

Las redes neuronales representan una herramienta muy importante en el

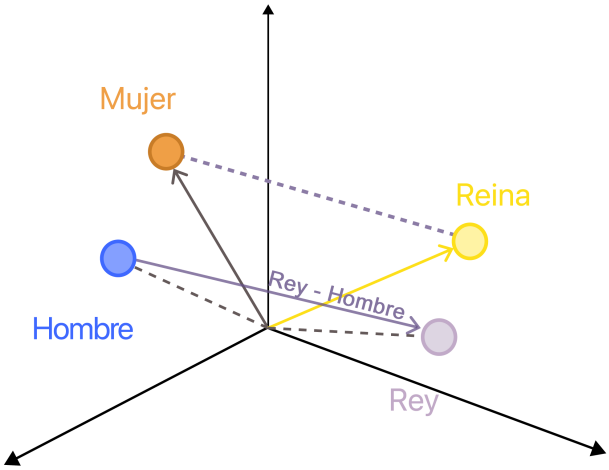

Figura 2.1: Aritmética de vectores presentando el ejemplo de `rey - hombre + mujer = reina`.

área de PLN, su nombre viene de que en sus inicios su arquitectura estaba inspirada en tratar de representar el cerebro humano. Cuando se comenzaron a utilizar para crear *word embeddings*, llegaron a representaciones aún mejores que las creadas por los humanos a través de la ingeniería de características. Las redes son capaces de procesar grandes volúmenes de datos de forma no supervisada, teniendo un gran éxito en tomar el contexto e interpretar las relaciones intrínsecas que existen entre las palabras.

En las primeras secciones introducimos los conceptos de redes neuronales y *embeddings*, necesarios para el desarrollo del proyecto. Introducimos los conceptos de distintas metodologías para crear *embeddings* como los métodos basados en conteo de contexto y en predicción de contexto. Presentamos también un conjunto de corpus y *embeddings* existentes para el idioma español.

En secciones siguientes, presentamos la definición formal del término sesgo y ponemos en contexto la importancia junto al impacto que tienen estos en las diferentes aplicaciones de aprendizaje automático y específicamente en el lenguaje natural, destacando cómo los corpus y *embeddings* se caracterizan por ser portadores de sesgos.

2.1. Redes neuronales

Las redes neuronales son redes conformadas por unidades pequeñas de cómputo llamadas neuronas. Cada neurona recibe un vector de valores que consiste en la suma ponderada de los valores de entrada y devuelve un único valor de salida (Jurafsky y Martin, 2009).

Una red está conformada por uno o más niveles de neuronas denominados

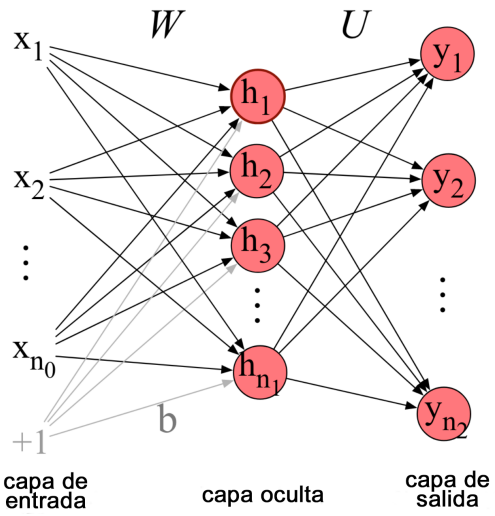

Figura 2.2: Una red *feed-forward* de 2 capas, con una capa oculta, una capa de salida y una capa de entrada (por lo general, no se cuenta la capa de entrada al enumerar las capas). Tomada de (Jurafsky y Martin, 2009).

capas. Cada capa le trasmite a otra información y pesos utilizando principios matemáticos. A la primera capa se le suele llamar capa de entrada, a la última capa de salida y a las intermedias capas ocultas. Las redes más sencillas, llamadas *feed-forward* son las que reciben como entrada de cada capa la salida de la anterior. Mostramos un ejemplo de estas en la figura 2.2.

En el contexto de este proyecto, si bien no utilizamos redes neuronales de forma directa, estas se encuentran por detrás de algunos métodos para la creación de los *word embeddings*. Estos se pueden obtener tomando los pesos de alguna de las capas de la red neuronal cuando se proporciona texto como entrada.

Una de las maneras más conocidas de creación de *word embeddings* es utilizar aprendizaje supervisado para entrenar un modelo sobre una tarea de pretexto, en la cual, quitamos una palabra conocida de una frase y le pedimos al modelo que complete esa frase. En el proceso de predicción de la palabra faltante, el modelo genera su representación vectorial. Este tipo de acercamiento es conocido como aprendizaje auto-supervisado ya que evita la necesidad del etiquetado manual de los datos de entrenamiento. (Jurafsky y Martin, 2009)

2.2. Métodos de creación de *word embeddings*

La creación de los *embeddings* se basa en la hipótesis distribucional de Harris (Harris, 1954), que plantea que el significado de una palabra está dado por las palabras que la rodean. En consecuencia, palabras con contexto similar tendrán representaciones similares.

La noción de similitud y relacionamiento entre palabras es muy amplia. Por ejemplo, podemos decir que "perro" y "gato" son similares porque comparten varias características, ya que ambos son animales domésticos. También existen palabras, como lo son "taza" y "café", que están relacionadas sin ser similares. Sus significados son muy diferentes (uno es un objeto, el otro el fruto de una planta) pero se suelen utilizar en contextos similares, "una taza de café".

A continuación presentamos distintas metodologías de construcción de *embeddings*, basadas en conteo y predicción del contexto.

2.2.1. Métodos basados en conteo del contexto

Existen representaciones basadas en conteo del contexto, que se basan en métricas como tf-idf (*term frequency - inverse document frequency*) o información mutua puntual positiva (PPMI) (Jurafsky y Martin, 2009).

Tf-idf resulta de utilidad cuando las dimensiones son documentos. Se utiliza en las matrices término-documento (matrices donde las filas representan palabras y las columnas documentos, cada celda contiene el número de veces que una palabra aparece en un documento dado) y para métodos como LSA (*Latent Semantic Analysis*). Este último es un método presentado en (Deerwester, Dumais, Furnas, Landauer, y Harshman, 1990) para la recuperación de información. Para representar los documentos utilizan una matriz término-documento que luego es analizada utilizando descomposición en valores singulares (SVD) para derivar el modelo de estructura latente que es después utilizado para la indexación y recuperación.

SVD (*Singular Value Decomposition*) El método de descomposición en valores singulares se emplea para descomponer una matriz en tres matrices más simples: una matriz de "valores singulares", y dos matrices de "vectores singulares". Esta descomposición se utiliza ampliamente en aplicaciones como reducción de dimensionalidad, análisis de datos y factorización de matrices.

Por otro lado, PPMI resulta de utilidad cuando las dimensiones son palabras, por ejemplo, en las matrices de frecuencia término-término, donde las filas y las columnas representan palabras, cada celda contiene el número de veces que una palabra en un corpus dado aparece cerca de otra palabra dada.

Tf-idf (*Term Frequency-Inverse Document Frequency*)

La métrica tf-idf es utilizada para evaluar el peso que tiene un término en un documento dentro de una colección de documentos.

El peso se calcula de la siguiente manera:

$$w_{t,d} = tf_{t,d} \times idf_t$$

El primer factor es la frecuencia de la palabra t en el documento d, es decir:

$$tf_{t,d} = count(t, d)$$

El segundo factor es utilizado para darle mayor importancia a palabras que ocurren en menor cantidad de documentos. Los términos que aparecen en menos ocasiones son útiles para discriminar los documentos en los que aparecen con otros documentos, mientras que términos que aparecen en muchos documentos aportan poca información. La ecuación es:

$$idf_t = \frac{N}{df_t}$$

siendo df_t la cantidad de documentos en los que aparece el término t.

Generalmente, ambas ecuaciones se utilizan agregándoles logaritmo en base 10 dada la gran cantidad de documentos y términos que se suelen utilizar, quedando

$$tf_{t,d} = \log_{10}(count(t,d) + 1)$$

$$idf_t = \log_{10} \frac{N}{df_t}$$

En la tabla 2.1 podemos ver un ejemplo de *embeddings* generados a partir del método tf-idf, donde cada fila de la matriz representa el *embedding* de la palabra correspondiente.

	Como gustéis (*As You Like It*)	Noche de Reyes (*Twelfth Night*)	Julio César (*Julius Caesar*)	Enrique V (*Henry V*)
batalla (*battle*)	0.074	0	0.22	0.28
bueno (*good*)	0	0	0	0
tonto (*fool*)	0.019	0.021	0.0036	0.0083
ingenio (*wit*)	0.049	0.044	0.018	0.022

Tabla 2.1: Las filas de la tabla corresponden a palabras, las columnas a novelas de Shakespeare y cada celda al cálculo del método tf-idf para cada palabra en la novela. Tomado de (Jurafsky y Martin, 2009).

PPMI (*Positive Pointwise Mutual Information*)

La información mutua puntual (PMI) es una métrica que intenta discriminar cuando la ocurrencia conjunta de dos variables se da únicamente por el azar. Se calcula con la siguiente fórmula, donde x e y representan dos variables aleatorias:

$$PMI(x;y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

El PMI vale 0 cuando las variables x e y son independientes, y toma su máximo valor cuando las variables están perfectamente asociadas. La PPMI, es una variación que lleva los valores negativos de la PMI a cero.

En la tabla 2.2 podemos ver un ejemplo de *embeddings* generados a partir del método PPMI, donde cada fila de la matriz representa el *embedding* de la palabra correspondiente.

	computadora (computer)	datos (data)	resultado (result)	pastel (pie)	azúcar (sugar)
cereza(cherry)	0	0	0	4.38	3.30
frutilla (stawberry)	0	0	0	4.10	5.51
digital (digital)	0.18	0.01	0	0	0
información (information)	0.02	0.09	0.28	0	0

Tabla 2.2: Las filas corresponden a palabras, las columnas a palabras de contexto y cada celda al cálculo del método PPMI para cada palabra con las palabras de contexto. Tomado de (Jurafsky y Martin, 2009).

2.2.2. Métodos basados en predicción del contexto

Existen otros métodos basados en predicción del contexto, en los cuales se obtienen los *embeddings* al optimizar una función objetivo para predecir palabras a partir de contextos.

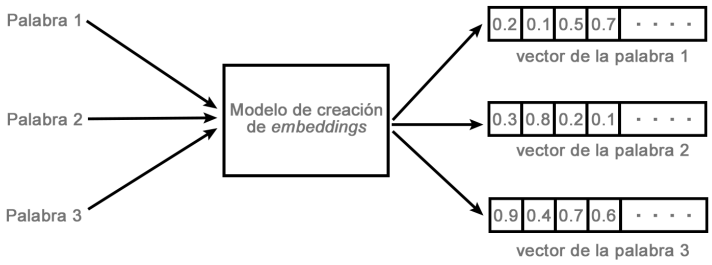

Figura 2.3: Una manera simple de visualizar la generación de los embeddings es pensar en que tomamos una palabra, se la damos a un modelo de creación de embeddings y este nos devuelve un listado de números reales que pasa a ser la representación vectorial de nuestra palabra.

A continuación presentamos dos métodos utilizados en la construcción y refinamiento de *embeddings* en este proyecto.

Word2vec

Word2vec (Mikolov, Sutskever, Chen, Corrado, y Dean, 2013) es un método que se utiliza para la creación de *embeddings*. La intuición detrás de este es que

en lugar de contar qué tan seguido aparece una palabra en el contexto de la otra, se utiliza una clasificación binaria que responde la pregunta "¿Es posible que la palabra A se encuentre cerca de la palabra B?". Para esto se utiliza una ventana que define cuáles son las palabras consideradas "cercanas" al objetivo. Si la palabra A está dentro de la ventana entonces la pregunta se responde de forma afirmativa.

Este tipo de modelos se consideran auto-supervisado porque no necesitan ningún tipo de etiquetado manual.

Word2vec tiene dos variantes, una se llama *Continuous Bag-of-words* (CBOW) y la otra *Skipgram with negative sampling* (SGNS).

Continuous Bag-of-words (CBOW) Esta variante, presentada en (Mikolov, Chen, Corrado, y Dean, 2013) se centra en predecir la palabra objetivo en base a su contexto, es decir, las palabras que la rodean en una oración considerando una ventana determinada. Una particularidad es que no considera el orden de las palabras en el contexto, por este motivo a esta variante se la llama *bag-of-words*.

Skipgram with negative sampling (SGNS) La segunda variante es similar a CBOW, pero en vez de predecir la palabra actual basándose en el contexto que la rodea, SGNS predice, para una palabra dada, las palabras que se encuentran antes y después en una ventana. El muestreo negativo hace referencia a que se utilizan palabras aleatorias del corpus, que no aparecen en el contexto, como ejemplos negativos y luego se entrenan clasificadores para las dos clases (positivo y negativo) según si la palabra está o no en el contexto. Este enfoque permite capturar relaciones semánticas y contextuales entre las palabras en una oración (Mikolov, Chen, y cols., 2013).

En la figura 2.4 mostramos las arquitecturas de Word2Vec.

FastText

FastText (Bojanowski, Grave, Joulin, y Mikolov, 2017) es una extensión de SGNS en la que el vector de cada palabra es representado como la sumatoria de los vectores de sus n-gramas.

Como consecuencia se puede obtener la representación de palabras con vectores ausentes, dado que se conocen los vectores de sus subsecuencias de palabras. A su vez, se obtienen representaciones similares para palabras que comparten subsecuencias (e.g., la palabra preadolescente es similar a joven, porque contiene la palabra adolescente). De acuerdo a los autores esto favorece principalmente la representación de vectores de palabras en idiomas morfológicamente ricos, como es el caso del idioma español.

En la figura 2.5 mostramos la arquitectura de Fasttext.

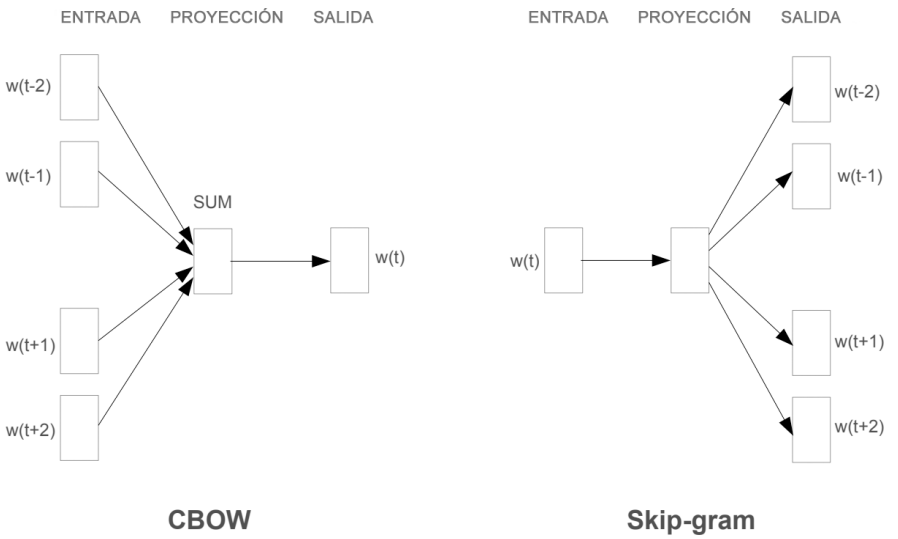

Figura 2.4: Arquitecturas de Word2Vec. La arquitectura CBOW predice la palabra actual basada en el contexto, y la de Skip-gram predice las palabras circundantes dada la palabra actual. Imagen tomada de (Mikolov, Chen, y cols., 2013)

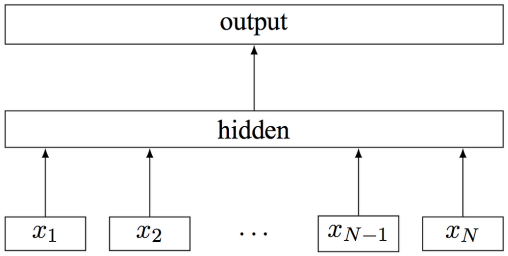

Figura 2.5: Arquitectura del modelo Fasttext para una oración con N características de n-gramas x_1, ..., x_N. Las características se incrustan y promedian para formar la capa oculta. Imagen tomada de (Joulin y cols., 2016)

Embeddings de contexto

Tanto los métodos de conteo como los de predicción del contexto mencionados generan *embeddings* estáticos, esto es, cada palabra del vocabulario se mapea a un único vector, independiente del contexto en el que aparezca esa palabra. Por ejemplo, el vector de "banco" como institución financiera y el de "banco" de sentarse serán el mismo.

Existen otros métodos que en lugar de generar *embeddings* estáticos, generan *embeddings* de contexto, es decir, que el vector que representa a la palabra depende del contexto en el que esta se encuentra. Tomando el ejemplo anterior, en un modelo de *embeddings* de contexto, los vectores de "banco" como institución y "banco" como mueble serían diferentes. Existen modelos de lenguaje como son BERT (Devlin, Chang, Lee, y Toutanova, 2018), GPT (Radford, Narasimhan, Salimans, Sutskever, y cols., 2018) y ELMo (Peters y cols., 2018) de los cuales se pueden extraer *embeddings* de contexto.

Como en este trabajo nos enfocamos en el análisis de palabras sueltas, que no se encuentran en contexto, las representaciones que usamos son estáticas. Además, usamos únicamente representaciones densas, esto significa que los vectores tienen muy pocas posiciones (o ninguna) con valor 0 y que existe una única representación para cada palabra del vocabulario, respectivamente. Estas representaciones densas, a diferencia de los vectores de ocurrencias o los vectores de características, capturan mejor características como la sinonimia entre las palabras (Jurafsky y Martin, 2009). Esto se da porque, en representaciones dispersas, la diferencia entre dos vectores es muy marcada ya que las dimensiones que representan a cada palabra no se relacionan entre sí.

Además de la construcción de *embeddings*, se puede realizar un ajuste fino a los modelos, llamado *fine-tuning*. El *fine-tuning* es un procedimiento en el cual se parte de un modelo pre-entrenado y, realizando pocos cambios, se lo ajusta a un determinado dominio. Esto es útil dado que, a veces, los modelos que tenemos pueden no ser óptimos para una tarea en específico. Por lo tanto, mediante este procedimiento, se ajustan los parámetros de nuestro modelo inicial para mejorar el desempeño para la tarea en particular que queremos realizar.

2.3. Corpus de texto y *embeddings* en español

En esta sección introducimos los corpus y conjuntos de *embeddings* utilizados en este proyecto. Además, presentamos otros recursos existentes hasta la fecha de comenzado el proyecto para el idioma español.

2.3.1. Corpus de texto

En esta sección presentamos algunos de los corpus disponibles más conocidos del idioma español. Estos corpus fueron creados a partir de textos o recursos de diferentes fuentes y recolectados en diferentes regiones del habla hispana. Los utilizamos para la construcción de *embeddings* y refinamiento de modelos.

Spanish Billion Word Corpus El *Spanish Billion Word Corpus* (SBWC) (Cardellino, 2019) consiste en un corpus de aproximadamente 1.5 mil millones de palabras sin anotar. Este conjunto de datos está público y fue recopilado por los autores de diversos sitios, intentando que fueran de lugares gratuitos con el fin de evitar problemas de derechos de autor.

Spanish Web Corpus El *Spanish Web Corpus*[1] es un corpus perteneciente a una familia de corpus creados a partir de textos de internet. Está formado por varios sub-corpus de las variedades del idioma español, cada uno recolectado de diferentes dominios web de su región. En particular, tiene un sub-corpus para el español uruguayo y uno para el español argentino. En este proyecto utilizamos estos dos sub-corpus para crear diversos corpus rioplatenses (ver sección 3.1.1).

Spanish Unannotated Corpora El *Spanish Unannotated Corpora* (SUC) (Cañete, 2019) está formado por la recopilación de las partes en idioma español de diferentes corpus disponibles en internet. Entre las diferentes fuentes que lo conforman se encuentran traducciones de documentos, subtítulos y la Wikipedia en español. Contiene 3 mil millones de palabras sin anotar.

Este es uno de los corpus más grandes (sino el más grande) que existe al momento de inicio de la escritura de este proyecto para el idioma español.

Small World of Words El *Small World of Words* (SWOW) (Cabana y cols., 2023) consiste en la recopilación de datos (respuestas hacia estímulos, datos demográficos, tiempos de respuestas, etc) con el principio psicológico de asociación libre de palabras. Proponen recopilar el significado de las palabras para las personas en el mundo actual. Intentan mapear el significado de las palabras con la mente humana. Recopilan datos a través de su página web[2], en la que se le presenta a los usuarios una tarea de asociación continuada de palabras. Este tipo de tarea consiste en proveer múltiples respuestas a una única palabra de estímulo (en este caso 3 respuestas para 18 palabras distintas), por ejemplo, si la palabra estímulo es "perro", a la persona que está llenando el formulario se le puede ocurrir "canino", "gato" y "lindo", entre otras cosas.

Tienen disponibles corpus de asociación libre para diversos idiomas, incluyendo al español y más específicamente al español rioplatense (idioma hablado en Uruguay y Argentina). Para este último, el corpus consta de 85714 líneas, donde cada una es un estímulo con tres asociaciones como respuestas.

2.3.2. Conjuntos de *embeddings*

En esta sección introduciremos algunos de los conjuntos de *embeddings* más conocidos del idioma español. La mayoría fueron entrenados a partir de los corpus presentados en la sección anterior.

[1] Spanish Web Corpus https://www.sketchengine.eu/estenten-spanish-corpus
[2] Proyecto Small World of Words https://smallworldofwords.org/uy

Fasttext Conjunto de vectores provisto por Fasttext (Grave, Bojanowski, Gupta, Joulin, y Mikolov, 2018) para el idioma español. Fasttext tiene vectores pre-entrenados para 157 idiomas, incluido el español. Los datos para la creación de esos vectores son de búsquedas comunes en internet y de Wikipedia.

Estos vectores fueron creados utilizando `fasttext` con CBOW. Son de dimensión 300 pero proveen un método para reducirles la dimensión en caso de ser necesario. Fasttext empezó a proveer datos para analogías para otros idiomas además del inglés, pero a la fecha de comenzado el proyecto, aún no hay para el español. Tienen un tamaño de vocabulario de 2 millones de palabras.

Spanish Unannotated Corpora Embeddings Consiste en un conjunto de vectores[3] que fueron entrenados utilizando el corpus SUC (ver 2.3.1) y la biblioteca `fasttext` con el modelo Skipgram. El modelo cuenta con 1.3 millones de vectores y se encuentran disponibles versiones con 10, 30, 100 y 300 dimensiones. Llamaremos a estos *embeddings* **SUC**.

Spanish Billion Word Embeddings Consiste en *embeddings* creados a partir del SBWC (Cardellino, 2019) utilizando el algoritmo de `word2vec`. Para la creación de los *embeddings* el autor realizó una limpieza del corpus quitando los caracteres no alfanuméricos, cambiando los números por la palabra "DIGITO" y los espacios múltiples por simples. Luego realizó *skipgram with negative sampling* y obtuvo vectores de dimensión 300. Llamaremos a estos *embeddings* **SBWE**. Tienen un tamaño de vocabulario de aproximadamente 1 millón de palabras.

Small World of Words El proyecto SWOW (Cabana y cols., 2023) tiene una colección de *embeddings* publicado para cada uno de sus corpus. Para los del español rioplatense se encuentra disponible un modelo con un vocabulario de 13168 palabras, donde la dimensión de cada vector es 400.

Este conjunto de *embeddings* fue creado utilizando el método presentado por (Steyvers, Shiffrin, y Nelson, 2005). Primero crearon el espacio de asociación de palabras (WAS) y luego utilizaron el enfoque de descomposición en valores singulares (SVD) para obtener un conjunto de vectores de las palabras estímulo. Cada vector es representado como los primeros 400 valores.

Al momento de la fecha de comienzo del proyecto, no encontramos *embeddings* realizados para el español rioplatense (ni creados a partir de textos tradicionales ni con asociaciones libres de palabras) además del SWOW, por lo que muchas de las pruebas que realizamos fueron con este modelo.

A continuación presentamos una tabla mostrando los *embeddings* existentes utilizados junto a su dimensión, tamaño de vocabulario (medido en cantidad de palabras), origen de los datos de construcción y método de construcción:

[3]Spanish Unannotated Corpora Embeddings disponible en https://github.com/BotCenter/spanishWordEmbeddings

Nombre	**Dimensión**	**Tamaño (palabras)**	**Origen de los datos de construcción**	**Método**
Fasttext	300	2 mill.	- Búsquedas comunes en internet - Wikipedia en español	CBOW
Spanish Unannotated Corpora Embeddings	300	1,3 mill.	- Traducciones de documentos - Subtítulos - Wikipedia en español	Skipgram
Spanish Billion Word Embeddings	300	1 mill.	- Wikipedia en español - Parlamento europeo en español - Otros proyectos	SGNS
Small World of Words	400	13.168	Asociación libre de palabras recopiladas a través de su página web	WAS y SVD

2.4. Sesgo en *word embeddings*

El término sesgo se refiere a un prejuicio, a favor o en contra, de una cosa, persona o grupo, comparado con otro[4]. En el área de IA, usualmente se hace referencia al sesgo algorítmico, este sesgo describe errores sistemáticos y repetibles en un sistema informático que generan resultados injustos, como privilegiar a un grupo arbitrario de usuarios sobre otros o cuando un algoritmo produce resultados que están sistemáticamente sesgados debido a suposiciones erróneas en el proceso de aprendizaje automático[5]. Si queremos poder medir y analizar el sesgo, es importante que sepamos decir cuándo un algoritmo está generando o no resultados "justos", para esto en el área se utiliza el concepto de *fairness*.

Existen diversas definiciones del concepto *fairness*, utilizaremos la definición del libro (Barocas y cols., 2019), donde se distinguen tres perspectivas fundamentales: el *fairness* individual, la igualdad justa de oportunidades y una visión intermedia. Cada una de estas perspectivas plantea un enfoque de cómo garantizar que todos tengamos igualdad de oportunidades en la sociedad. El *fairness* individual se centra en el trato igualitario de individuos que comparten similitudes relevantes para una tarea o un objetivo específico. En este contexto, la similitud se define en función de las características consideradas pertinentes para la tarea en cuestión. La equidad individual es una noción comparativa que se pregunta si existen diferencias en el trato de personas similares. No se ocupa directamente de cómo se trata a los miembros de diferentes grupos, sino que compara a todas las personas como individuos. Por otro lado, tenemos la "igualdad justa de oportunidades" que se basa en la idea de que la única razón válida para que las personas experimenten resultados diferentes en sus vidas es que posean diferentes habilidades o ambiciones. Cualquier factor ajeno (por ejemplo, discriminación institucionalizada, brechas en la educación, dificultades económicas heredadas, estereotipos y prejuicios sociales, y acceso limitado a re-

[4]Traducido de Oxford English Dictionary https://languages.oup.com

[5]Traducción propia a partir de lo publicado en Florida State University Libraries - Algorithmic Bias https://guides.lib.fsu.edu/algorithm

cursos, entre otros) que obstaculice a que las personas igualmente merecedoras alcancen el éxito se consideraría injusto para esta perspectiva. Por último, tenemos un punto medio entre las dos perspectivas anteriores que se preocupa por la equidad en la toma de decisiones, pero también es sensible a las dinámicas que pueden perpetuar la desventaja en la sociedad. Sostiene que los tomadores de decisiones (como el gobierno, empresas, instituciones educativas, etc) tienen la obligación de evitar perpetuar la injusticia. En particular, deberían considerar las oportunidades pasadas de los individuos al evaluarlos, reconociendo que el rendimiento puede verse afectado por la falta de oportunidades. Esta perspectiva busca equilibrar el *fairness* individual con una comprensión más amplia de la justicia social.

Existen diversos estudios que prueban cómo los modelos de aprendizaje automático tienden a reproducir (Bolukbasi y cols., 2016a) (Bolukbasi y cols., 2016b) (Caliskan, Bryson, y Narayanan, 2017) e incluso incrementar (Barocas y cols., 2019) (Zhao y cols., 2017) los sesgos presentes en la sociedad. Estos sesgos son introducidos a los modelos en diferentes partes de su proceso de creación, comenzando desde la recolección de los datos que serán utilizados para su entrenamiento (Brunet, Alkalay-Houlihan, Anderson, y Zemel, 2019).

Consideramos que es esperable que los modelos de aprendizaje automático presenten sesgos, dado que estos, al igual que los estereotipos, forman parte de nuestra vida cotidiana. Si los modelos no presentaran ningún tipo de sesgo sus predicciones probablemente no serían correctas ya que estos no estarían representando la realidad. Por ejemplo, si quisiéramos hacer un modelo para que sea implementado únicamente en Uruguay, los sesgos regionales del país harían que el modelo se adecúe a esa realidad, más que si el modelo presentara sesgos de España.

La mayoría de los problemas de los sesgos reportados en los trabajos mencionados anteriormente surge cuando, a partir de la realidad sesgada por naturaleza, se crea un modelo que genera decisiones discriminatorias o injustas en las aplicaciones del mundo real. Podemos llegar a esta situación por diversos motivos, entre ellos: si el modelo aprende prejuicios y los utiliza para la toma de decisiones, si refuerza estereotipos y los utiliza para representar a ciertos grupos o si genera categorías sub-representadas y otras sobre-representadas acorde al dominio en el que se quiere aplicar. Al utilizar este modelo en un sistema, algunas clases terminan obteniendo algún tipo de ventaja por sobre otras, sobre todo si el sistema se utiliza sin tener cuidado. Si se entrena un modelo con datos bancarios históricos y luego se quiere utilizar ese modelo para aprobar préstamos, los datos históricos seguramente presenten tendencia a identificar a los hombres como mejores candidatos, porque históricamente han tenido mayor poder adquisitivo. Si se quiere utilizar un modelo en un sistema médico, si en lugar de construir otro modelo con datos médicos se utiliza el modelo creado a partir de datos bancarios, ya sea por cuestiones de tiempo, costos, falta de datos u otros recursos, ¿qué tipo de decisiones generaría en ese caso y a qué pacientes priorizaría?

En particular, los recursos lingüísticos, como corpus y *embeddings*, son claros portadores de sesgos ya que el idioma es un reflejo de la sociedad y cultura en la

que se desarrolla. Esto acarrea no solamente las relaciones que queremos representar en nuestros modelos (e.g.: los jubilados son usualmente personas mayores, el limón es amargo o los elefantes son grandes mamíferos), sino que también relaciones no deseables, que podrían generar resultados injustos y peligrosos (e.g.: racismo, machismo o discriminación contra la comunidad LGBTIQ+[6]) que no queremos replicar. Existen idiomas con características gramaticales, vocabulario y expresiones que favorecen a ciertos grupos por sobre otros. Como es el caso del idioma español, que al ser un idioma con género gramatical, tiende a favorecer al género masculino por sobre el femenino en ciertos contextos, además de que generalmente se utiliza el masculino como neutro o genérico.

Los recursos lingüísticos son también portadores de sesgos históricos (por relaciones históricas de poder perpetuadas en la lengua) y sub-representación de algunos grupos. Esto último se debe a que las lenguas predominantes tienen mayor cantidad de recursos y el uso de estos lleva a que las lenguas minoritarias no siempre se encuentren representadas en los modelos.

Dada la diversidad que presenta el lenguaje natural, resulta clara la importancia de tener recursos que abarquen diferentes idiomas, así como sus variantes. Además, es importante el análisis de los sesgos que puedan existir en estos conjuntos, ya que los sesgos particulares no necesariamente se evidencian cuando se trabaja sobre reportes de análisis que fueron realizados en un corpus más general del idioma.

En el caso del idioma español, existe una gran cantidad de variantes que no se encuentran representadas actualmente. Además, cabe destacar que varios de los corpus que existen para el español son traducciones de recursos en inglés.

2.4.1. Casos de impacto social del sesgo en los algoritmos

A lo largo de los últimos años, se han identificado sesgos en el aprendizaje automático en diversas aplicaciones. Existen varios estudios que ilustran la presencia de sesgos en diferentes contextos y su impacto en la sociedad.

Enseñanza En el ámbito de la enseñanza existen diferentes aplicaciones posibles para algoritmos de inteligencia artificial, para corrección automática de tareas, generación de resúmenes, calificación de profesores, entre otros. Si bien supone un gran avance y apoyo en un área fundamental de la sociedad, la generación de resultados injustos por parte de los algoritmos puede impactar directamente en la vida de estudiantes y trabajadores de centros educativos.

El sistema *Value-Added Model* (VAM), implementado en Estados Unidos en 2010, fue empleado para evaluar la efectividad de los docentes calculando su "valor agregado" en algunas escuelas y liceos. Este valor se calculaba comparando las notas de los alumnos en los exámenes del año actual, contra las notas de los mismos alumnos en exámenes de años anteriores. Luego, teniendo en cuenta si

[6]Acrónimo para: lesbianas, gays, transgéneros, bisexuales, intersex, queer. El "+" representa identidades de género y sexualidades minoritarias que no están explícitamente incluidas en el término.

los alumnos mejoraron o si empeoraron sus promedios, tomaban decisiones de retención de los profesores o la asignación de bonificaciones por buenas metodologías de enseñanza. En (O'Neil, 2017) se analizan las repercusiones provocadas al utilizar las clasificaciones generadas por este modelo para tomar decisiones tan importantes. Hubo muchos casos en donde despidieron a profesores porque el modelo dio una calificación mala a pesar de haber tenido una calificación muy buena por sus superiores.

Redes sociales Dentro de las redes sociales se utilizan diversos algoritmos de inteligencia artificial, para recomendaciones, anuncios, etiquetado automático de imágenes y videos, entre otros. Inicialmente estas aplicaciones pretendían tener un uso social o recreativo, pero se han convertido en una parte íntegra de la vida de las personas, por lo que el impacto del sesgo dentro de estas puede generar grandes repercusiones.

Existen casos conocidos de sesgo en los algoritmos de etiquetado de imágenes de Google[7] y Facebook[8]. En 2015 se viralizó un caso en el que Google Photos[9] etiquetó como gorilas a personas afrodescendientes (Zhang, 2015). En 2020 se viralizó otro incidente, por parte de Facebook, en el que el etiquetado en los videos reconocía a las personas afrodescendientes como primates (Mac, 2021).

Una auditoría al algoritmo de recomendación de Youtube[10] realizada en 2023 encontró discriminación hacia inmigrantes en la manera en que son retratados en los videos recomendados (Eticas, 2023).

Sistema judicial Dentro del sistema penal se han utilizado algoritmos de inteligencia artificial en diferentes aplicaciones, desde la identificación de criminales por imágenes, hasta el dictamen de sentencias. Dada la sensibilidad propia del ámbito de aplicación, si los algoritmos utilizados generan resultados discriminatorios o injustos, esto puede suponer graves consecuencias.

En una auditoría realizada por la agencia de noticias sin fines de lucro ProPublica (Angwin, Larson, Mattu, y Kirchner, 2019) se encontró sesgo racial en el sistema COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). Este sistema tenía un algoritmo que calculaba la sentencia que se le debía de adjudicar a un criminal, en base a su probabilidad de reincidencia en el sistema judicial, llegando a dictaminar sentencias mayores por delitos menores a personas afrodescendientes, en comparación con las dictaminadas a personas blancas por delitos más graves.

Sistemas médicos Existen varias aplicaciones dentro de la medicina que se pueden beneficiar del uso de algoritmos de inteligencia artificial, como diagnósticos, segmentación de imágenes y predicción de enfermedades, entre otros. Si bien

[7] Google https://www.google.com
[8] Facebook https://www.facebook.com
[9] Google Photos https://photos.google.com
[10] Youtube https://www.youtube.com

su uso podría acelerar procesos se deben utilizar con mucho cuidado por su gran impacto.

Varios estudios (Williams, Lawrence, y Davis, 2019) (Lee, Liang, y Shi, 2021) han demostrado que el sistema de triaje médico de salas de emergencia utilizado en Estados Unidos, presenta racismo contra personas afrodescendientes. Se ha probado que para un mismo caso de atención médica, el tiempo de espera hasta ser atendida es menor para una persona blanca.

Selección de personal Hoy en día, un gran número de empresas utilizan algoritmos de inteligencia artificial en las diferentes etapas del proceso de selección de personal, para la definición del público al cual se le muestran los anuncios o el pre-filtrado de currículums de forma automática. Estos algoritmos influyen en gran medida en la decisión final para puestos de trabajo, pudiendo generar resultados injustos entre los candidatos.

En (Lambrecht y Tucker, 2019) se analiza el sesgo de género en anuncios de trabajos en el área de STEM (Ciencia, Tecnología, Ingeniería y Matemática). En este estudio se presentan diferentes factores que llevan a la existencia de sesgo en los algoritmos de presentación de anuncios de trabajo. En particular, se evidencia sesgo totalmente ajeno a la creación de la aplicación. En este caso, mostrarle anuncios publicitarios a un hombre es más barato que a una mujer, independientemente de la probabilidad que exista de que la persona haga clic en el anuncio, dado que las mujeres ocupan entre un 70 % y 80 % del mercado, por lo cual los precios de los anuncios publicitarios apuntados hacia las mujeres son más competitivos. Un algoritmo que quiere mostrar la mayor cantidad de anuncios reduciendo los costos, termina mostrándole muchos más anuncios a hombres que a mujeres.

En 2021 (Imana, Korolova, y Heidemann, 2021) se realiza una auditoría de caja negra, sobre las plataformas Facebook y LinkedIn[11], para buscar sesgo en los algoritmos que presentan anuncios publicitarios de ofertas laborales. La auditoría confirma la presencia de sesgo de género en el algoritmo de Facebook. La plataforma le mostraba más anuncios de trabajos estereotipados a lo femenino y una menor cantidad de anuncios técnicos a la audiencia femenina, independientemente de si se trataba de especificar una audiencia balanceada o de las habilidades técnicas de los candidatos.

Se han encontrado casos de sesgo en los algoritmos de selección y evaluación de candidatos en procesos de contratación. Estos algoritmos pueden basarse en características específicas y datos históricos, lo que puede resultar en sesgos de género, raza u otros factores que influyen en las decisiones de contratación. Un estudio realizado a Amazon[12] (Dastin, 2022) mostró que el algoritmo que utilizaban presentaba sesgos de género al favorecer a candidatos masculinos para cargos técnicos.

[11] LinkedIn https://www.linkedin.com
[12] Amazon https://www.amazon.com

Selección de precios y sistemas bancarios Observamos algoritmos utilizados en el área económica para la selección de precios y para la aprobación de créditos o préstamos bancarios. Estos algoritmos pueden influir en el mercado, guiando a consumidores por las tendencias y generar sesgo de género por oportunidades bancarias diferentes.

Un caso visto en (O'Neil, 2017) son los algoritmos de fijación de precios dinámicos. Estos algoritmos son utilizados por compañías para establecer precios de forma dinámica según el comportamiento de los consumidores, lo que puede resultar en precios discriminatorios o injustos.

Por otro lado, los algoritmos utilizados para evaluar el riesgo crediticio pueden ser opacos y sesgados, lo que puede tener un impacto negativo en las personas que intentan acceder a créditos o préstamos. En (Telford, 2019) se evidencia que la tarjeta de crédito de Apple[13] presentaba sesgos dándoles mayor línea de crédito a hombres que a mujeres aunque la mujer tenga mejor calificación crediticia.

Como hemos mencionado, hay numerosos ejemplos que demuestran la presencia de sesgos en los modelos de inteligencia artificial. Es crucial analizar y comprender estos sesgos, ya que su existencia puede tener efectos perjudiciales en diversos aspectos de la sociedad. Sin embargo, también es importante reconocer que el objetivo no es eliminar el sesgo de los modelos, ya que este es inherente a los datos y necesario para que un modelo pueda aprender a resolver alguna tarea en específico. Lo fundamental es desarrollar metodologías y herramientas para detectar, analizar y mitigar (cuando es necesario) los sesgos en los modelos de inteligencia artificial. De esta manera, podremos tomar decisiones informadas sobre dónde y cómo aplicar estos modelos, pudiendo considerar las implicaciones éticas y sociales de su aplicación.

2.4.2. Análisis de sesgo en *embeddings* y modelos de lenguaje

Trabajos anteriores, como por ejemplo (Barocas y cols., 2019), demostraron que si bien los modelos de aprendizaje automático son replicadores de sesgos y que su uso sin las correctas precauciones puede impactar directamente a las personas de forma negativa, también pueden utilizarse como método de análisis y mitigación del sesgo.

Existen trabajos previos (Friedman y Nissenbaum, 1996) (Bolukbasi y cols., 2016b) (Zhao y cols., 2017) que analizan el sesgo en *embeddings* y modelos del lenguaje. La mayoría de estos estudios y, en general, los más exhaustivos, fueron desarrollados para el idioma inglés. Aún así, existen varios trabajos previos que analizan los modelos de vectores de palabras más conocidos del español como *Spanish Billion Word Embeddings* y FastText, y los extraídos de los modelos de lenguaje como BETO, Berta y Robertuito (Zhou y cols., 2019) (Garrido-Muñoz, Martínez-Santiago, y Montejo-Ráez, 2022). Pero estos últimos se enfocan prin-

[13] Apple https://www.apple.com

cipalmente en la detección del sesgo de género y utilizan un idioma español más "neutro".

A la fecha de inicio de este proyecto no hemos encontrado un trabajo que haga un análisis de sesgos sobre vectores de palabras en español que haya sido recolectado en una región específica y se enfoque en el análisis de sesgos presentes en esa región. En este trabajo realizamos un análisis de sesgos regionales y culturales sobre modelos de vectores de palabras entrenados a partir de un corpus rioplatense y el SWOW (Cabana y cols., 2023) (creado a partir de asociaciones libres de palabras), con la intuición de que estos sean característicos de la región del Río de la Plata, así como también la comparación con otros modelos del español de los sesgos más tradicionales de género, racial, etc. Para realizar el análisis de sesgo utilizamos el método y métricas presentadas en (Zhou y cols., 2019) (ver sección 5.1.1) para modelos monolingües de lenguajes con género y una variación del método presentado en (Grand, Blank, Pereira, y Fedorenko, 2022) (ver sección 5.1.3).

2.4.3. *Data statements*

Los *data statements* son declaraciones que tienen como objetivo proporcionar transparencia sobre los conjuntos de datos que describen. Sirven como herramienta para fomentar la responsabilidad en el desarrollo de sistemas y para facilitar la investigación científica, abordando en particular los sesgos. Estas declaraciones fueron introducidas por (Bender y Friedman, 2018) y se desarrollaron inicialmente para ser utilizadas sobre conjuntos de datos de lenguaje para sistemas de procesamiento de lenguaje natural, aunque hoy en día pueden ser utilizadas para todo tipo de conjunto de datos.

Estas declaraciones proporcionan información sobre las características de los conjuntos, entre los datos proporcionados en general se incluye el proceso de recopilación de los datos, la composición de los datos y su curación, las técnicas de mitigación de sesgo utilizadas, las consideraciones éticas tomadas en el desarrollo del conjunto y las métricas de evaluación utilizadas. En el caso de las aplicaciones de lenguaje natural se incluyen además características particulares, e.g.: contexto del habla, características demográficas de los hablantes y de los anotadores, entre otros[14].

Esta información resulta de gran interés, ya que puede utilizarse para mitigar los daños causados por sesgos en los conjuntos de datos. Por ejemplo, cuando nos encontramos en un caso de algoritmos descontextualizados (cuando el contexto del conjunto de entrenamiento no se corresponde con el contexto de implementación del sistema).

Presentamos un *data statement* desarrollado con la plantilla que ofrece el Tech Policy Lab de la Universidad de Washington[15] para el modelo SWOW en el anexo A.8.

[14] Data Statements https://techpolicylab.uw.edu/data-statements

[15] Plantilla de Overleaf para data statement https://www.overleaf.com/read/kqftjwzvhmwx

Capítulo 3

Creación de *embeddings*

En este capítulo detallamos el proceso que realizamos para la creación de *embeddings*. Presentamos primero el preprocesamiento que realizamos a dos corpus, siguiendo esta sección presentamos la selección de hiper-parámetros, luego el entrenamiento de los *embeddings* y por último el proceso de *fine-tuning*.

3.1. Preprocesamiento

En esta sección presentamos el preprocesamiento realizado a dos corpus: textos y asociaciones libres de palabras de la región rioplatense. El resultado de esta sección son 6 corpus: corpus rioplatense base, corpus rioplatense lower, corpus rioplatense cased, corpus rioplatense SBWC, corpus rioplatense SUC y corpus SWOW, los cuales se describen a continuación.

3.1.1. Corpus rioplatense

El corpus rioplatense está formado por textos de diversas temáticas de Argentina y Uruguay. Para su creación partimos de los sub-corpus de español uruguayo y español argentino del *Spanish Web Corpus* (ver sección 2.3.1). El archivo original del sub-corpus argentino contiene poco más de 116 millones de líneas y del sub-corpus uruguayo tiene más de 261 millones de líneas. Por lo tanto, previo a cualquier preprocesamiento obtuvimos un corpus de aproximadamente 377 millones de líneas. Cabe destacar que en cada línea hay una palabra, no una oración. El inicio y fin de oración están indicados por los tokens "$< s >$" y "$< /s >$". Siguiendo esta estructura, recuperamos las oraciones de los textos, obteniendo un corpus final, al que llamaremos **corpus rioplatense base**, de más de 305 millones de palabras. Luego, eliminamos los espacios en blanco del inicio y fin de cada línea.

Una vez que obtuvimos el **corpus rioplatense base**, le aplicamos preprocesamientos específicos según la tarea que queríamos realizar con cada uno, obteniendo un total de 4 corpus procesados.

Para el primer corpus, al que llamaremos **corpus rioplatense lower**, quitamos un listado predefinido de *stop words*, ya que la información que aportan no es relevante, dado que tienden a aparecer reiteradas veces en muchos contextos diferentes, provocando que no aporten un significado en específico. También borramos los espacios múltiples, dejando solo uno en su lugar, decidimos mantener la puntuación y pasamos todo el texto a minúscula. Para el segundo corpus utilizamos el mismo preprocesamiento pero en lugar de pasar el corpus a minúscula dejamos las mayúsculas existentes, a este corpus lo llamaremos **corpus rioplatense cased**. Utilizamos estos corpus para crear *embeddings* (ver 3.2).

Para el tercer corpus no quitamos las *stop words*, mantuvimos las mayúsculas y quitamos todos los caracteres no alfanuméricos. A este corpus lo llamaremos **corpus rioplatense SBWC** y lo utilizamos para realizarle *fine-tuning* al conjunto de *embeddings* SBWE (ver 3.3.1).

Por último, creamos un corpus al que llamaremos **corpus rioplatense SUC**, que toma el corpus rioplatense base y le realiza el mismo preprocesamiento que el realizado para el SUC (ver 2.3.1) en la creación de los *embeddings* del mismo nombre[1]. Consistió en pasar el corpus a minúsculas, quitar las URLs, los listados (i.e., "1.", "2.", "a.") y reemplazar los espacios múltiples por uno solo. Utilizamos este corpus para realizarle *fine-tuning* a los *embeddings* SUC (ver 3.3.2).

En la siguiente tabla resumimos los corpus generados y el uso que le dimos a cada uno:

	Creación	**Uso**
Corpus rioplatense base	Spanish Web Corpus Argentina + Spanish Web Corpus Uruguay + eliminación de etiquetas HTML y de espacios	Creación de corpus más específicos
Corpus rioplatense lower	Corpus rioplatense base + eliminación de etiquetas HTML y de espacios + pasar todo el texto a minúsculas + eliminación *stop words*	Creación de nuevos *embeddings*
Corpus rioplatense cased	Corpus rioplatense base + eliminación de etiquetas HTML y de espacios + eliminación *stop words*	Creación de nuevos *embeddings*
Corpus rioplatense SBWC	Corpus rioplatense base + eliminación de etiquetas HTML y de espacios + eliminación de todo carácter no alfanumérico	Fine-tuning de Spanish Billion Word Embeddings
Corpus rioplatense SUC	Corpus rioplatense base + eliminación de etiquetas HTML y de espacios + pasar todo el texto a minúsculas + eliminación de URLs y listados	Fine-tuning de *embeddings* del Spanish Unannotated Corpora

[1]Embeddings de Fasttext a partir del corpus SUC https://github.com/dccuchile/spanish-word-embeddings#fasttext-embeddings-from-suc

3.1.2. Corpus de asociación libre de palabras

Para la creación de *embeddings* con el corpus provisto por SWOW, a diferencia del corpus rioplatense, no realizamos preprocesamiento de los datos, dado que el proyecto ya los provee preprocesados 2.3.1. Definimos una oración como la concatenación de la palabra estímulo con sus palabras respuestas separadas por un espacio, por ejemplo, si para la palabra estímulo "bar" algún participante respondió con las palabras "abierto", "cerveza" y "noche", una oración sería: "bar abierto cerveza noche". Esto genera que al momento de entrenar el modelo estas palabras se encuentren unas en el contexto de las otras.

3.2. Construcción de nuevos *embeddings*

En esta sección presentamos los *embeddings* que entrenamos, destacando el origen de los corpus que empleamos y los hiper-parámetros que configuramos para su creación.

En nuestra elección, optamos por utilizar el modelo `word2vec` proporcionado por la biblioteca `gensim`. A lo largo de nuestra investigación, generamos *embeddings* a partir de diferentes corpus y ajustamos diversos hiper-parámetros para obtener los mejores resultados posibles.

Sin embargo, es importante mencionar que al trabajar con el corpus SWOW, como se detalla en la sección 2.3.1, nos enfrentamos a resultados no tan favorables. Estos *embeddings* no lograron capturar adecuadamente el significado de palabras básicas en el idioma español. En vista de que el proyecto SWOW ya ofrece *embeddings* de buena calidad, tomamos la decisión de no continuar nuestra investigación ni realizar un análisis más profundo en este ámbito, en su lugar, utilizamos estos *embeddings* para el análisis posterior.

Por otro lado, al generar *embeddings* a partir del corpus rioplatense lower, tal como se describe en la sección 3.1.1, creamos múltiples conjuntos de *embeddings*. De entre estos, seleccionamos 3 para su análisis en la sección 4, a los cuales denominamos RP-1, RP-2 y RP-3. Para su creación, aplicamos los hiper-parámetros óptimos que obtuvimos mediante el proceso de *random search* detallado en la sección 3.4.1.

3.3. *Fine-tuning* de *embeddings*

Al realizar *fine-tuning* (ver 2.2.2) sobre los modelos de vectores tuvimos en cuenta el preprocesamiento utilizado para la creación de ese modelo y aplicamos, en la medida de lo posible, el mismo preprocesamiento al corpus que se utilizó para realizar el *fine-tuning*, en este caso el corpus a utilizar es el rioplatense (ver sección 3.1.1).

En esta sección presentamos los *fine-tuning* que realizamos a los *embeddings* SUC y SBWE.

3.3.1. SBWE

Para el *fine-tuning* del SBWE utilizamos un corpus al que llamamos corpus rioplatense SBWC (ver sección 3.1.1). Este corpus toma como base la unión de dos corpus conformados por textos rioplatenses y tiene el mismo preprocesamiento que el realizado al SBWC utilizado para los *embeddings* del SBWE (ver 2.3.2), a excepción del procesamiento de números, los cuales son reemplazados por el token "DIGITO" en el SBWC pero en nuestro corpus decidimos mantenerlos intactos. El texto de este corpus proviene, en gran parte, de entradas de blogs de internet, las cuales incluyen entre otras cosas su fecha de publicación. Por este motivo, decidimos conservar los números, ya que consideramos que la información numérica provista es relevante dada la naturaleza del corpus.

Utilizamos el modelo `word2vec` provisto por la biblioteca `gensim` para el *fine-tuning*, ya que esta misma biblioteca fue utilizada para su creación.

El proceso que realizamos fue el siguiente:

- Definimos el vocabulario del nuevo modelo partiendo de un vocabulario generado con el corpus rioplatense.
- Extendimos con el vocabulario del modelo que vamos a ajustar.
- Fijamos los pesos de la intersección entre ambos vocabularios utilizando los ya existentes del modelo a ajustar.
- Le especificamos al modelo que estos pesos deben ser actualizados durante el entrenamiento.
- Volvimos a entrenar el modelo, utilizando un iterador de sentencias creado con el corpus rioplatense.

3.3.2. SUC

Para el *fine-tuning* del SUC utilizamos el corpus rioplatense SUC (ver sección 3.1.1) y la biblioteca `fasttext` con el método *train_supervised*, ya que esta misma biblioteca fue utilizada para su creación.

El proceso que realizamos fue el siguiente:

- Definimos el corpus rioplatense SUC para el entrenamiento.
- Definimos los hiper-parámetros a utilizar.
- Corrimos la función *train_supervised*, que se encarga de realizar el entrenamiento, pasándole los parámetros definidos previamente.
- Por último, guardamos los *embeddings* nuevos.

Los resultados experimentales obtenidos con este entrenamiento y el anterior se presentan a detalle en la sección 4.

3.4. Selección de hiper-parámetros para la construcción y *fine-tuning* de *embeddings*

En esta sección presentamos la selección de hiper-parámetros y el método de búsqueda que realizamos para encontrarlos. Luego los utilizamos para la creación de *embeddings* de los corpus que presentamos en la sección anterior.

3.4.1. Método de selección

Existen diversos métodos para la selección de hiper-parámetros, uno de los más básicos es *grid search*. Este método consiste en definir un listado de posibles valores para cada uno de los parámetros a ajustar y entrenar el modelo con cada una de las combinaciones posibles.

En este trabajo, para la selección de hiper-parámetros utilizamos una variante del *grid search* llamada *random search*. Este método es muy similar al *grid search*, exceptuando que, en lugar de probar todas las combinaciones posibles, se seleccionan de forma aleatoria k combinaciones con las cuales se entrenan k modelos diferentes, restringiendo el espacio y acotando el tiempo de búsqueda.

3.4.2. Hiper-parámetros seleccionados

A continuación detallamos los hiper-parámetros[2] que seleccionamos para ajustar en la creación de *embeddings*, así como los posibles valores que elegimos para cada uno.

Cantidad de ejemplos negativos Especifica la cantidad de ejemplos negativos utilizada para el algoritmo Skipgram. Definimos valores dentro del rango de 5 y 8 ejemplos negativos por palabra, dado que estudios anteriores (Mikolov, Sutskever, y cols., 2013) demuestran que valores dentro de este rango presentan mejores resultados, teniendo en cuenta el tamaño del corpus que utilizamos.

Tamaño de *embeddings* Indica la dimensión de los *embeddings* que devolverá el modelo. Los posibles valores que elegimos fueron 200, 300 y 400. Se probó que la mejor dimensión para los *embeddings* en cuanto tiempo de creación y exactitud es 300 (Pennington, Socher, y Manning, 2014). Decidimos probar también con dimensión 200 y 400 para observar si tenían mejor o peor resultado para los fines de nuestro proyecto. También decidimos probar con dimensión 400 porque los *embeddings* de SWOW (ver 2.3.2) son de esta dimensión.

Tamaño de ventana Hace referencia a cuántas palabras, de las que rodean a la palabra actual en la oración, serán consideradas como palabras dentro de su contexto. Cuanto más bajo es el valor de la ventana las representaciones de las palabras que se obtienen tienden a ser más sintácticas (Mikolov, Sutskever, y

[2]Hiper-parámetros ajustables de la implementación de Word2Vec de gensim https://radimrehurek.com/gensim/models/word2vec.html#gensim.models.word2vec.Word2Vec

cols., 2013), mientras que a valores más altos se obtienen representaciones más semánticas. También es importante tener en cuenta el tamaño de las oraciones, ya que si la ventana supera el tamaño de las oraciones comienza a tomar valores del relleno. Como posibles valores para este hiper-parámetro seleccionamos 14 y 15, dado que es el promedio de palabras en cada oración (ver anexo A.1).

Tasa inicial de aprendizaje Corresponde a la velocidad con la que el modelo aprende. Cuanto más chico el valor, se necesitan más épocas para aprender. Optamos por una tasa inicial de aprendizaje pequeña (entre 0.001 y 0.0001) para no tener cambios abruptos en el aprendizaje.

Tasa mínima de aprendizaje Representa el valor de la tasa mínima de aprendizaje. Optamos por 0.00001 porque representa el mínimo de los posibles valores para la tasa inicial de aprendizaje.

Softmax jerárquico Indica si el modelo devuelve el resultado utilizando softmax jerárquico o no. El softmax jerárquico es un softmax computacionalmente eficiente. Definimos este valor como 0 (no se utiliza softmax jerárquico) debido a que este método es una alternativa al *Negative Sampling* y decidimos usar este último.

Tamaño máximo de vocabulario Limita la RAM durante la construcción del vocabulario. Si hay más palabras únicas que las que puede almacenar el tamaño asignado, se quitan las palabras infrecuentes. Decidimos no limitar el tamaño del vocabulario dado que el poder de cómputo nos fue suficiente. Además, nuestro proyecto se basa en el análisis de palabras que pertenecen al dialecto rioplatense, donde quizás algunas palabras aparecen pocas veces, por lo que no quisimos que se perdiera información quitando palabras infrecuentes.

Ocurrencias mínimas para incluir al vocabulario Indica las ocurrencias mínimas que tiene que tener una palabra en los datos de entrenamiento para que sea incluida en el vocabulario del modelo. Incluimos las palabras al vocabulario cuando ocurrían como mínimo 2 veces.

Algoritmo de entrenamiento Indica si el algoritmo de entrenamiento a utilizar es CBOW (0) o Skipgram (1). Hicimos experimentos utilizando ambos valores.

Épocas Representa la cantidad de iteraciones a realizar durante el entrenamiento del modelo. Decidimos utilizar el mínimo valor posible y que a su vez genere buenos resultados, para optimizar el uso de los recursos de entrenamiento. Para elegir este valor entrenamos diferentes modelos utilizando el corpus rioplatense, generando y evaluando sus *embeddings* al finalizar cada época. A

partir de los resultados obtenidos de los diferentes entrenamiento pudimos observar que el desempeño de los *embeddings* generados comenzó a estancarse a partir de 10 épocas, este valor fue el que elegimos.

Umbral de reducción Define el umbral a partir del cual se reducen aleatoriamente la ocurrencia de palabras con frecuencia alta. Utilizamos el valor 0 para no reducir ninguna palabra.

Otros Hay otros hiper-parámetros que el modelo permite cambiar pero decidimos utilizar sus valores por defecto para reducir el espacio de búsqueda de estos en el *random search*.

Capítulo 4

Análisis de la calidad de los *embeddings*

Los conjuntos de *embeddings* son una herramienta fundamental para poder representar datos en aplicaciones de procesamiento del lenguaje natural. Una vez que generamos un conjunto de *embeddings*, es importante que evaluemos su calidad y midamos qué tan bien logran representar la información y capturar las relaciones del lenguaje para asegurar que sean correctos y eficientes para la aplicación que les queremos dar.

En este capítulo introducimos las pruebas realizadas para evaluar la calidad de los *embeddings* creados utilizando los diferentes enfoques vistos en el capítulo 3; luego presentamos los resultados obtenidos para cada conjunto. Para esto definimos una serie de tareas:

4.1. Obtención de palabras cercanas

Esta prueba consistió en seleccionar palabras y obtener las palabras más cercanas a cada una, utilizando como medida de cercanía la distancia coseno entre los vectores que las representan. Luego analizamos las palabras que fueron obtenidas, así como su valor de cercanía (en un rango de 0 a 1, donde 1 representa la mayor cercanía). Este último valor puede servir como un buen indicador, en los casos en que la palabra más cercana recuperada por el modelo no tiene sentido en el lenguaje, si su valor de cercanía es muy bajo podemos desestimar considerándolo como un resultado irrelevante. Los resultados obtenidos de este tipo de pruebas reflejan parte del significado que les otorgan los modelos a las palabras y cómo las asocian.

Utilizamos esta prueba como método empírico y fácil de entender, ya que no representa la necesidad de realizar tareas complejas.

Para realizar esta prueba elegimos 19 palabras para analizar, 15 las seleccionamos al azar (duda, constante, jefa, trineo, fresco, droga, lectura, frío, habano, billete, sierra, agenda, esclavitud, pingüino y policía) y 4 las seleccionamos ma-

nualmente para analizarlas en profundidad (champión, religión, barrilete y medicamento). Luego, obtuvimos las 10 palabras más cercanas para cada modelo de *embeddings*. El modelo SUC y su *fine-tuning* con el corpus rioplatense SUC (F-SUC), el modelo SBWE y su *fine-tuning* con el corpus rioplatense SBWC (F-SBWE), los mejores tres modelos creados a partir del corpus rioplatense (RP-1, RP-2 y RP-3) y el modelo SWOW.

Dentro de las 19 palabras previamente mencionadas, elegimos 4 palabras para entrar en detalle: "champión", "religión", "barrilete" y "medicamento". Estas palabras presentaban algún tipo de particularidad. En primer lugar, la palabra "champión" es uruguaya, por lo que nos resultó interesante observar cómo se comportarían modelos entrenados en corpus con español genérico y rioplatense. La palabra "religión", en Argentina, que es un país católico, suponemos que estará más asociada al catolicismo que a otras religiones. Luego, la palabra "barrilete" es interesante porque tiene distinto significado si se está hablando en Uruguay o en otro lugar de latinoamérica. Por último, elegimos la palabra "medicamento", dado que consideramos que es una palabra básica y común en español sin importar el dialecto. A priori, esperamos que las palabras cercanas tengan sentido, de no tenerlo, podríamos asumir que los *embeddings* no logran capturar el español.

A priori consideramos que si alguna de las palabras que sea más bien rioplatense (como "champión" que es una palabra uruguaya) no es conocida por algún modelo, el *fine-tuning* de este modelo debería conocerla.

Una vez recuperadas las palabras más cercanas, las observamos sin tener en cuenta la distancia coseno y comparamos esas palabras con lo que creíamos que debía ser la palabra más cercana en la realidad. Por ejemplo, si obtenemos para la palabra "champión" a "zapatilla" es un buen indicio de que los *embeddings* capturan la esencia del dialecto. En cambio, si obtenemos una palabra que no tiene sentido en el dialecto, observamos el valor de cercanía de ambas palabras y nos encontramos con dos escenarios. Si el nivel de cercanía es alto (mayor que 0.5), nos indica que el significado que el modelo entiende de la palabra es diferente al del español rioplatense. Si el nivel de cercanía es bajo (menor que 0.5), esto podría decirnos que el modelo no conoce el significado de la palabra.

Los resultados obtenidos buscando las palabras más similares de los *embeddings* creados fueron para algunas palabras buenos y para otras malos. Para "religión" y "medicamento", las palabras más cercanas se asemejaron bastante a la palabra base. En otros *embeddings* creados, no plasmados en el proyecto, las palabras más cercanas retornaron palabras inexistentes o palabras muy distantes a lo que se creyó correctas. En estos casos, pudimos observar que hubo una correlación entre esta prueba y las otras, dado que para los *embeddings* que dieron malos resultados en esta prueba, dieron también resultados malos en otras pruebas.

En las tablas 4.1, 4.2, 4.3 y 4.4 se muestran las comparativas de las palabras más cercanas a "champión", "religión", "barrilete" y "medicamento" entre los modelos RP-1, RP-2, RP-3, SWOW, SUC, F-SUC, SBWE y F-SBWE.

Modelos							
RP-1		RP-2		RP-3		SWOW	
limpita	0.9873	limpita	0.9860	calzo	0.9356	championes	0.9311
huevito	0.9870	huevito	0.9852	limpita	0.9354	calzado	0.9060
pisarlo	0.9866	pisarlo	0.9847	huevito	0.9346	zapatillas	0.8971
calzo	0.9863	rayita	0.9838	bobeando	0.9246	zapatilla	0.8931
camionetita	0.9861	camionetita	0.9836	gorrita	0.9243	zapato	0.8540
SUC		F-SUC		SBWE		F-SBWE	
lampión	0.8367	lampión	0.8367	-		saguaypicida	0.2700
campión	0.7738	campión	0.7738	-		escatologicos	0.2563
rompión	0.7356	rompión	0.7356	-		lepromina	0.2546
pipión	0.7055	pipión	0.7055	-		Proessdorf	0.2530
espión	0.6950	espión	0.6950	-		Zisuela	0.2465

Tabla 4.1: Listado de las 5 palabras más cercanas a la palabra **champión** obtenidas para cada uno de los modelos (RP-1, RP-2, RP-3, SWOW, SUC, F-SUC, SBWE y F-SBWE).

Modelos							
RP-1		RP-2		RP-3		SWOW	
religiosa	0.9083	religiosa	0.9080	religiosa	0.8745	religioso	0.9037
religiones	0.8922	religiones	0.8905	credo	0.8728	catolicismo	0.8609
religioso	0.8675	credo	0.8663	cristianismo	0.8541	católico	0.8527
creencias	0.8667	religioso	0.8659	religiones	0.8538	santos	0.8238
credo	0.8639	creencias	0.8642	religioso	0.8392	iglesia	0.8209
SUC		F-SUC		SBWE		F-SBWE	
irreligión	0.7271	irreligión	0.7271	creencias	0.7758	creencias	0.7758
religiosa	0.7220	religiosa	0.7220	credo	0.7615	credo	0.7615
creencias	0.7218	creencias	0.7218	religiones	0.7166	religiones	0.7166
religiones	0.7100	religiones	0.7100	politeísta	0.7053	politeísta	0.7053
religios	0.7002	religios	0.7002	cristianismo	0.7015	cristianismo	0.7015

Tabla 4.2: Listado de las 5 palabras más cercanas a la palabra **religión** obtenidas para cada uno de los modelos (RP-1, RP-2, RP-3, SWOW, SUC, F-SUC, SBWE y F-SBWE).

Modelos							
RP-1		RP-2		RP-3		SWOW	
agite	0.9418	agite	0.9426	equilibrista	0.8355	cometa	0.7146
llorona	0.9414	balero	0.9423	barriletes	0.8320	avioneta	0.5267
estrellita	0.9406	llorona	0.9414	arcoiris	0.8295	aéreo	0.5219
bichito	0.9396	bichito	0.9402	lucecita	0.8272	volar	0.5163
viajera	0.9395	arcoiris	0.9393	luciérnaga	0.8262	voladora	0.5129
SUC		F-SUC		SBWE		F-SBWE	
barriletes	0.8058	barriletes	0.8058	pandorga	0.7058	pandorga	0.7058
trilete	0.6127	trilete	0.6127	volantín	0.6993	roncador	0.6314
carrilet	0.5657	carrilet	0.5657	rehilete	0.6397	cinchón	0.6286
volantín	0.5501	volantín	0.5501	roncador	0.6314	papalote	0.6285
sedal	0.5480	sedal	0.5480	cinchón	0.6286	piola	0.6265

Tabla 4.3: Listado de las 5 palabras más cercanas a la palabra **barrilete** obtenidas para cada uno de los modelos (RP-1, RP-2, RP-3, SWOW, SUC, F-SUC, SBWE y F-SBWE).

Modelos							
RP-1		RP-2		RP-3		SWOW	
fármaco	0.8970	fármaco	0.8984	fármaco	0.8581	remedio	0.9602
medicación	0.8796	medicación	0.8799	medicación	0.8433	fármaco	0.9452
medicamentos	0.8673	medicamentos	0.8683	medicamentos	0.8393	medicación	0.9416
antibióticos	0.8469	antibióticos	0.8471	fármacos	0.7877	farmacia	0.9225
fármacos	0.8377	fármacos	0.8343	antibióticos	0.7612	fármacos	0.9195
SUC		F-SUC		SBWE		F-SBWE	
medicamen	0.8390	medicamen	0.8390	fármaco	0.8657	fármaco	0.8657
medicamentosa	0.8388	medicamentosa	0.8388	medicamentos	0.7787	medicamentos	0.7787
medicamentoso	0.8364	medicamentoso	0.8364	fármacos	0.7730	fármacos	0.7730
fármaco	0.8339	fármaco	0.8339	antidepresivo	0.7506	antidepresivo	0.7506
medicamente	0.8296	medicamente	0.8296	oseltamivir	0.7451	oseltamivir	0.7451

Tabla 4.4: Listado de las 5 palabras más cercanas a la palabra **medicamento** obtenidas para cada uno de los modelos (RP-1, RP-2, RP-3, SWOW, SUC, F-SUC, SBWE y F-SBWE).

Observamos que el modelo del SWOW fue muy bueno con las palabras más cercanas, logró capturar la esencia de las palabras, retornando sinónimos o pa-

labras muy similares.

Los resultados de RP-1, RP-2 y RP-3 de las palabras "champión" (ver tabla 4.1) y "barrilete" (ver tabla 4.3) no fueron buenos. Las palabras difirieron mucho de la palabra base y el nivel de cercanía devuelto fue muy alto en todas las palabras cercanas, esto quiere decir que el modelo interpreta estas palabras como similares cuando no lo son, indicándonos que estos modelos no logran capturar el significado real de la palabra base. Las únicas palabras que podrían tener sentido fueron "pisarlo" para "champión" y "viajera" o "equilibrista" para "barrilete". Por otro lado, en palabras como "religión" (ver tabla 4.2) y "medicamento" (ver tabla 4.4), lograron capturar el significado a la perfección.

Pudimos observar un comportamiento particular en los modelos SUC y F-SUC. Las palabras similares devueltas, en lugar de tener significados parecidos, tuvieron una morfología similar. Esto puede deberse a que los modelos SUC y F-SUC fueron entrenados con `fasttext` y este método de creación de *embeddings* captura relaciones basadas en la estructura de las palabras y sus componentes de sub-palabras en lugar de basarse en su significado (Bojanowski y cols., 2017). Para las palabras cercanas a "champión" (ver tabla 4.1) observamos que devuelve palabras con la misma terminación "ión". Para "religión" (ver tabla 4.2) obtuvimos palabras morfológicamente similares pero también mal escritas como, por ejemplo, "religios". Para la palabra "barrilete" (ver tabla 4.3) ocurrió algo similar a "champión", obtuvimos palabras con morfología similar pero significado diferente a excepción de la palabra más cercana que dio el plural y "volantín" pero que fue la cuarta más cercana estando a un nivel de cercanía no muy alto. Por último, para "medicamento" (ver tabla 4.4) sucedió similar al resto pero la palabra más cercana no es correcta, siendo esta "medicamen".

Para los modelos SBWE y F-SBWE pudimos ver algo interesante respecto al *fine-tuning*. Si bien observamos que las palabras similares entre SBWE y F-SBWE son muy parecidas, esto no ocurrió para "champión" (ver tabla 4.1). Para esta palabra, el modelo SBWE no retornó una palabra similar dado que fue una palabra desconocida mientras que en F-SBWE al haber hecho *fine-tuning* con el corpus rioplatense SBWC, adhirió esta palabra a su vocabulario y nos devolvió palabras. Estas no fueron similares a "champión" y el nivel de cercanía fue muy bajo, pero nos dio una pauta de que el modelo adhirió la palabra a su vocabulario. Luego, para las palabras similares a "religión" (ver tabla 4.2), "barrilete" (ver tabla 4.3) y "medicamento" (ver tabla 4.4) los resultados fueron palabras que tienen sentido en la lengua española y con un alto nivel de similitud.

En los resultados obtenidos para la palabra "barrilete" (ver tabla 4.3) pudimos observar algo interesante. En el SWOW, dado que fue un modelo entrenado mayoritariamente con palabras de Uruguay y Argentina, la palabra más cercana fue "cometa". En el SUC (corpus de Chile), aunque no fue la primera palabra, el resultado fue "volantín" que es la forma más común de decirle al barrilete en Chile y por último, en el SBWE observamos "pandorga" como la palabra más cercana que es otra forma de decirle al barrilete en algunas regiones.

4.2. Correlación de Spearman entre similitud devuelta por *embeddings* y juicios humanos

El coeficiente de correlación de Spearman [1] se utiliza para medir la relación entre dos variables aleatorias numéricas. Tiene un rango entre -1 y 1 en el cual los extremos indican una relación negativa perfecta y positiva perfecta respectivamente, y el valor 0 indica que no hay relación entre las variables.

Para interpretar las métricas utilizamos como convención que si el resultado de Spearman supera el umbral de 0.5, esto indica la existencia de correlación entre la similitud devuelta por los *embeddings* y los juicios humanos. Un valor que supere el umbral de 0.7 indica una correlación alta y si supera el umbral de 0.9 consideramos que la asociación es muy alta. Además del resultado obtenido tenemos que tener en cuenta el p-valor asociado al coeficiente de correlación de Spearman, si este valor es significativo entonces aceptamos la correlación, de lo contrario consideramos que la correlación no existe, o análogamente, consideramos el valor como 0. Se toma como umbral de significación el valor 0.05, es decir, que aceptamos el resultado cuando el p-valor está por debajo de este umbral.

Para las pruebas de correlación de Spearman utilizamos 4 conjuntos de similitud (*Multi-SimLex*, *abstract*, *concrete* y RG-65) para medir la correlación entre estos conjuntos de datos anotados y la similitud devuelta por los modelos.

Multi-SimLex

El conjunto de datos *Multi-SimLex* (MSL)[2] (Vulić y cols., 2020) corresponde a la relación léxica de similitud semántica de pares de palabras. Consiste en 1888 pares de palabras que fueron anotadas por 10 personas, hablantes nativas del español. Utilizamos el promedio del valor anotado por los anotadores para cada par de palabras.

Juicios de relacionamiento semántico en hablantes rioplatenses para palabras abstractas y concretas

Estos datos provienen del estudio (De Deyne, Cabana, Li, Cai, y McKague, 2020), y consisten en juicios de relacionamiento semántico (qué tan relacionadas están las palabras) para un conjunto de 3321 pares de palabras abstractas (acuerdo-alma) y 3321 pares de palabras concretas (animal-camello). Estos datos fueron obtenidos a partir de una tarea realizada en línea por hablantes nativos del español rioplatense. Una característica de estos conjuntos es que entre ellos hay muy pocos pares con alto relacionamiento, la mayoría de los pares tienen un relacionamiento más bien bajo.

[1] Coeficiente de correlación de Spearman de la librería `Scipy` https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

[2] Multi-SimLex https://multisimlex.com

RG-65

El conjunto de datos RG-65 fue creado por (Rubenstein y Goodenough, 1965). Es una colección de 65 pares de palabras anotadas por 51 personas en una escala del 0 al 4 (cuanto más grande el número, mayor similitud). Este valor es el promedio de lo anotado por todos los participantes. El conjunto original se realizó en inglés, pero en (Camacho Collados, Pilehvar, y Navigli, 2015) realizaron la traducción al español y este último fue el que utilizamos para las pruebas.

En el estudio original SWOW-RP (Cabana y cols., 2023), se muestra que los *embeddings* derivados de la asociación libre de palabras muestran una pequeña ventaja frente a los obtenidos a partir de corpus textuales cuando son evaluados en estos datos.

A priori, consideramos que los *embeddings* rioplatenses o los *embeddings* a los que le realizamos *fine-tuning* deberían dar mejores valores de correlación en las pruebas realizadas contra *abstract* y *concrete* que con el conjunto MSL dado que los primeros fueron anotados por personas de Argentina o Uruguay y el último fue anotado por hablantes nativos del español de cualquier parte del mundo.

Antes de obtener la correlación entre la similitud devuelta por los modelos y la similitud anotada en los conjuntos de datos, realizamos un preprocesamiento a estos últimos, dado que hubo una gran cantidad de palabras que no fueron encontradas para poder calcular su similitud con otra. Por ejemplo, para los pares de palabras del MSL, todos los *embeddings* tuvieron muy pocas palabras no encontradas a excepción del SWOW que tuvo casi un 20 % de palabras no encontradas. Por otro lado, para el conjunto de datos RG-65 no se encontraron algunas palabras en SWOW y otras tampoco en RP-1, RP-2 y RP-3, con un total de 31 % palabras no encontradas.

El preprocesamiento consistió en obtener las palabras desconocidas de todos los conjuntos de datos para cada modelo y luego eliminar de todos los conjuntos estas palabras. De esta forma, las comparaciones que presentaremos a continuación, las realizamos comparando para cada modelo, el mismo conjunto de datos. También realizamos un análisis reducido utilizando todas las palabras conocidas para cada modelo, aunque estas no se encontraran en el vocabulario de algún otro, ver anexo A.2.

El numero final de pares de palabras para cada conjunto de datos fue:

- MSL: pasó de 1793 a 1339.
- *Abstract*: pasó de 3321 a 3160.
- *Concrete*: pasó de 3321 a 3240.
- RG-65: pasó de 65 a 45.

En las gráficas 4.1, 4.2, 4.3 y 4.4 mostramos la comparación de los resultados obtenidos en el cálculo de la correlación de Spearman comparando los 8 *embeddings* contra los conjuntos MSL, *abstract*, *concrete* y RG-65.

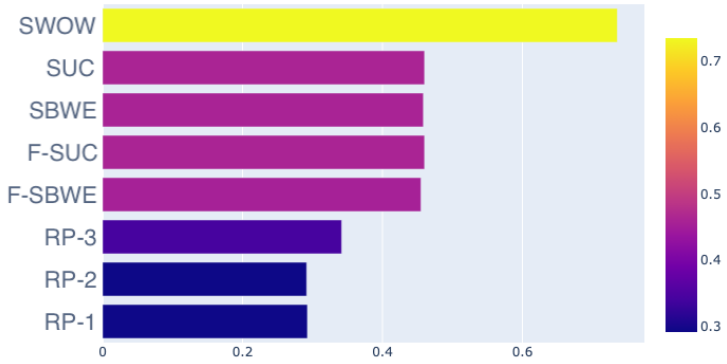

Figura 4.1: Correlación de Spearman con MSL

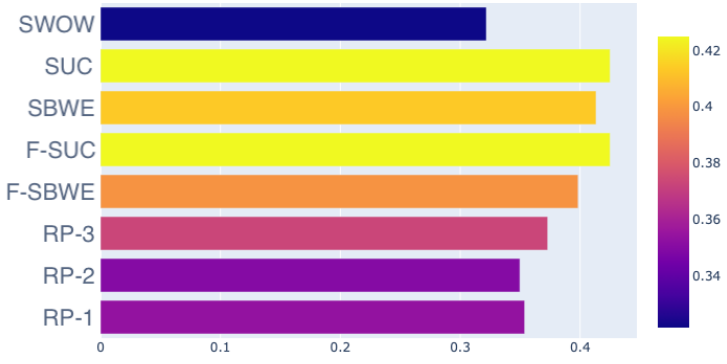

Figura 4.2: Correlación de Spearman con *Abstract*

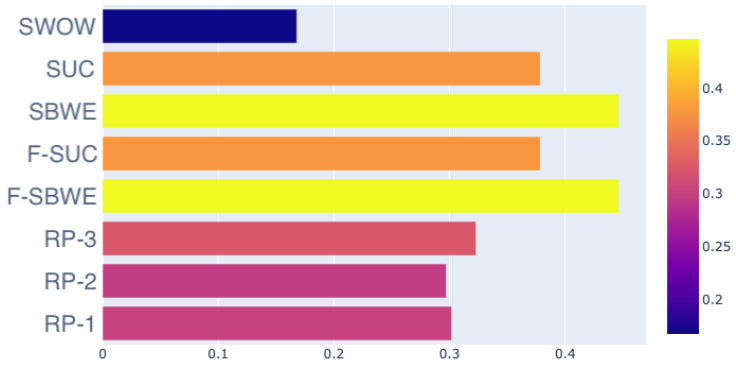

Figura 4.3: Correlación de Spearman con *Concrete*

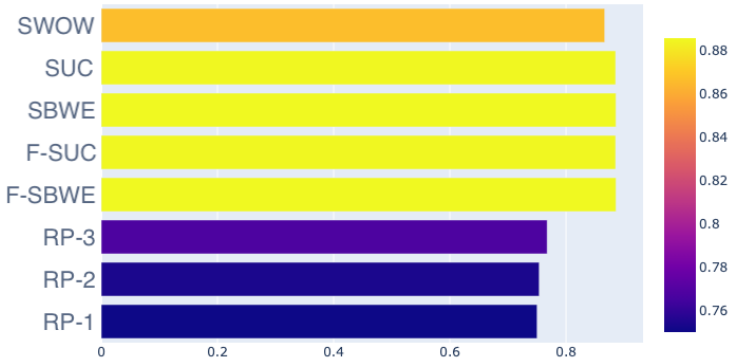

Figura 4.4: Correlación de Spearman con RG-65

En primer lugar, al analizar los resultados obtenidos para los *embeddings* RP-1, RP-2 y RP-3, observamos una baja correlación con el conjunto MSL, mientras que para los conjuntos *abstract* y *concrete* los resultados fueron mejores, aunque aún inferiores al SWOW. Además, notamos que los modelos están correlacionados con el conjunto RG-65, aunque no tanto como lo están otros *embeddings*. También observamos que el modelo RP-3 obtuvo ligeramente mejores resultados en general que RP-1 y RP-2.

Por otro lado, el modelo SWOW mostró resultados de correlación excelentes con los conjuntos MSL y RG-65, superando el umbral de 0.7, lo que indica una correlación significativa. Sin embargo, en los otros dos conjuntos (*abstract* y *concrete*), el SWOW obtuvo los peores resultados en comparación con los demás modelos.

En cuanto a los modelos SUC y F-SUC, aunque la correlación con MSL, *abstract* y *concrete* no superó 0.5, los resultados fueron muy cercanos, lo que sugiere cierta correlación entre el modelo y los datos. Por otro lado, la correlación con RG-65 fue excelente, siendo estos *embeddings* junto con los modelos SBWE

y F-SBWE los que obtuvieron la mayor correlación, con un valor cercano a 0.9, lo que indica una correlación muy alta.

Por último, los resultados del SBWE y F-SBWE fueron similares a los del SUC y F-SUC en MSL y RG-65, ligeramente más bajos en *abstract* y ligeramente más altos en *concrete*. Resulta interesante destacar que la correlación de SBWE es mayor que la del F-SBWE en el conjunto *abstract*, lo que difiere de nuestras expectativas iniciales.

Los resultados obtenidos ofrecen diversas conclusiones. En el caso del modelo SWOW, observamos que logró capturar la similitud semántica entre palabras en español genérico al obtener buenos resultados con los conjuntos MSL y RG-65. Sin embargo, notamos que no existe una correlación significativa con los conjuntos *abstract* ni *concrete*, lo que indica que, a priori, no podemos afirmar que tenga la capacidad de diferenciar adecuadamente palabras del español rioplatense.

En cuanto a los modelos RP-1, RP-2 y RP-3, los resultados son menos concluyentes. La baja correlación con MSL, *abstract* y *concrete*, y la alta correlación con RG-65, dificultan extraer conclusiones claras. No obstante, dado que RG-65 es un conjunto de palabras pequeño (45 palabras), no podemos afirmar que estos modelos logren capturar adecuadamente el español genérico ni su variante rioplatense.

Por otro lado, los modelos SUC y F-SUC muestran un leve indicio de capturar el español genérico debido a su baja correlación con MSL y su alta correlación con RG-65. Sin embargo, no podemos decir que estos modelos logren capturar el español rioplatense debido a su baja correlación con *abstract* y su falta de correlación con *concrete*.

En cuanto a los modelos SBWE y F-SBWE, presentan resultados similares a los modelos SUC y F-SUC en relación con el español genérico. Sin embargo, estos dos últimos logran capturar levemente el español rioplatense al estar casi correlacionados tanto con *abstract* como con *concrete*.

En resumen, el análisis que realizamos sugiere que el modelo SWOW logra capturar la similitud semántica entre palabras en español genérico, pero no en el español rioplatense. Los modelos RP-1, RP-2 y RP-3 no logran capturar adecuadamente el español en ninguna variante. Los modelos SUC, F-SUC, SBWE y F-SBWE lograron capturar de manera leve el español genérico, y los dos últimos también capturaron de manera leve el español rioplatense.

4.3. Prueba de analogías

Una forma de analizar la capacidad que tienen los *embeddings* de capturar los significados relacionales, consiste en utilizar el modelo del paralelogramo (desarrollado por Rumelhart y Abrahamson en 1973) para resolver problemas de analogía simples del tipo “v es a w como u es a qué”.

En este tipo de problemas se toma una pregunta del estilo “manzana : árbol :: uva : ?”. El modelo del paralelogramo suma el vector que va desde manzana hasta árbol con el vector de uva y devuelve la palabra más cercana al punto

obtenido.

Trabajos anteriores han demostrado que estas pruebas resultan de gran utilidad para vectores creados con `word2vec` (Mikolov, Yih, y Zweig, 2013), que logran obtener resultados como que "rey - hombre + mujer" es un vector cercano a "reina", o "París - Francia + Italia" es un vector cercano a "Roma".

Existen limitaciones conocidas para estas pruebas, `word2vec` puede obtener buenos resultados utilizando el método del paralelogramo para palabras frecuentes, distancias cortas o relaciones simples, pero para relaciones no tan simples, el algoritmo suele devolver como resultado una de las tres palabras de entrada o una de sus variantes morfológicas. Por lo que se considera que no puede modelar adecuadamente analogías complejas (Jurafsky y Martin, 2009).

Decidimos realizar esta prueba debido a que la mayoría de los *embeddings* con los que trabajamos utilizan algoritmos de `word2vec` o extensiones. Para realizar la prueba lo primero que hicimos fue definir las entradas para el algoritmo. Incluimos analogías muy similares a las presentadas en (Jurafsky y Martin, 2009) que creímos deberían tener buenos resultados y dar un primer indicio de la capacidad de capturar relaciones simples en los modelos de *embeddings*, estas fueron: "hombre es a rey como mujer es a", "saltar es a saltando como correr es a" y "francia es a parís como uruguay es a". También incluimos analogías variantes de la presentada en (Bolukbasi y cols., 2016a) enfocadas en detectar la presencia de estereotipos de género en los vectores, estas fueron: "hombre es a desarrollador como mujer es a", "mujer es a limpiadora como hombre es a" y "hombre es a fuerte como mujer es a". Por último, decidimos incluir analogías con palabras rioplatenses (con relaciones sencillas de sinonimia) para observar si alguno de los modelos lograban capturar el significado, estas fueron: "zapato es a champión como calcetín es a" y "pancho es a salchicha como sándwich es a".

v es a w como u es a (w - v + u)	**resultado esperado**	**SUC**	**F-SUC**	**SBWE**	**F-SBWE**
rey - hombre + mujer	reina	reina	reina	reina	reina
saltando - saltar + correr	corriendo	corriendo	corriendo	corriendo	corriendo
parís - francia + uruguay	montevideo	montevideo	montevideo	Malargue	Malargue
desarrollador - hombre + mujer	desarrolladora	desarrolladora	desarrolladora	TOra	desarrolladores
limpiadora - mujer + hombre	limpiador	limpiador	limpiador	barrendero	barrendero
fuerte - hombre + mujer	fuerte	fuerta	fuerta	fuertes	fuertes
champión - zapato + calcetín	media	lampión	lampión	-	Gaillardo
salchicha - pancho + sándwich	refuerzo	hamburguesa	hamburguesa	sandwich	sandwich

Tabla 4.5: Resultados obtenidos en la prueba de analogías para los modelos SUC, F-SUC, SBWE y F-SBWE.

v es a w como u es a (w - v + u)	**resultado esperado**	**RP-1**	**RP-2**	**RP-3**	**SWOW**
mujer - hombre + rey	reina	princesa	reina	princesa	reina
saltando - saltar + correr	corriendo	corriendo	corriendo	corriendo	-
parís - francia + uruguay	montevideo	uruguayo	uruguayo	montevideo	uruguayo
desarrollador - hombre + mujer	desarrolladora	labs	labs	developer	-
limpiadora - mujer + hombre	limpiador	panadero	panadero	panadero	-
fuerte - hombre + mujer	fuerte	fuertes	débil	fuertes	poderosa
champión - zapato + calcetín	media	ltera	ltera	centrifugar	medias
salchicha - pancho + sándwich	refuerzo	tocino	tocino	patata	fiambre

Tabla 4.6: Resultados obtenidos en la prueba de analogías para los modelos RP-1, RP-2, RP-3 y SWOW.

Al igual que en las pruebas de cercanía de palabras, podemos corroborar que el *fine-tuning* en el modelo SBWE logró incorporar características del dialecto (al pasar a conocer la palabra "champión"), también mejora al capturar el significado de la palabra "desarrollador", creemos que esto es porque "desarrollador" podría ser considerada una palabra más "moderna" que también se incorporó al vocabulario del modelo en el ajuste.

Observamos que para analogías entre palabras relativamente sencillas (como femenino-masculino, capital-ciudad-de y verbo-gerundio), todos los modelos parecen devolver resultados que indican que son capaces de capturar significados relacionales. De igual manera, algunos devuelven mejores valores que otros, por ejemplo, RP-1 y RP-3 devuelve "princesa" cuando debería devolver "reina". También observamos que en lugar de obtener "Montevideo", en SBWE y F-SBWE los modelos devuelven "Malargue" y en RP-1, RP-2 y SWOW devuelve "uruguayo". Para los primeros dos, el resultado se debe a que la analogía se buscó con las palabras en minúscula, al hacerlo con mayúscula, el resultado fue también "Montevideo".

Por otro lado, observamos que para la analogía de "champión", el único modelo que devuelve algo coherente es el modelo SWOW con "medias". Creemos que esto se debe a que estas palabras son de la región rioplatense y que los otros modelos no tienen el significado de estas palabras.

Capítulo 5

Análisis de sesgo de *embeddings*

Como vimos en el capítulo 2.4 es importante evaluar la calidad de los *embeddings* en cuanto a qué tan efectivos son como representación de la realidad; sin embargo, la efectividad de los *embeddings* por si sola no es suficiente para su aplicación en un sistema, es igual de relevante considerar los sesgos presentes (subyacentes) en los *embeddings* para garantizar la idoneidad de su uso en una determinada aplicación.

En este capítulo definimos las métricas utilizadas para cuantificar y medir el sesgo de los *embeddings*; luego, presentamos los resultados experimentales obtenidos de su aplicación en diferentes modelos. En particular, analizamos la existencia de sesgos regionales inherentes a los modelos creados o ajustados utilizando datos de la región rioplatense, en comparación con modelos que para su desarrollo utilizaron datos de otras regiones del habla hispana.

5.1. Medidas de sesgo

En esta sección presentamos las pruebas utilizadas para medir el sesgo en los *embeddings*.

5.1.1. Cuantificación del sesgo con MWEAT

Una de las pruebas que realizamos para medir el sesgo en un conjunto de vectores fue el MWEAT (*Modified Word Embedding Association Test*), una modificación del WEAT (*Word Embedding Association Test*) propuesta por (Zhou y cols., 2019).

El WEAT nos da un valor de asociación entre dos conjuntos de conceptos objetivo y dos conjuntos de atributos. Mientras que el MWEAT se utiliza para cuantificar el sesgo en conjuntos monolingües de *embeddings*. Esta prueba consiste en, dados dos conjuntos de conceptos objetivos (X e Y) y dos conjuntos de

atributos (A y B), computar:

$$|| \sum_{\vec{x} \in X} s(\vec{x}, A, B)| - | \sum_{\vec{y} \in Y} s(\vec{y}, A, B)||$$

donde el valor de la primera sumatoria representa la asociación del conjunto de conceptos X con los conjuntos de atributos, y el valor de la segunda sumatoria la asociación del conjunto de conceptos Y con los conjuntos de atributos.

Una vez calculado el valor del MWEAT, calculamos el p-valor del test de permutaciones. Para esto, se realizan N muestreos aleatorios de los conjuntos de atributos, generando N nuevos pares de conjuntos. Luego, para cada uno de estos pares se vuelve a calcular el valor del MWEAT, generando una distribución de valores, sobre la cual se calcula el p-valor como la proporción de resultados en los que el MWEAT calculado es mayor o igual al valor del MWEAT obtenido originalmente.

La interpretación de esta prueba es que cuanto mayor es el valor obtenido, mayor es el sesgo que presenta el conjunto de *embeddings*.

5.1.2. Distancia a subespacios de palabras

Otra prueba realizada para medir el sesgo de los *embeddings* fue definir pares de palabras que representan los dos extremos de la categoría del tipo de sesgo a analizar. Estos dos conjuntos están formados tomando para cada palabra de un extremo, la palabra que consideramos se corresponde al extremo opuesto (e.g., para definir los polos femenino-masculino, dos posibles conjuntos serían {mujer, chica, muchacha} y {hombre, chico, muchacho}).

Una vez definidos estos dos conjuntos, medimos, para una palabra dada, la similitud a cada uno de los polos. Para computar esta distancia utilizamos como métrica la similitud coseno promedio. La idea detrás de esto es que, si la distancia de la palabra con ambos polos es igual, esa palabra no se encuentra sesgada para la categoría definida por los polos. En cambio, si la distancia es mayor hacia uno de los dos polos, este desvío representa un sesgo a favor del conjunto representado por el polo.

5.1.3. Proyección semántica

Otro método utilizado para medir el sesgo, conceptualmente similar al anterior, es el método de proyección semántica propuesto en (Grand y cols., 2022). Este método consiste en representar un subespacio semántico como una línea y luego utilizarla como referencia para comparar diferentes atributos bajo ese subespacio. Para poder representar el subespacio y los atributos gráficamente, se utilizan tres dimensiones seleccionadas de forma arbitraria a través de la aplicación del método de análisis de componentes principales o PCA[1] (*Principal Component Analysis*) sobre los *embeddings*.

[1]Documentación de PCA de la librería Scikit-learn https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

El estudio original de proyección semántica, comprueba que este método es eficiente para estimar el conocimiento humano sobre la estructura de categorías bajo diferentes contextos. En particular, demuestra que se obtienen mejores resultados cuando la representación de los subespacios se computa como la diferencia entre dos polos opuestos, y cuando la métrica de similitud es la proyección ortogonal sobre el subespacio y no la distancia a este. Para la implementación del método se decidió utilizar este acercamiento.

Lo primero que debemos hacer es dibujar la recta que va a representar el subespacio semántico, para esto es necesario definir los dos extremos. Para nuestro análisis utilizamos los mismos conjuntos de pares de atributos opuestos de la sección anterior. Cada extremo del subespacio está entonces representado por el promedio de los vectores de cada conjunto.

Luego, para cada palabra, calculamos su proyección ortogonal sobre el subespacio semántico. La idea detrás de esto es que, si la proyección de la palabra se acerca más hacia un extremo, podríamos decir que está más sesgada hacia él en ese subespacio. E.g., nos creamos un subespacio semántico que representa el atributo "tamaño" utilizando los extremos "chico" y "grande", si proyectamos los vectores "perro" y "ballena" sobre ese subespacio, entonces podemos decir que el "perro" es más pequeño que la "ballena" porque se encuentra más hacia el extremo de "chico" que la "ballena". Es importante destacar que esto no necesariamente nos dice que el "perro" sea un animal pequeño, solo nos dice que es más pequeño que la "ballena".

Una de las principales fortalezas de esta prueba es que al reducir a 3 la dimensionalidad de los vectores, permite visualizarlos gráficamente y analizarlos más fácilmente que otras pruebas.

5.2. Resultados experimentales

En esta sección, presentaremos los resultados experimentales obtenidos de la aplicación de las pruebas descritas en la sección anterior para medir el sesgo presente en diferentes modelos de *embeddings* del idioma español. En particular, utilizamos algunos modelos entrenados utilizando datos obtenidos en la región del Río de la Plata, además de otros entrenados con datos obtenidos de internet que generalmente utilizan un vocabulario representado por un español más neutro. Estos *embeddings* fueron introducidos en el capítulo 3.

Para poder realizar un análisis de sesgo, lo primero que debemos hacer es definirnos el subespacio bajo el cual queremos analizarlo. Existen diversas formas de definirlo, en este proyecto decidimos basarnos en trabajos anteriores (Grand y cols., 2022) que han demostrado que los resultados más relevantes surgen a partir de la definición de este como dos conjuntos de palabras "opuestas" que representan los extremos o polos del subespacio.

Una vez definido el subespacio de estudio debemos definir los atributos que queremos estudiar bajo este. Esto consiste en definir conjuntos de palabras o pares de palabras que puedan encontrarse sesgadas en el subespacio.

En este proyecto seleccionamos para su análisis los subespacios de género,

raza, y representación de colonizados y colonizadores. Consideramos pertinente el análisis de los primeros dos, dado que se encuentran dentro de las recomendaciones éticas para la inteligencia artificial (UNESCO, 2022) establecidas por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. En particular, estas recomendaciones enfatizan la importancia del control en la representación desigual en ciertas profesiones y actividades, por lo que consideramos que, a pesar de no ser un análisis innovador, los atributos de profesiones debían ser incluidos como parte de las pruebas.

Definimos el **subespacio de género** como los polos {femenino - masculino}, representando al atributo "femenino" como el conjunto {"abuela", "chica", "doña", "ella", "esposa", "femenino", "gurisa", "hermana", "hija", "madre", "muchacha", "mujer", "niña", "novia", "señora", "tía"} y al atributo "masculino" como el conjunto {"abuelo", "chico", "don", "él", "esposo", "masculino", "gurí", "hermano", "hijo", "padre", "muchacho", "hombre", "niño", "novio", "señor", "tío"}.

El **subespacio de raza** como los polos {raza afrodescendiente - raza blanca}. Donde el atributo "raza afrodescendiente" está conformado por el conjunto {"negra", "negro", "afrodescendiente", "africano", "moreno", "oscuro"} y el atributo "raza blanca" está conformado por el conjunto {"blanca", "blanco", "caucásico", "europeo", "nórdico", "rubio"}.

Por último, el **subespacio de representación de colonizados y colonizadores** como {colonizado - colonizador}. Donde el atributo "colonizado" está conformado por el conjunto {"indígena", "nativo", "sometido", "despojado", "oprimido", "dominado", "desplazado", "originario", "colonizado"} y "colonizador" conformado por el conjunto {"invasor", "colonizador", "conquistador", "ocupante", "opresor", "esclavista", "colono", "dominador", "saqueador"}.

Utilizamos los conjuntos de atributos {profesiones} y los cuatro conjuntos de polaridad y visibilidad que creamos según la categorización propuesta en (Garrido-Muñoz y cols., 2022), en donde se menciona que los resultados obtenidos del análisis de sesgo utilizando estos cuatro conjuntos son más significativos en comparación con otras categorías.

Además de analizar el sesgo con una visión más "clásica", decidimos enfocar parte del análisis hacia posibles sesgos regionales, realizando una comparación y análisis de sesgos que podrían deberse únicamente a la región de recolección de datos. Al utilizar un modelo basado en asociación libre de palabras dentro de los modelos, también podemos analizar los sesgos inherentes a este tipo de recolección de datos en comparación con los *embeddings* generados a partir de contexto extraído de grandes colecciones de textos.

5.2.1. Cuantificación del sesgo con MWEAT

En el experimento original realizado por (Zhou y cols., 2019), emplearon un conjunto de *embeddings* de Fasttext[2] para llevar a cabo las pruebas.

[2]Embeddings en español de Fasttext https://dl.fbaipublicfiles.com/arrival/vectors/wiki.multi.es.vec

Lo primero que hicimos fue utilizar los conjuntos de atributos y conceptos de la investigación original (ver Anexo A.3), para compararnos contra sus resultados. Para esto los dos conjuntos de conceptos objetivo son profesiones en su forma masculina y femenina respectivamente, y los dos conjuntos de atributos son palabras que representan a hombre y a mujer respectivamente.

En la tabla 5.1 mostramos los resultados obtenidos por cada modelo de *embeddings* en relación con la función MWEAT utilizando las profesiones y el género.

Modelo	MWEAT	P-valor
Fasttext	3.6605	0.000
RP-3	**2.1108**	0.000
SBWE	2.1212	0.014
RP-2	25.2377	0.000
RP-1	25.7560	0.000
F-SBWE	50.7722	0.000
SUC	75.1728	0.000
F-SUC	75.1728	0.000

Tabla 5.1: Resultados del cálculo de la métrica MWEAT para cada modelo, comparando profesiones en los subespacios del **género** (mujer-hombre)

Como mencionamos previamente en la sección 5.1.1, cuanto mayor el resultado del MWEAT, mayor es el sesgo que presentan los modelos de *embeddings* para las profesiones en el espacio del género binario (mujer-hombre). En los resultados que presentamos en la tabla 5.1 podemos observar que el modelo RP-3 es el que presenta menor sesgo seguido por el modelo SBWE, ambos menores incluso que el modelo Fasttext. Es importante observar que el valor de MWEAT reportado en esta tabla para el modelo de Fasttext difiere mínimamente del valor presentado en (Zhou y cols., 2019). Para validar este resultado, comparamos el valor obtenido con el resultado de la prueba que se encuentra disponible en el repositorio de la investigación original[3] y obtuvimos el mismo valor. Entendemos que esta diferencia puede deberse a cambios en los *embeddings* de Fasttext, dado que su conjunto de vectores se sigue actualizando desde la publicación del paper.

Pudimos observar que el resto de los modelos devuelven un valor de MWEAT bastante malo en comparación a los primeros dos.

Para el modelo SWOW, fue necesario realizar un preprocesamiento de las

[3]Repositorio de "Examining Gender Bias in Languages with Grammatical Gender" https://github.com/shaoxia57/Bias_in_Gendered_Languages/blob/master/bias_emnlp19.ipynb

profesiones, ya que encontramos que este modelo carecía de muchas de ellas. Inicialmente, contábamos con 60 pares de profesiones con género sintáctico, pero tras el preprocesamiento, se redujeron a solo 18 (ver Anexo A.3 para ver los pares de palabras). Con el objetivo de obtener resultados más realistas, repetimos la evaluación con el modelo Fasttext utilizado en el experimento, pero esta vez utilizando únicamente las 18 profesiones filtradas. Presentamos el resultado obtenido en la tabla 5.2.

Modelo	MWEAT	P-valor
Fasttext	0.95375	0.000
SWOW	**0.00002**	0.011

Tabla 5.2: Resultados del cálculo de MWEAT para los modelos Fasttext y SWOW con un conjunto reducido de pares de profesiones en los subespacios del **género** (mujer-hombre)

En la tabla 5.2 podemos ver que el modelo SWOW exhibe considerablemente menos sesgo en comparación con el modelo de Fasttext. Sin embargo, es importante destacar que obtuvimos este resultado a través de una reducción significativa en la cantidad de los ejemplos que utilizamos. Si pudiéramos contar con el mismo número de ejemplos que en el experimento original (60 pares de profesiones), tendríamos mayor sustento para afirmar que los *embeddings* de SWOW son realmente los que presenta la menor cantidad de sesgo en las profesiones en el espacio del género.

Posteriormente realizamos el mismo experimento pero, en lugar de utilizar el género como atributo, utilizamos el subespacio de raza. En la tabla 5.3 presentamos los resultados obtenidos por cada modelo de *embeddings* en relación con la función MWEAT utilizando el mismo conjunto de profesiones y la raza. Observamos que todos los modelos muestran mayor sesgo que Fasttext, aunque el modelo SBWE obtuvo una diferencia ínfima en comparación a este.

En la tabla 5.4 presentamos los resultados comparando Fasttext y SWOW con el conjunto de profesiones reducido (ver Anexo A.3 para ver los pares de palabras). En esta tabla observamos resultados similares a los que obtuvimos anteriormente frente al SWOW, el valor es mejor que Fasttext y si tuviéramos más datos podríamos afirmar que es el que tiene menos sesgo.

Por último, realizamos el experimento para el subespacio de representación de **colonizadores** y **colonizados**. En la tabla 5.5 podemos observar que el SBWE es el modelo que obtiene el mejor resultado.

Modelo	MWEAT	P-valor
Fasttext	**0.3283**	0.205
SBWE	0.6192	0.028
RP-2	3.5023	0.448
RP-3	4.6593	0.373
RP-1	4.7681	0.303
F-SBWE	10.1845	0.028
SUC	13.9815	0.005
F-SUC	13.9815	0.004

Tabla 5.3: Resultados del cálculo de MWEAT para cada modelo, comparando pares de profesiones en los subespacios definidos para **raza** (raza blanca-raza afrodescendiente)

Modelo	MWEAT	P-valor
Fasttext	0.0299036	0.868
SWOW	**0.0000057**	0.729

Tabla 5.4: Resultados del cálculo de MWEAT para los modelos Fasttext y SWOW con un conjunto reducido de pares de profesiones en los subespacios definidos para **raza** (raza blanca-raza afrodescendiente)

Modelo	MWEAT	P-valor
Fasttext	1.7686	0.000
SBWE	**1.4106**	0.000
RP-3	11.1814	0.008
RP-2	18.3502	0.004
RP-1	18.8834	0.005
F-SBWE	26.4794	0.000
SUC	31.7956	0.000
F-SUC	31.7956	0.000

Tabla 5.5: Resultados del cálculo de MWEAT para cada modelo, comparando pares de profesiones en los subespacios definidos para **colonización** (colonizado-colonizador)

Modelo	MWEAT	P-valor
Fasttext	0.5643881	0.000
SWOW	**0.0000035**	0.493

Tabla 5.6: Resultados del cálculo de MWEAT para los modelos Fasttext y SWOW con un conjunto reducido de pares de profesiones en los subespacios definidos para **colonización** (colonizado-colonizador)

En la tabla 5.6 mostramos los resultados comparando Fasttext y SWOW para el concepto de colonización. Al igual que en los conjuntos de atributos de género y raza, en los conjuntos de atributos para colonización, el modelo con menor sesgo es el SWOW.

En resumen, al analizar los resultados de todos los conjuntos, podemos observar consistentemente que los modelos SUC y F-SUC presentaron los mayores niveles de sesgo. Por otro lado, el modelo SBWE mostró el menor sesgo en la mayoría de los conjuntos de atributos. Es interesante notar que, a pesar de que el modelo SBWE fue el menos sesgado en casi todos los conjuntos, el modelo F-SBWE demostró ser significativamente peor, presentando niveles de sesgo casi tan altos como los observados en SUC o F-SUC en todos los conjuntos que analizamos. También pudimos observar que el modelo SWOW parece estar muy poco sesgado pero dada la poca cantidad de datos para las pruebas, no podemos afirmarlo.

También realizamos el experimento con todos los modelos pero utilizando el conjunto reducido de profesiones (ver Anexo A.3) en lugar del conjunto entero. Vimos que el modelo SWOW obtuvo el menor nivel de sesgo comparado a todos los otros modelos para todos los atributos, y que el resto se mantuvo en la misma posición que las pruebas realizadas con el conjunto entero de profesiones. Sería muy interesante poder observar al modelo SWOW con el conjunto entero de profesiones (si lo tuviera) para poder o no afirmar que es el modelo con menor nivel de sesgo en base a esta prueba.

5.2.2. Distancia a subespacios de palabras

Lo primero que hicimos fue seleccionar dos conjuntos de palabras que representan los extremos del género binario, conformando el subespacio de género. Una vez definimos estos subespacios, calculamos el vector promedio dentro de cada uno de los conjuntos de *embeddings* a analizar. Luego, calculamos la distancia (similitud coseno) entre el vector promedio de cada subespacio y otras palabras por fuera de los subespacios para cuantificar su relación semántica, como explicamos en la subsección 5.1.2.

Como nuestro objetivo es analizar si estos vectores pueden acarrear algún tipo de prejuicio en el subespacio del género, decidimos seleccionar profesiones como palabras de análisis, teniendo en cuenta que, en idioma español, la mayoría de las profesiones tienen el género incorporado en su gramática. Si

bien conformamos un conjunto de más de 200 profesiones (ver Anexo A.4), por simplicidad vamos a analizar las 4 profesiones {"actor/actriz/actuación", "doctor/doctora/medicina", "enfermero/enfermera/enfermería", "maestro/maestra/enseñanza"} que en particular, cumplen que tanto su forma femenina y su forma masculina, como la carrera profesional que representan se encuentran en el vocabulario de todos los *embeddings* a analizar.

En una primera instancia, realizamos el análisis utilizando la forma femenina de las profesiones para calcular la distancia al subespacio femenino y la forma masculina para la distancia al subespacio masculino (ver Figura 5.1).

A priori, al descomponer las profesiones en su forma masculina y femenina, podemos esperar que la forma femenina siempre sea más marcadamente femenina de lo que la forma masculina es considerada masculina, esto debido a que en el idioma español el género masculino es muchas veces utilizado como genérico o neutro. Observando la gráfica de la figura 5.1, vemos que esta tendencia se cumple. Dado que la mayoría de los *embeddings* que estamos analizando son *embeddings* de contexto, tiene sentido que muchas veces la palabra masculina esté presente en un contexto en el que se habla también de mujeres, en cambio, muy rara vez vamos a encontrar en algún texto una referencia al masculino en términos de una profesión en su forma femenina.

Los modelos rioplatenses (RP-1, RP-2 y RP-3) son los únicos *embeddings* de contexto que tienen mayor asociación masculina para el par doctor-doctora. Que los resultados de estos tres modelos sea parecido no es de extrañar ya que los tres utilizaron el mismo corpus para su creación y difieren únicamente en algunos hiper-parámetros de entrenamiento.

Destacando entre los modelos vemos que el SWOW devuelve resultados marcadamente diferentes en comparación con el resto de los modelos. Como vimos anteriormente, este es el único de los modelos que no fue creado a partir de colecciones de texto, sino que se creó representando el contexto de las palabras como las palabras devueltas de un test de asociación libre. De igual manera, en general devuelve mayor asociación femenina, pero un caso particular es para el par enfermero-enfermera, donde la palabra enfermero tiene mayor asociación masculina que la palabra enfermera asociación femenina. Este resultado es interesante ya que la enfermería es una de las profesiones más estereotipadas a lo femenino, si recordamos el caso de (Prates, Avelar, y Lamb, 2020) donde la traducción de la palabra *doctor* en inglés al español es el masculino "doctor" pero de la palabra *nurse* es "enfermera" y no "enfermero".

En una segunda instancia, decidimos realizar la misma prueba, pero esta vez tomando el nombre de la profesión para calcular tanto la distancia al subespacio **femenino** como la distancia al subespacio **masculino** (ver Figura 5.2). A priori, consideramos que este segundo acercamiento podría ser interesante ya que quitamos el factor del género masculino siendo utilizado como neutro en las profesiones, porque dejamos de tener la posibilidad de una forma masculina y otra femenina para cada palabra.

En la gráfica podemos observar cómo la palabra "enseñanza" pasó a estar más inclinada hacia el subespacio **masculino** que la tendencia vista para el par maestro-maestra, destacándose en los modelos RP-1, RP-2, RP-3 y SWOW. Este

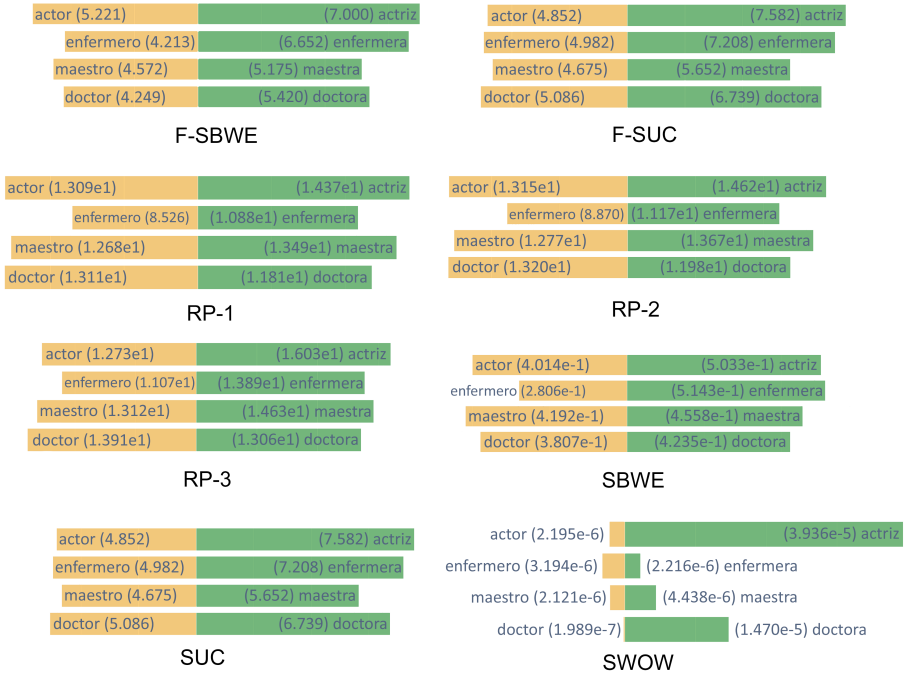

Figura 5.1: Resultados obtenidos para los *embeddings* F-SBWE, F-SUC, RP-1, RP-2, RP-3, SBWE, SUC y SWOW. En verde se muestra la distancia obtenida entre la forma femenina de las profesiones y el vector que representa el subespacio **femenino**. En naranja, se muestra la distancia entre la forma masculina y el vector del subespacio **masculino**. Se puede observar que la mayoría de los modelos devuelven resultados similares, sin tener una inclinación particular hacia ninguno de los extremos del género. En este punto se diferencia el modelo SWOW, en el cual se destaca una inclinación hacia el extremo femenino, en especial para las profesiones actriz y doctora.

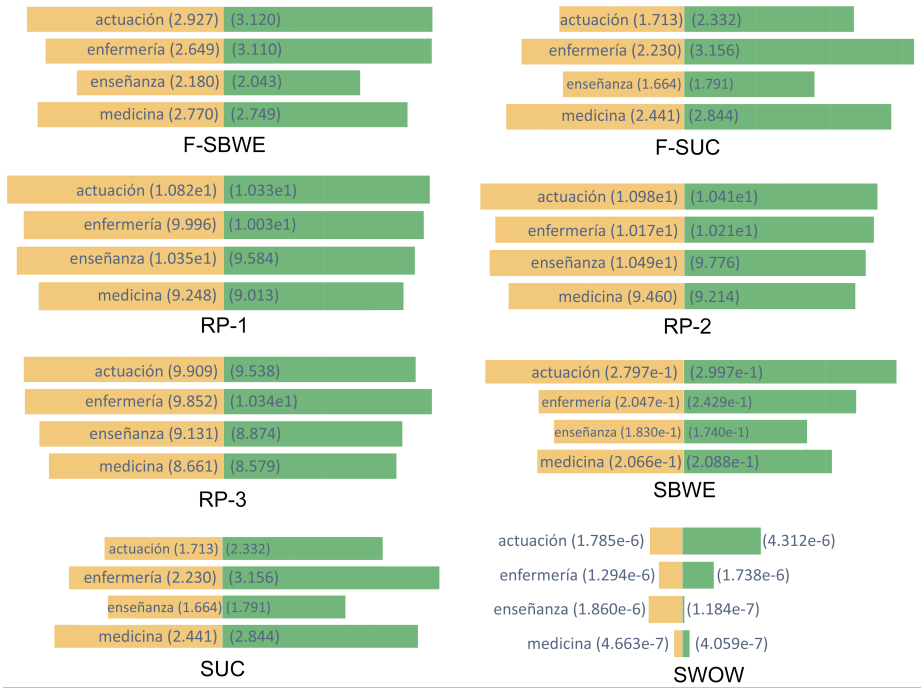

Figura 5.2: Resultados obtenidos para los *embeddings* F-SBWE, F-SUC, RP-1, RP-2, RP-3, SBWE, SUC y SWOW. En verde se muestra la distancia obtenida entre la profesión y el vector que representa el subespacio **femenino**. En naranja, se muestra la distancia entre la profesión y el vector del subespacio **masculino**. La diferencia más destacable en comparación con el análisis que utiliza las formas femenina y masculina se puede ver en el SWOW, donde "enfermería", que antes se inclinaba al masculino pasa a ser marcadamente femenina y "medicina" que deja de ser marcadamente femenina.

tipo de asociación puede no deberse necesariamente a que la profesión enseñanza se asocie más al género masculino, sino a que la enseñanza y educación están históricamente más asociadas al género masculino, quienes tenían acceso a la educación antes de que a las mujeres se les diera acceso a esta.

En la gráfica anterior (ver Figura 5.1) habíamos observado que para el SWOW se generaba un fenómeno "anti-estereotípico" para los pares enfermero-enfermera y doctor-doctora, ahora nos encontramos con que "enfermería" pasó a ser más femenina y "medicina" más masculina, encajando dentro de los estereotipos clásicos de sesgo.

Además, la palabra "medicina" es más masculina y "enfermería" más femenina en todos los modelos rioplatenses (RP-1, RP-2 y RP-3) y en el F-SBWE, esto tiene sentido porque el *fine-tuning* utiliza el mismo corpus para ajustarse que el corpus utilizado para entrenar a los modelos rioplatenses. Si observamos los resultados obtenidos en estos 4 modelos podemos ver que esta tendencia se mantiene para la mayoría de las asociaciones. Esto nos indica que el ajuste realizado al modelo SBWE, que no presenta las mismas asociaciones que estos modelos, logró incorporar características del corpus rioplatense.

Por otro lado, en los modelos SUC y F-SUC vemos que las 4 profesiones están más asociadas al subespacio **femenino** que al **masculino**, siendo consistentes con la gráfica anterior (ver Figura 5.1).

5.2.3. Categorías de visibilidad y polaridad

En el estudio realizado por (Garrido-Muñoz y cols., 2022) para analizar modelos de lenguaje, proponen utilizar categorías de visibilidad y polaridad como atributos para medir distancias a subespacios de palabras, habiendo demostrado tener resultados relevantes cuando queremos realizar un análisis de sesgo bajo estos subespacios. La categorización propuesta consiste en formar dos categorías de adjetivos, una para atributos visibles (e.g.: "alto/a", "flaco/a") y otra para atributos invisibles (e.g.: "inteligente", "inseguro/a"). A su vez estas dos categorías se subdividen en los polos positivo y negativo, según si el atributo usualmente es utilizado con connotación positiva o negativa. Por ejemplo, para los atributos visibles "lindo/a" se utiliza con connotación positiva y "feo/a" con connotación negativa. Siguiendo esta categorización se forman un total de cuatro conjuntos.

Para construir los conjuntos (Visible+, Visible-, Invisible+, Invisible-) fuimos seleccionando atributos visibles o invisibles, formando pares de palabras opuestas o antónimos entre sí para representar la polaridad. En este proceso generamos varios duplicados que después eliminamos. Llegamos a un total de 37 palabras en el conjunto Visible+, 34 en el conjunto Visible-, 105 en el conjunto Invisible+ y 85 en el conjunto Invisible-.

La mayoría de los atributos a analizar son palabras con género gramatical. En la prueba en el espacio del género nos interesaba calcular las distancias a los subespacios descomponiendo los atributos en sus formas femenina y masculina. Ahora, queremos conservar el concepto de cada atributo sin la carga del género, es decir, utilizar una única palabra que represente el concepto "lindo/a" en

lugar de utilizar “lindo” y “linda”. En el primer caso, la palabra no tiene género gramatical, por lo que utilizamos la similitud coseno promedio análogo a la prueba anterior; en el segundo caso, cuando la palabra tiene género gramatical, utilizamos una alternativa para la métrica de distancia basada en (Zhou y cols., 2019). Para calcular esta distancia primero descomponemos la palabra en sus formas femenina y masculina; luego, calculamos la similitud coseno promedio de cada forma de género a los subespacios de análisis y por último, computamos la distancia como la diferencia en valor absoluto entre estos dos vectores.

Género

En la figura 5.3 se muestran los resultados obtenidos de la prueba realizada para el espacio de género definidos por los subespacios **femenino** y **masculino**.

Lo primero que podemos observar es que casi todos los modelos presentan valores muy similares para el subespacio **femenino**, con tendencia hacia los atributos invisibles. En particular, para los modelos RP-1 y RP-2 observamos una inclinación mayor del subespacio **masculino** por los invisibles negativos. Por otro lado, para el SWOW vemos que el subespacio **femenino** está muy inclinado a atributos visibles positivos y para el masculino está muy poco representado por lo que no podemos realizar ningún análisis.

En el SWOW para el subespacio de género seleccionado, las asociaciones a los atributos de esta prueba fueron muy pequeñas, con valores cercanos al 0. Además, la representación del subespacio **masculino** en comparación con el femenino, en factor micro, es prácticamente nula. Este resultado se condice con los resultados que obtuvimos para los atributos de profesiones, en los cuales desglosamos las profesiones en sus formas masculina y femenina, y pudimos observar mayores valores de asociación hacia el subespacio **femenino**.

En los modelos F-SBWE, SBWE, F-SUC, SUC, RP-3 y SWOW, el subespacio **femenino** toma valores más altos de asociación en todos los polos que el subespacio **masculino**, esto se puede ver claramente ya que la figura verde sobresale de la figura rosada en las gráficas. Pensamos que podría deberse a que haya mayor representación del subespacio **femenino** que del **masculino**.

Una particularidad que pudimos identificar, es que si bien se suelen asociar atributos positivos visibles con lo **femenino** (Garrido-Muñoz y cols., 2022), en los resultados de los modelos existe una inclinación hacia atributos invisibles tanto positivos como negativos (con la excepción del SWOW).

Por otro lado, al explorar el subespacio **masculino**, es común, en la sociedad, vincularlo con atributos visibles negativos; sin embargo, observamos que los resultados de los modelos tienden a enfocarse en atributos invisibles.

En el resto de los modelos (RP-1 y RP-2), los valores e inclinaciones son muy similares en cuanto a la representación de cada subespacio. Comparten las mismas particularidades que el resto de los modelos de ser “anti-estereotípicos”.

Por otro lado, el SWOW presenta resultados interesantes, teniendo una clara y fuerte inclinación por atributos visibles positivos para el subespacio **femenino**. Esto podría indicarnos que este modelo contiene sesgo de género, aunque resulta extraño dado que el SWOW fue mayoritariamente anotado por mujeres (82 %).

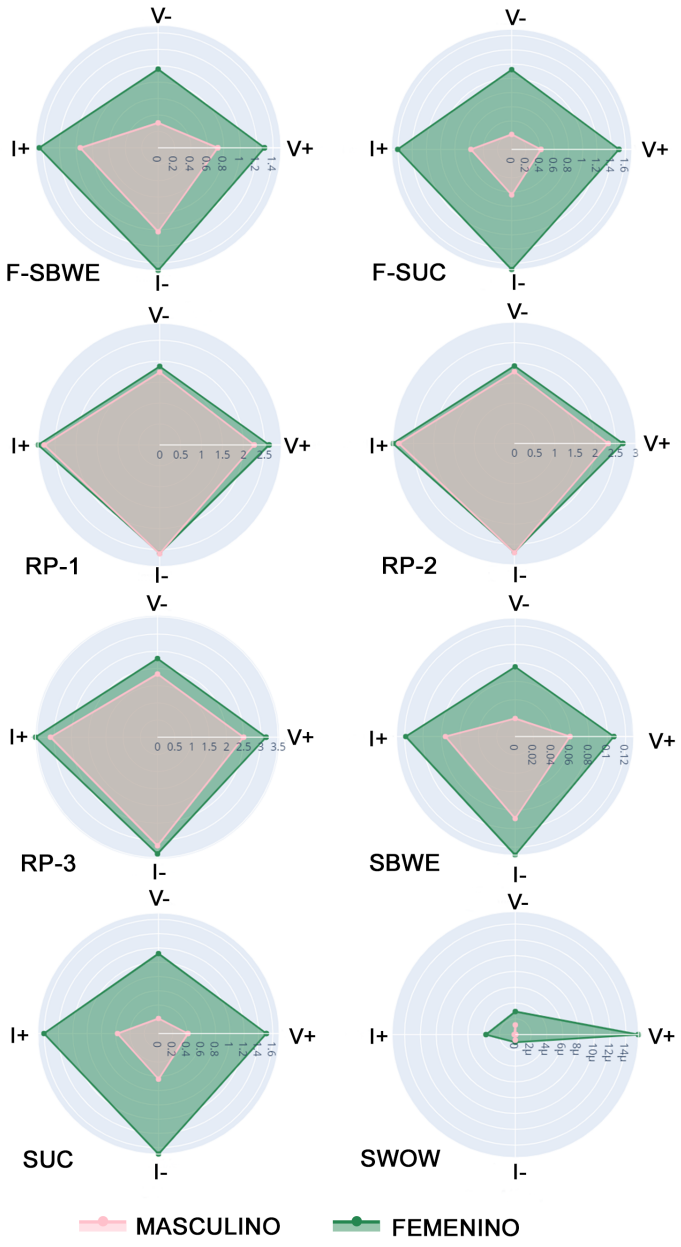

Figura 5.3: En esta figura se muestran, para cada modelo analizado, los resultados de comparar los subespacios **masculino** y **femenino** contra los conjuntos de atributos visible+, visible-, invisible+ e invisible-. En cada gráfica se muestra el promedio de los valores obtenidos al calcular la distancia entre el vector que representa a cada subespacio y los vectores de cada palabra del polo.

La gráfica con los valores de distancia obtenidos para cada una de las palabras de los cuatro polos se encuentra en el anexo A.5.

Raza

En la figura 5.4 se muestran los resultados obtenidos de la prueba realizada para el espacio de raza definido por los subespacios **raza blanca** y **raza afrodescendiente**.

Lo primero que podemos observar es que los resultados parecen ser muy similares entre los distintos modelos, casi separándose por dos clases de equivalencia.

A primera vista, tanto para el subespacio de **raza afrodescendiente** como de **raza blanca** la mayoría de los modelos presentan asociaciones muy similares hacia todos los polos. Si bien las asociaciones difieren en magnitud entre los subespacios, dentro de cada subespacio no se marca una desviación muy importante hacia ningún conjunto de atributos. Estos casos podrían significar que los modelos no presentan un sesgo marcado en ninguno de los subespacios, por lo que no podemos decir que un modelo asocia a los atributos de una raza más hacia un polo de atributos ya que en magnitud la asociación hacia el polo opuesto es muy similar. Tampoco presentan mayores diferencias comparando uno con el otro, por más que la asociación de la **raza blanca** es mayor para atributos invisibles negativos que la asociación de **raza afrodescendiente**, esto se cumple de igual manera para los atributos invisibles positivos. Este tipo de resultado se puede dar porque probablemente exista mayor representatividad de los atributos de la **raza blanca** dentro de los datos y se puede ver claramente ya que la figura verde sobresale de la figura rosada en las gráficas. Como caso particular el SWOW es al revés, insinuando que este modelo tiene una mayor representación de los atributos de la **raza afrodescendiente** que de la **raza blanca**.

En la **raza blanca**, por un lado, tenemos a los modelos F-SBWE, F-SUC, SBWE y SUC con una inclinación más marcada hacia los atributos invisibles negativos, y por otro lado a RP-1, RP-2 y RP-3, con una inclinación más marcada hacia los atributos invisibles positivos. Luego, para la **raza afrodescendiente**, tenemos al F-SUC y SUC con inclinación a atributos visibles positivos, a RP-1, RP-2 y RP-3 con inclinación a atributos invisibles positivos y a SBWE y F-SBWE con inclinación a invisibles negativos.

El modelo SWOW para la **raza blanca** no podemos analizarlo por falta de representación y presenta una muy fuerte inclinación de **raza afrodescendiente** a atributos visibles positivos.

La gráfica con los valores de distancia obtenidos para cada una de las palabras de los cuatro polos se encuentra en el anexo A.6.

Representación de colonizadores y colonizados

Por último, decidimos incluir una prueba sobre un espacio no tan usual como el del género y de la raza, donde también podría haber sesgo. Generamos dos

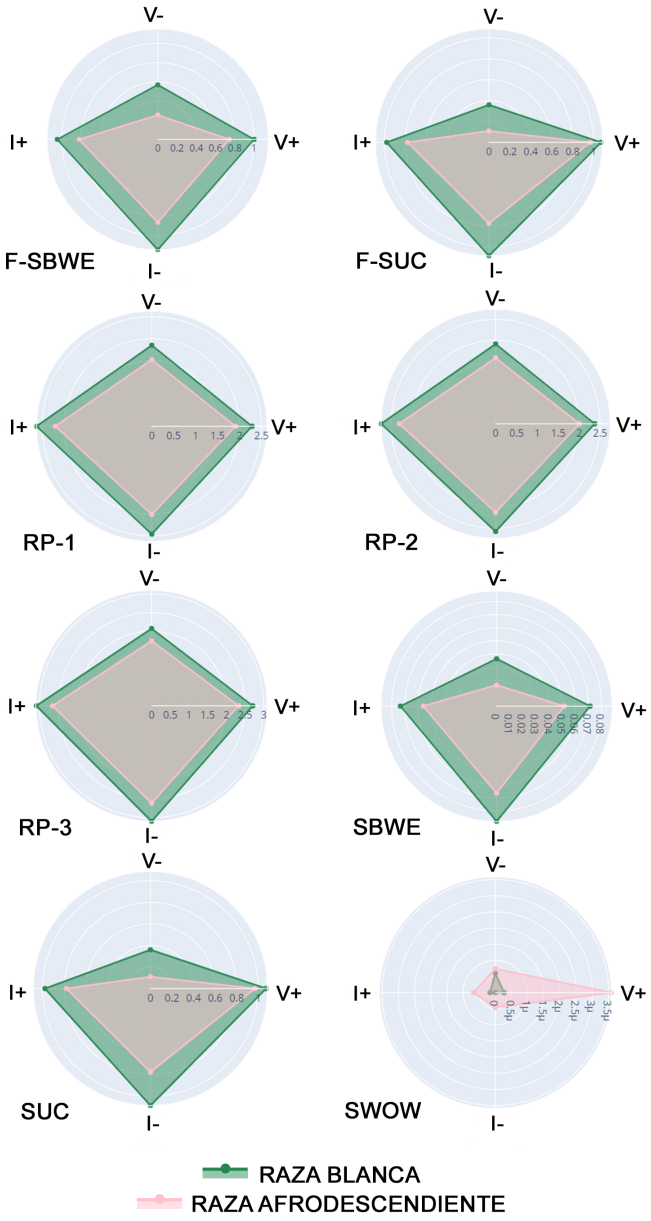

Figura 5.4: En esta figura se muestran, para cada modelo analizado, los resultados de comparar los subespacios **raza blanca** y **raza afrodescendiente** contra los conjuntos de atributos visible+, visible-, invisible+ e invisible-. En cada gráfica se muestra el promedio de los valores obtenidos al calcular la distancia entre el vector que representa a cada subespacio y los vectores de cada palabra del polo.

conjuntos de palabras que representan los subespacios **colonizado** y **colonizador**.

En el modelo SWOW la mayoría de las palabras que consideramos clave para la representación de estos dos subespacios no se encuentran presentes. Por lo que no podemos asegurar que los resultados de la prueba sobre este conjunto representen de forma precisa los sesgos asociados a estos grupos.

De igual forma, es importante destacar que, aunque los términos explícitos que representan a los subespacios de estudio no se encuentren en el modelo, sigue siendo posible que existan sesgos relacionados con los grupos colonizador/colonizado. La ausencia de los términos no implica ausencia de sesgo, sino que el sesgo puede estar menos definido o integrado contextualmente a los *embeddings*.

Pensamos en buscar atributos o palabras asociadas que pudieran capturar indirectamente los conceptos de estos subespacios, para suplir la falta de definiciones específicas. Consideramos que en este caso no habían otras palabras que cumplan con esto, dado que estos conceptos en particular se abarcan en unas pocas palabras clave. Fuera de estos, la mayoría de las palabras que encontramos (e.g.: utilizar como palabras asociadas a "colonizador" las palabras "extranjero" u "ocupa") tienen significados en español que pueden afectar significativamente la naturaleza del subespacio.

En la figura 5.5 se muestran los resultados obtenidos para esta prueba. Lo primero que podemos observar es que la mayoría de los modelos presentan una inclinación hacia los atributos visibles positivos para ambos subespacios. El modelo SWOW es el que difiere del resto en este punto, ya que solo cumple esta tendencia para el subespacio **colonizador**, pero en el caso del subespacio **colonizado** las asociaciones más grandes se dan hacia el polo de atributos visibles negativos. Esto puede ser un indicio de que este modelo se encuentra más sesgado que el resto en estos subespacios.

En los modelos RP-1, RP-2 y RP-3, el subespacio **colonizado** toma valores más altos de asociación en todos los polos que el subespacio **colonizador**, esto se puede ver claramente ya que la figura verde sobresale de la figura rosada en las gráficas. En el resto de los modelos esto es al revés, pensamos que podría deberse a que haya mayor representación del subespacio **colonizado** en los modelos rioplatenses que en el resto de los modelos.

Se puede ver que el promedio de asociación en cada polo tenía valores muy bajos en el modelo SBWE (por debajo del 0.3), al observar F-SBWE vemos que estos valores aumentan pero mantienen la misma tendencia con mayor magnitud. En el caso de SUC y F-SUC no se perciben cambios. Esto sugiere que el *fine-tuning* apenas afecta la asociación promedio de palabras en cada polo en comparación con los modelos no ajustados.

Durante la colonización, los colonizadores utilizaron la lengua como herramienta de conquista, perpetuando en el texto su visión de los nativos colonizados y de sí mismos. Representando la relación entre ambos como una relación de poder en la que el polo colonizador es el dominante (Shakib, 2011). Cuando comparamos los resultados obtenidos, observamos que todos los modelos tienen un comportamiento que se adecúa a esto, mostrando al subespacio **coloniza-**

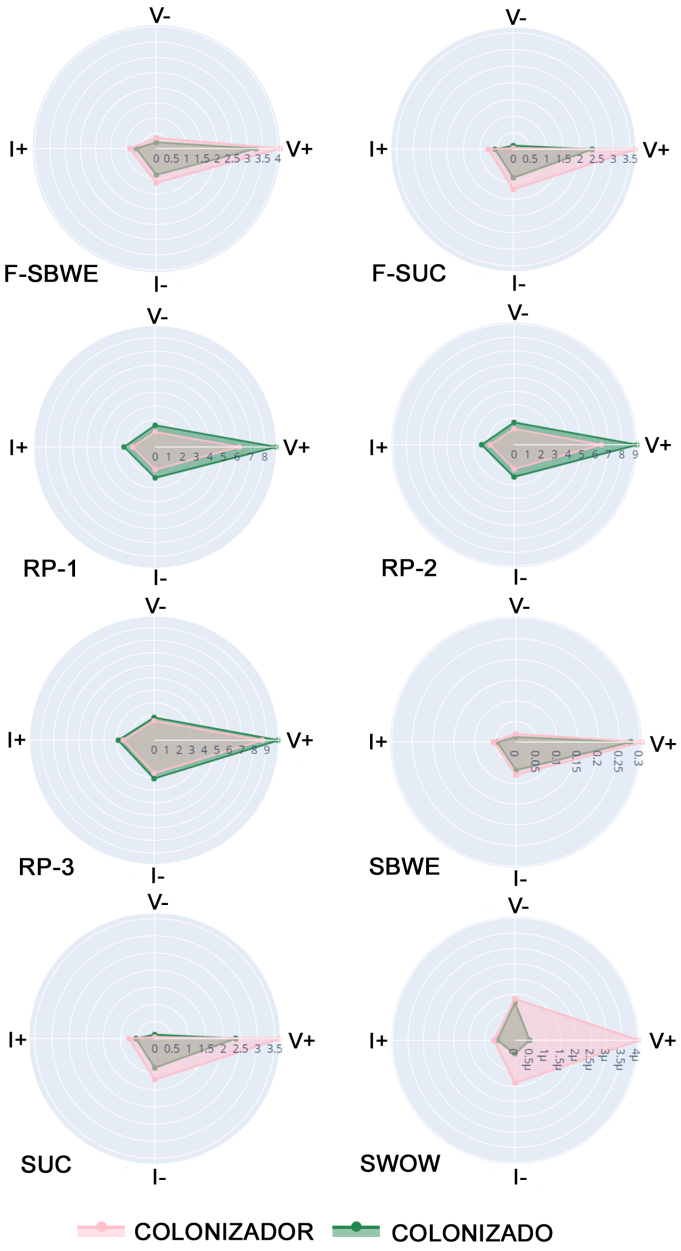

Figura 5.5: En esta figura se muestran, para cada modelo analizado, los resultados de comparar los subespacios **colonizador** y **colonizado** contra los conjuntos de atributos visible+, visible-, invisible+ e invisible-. En cada gráfica se muestra el promedio de los valores obtenidos al calcular la distancia entre el vector que representa a cada subespacio y los vectores de cada palabra del polo.

dor con mayor inclinación hacia atributos positivos visibles y en general mayor asociación a ese polo que el subespacio **colonizado**.

Por otro lado, todos los modelos menos el SWOW, en el subespacio **colonizado**, también se inclinan a atributos visibles positivos. Esto no condice con los estereotipos que suelen haber frente a este subespacio. Para el SWOW observamos que tiene una inclinación hacia atributos visibles negativos lo cual se alinea con la idea de que los colonizadores están representados en una posición de dominancia.

La gráfica con los valores de distancia obtenidos para cada una de las palabras de los cuatro polos se encuentra en el anexo A.7.

5.2.4. Proyección semántica

Para realizar esta prueba utilizamos las 4 profesiones {"actor/actriz", "doctor/doctora", "enfermero/enfermera", "maestro/maestra"} seleccionadas para las pruebas de distancia a palabras.

En la figura 5.6 podemos ver que los modelos entrenados con `fasttext` y `word2vec` (SUC, F-SUC, SBWE, F-SBWE, RP-1, RP-2, RP-3) representan hacia un lado del espacio todas las palabras en su forma femenina y hacia el otro las palabras en su forma masculina. Al observar la proyección de cada punto sobre la recta que representa el género masculino-femenino esto resulta más claro de ver; podríamos definir una división en la recta que cumpla que todos los puntos femeninos se encuentren en una de las semirectas resultantes y todos los puntos masculinos en la otra. El SWOW es el único modelo que no tiene este comportamiento, en su lugar "ordena" las proyecciones de los vectores masculinos y femeninos a lo largo de la recta, manteniendo las formas masculina y femenina de cada profesión contiguas.

En los modelos SBWE y F-SBWE podemos observar que todas las profesiones femeninas están más cerca del punto que representa al subespacio **femenino** y todas las profesiones masculinas se encuentran más cerca del subespacio **masculino**. Luego del *fine-tuning* (comparando SBWE contra F-SBWE) vemos que "actor" se desplaza hacia el subespacio **femenino**. También observamos que la profesión más femenina es "actriz" y la más masculina es "maestro". Además, vemos que la profesión "maestra" es lo más masculino dentro de lo femenino y "enfermero" lo más femenino dentro de lo masculino para el caso de SBWE y luego del *fine-tuning*, "actor" pasa a ser la profesión más femenina dentro de las formas masculinas.

En los modelos SUC y F-SUC observamos algo muy similar al F-SBWE. Se mantiene la separación entre las profesiones junto a sus polos y "actriz" y "maestro" como profesión más femenina y masculina respectivamente. En estos modelos, la profesión "enfermero" se acercó más al polo femenino que al masculino.

Los resultados del modelo RP-1 muestran resultados un poco diferentes. Se sigue manteniendo la separación entre polos, pero "enfermero" y "actor" están más cerca del polo femenino. También observamos que la palabra más masculina es "doctor" y la más femenina "actriz".

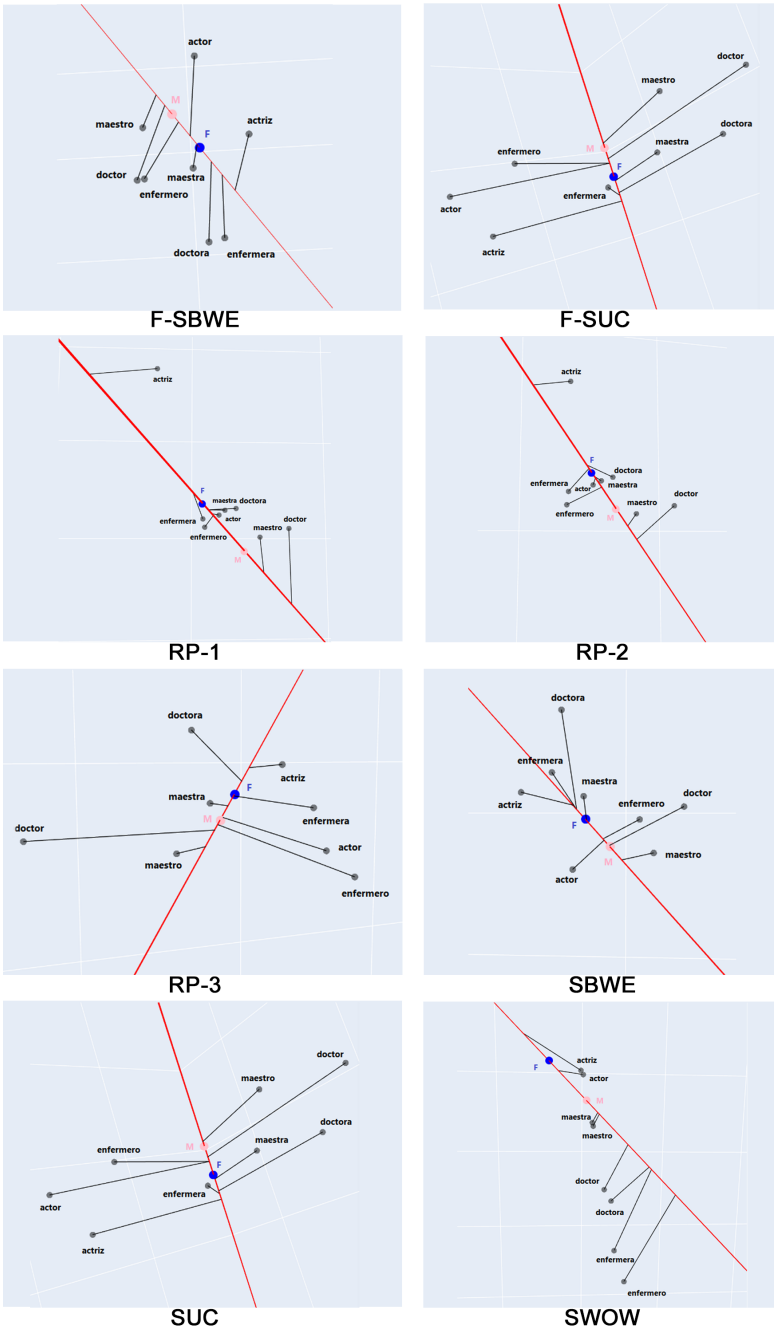

Figura 5.6: Resultados de realizar proyección semántica comparando los subespacios **femenino** y **masculino** contra un conjunto reducido de profesiones.

El modelo RP-2 presenta resultados similares al del modelo RP-1 salvo por una pequeña diferencia. La profesión "doctora" se mantiene más cerca del polo femenino pero se encuentra más lejos del polo masculino comparado con el modelo RP-1, es decir, en este modelo la profesión "doctora" es aún más femenina que en el modelo anterior.

En el modelo RP-3 observamos que una vez más las profesiones femeninas están más cerca del polo femenino y las masculinas más cerca del polo masculino. Vemos que los datos presentan las mismas particularidades que los modelos SBWE y F-SBWE frente a la palabra más femenina "actriz", la más masculina "maestro", la más femenina dentro de las masculinas "actor" y la más masculina dentro de las femeninas "maestra".

En el modelo SWOW observamos, como modelos anteriores, que la profesión "actriz" es la más femenina, "enfermero" la más masculina, "actor" la más femenina dentro de las masculinas y "enfermera" la más masculina dentro de las femeninas. Además, en la gráfica de proyecciones de este modelo podemos ver que la mayoría de las profesiones a lo largo de la recta cumplen que, si tomamos las proyecciones de su forma femenina y masculina, la forma femenina se encuentra más hacia el extremo femenino que la masculina. Exceptuando a la profesión "doctor/doctora", en este caso la palabra "doctor" se encuentra más hacia el extremo femenino que la palabra "doctora"; sin embargo, si observamos también los puntos que representan "femenino" y "masculino", podemos ver que la distancia entre "femenino" y "doctora" es mayor que la distancia entre "masculino" y "doctor".

5.2.5. Observación general

En la mayoría de las pruebas tuvimos que limitar el vocabulario para mantener una comparación justa entre los modelos, en particular, la representación femenina en el vocabulario de SWOW es muy pobre, no encontrándose palabras como "ingeniera" o "arquitecta". En este ejemplo el sesgo es la invisibilización de estas profesiones femeninas, pero se podría dar el otro extremo en el que sí tenemos la representación femenina pero esta es mala o injusta. Este tipo de resultado muchas veces no se puede ver en un análisis cuantitativo, opinamos que debería de tenerse en cuenta la representación de ciertos grupos en el vocabulario de los *embeddings*, más allá de las métricas obtenidas, al momento de juzgar un modelo para una aplicación dada.

Capítulo 6

Conclusiones y trabajo futuro

Para la realización de este proyecto nos planteamos analizar el sesgo en representaciones vectoriales de palabras en español, haciendo foco a sesgos regionales que se pudieran dar en el Río de la Plata. Dado que no encontramos recursos más allá de los del proyecto SWOW (Cabana y cols., 2023) disponibles en el dialecto rioplatense, como primera parte, creamos diferentes corpus rioplatenses a partir de corpus de Argentina y Uruguay, difiriendo en su preprocesamiento. Luego los utilizamos para la creación de nuevos modelos de *word embeddings* rioplatenses y para el *fine-tuning* de *word embeddings* ya existentes del idioma español.

Como segunda parte del proyecto creamos pruebas de referencia de calidad, que luego utilizamos para medir la corrección de los modelos creados y también compararlos con otros modelos de *word embeddings* en español, dentro de los cuales destacamos el SWOW, un corpus rioplatense creado utilizando el principio psicológico de la asociación libre.

Por último, nos propusimos realizar un análisis de sesgo para todos los modelos de *word embeddings*. Con este fin generamos pruebas específicas para abordar el estudio de sesgo en representaciones vectoriales de palabras en español, adaptando pruebas conocidas ya existentes del idioma inglés.

Como parte de esto, tuvimos que definir los subespacios de estudio bajo los cuales realizar los análisis. Para esto no encontramos conjuntos de definición para el español, por lo que partimos de las traducciones del inglés de listados de palabras de trabajos previos y los agrandamos. Además, generamos otros desde cero.

Por otra parte, desarrollamos *notebooks* individuales que encapsulan cada una de las pruebas para detectar y cuantificar el sesgo. Estas pruebas incluyen evaluaciones tanto gráficas como numéricas, permitiendo una comprensión más profunda y visualmente clara de los patrones de sesgo presentes en los datos, y podrían ser utilizadas para trabajos futuros de forma sencilla.

En particular, las pruebas de distancia a palabras para subespacios de análisis de sesgo, siguen los pasos del proyecto `Responsibly`[1] ya que utilizan métricas similares, pero no pudimos hacer uso de este dado que no se está manteniendo y no funciona con versiones nuevas de `python`.

Realizamos un *data statement* siguiendo la guía de (Bender y Friedman, 2018) a los *word embeddings* del proyecto del SWOW en español rioplatense para que las personas que lo utilicen a futuro sean conscientes de las limitaciones que presenta. Realizamos el *data statement* en español.

Uno de los objetivos principales que nos planteamos fue el análisis de sesgos inherentes al dialecto y cultura presentes en la región del Río de la Plata. Creemos que este objetivo se pudo cumplir de forma parcial, en primer lugar, porque los modelos estudiados que fueron creados a partir de datos recolectados en la región presentaban características muy diferentes, por motivos totalmente ajenos a la región de recolección de los datos (SWOW); en segundo lugar, por la calidad alcanzada de los modelos creados para este trabajo (RP-1, RP-2 y RP-3), estos dos factores dificultaron la obtención de resultados claros como para poder formular una definición precisa en el análisis.

Destacamos que, cuando llevamos a cabo el proceso de *fine-tuning*, este permitió que los modelos de lenguaje en español reconocieran palabras específicas del dialecto rioplatense, como la palabra uruguaya "champión". Este aspecto es relevante, porque nos muestra que los modelos pueden adaptarse para incorporar particularidades lingüísticas regionales.

A pesar de esto, aunque creemos haber generado los recursos necesarios para el análisis de sesgos regionales del Río de la Plata, logrando adaptar los modelos al dialecto rioplatense, nuestro análisis de sesgo no arrojó resultados significativos ni reveló sesgos de relevancia en relación con el dialecto y la cultura de la región.

Planteamos los siguientes ítems como trabajo a futuro:

- Analizar si la eliminación de las *stop words* afectó la calidad de los *embeddings* que creamos o en el *fine-tuning* que les realizamos.
- Definición de diferentes espacios de estudio para abarcar en mayor horizontalidad el sesgo en los modelos.
- Poder utilizar estas pruebas para que personas idóneas en el área del lenguaje natural puedan realizar análisis de modelos, y que puedan identificar patrones, tendencias o relaciones significativas en los resultados.
- Automatizar las diferentes pruebas creadas para que personas sin conocimiento técnico puedan proporcionar un modelo de *word embeddings* y analizar los resultados devueltos sin necesidad de preocuparse por la implementación. Este punto se podría ampliar también a incluir en la automatización la creación de embeddings, pasando a ser necesario proporcionar únicamente el corpus de entrada.

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

Referencias

Angwin, J., Larson, J., Mattu, S., y Kirchner, L. (2019). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. *URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*.

Barocas, S., Hardt, M., y Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org. (http://www.fairmlbook.org)

Bender, E. M., y Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. Descargado de https://aclanthology.org/Q18-1041 doi: 10.1162/tacl_a_00041

Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., y Kalai, A. (2016b). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., y Kalai, A. T. (2016a). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., y Zemel, R. (2019). Understanding the origins of bias in word embeddings. En *International conference on machine learning* (pp. 803–811).

Cabana, Á., Zugarramurdi, C., Valle-Lisboa, J. C., y De Deyne, S. (2023). The"small world of words"free association norms for rioplatense spanish. *Behavior Research Methods*, 1–18.

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

Camacho Collados, J., Pilehvar, M. T., y Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets..

Cardellino, C. (2019, August). *Spanish Billion Words Corpus and Embeddings.* Descargado de https://crscardellino.github.io/SBWCE/

Cañete, J. (2019, mayo). *Compilation of large spanish unannotated corpora.* Zenodo. Descargado de https://doi.org/10.5281/zenodo.3247731 doi: 10.5281/zenodo.3247731

Dastin, J. (2022). Amazon scraps secret ai recruiting tool that showed bias against women. En *Ethics of data and analytics* (pp. 296–299). Auerbach Publications.

De Deyne, S., Cabana, Á., Li, B., Cai, Q., y McKague, M. (2020). A cross-linguistic study into the contribution of affective connotation in the lexico-semantic representation of concrete and abstract concepts. En *Cogsci.*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., y Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eticas. (2023). *Auditing social media: Portrayal of migrants on youtube.* Descargado de https://eticas.tech/wp-content/uploads/2023/04/ETICAS-_-Auditing-Social-Media_-Portrayal-of-Migrants-on-Youtube.pdf

Friedman, B., y Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Garrido-Muñoz, I., Martínez-Santiago, F., y Montejo-Ráez, A. (2022). Maria and beto are sexist: evaluating gender bias in large language models for spanish.

Grand, G., Blank, I. A., Pereira, F., y Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, *6*(7), 975–987.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., y Mikolov, T. (2018). Learning word vectors for 157 languages. En *Proceedings of the international conference on language resources and evaluation (lrec 2018).*

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Imana, B., Korolova, A., y Heidemann, J. (2021). Auditing for discrimination in algorithms delivering job ads. En *Proceedings of the web conference 2021* (pp. 3767–3778).

Joulin, A., Grave, E., Bojanowski, P., y Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jurafsky, D., y Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd edition.* Prentice Hall, Pearson Education International.

Khurana, D., Koli, A., Khatter, K., y Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), 3713–3744.

Lambrecht, A., y Tucker, C. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, *65*(7), 2966–2981.

Lee, D.-C., Liang, H., y Shi, L. (2021). The convergence of racial and income disparities in health insurance coverage in the united states. *International journal for equity in health*, *20*(1), 1–8.

Mac, R. (2021). Facebook apologizes after ai puts 'primates' label on video of black men. *The New York Times*, *3*(9), 2021. Descargado de https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., y Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mikolov, T., Yih, W.-t., y Zweig, G. (2013). Linguistic regularities in continuous space word representations. En *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Osgood, C. E., Suci, G. J., y Tannenbaum, P. H. (1957). *The measurement of meaning* (n.º 47). University of Illinois press.

Pennington, J., Socher, R., y Manning, C. D. (2014). Glove: Global vectors for word representation. En *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., y Zettlemoyer, L. (2018). *Deep contextualized word representations.*

Prates, M. O., Avelar, P. H., y Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*, 6363–6381.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., y cols. (2018). Improving language understanding by generative pre-training.

Rubenstein, H., y Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Shakib, M. K. (2011). The position of language in development of colonization. *Journal of Languages and Culture*, *2*(7), 117–123.

Steyvers, M., Shiffrin, R. M., y Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory.

Telford, T. (2019). Apple card algorithm sparks gender bias allegations against goldman sachs. *Washington Post*, *11*.

UNESCO. (2022). Recomendación sobre la ética de la inteligencia artificial. *SHS-2021/SANS COTE*, 43. Descargado de https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., ... others (2020). Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, *46*(4), 847–897.

Williams, D. R., Lawrence, J. A., y Davis, B. A. (2019). Racism and health: evidence and needed research. *Annual review of public health*, *40*, 105–125.

Zhang, M. (2015). Google photos tags two african-americans as gorillas through facial recognition software. *Forbes*. Descargado de https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=4bc20345713d

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., y Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., y Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. *arXiv preprint arXiv:1909.02224*.

Glosario

asociación libre de palabras Tarea en la cual el sujeto a partir de una palabra en texto que sirve como estímulo responde la primera palabra que se le viene a la mente. 14

contexto Entorno en el cual tiene sentido una palabra. En este proyecto el entorno está represento por las palabras que la rodean. 6

corpus Documento o colección de documentos de texto. 6

F-SBWE *Fine-tuning* para ajustar al modelo SBWE al dominio rioplatense con el método `word2vec`. 32

F-SUC *Fine-tuning* para ajustar al modelo SUC al dominio rioplatense con el método `fasttext`. 32

n-grama Subsecuencia de n caracteres de una palabra. Los trigramas de Perro son: Per, err, rro. 11

pesos En una red neuronal, se les llama pesos de la red a los coeficientes de la función que aplicada a la salida de una capa anterior, forma la entrada de la capa siguiente de la red. 7

RP-1, RP-2 y RP-3 Modelos de *word embeddings* creados a partir del corpus rioplatense lower, este corpus está preprocesado quitando las stop words, puntuaciones y pasando todo el texto a minúscula. 25, 32

SBWE Modelo de *word embeddings* generado a partir del Spanish Billion Word Corpus con el método *skip-gram* de `word2vec`. 15

softmax Es una función matemática utilizada comúnmente en problemas de clasificación en el área de aprendizaje automático y estadística. Su función principal es convertir un vector de números en una distribución de probabilidad. La función de Softmax se define como:

$$\sigma : R^K \to [0,1]^K$$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \forall j \in 1, ..., K$$

28

stop words Palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural. Por ejemplo: el, la, etc. 24

SUC Modelo de *word embeddings* generado a partir del Spanish Unannotated Corpora con el método `fasttext`. 14

SWOW Modelo de *word embeddings* creado a partir del corpus de asociación libre de palabras del proyecto SWOW-RP (Small World of Words Rioplatense). 14

Anexo A

Anexo 1

A.1. Tamaño de ventana

Mostramos a continuación una gráfica donde observamos el tamaño de ventana óptimo a colocar como hiper-parámetro en la creación de *embeddings*. Truncamos la gráfica con tamaño máximo de oración 120 dado que las oraciones con mayor cantidad de palabras eran muy pocas y generaban ruido para observar la parte de la gráfica más interesante.

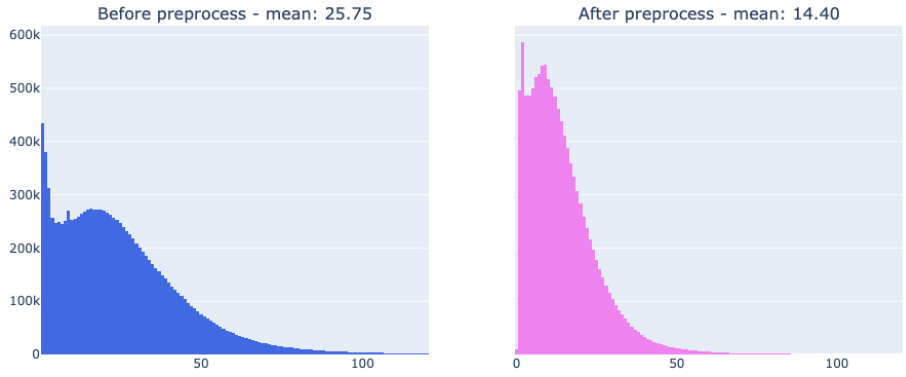

Figura A.1: Tamaño de ventana

A.2. Análisis de calidad de *embeddings*

En este anexo presentamos la comparación entre los 8 *embeddings* contra MSL, *abstract*, *concrete* y RG-65, pero a diferencia de la sección 4.2, la realizamos con todas las palabras, no solo las conocidas por todos los modelos:

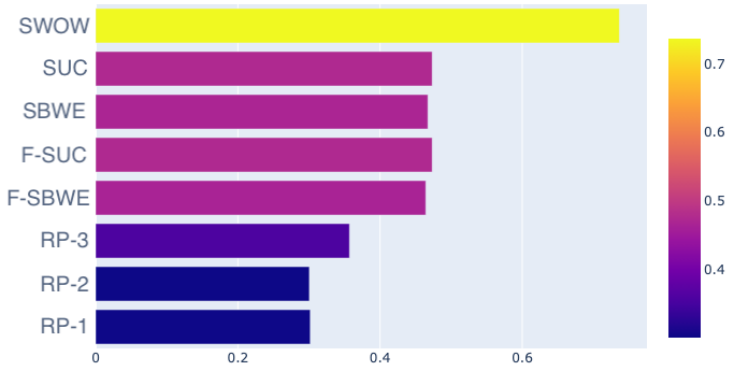

Figura A.2: Correlación de Spearman con MSL

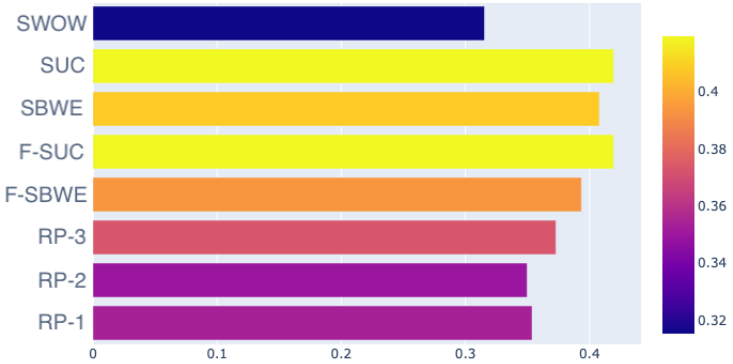

Figura A.3: Correlación de Spearman con Abstract

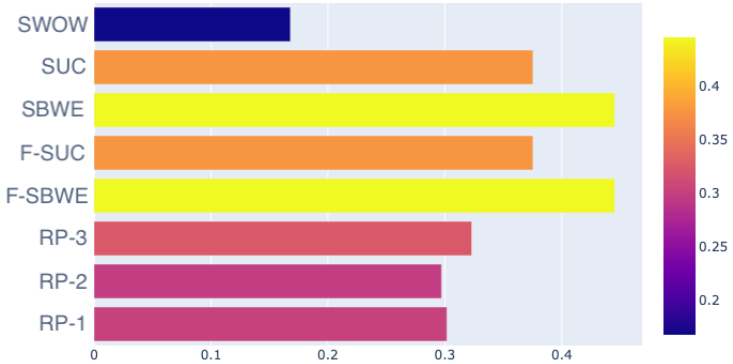

Figura A.4: Correlación de Spearman con Concrete

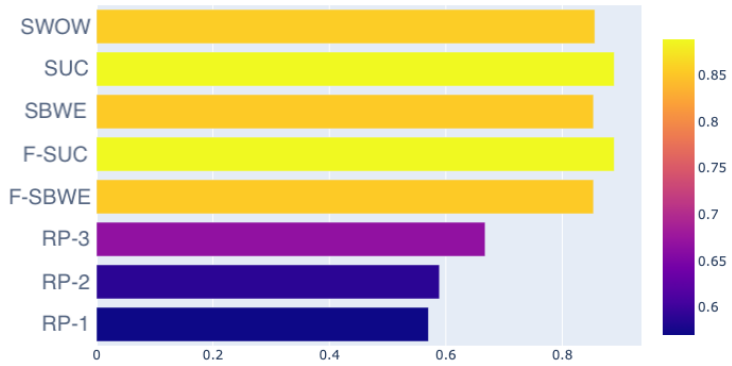

Figura A.5: Correlación de Spearman con RG-65

En primer lugar observamos que los resultados obtenidos para los corpus RP-1, RP-2 y RP-3 con MSL fueron muy malos, mientras que en los conjuntos *abstract* y *concrete* fueron bajos pero más altos que el SWOW. Por otro lado, pudimos observar que en el conjunto RG-65, si bien no tan alta como en otros *embeddings*, existe una correlación con RP-1, RP-2 y RP-3. También observamos que, aunque sea por poco, el RP-3 obtuvo mejores resultados en los 4 conjuntos de datos que RP-1 y RP-2.

Luego, pudimos observar que el SWOW obtuvo resultados de correlación excelentes con los conjuntos MSL y RG-65, superando la barrera de 0.7, indicando una correlación alta. Por otro lado, en los otros dos conjuntos el SWOW fue el que obtuvo peores resultados.

Para los modelos SUC y F-SUC, si bien en MSL, *abstract* y *concrete* la correlación no fue por arriba de 0.5 indicándonos una correlación entre modelo y datos, los resultados fueron muy cercanos. Por otro lado, la correlación con RG-65 fue excelente, fueron los *embeddings* con mayor correlación, con un valor muy cercano 0.9, indicándonos una correlación muy alta.

Por último, los resultados del SBWE y F-SBWE son similares a los del SUC y F-SUC en MSL, un poco más bajos en *abstract* y RG-65 y un poco más altos en *concrete*. Algo interesante a observar, es que la correlación de SBWE es mayor a la del F-SBWE para el conjunto *abstract*, dando resultados diferentes de lo que creíamos a priori.

A.3. Cuantificación del sesgo con MWEAT

A continuación presentamos los conjuntos de atributos de "femenino" y "masculino": los conjuntos son {"mujer", "niña", "madre", "hija", "chica", "femenino", "hermana", "tía", "María"} y {"hombre", "niño", "padre", "hijo", "chico", "masculino", "hermano", "tío", "Juan"} respectivamente.

A continuación presentamos los 60 pares de profesiones con género sintáctico utilizados para la cuantificación del sesgo: abogado/a, administrador/a, ani-

mador/a, arquitecto/a, asesor/a, astrónomo/a, autor/a, barbero/a, bibliotecario/a, camarero/a, cazador/a, científico/a, cocinero/a, conductor/a, consejero/a, conservador/a, constructor/a, coreógrafo/a, desarrollador/a, director/a, diseñador/a, editor/a, educador/a, empleado/a, encargado/a, enfermero/a, entrenador/a, escritor/a, estadístico/a, farmacéutico/a, grabador/a, herrero/a, ingeniero/a, investigador/a, jefe/a, juez/a, lector/a, locutor/a, maestro/a, matemático/a, mecánico/a, médico/a, músico/a, obrero/a, operador/a, profesor/a, promotor/a, psicólogo/a, químico/a, redactor/a, reportero/a, secretario/a, trabajador/a, técnico/a, vendedor/a, veterinario/a.

También presentamos los 18 pares de profesiones con género sintáctico que nos quedamos luego del preprocesamiento para el modelo SWOW: director/a, educador/a, empleado/a, encargado/a, enfermero/a, jefe/a, maestro/a, matemático/a, mecánico/a, médico/a, músico/a, profesor/a, psicólogo/a, químico/a, secretario/a, técnico/a, trabajador/a, veterinario/a.

A continuación mostramos la tabla comparativa entre modelos con los 18 pares de profesiones y los conjuntos de atributos de género “femenino” y “masculino”.

Modelo	MWEAT	P-valor
Fasttext	0.953753	0.0
SWOW	**0.000029**	0.014
RP-3	0.382857	0.0
SBWE	0.628755	0.0
RP-1	7.658878	0.0
RP-2	7.721414	0.0
F-SBWE	14.868805	0.022
SUC	20.726119	0.0
F-SUC	20.726119	0.0

A continuación mostramos la tabla comparativa entre modelos con los 18 pares de profesiones y los conjuntos de atributos de raza “afrodescendiente” y “blanca”.

MWEAT	Modelo	P-valor
0.0299036	Fasttext	0.852
0.0000057	**SWOW**	0.715
0.2535841	SBWE	0.123
0.7170367	RP-1	0.767
1.4424213	RP-2	0.578
1.9524711	RP-3	0.541
3.5464655	F-SBWE	0.156
4.9954253	SUC	0.0
4.9954253	F-SUC	0.0

A continuación mostramos la tabla comparativa entre modelos con los 18 pares de profesiones y los conjuntos de atributos de colonización "colonizado" y "colonizador".

MWEAT	Modelo	P-valor
0.5643881	Fasttext	0.001
0.0000035	**SWOW**	0.494
0.0354153	SBWE	0.003
1.1172842	RP-3	0.638
1.2615111	RP-2	0.67
1.3605978	RP-1	0.632
9.3656214	F-SBWE	0.001
10.7229625	SUC	0.005
10.7229625	F-SUC	0.002

A.4. Distancia a subespacios de palabras

A continuación presentamos todo el conjunto de profesiones utilizado: abogado/a, actor/actriz, administrador/a, administrativo/a, agente, agricultor/a, agrónomo/a, albañil, analista, animador/a, antropólogo/a, arqueólogo/a, arquitecto/a, artesano/a, artista, asesor/a, asistente, astronauta, astrólogo/a, astrónomo/a, atleta, auditor/a, azafato/a, bailarín/bailarina, banquero/a, barbero/a, biólogo/a, bombero/a, botánico/a, camarero/a, cantante, carpintero/a, carte-

ro/a, chef, científico/a, cirujano/a, coach, cocinero/a, compositor/a, conductor/a, conserje, consultor/a, contador/a, coreógrafo/a, cosmetólogo/a, crítico/a, decorador/a, dentista, deportista, dermatólogo/a, desarrollador/a, detective, diplomático/a, director/a, diseñador/a, docente, doctor/a, ecologista, economista, editor/a, ejecutivo/a, electricista, embajador/a, emprendedor/a, empresario/a, enfermero/a, entrenador/a, enólogo/a, escribano/a, escritor/a, escultor/a, estadístico/a, esteticista, estilista, etnógrafo/a, fabricante, farmacéutico/a, ferroviario/a, filósofo/a, florista, fontanero/a, fotoperiodista, fotógrafo/a, físico/a, genetista, gerente, geógrafo/a, geólogo/a, grabador/a, guardia, guionista, heladero/a, historiador/a, horticultor/a, ilustrador/a, impresor/a, informático/a, ingeniero/a, inspector/a, inventor/a, investigador/a, jardinero/a, jefe/a, joyero/a, juez/jueza, limpiador/a, locutor/a, luthier, líder, maestro/a, malabarista, maquillador/a, marinero/a, mariscador/a, masajista, matemático/a, mecánico/a, meteorólogo/a, microbiólogo/a, minero/a, modelador/a, modelista, modelo, modista, montador/a, musicoterapeuta, médico/a, músico/a, narrador/a, negociador/a, notario/a, nutricionista, oceanógrafo/a, oficial, oftalmólogo/a, operador/a, óptico/a, optometrista, paleontólogo/a, panadero/a, payaso/a, percusionista, periodista, pescador/a, piloto, pintor/a, poeta, policía, portero/a, presidente/a, productor/a, profesor/a, programador/a, proveedor/a, psicoanalista, psicólogo/a, psiquiatra, publicista, químico/a, radioterapeuta, radiólogo/a, realizador/a, recepcionista, redactor/a, reportero/a, repostero/a, representante, restaurador/a, sabio/a, sanitario/a, sastre/sastra, secretario/a, senador/a, serígrafo/a, sexólogo/a, sicólogo/a, sociólogo/a, soldado, soldador/a, soñador/a, subastador/a, supervisor/a, taquillero/a, taxidermista, tecnólogo/a, telefonista, tenista, terapeuta, tesorero/a, teólogo/a, topógrafo/a, torero/a, trabajador/a, traductor/a, traumatólogo/a, técnico/a, ufólogo/a, urbanista, urologista, vaquero/a, vendedor/a, ventilador/a, verificador/a, veterinario/a, viajero/a, videógrafo/a, vidriero/a, vigilante, viticultor/a, voluntario/a, webmaster, xilógrafo/a, youtuber, zapatero/a, zoofarmacéutico/a, zoólogo/a.

A.5. Atributos de visibilidad y polaridad para el género binario (femenino-masculino)

A continuación mostramos las gráficas con los resultados palabra a palabra de las pruebas realizadas utilizando atributos de visibilidad y polaridad para los subespacios de género.

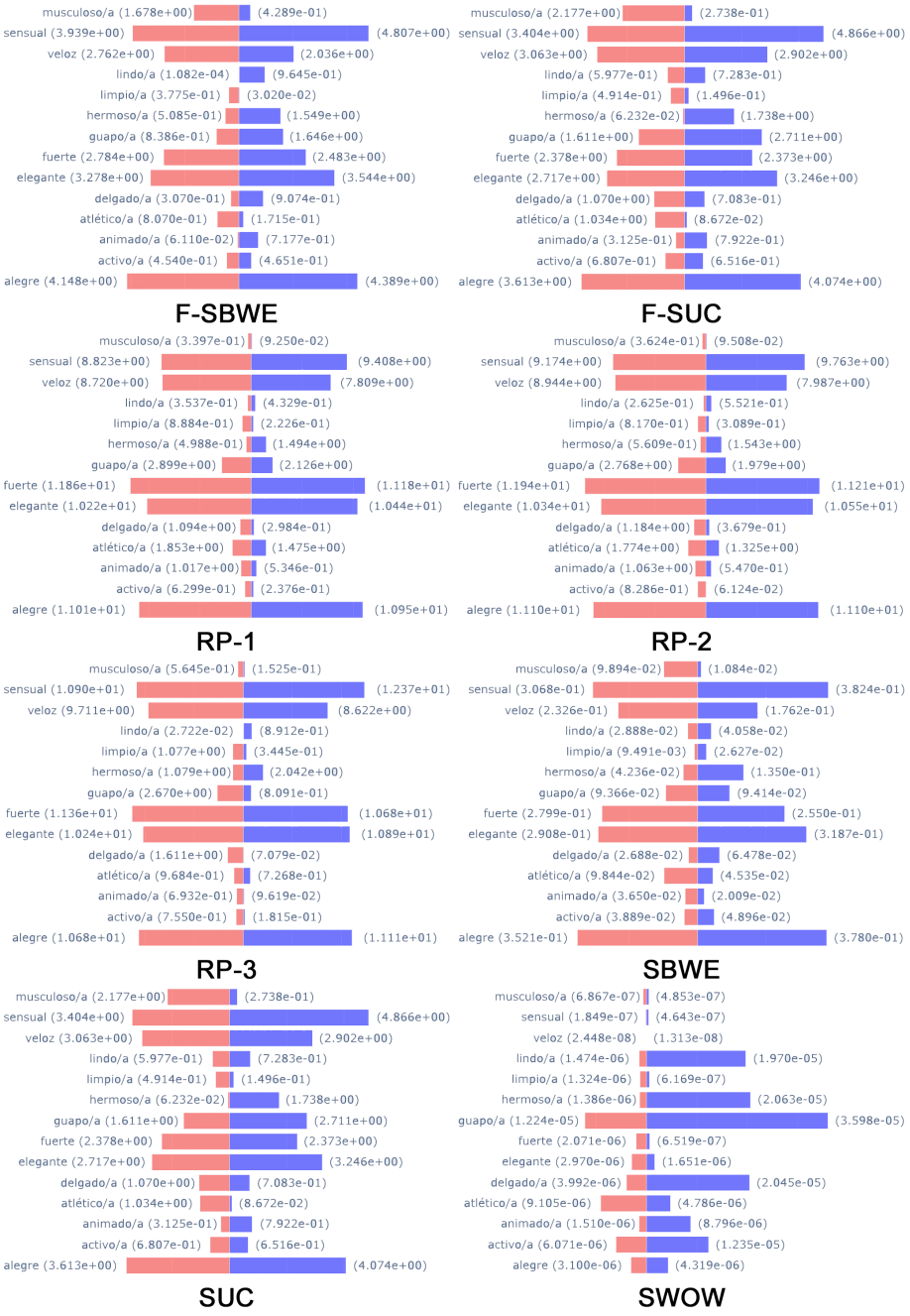

Figura A.6: Resultados de atributos visibles positivos para los diferentes modelos según los subespacios de estudio **masculino** (en rojo) y **femenino** (en azul).

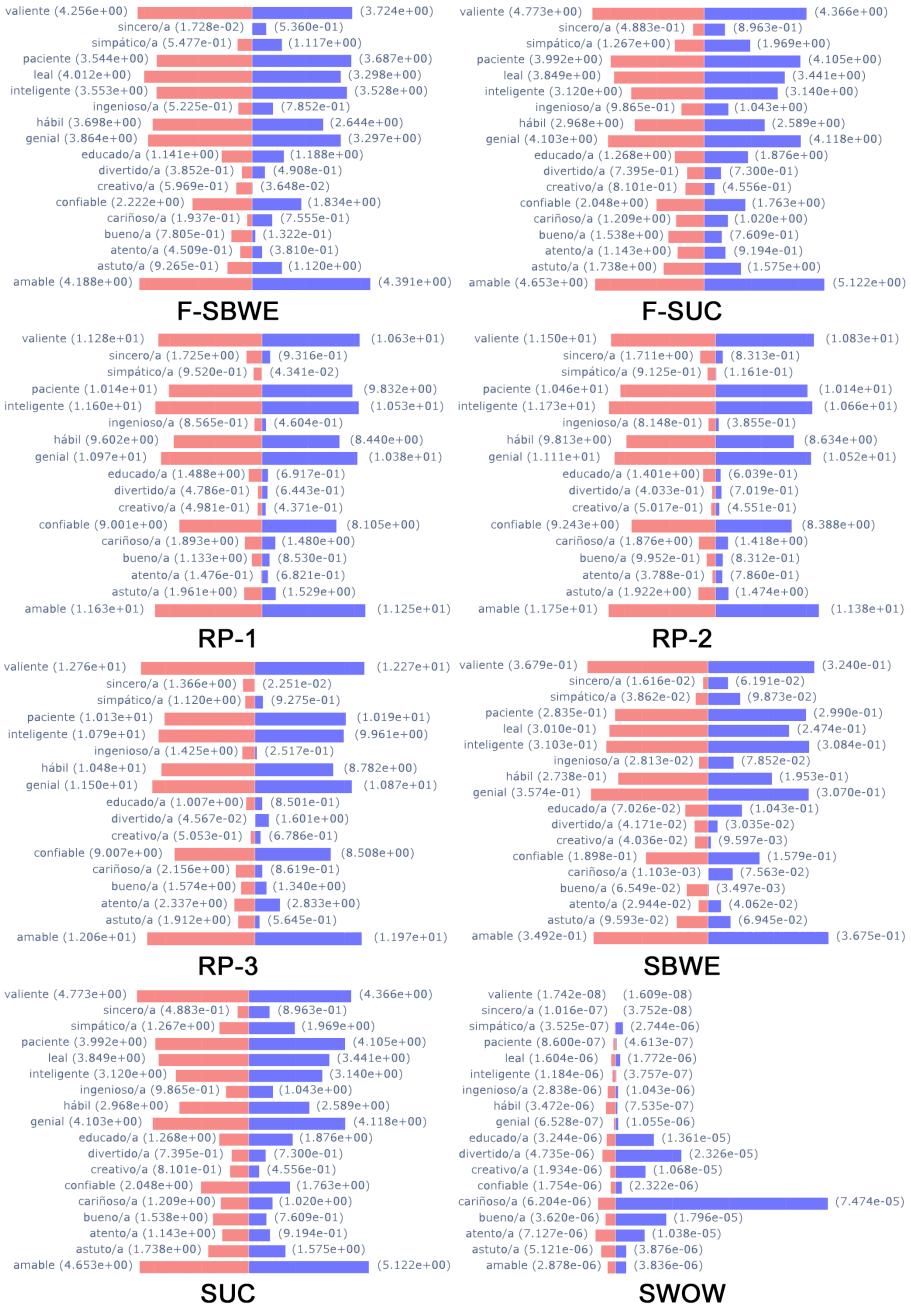

Figura A.7: Resultados de atributos invisibles positivos para los diferentes modelos según los subespacios de estudio **masculino** (en rojo) y **femenino** (en azul).

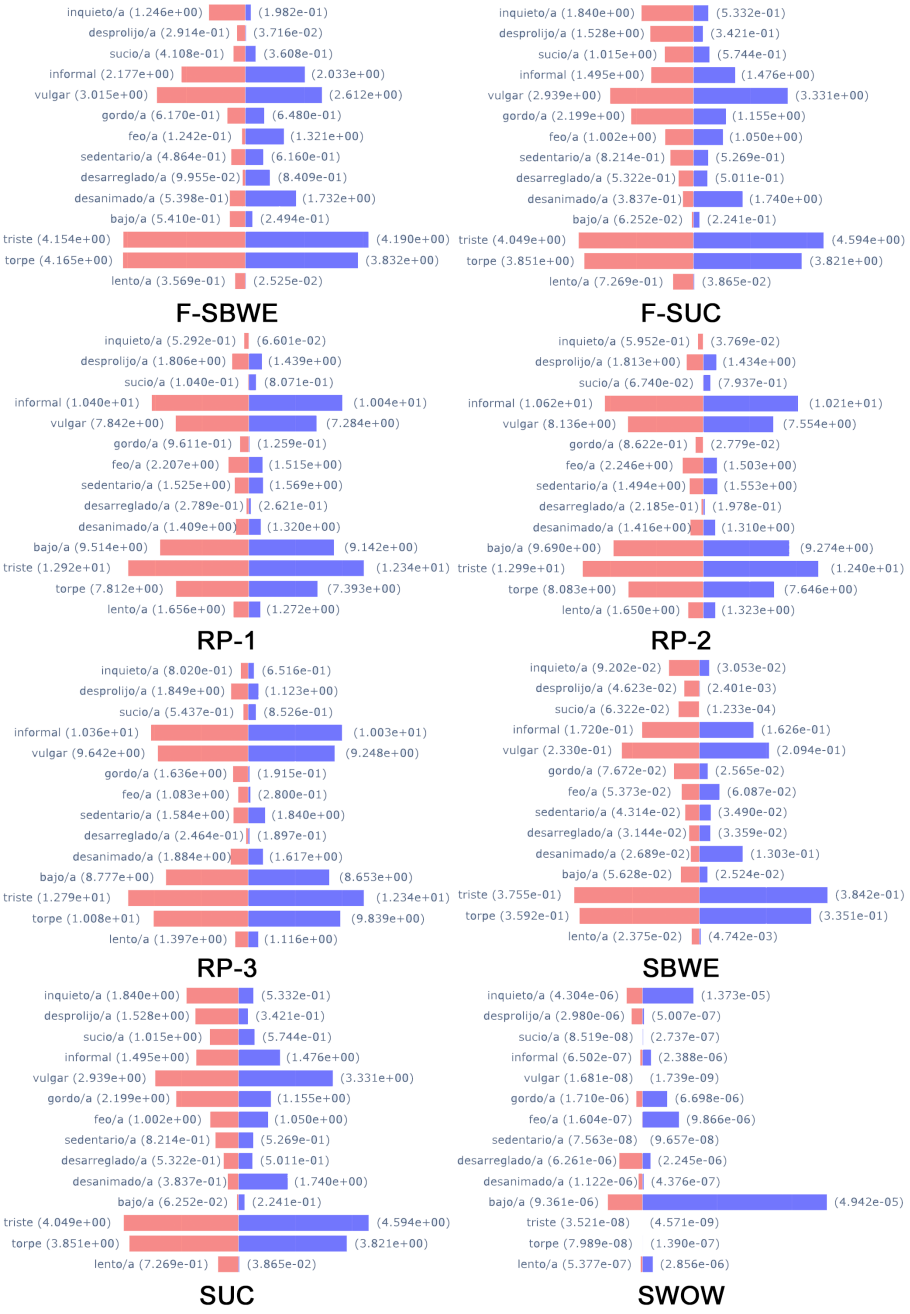

Figura A.8: Resultados de atributos visibles negativos para los diferentes modelos según los subespacios de estudio **masculino** (en rojo) y **femenino** (en azul).

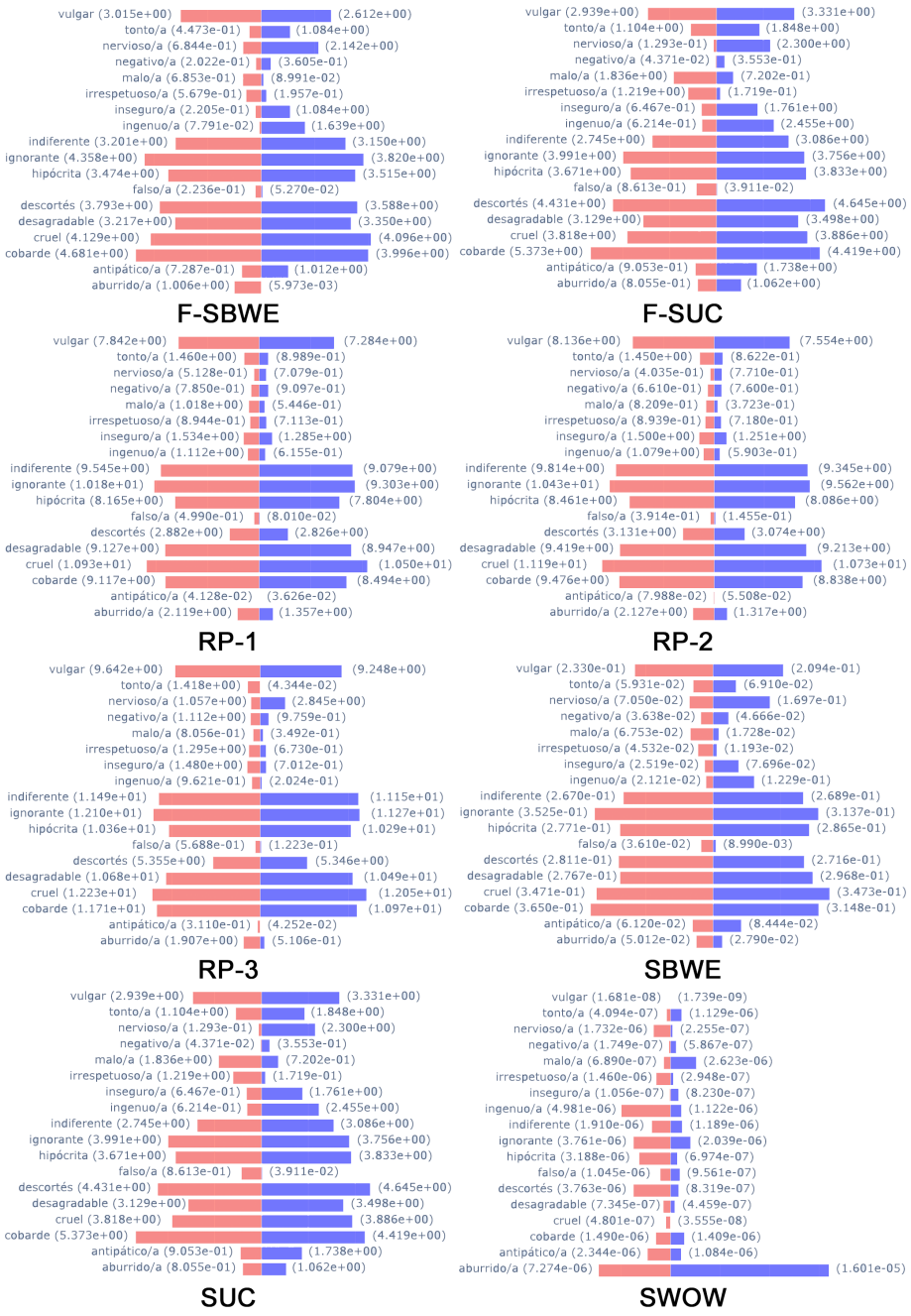

Figura A.9: Resultados de atributos invisibles negativos para los diferentes modelos según los subespacios de estudio **masculino** (en rojo) y **femenino** (en azul).

A.6. Atributos de visibilidad y polaridad para la raza (blanca-negra)

A continuación mostramos las gráficas con los resultados palabra a palabra de las pruebas realizadas utilizando atributos de visibilidad y polaridad para los subespacios de raza.

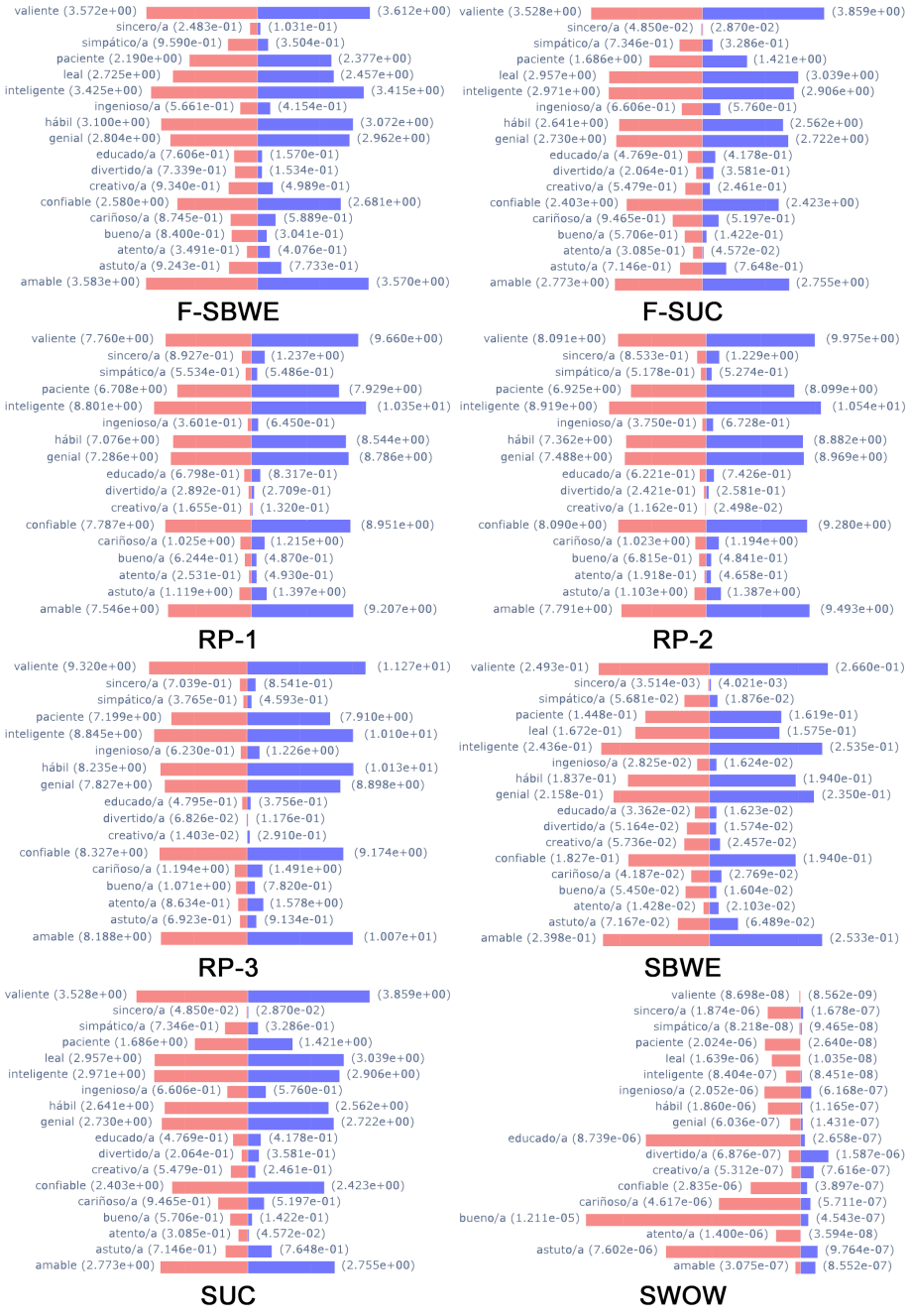

Figura A.10: Resultados de atributos visibles positivos para los diferentes modelos según los subespacios de estudio **raza negra** (en rojo) y **raza blanca** (en azul).

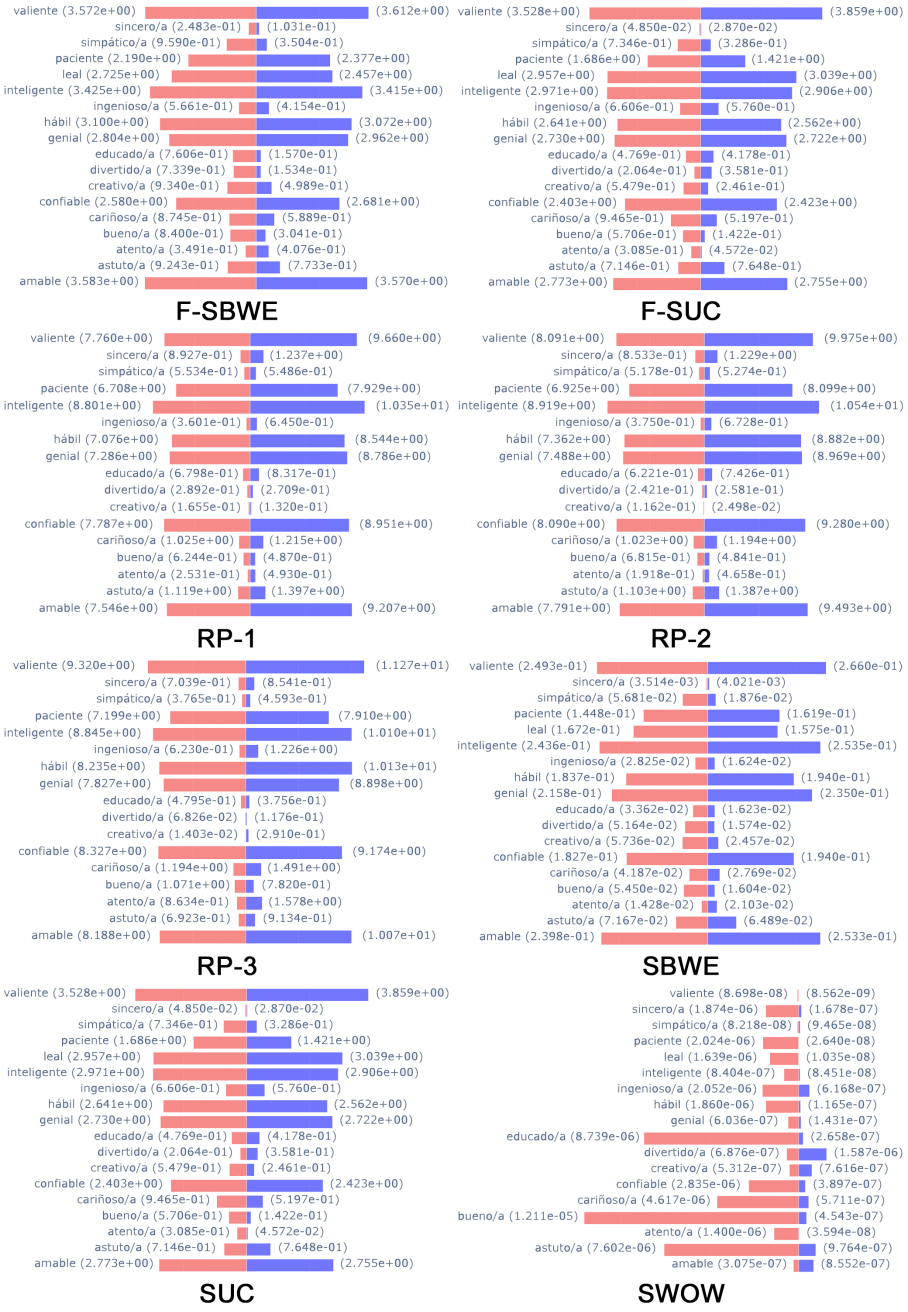

Figura A.11: Resultados de atributos invisibles positivos para los diferentes modelos según los subespacios de estudio **raza negra** (en rojo) y **raza blanca** (en azul).

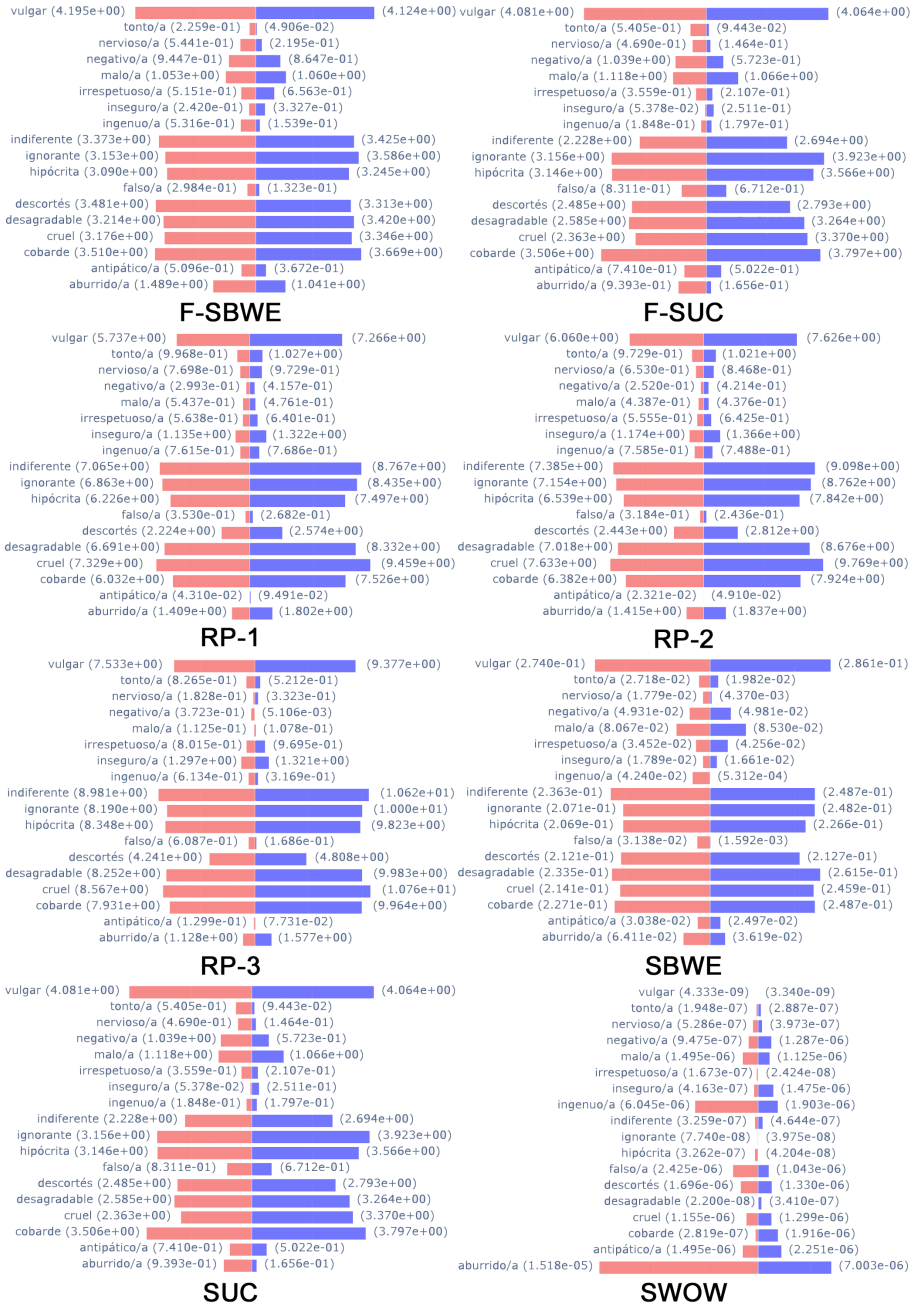

Figura A.12: Resultados de atributos visibles negativos para los diferentes modelos según los subespacios de estudio **raza negra** (en rojo) y **raza blanca** (en azul).

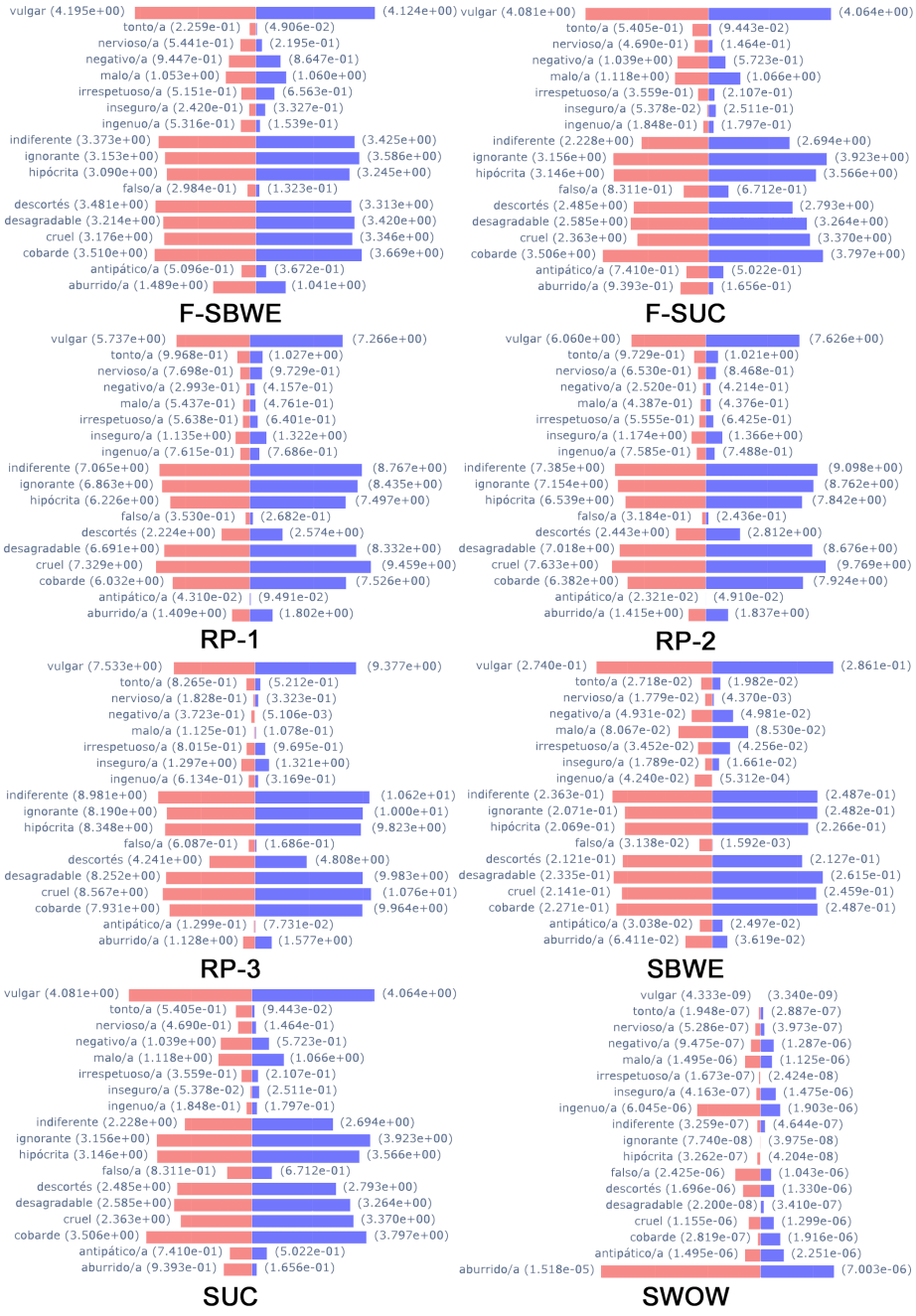

Figura A.13: Resultados de atributos invisibles negativos para los diferentes modelos según los subespacios de estudio **raza negra** (en rojo) y **raza blanca** (en azul).

A.7. Atributos de visibilidad y polaridad para la representación de colonizados y colonizadores

A continuación mostramos las gráficas con los resultados palabra a palabra de las pruebas realizadas utilizando atributos de visibilidad y polaridad para los subespacios colonizador y colonizado.

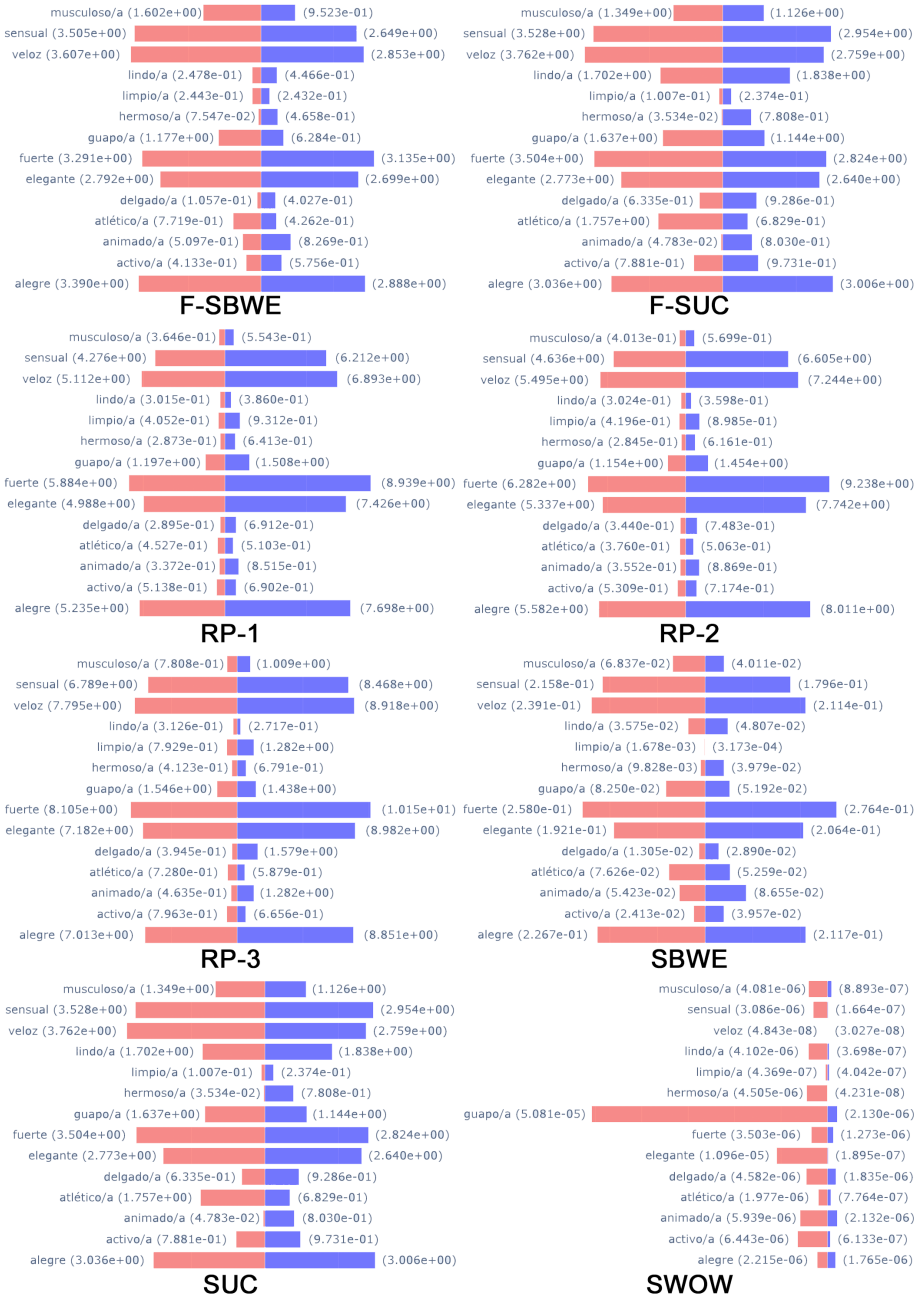

Figura A.14: Resultados de atributos visibles positivos para los diferentes modelos según los subespacios de estudio **colonizador** (en rojo) y **colonizado** (en azul).

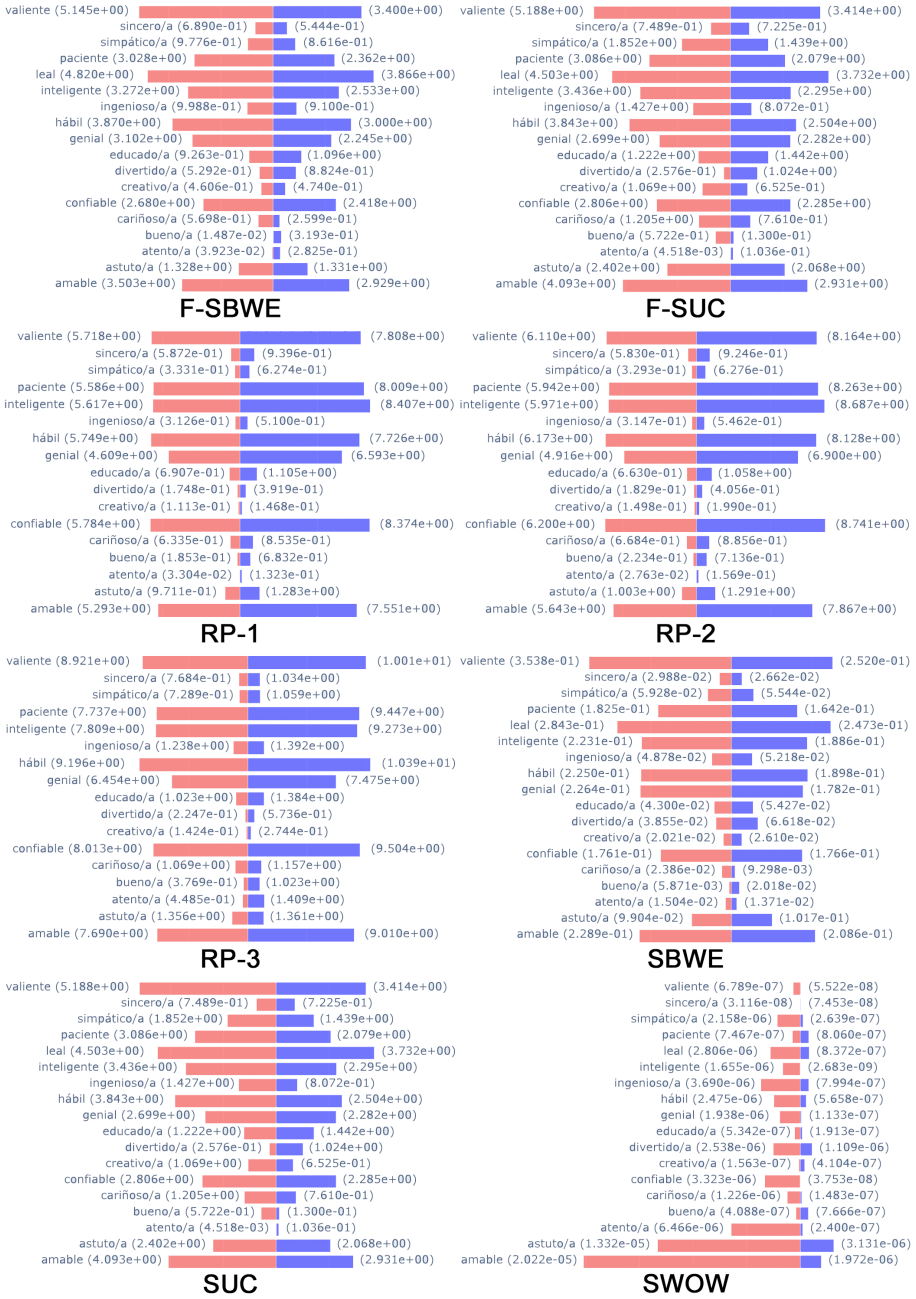

Figura A.15: Resultados de atributos invisibles positivos para los diferentes modelos según los subespacios de estudio **colonizador** (en rojo) y **colonizado** (en azul).

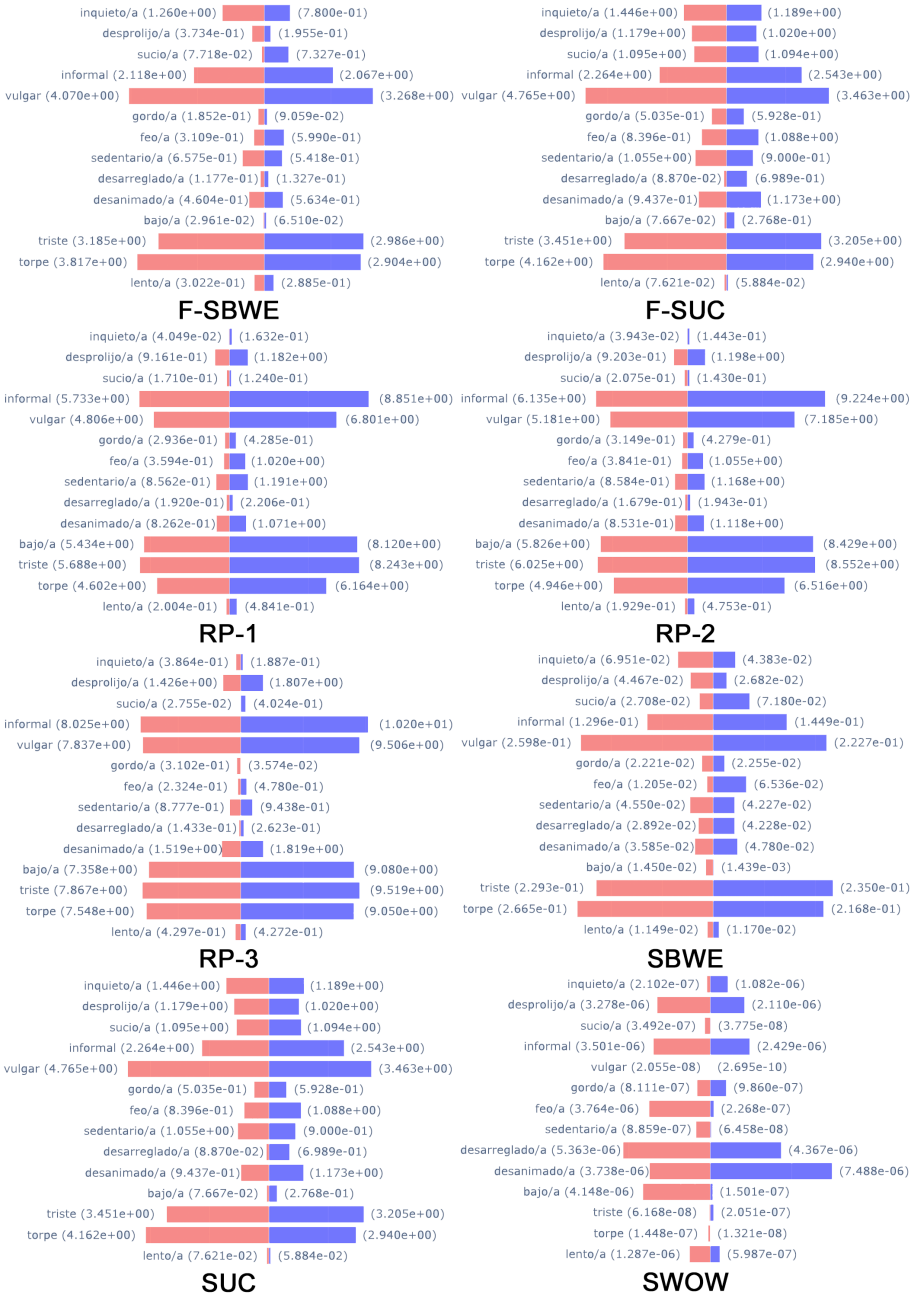

Figura A.16: Resultados de atributos visibles negativos para los diferentes modelos según los subespacios de estudio **colonizador** (en rojo) y **colonizado** (en azul).

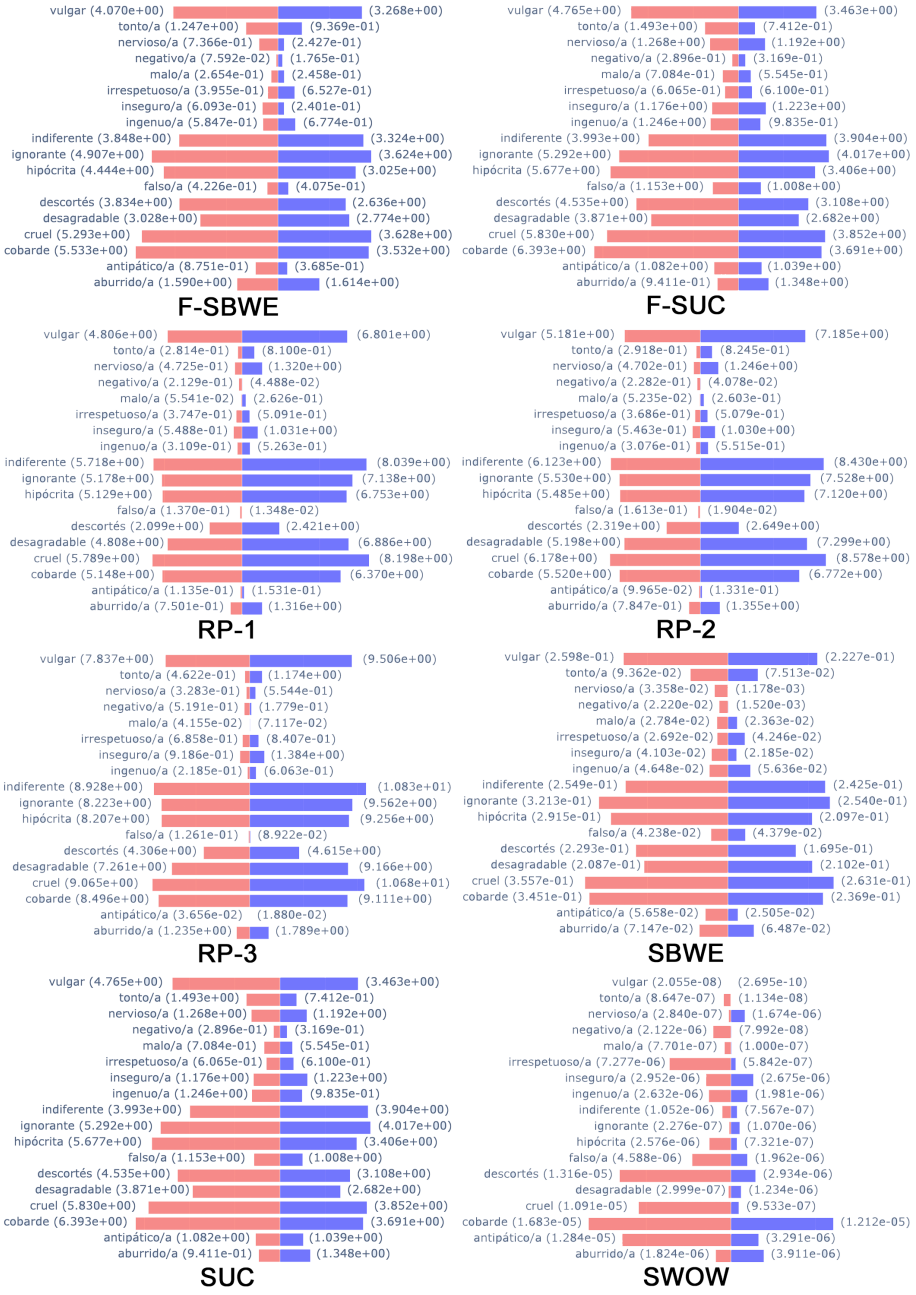

Figura A.17: Resultados de atributos invisibles negativos para los diferentes modelos según los subespacios de estudio **colonizador** (en rojo) y **colonizado** (en azul).

A.8. Data statement

A continuación presentamos el *data statement* que realizamos para los *embeddings* presentados en el proyecto SWOW.

Declaración de datos para "Rioplatense Spanish Small World of Words"

1 Cabezal

Título del conjunto de datos: swow.embedding.was.26-04-2022.vec.

Curador(es) del conjunto de datos: Álvaro Cabana, Camila Zugarramurdi, Juan Valle Lisboa y Simon De Deyne.

Versión del conjunto de datos: 26 de Abril, 2022.

Cita del conjunto de datos: Cabana, Á., Zugarramurdi, C., Lisboa, J. V., & De Deyne, S. (2022, September 19). The "Small World of Words" Free Association Norms for Rioplatense Spanish. https://doi.org/10.3758/s13428-023-02070-z.

Autores de la declaración de datos: María Fernanda Cánepa y Sebastián Lagomarsino.

Versión de la declaración de datos: 5 de Agosto, 2023.

2 Resumen

Los Word embeddings del proyecto SWOW-RP (Cabana, Zugarramurdi, Valle-Lisboa, & De Deyne, 2023) fueron creados a partir de un conjunto de datos recolectado a través de la técnica de asociación libre de palabras. Consiste de un vocabulario de 13,168 palabras formando *word embeddings* de dimensión 400. Fueron creados para ser parte de un proyecto aún más grande (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019), agregando información para el español rioplatense.

3 Justificación de la curación

Los conjuntos de datos creados a través de tareas de asociación de palabras se han utilizado para dos propósitos principales. Primero, como herramienta para medir propiedades léxicas en experimentos psicolingüísticos; segundo, para explorar y modelar la estructura semántica de la memoria. El proyecto "Small World of Words" (SWOW) tiene como objetivo recopilar normas de asociación de palabras para varios idiomas (De Deyne et al., 2019). El conjunto SWOW-RP busca ampliar la cobertura lingüística de este proyecto agregando el español rioplatense.

La recolección de datos del SWOW-RP fue realizada desde el 2014 hasta el 2022, fecha de publicación del conjunto. La modalidad de recolección fue escrita y asíncrona a través de un formulario web en el que los participantes tomaron parte del juego de asociación libre. Para su participación se les proporcionaron indicaciones claras y por escrito de las reglas del juego. La investigación está dirigida a personas originarias de la región del Río de la Plata, principalmente Uruguay y Argentina.

4 Variedad lingüística

Español rioplatense (es-UY y es-AR), variedad del español hablada en la región del Río de la Plata, en particular en las capitales de Uruguay (Montevideo) y Argentina (Buenos Aires).

5 Demografía de los hablantes

Hubo 67,525 participantes de los cuales el 82% se identificaron como mujeres, 17% como hombres y el restante 1% como otro. Los participantes tenían una

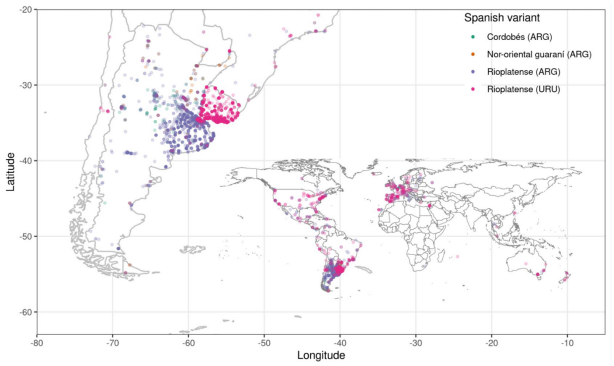

Figure 1: Localización geográfica de los participantes. Tomado de (Cabana et al., 2023).

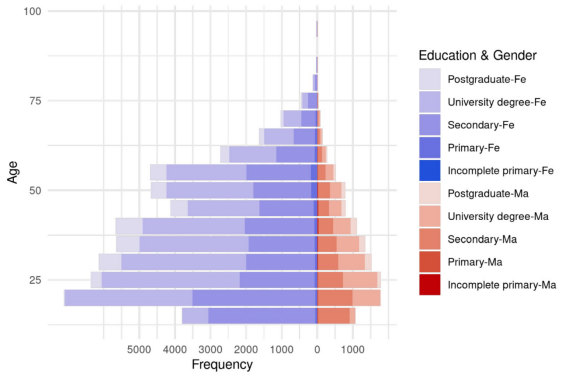

Figure 2: Distribución de edad, nivel educativo y género de los participantes. Tomado de (Cabana et al., 2023).

edad promedio de 38.3 años (con desviación estándar de 15.2, mínimo 5 y máximo 99).

La Figura 1 ilustra la ubicación geográfica de la mayoría de los participantes.

Sus niveles de educación alcanzados máximos y las distribuciones de edades por género se muestran en la Figura 2. Por lo tanto, más del 90% de los participantes declararon ser hablantes nativos de español rioplatense de los cuales: variantes del rioplatense argentino (49% de los participantes), guaraní no oriental (0.8%), cordobés (1.4%) y rioplatense uruguayo (43%), así como la opción "Otro" (5%). La mayoría de los participantes habían completado la educación secundaria (41%) o educación superior (48%).

6 Preprocesamiento y formato de datos

El preprocesamiento y formateo de los datos consistió en los siguientes pasos:

Limpieza inicial: consistió en quitar la puntuación, comillas, etiquetas, guiones y espacios extra innecesarios de las respuestas.

Cambio de respuestas: las respuestas ocasionalmente tenían alguna de las respuestas "No sé", "?" o "No más respuestas". Guardaron esas respuestas como si las personas hubieran cliqueado los botones específicos para eso.

Manejo de respuestas repetidas: en las respuestas que estuvo la misma respuesta varias veces, se guardó solo la primera respuesta y se guardó el resto como respuestas desconocidas.

Manejo de respuestas con tres palabras: algunos participantes escribieron tres palabras separadas por comas o espacios en lugar de responder en cada casilla. Para aquellos que tuvieron este comportamiento en más del 30% de sus respuestas, separaron la primera respuesta en las tres respuestas descartando lo que hayan escrito en la segunda y tercera respuesta.

Normalización de capitalización: si más del 15% de las respuestas de un participante eran todas con mayúsculas o comenzaban con mayúscula, todas las respuestas fueron pasadas a minúsculas. Los nombres propios o acrónimos fueron capitalizados como corresponde.

Normalización de ortografía: se corrigió la gramática de las palabras mal escritas. Para las palabras que tienen diferentes formas de escribirse, se utilizó la respuesta más frecuente como forma canónica.

Además de dar los datos preprocesados, los autores incluyeron los datos crudos, sin ningún tipo de preprocesamiento ni formateo.

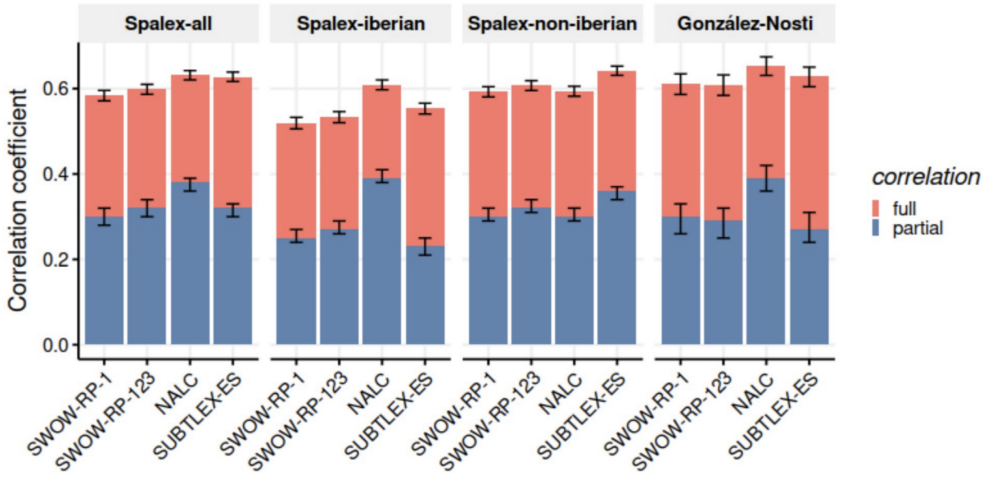

Figure 3: Correlación de Pearson entre el logaritmo de los tiempos de reacción en tareas de decisión léxica y las estimaciones de frecuencia obtenidas a partir de SWOW-RP y otras normas de frecuencia o asociación libre. Los valores en la gráfica se muestran con el signo invertido por legibilidad. Tomado de (Cabana et al., 2023).

7 Calidad de recolección

Como la recolección de datos se basa en un formulario web, la calidad de los datos depende mucho de las respuestas de los participantes. Las complicaciones principales son que los participantes pueden dejar respuestas incompletas, responder de forma "errónea" o ingresar errores ortográficos. Estas complicaciones fueron atacadas durante el proceso de curación manual al que se sometieron los datos luego de su recolección, pero algunos errores podrían persistir.

8 Limitaciones

Pudimos observar ausencia de palabras referentes a identidad de género, etnia y orientación sexual de las personas, no encontrando términos tales como "transgénero", "afrodescendiente" o "bisexual".

Se debería de tener precaución al utilizar este modelo para aplicaciones en contextos que necesiten o utilicen estos datos, ya que podría resultar en discriminación involuntaria a los grupos que no se encuentran representados.

También destacamos una representación no equitativa de las profesiones en su forma femenina, encontrándose ausentes algunas tales como "abogada", "ingeniera" o "arquitecta".

Esta carencia resulta relevante si queremos utilizar este modelo en aplicaciones relacionadas con búsqueda laboral o selección de personal, ya que las mujeres se verían en una posición de desfavorable al no estar reflejadas en el vocabulario.

Si bien describimos dos problemas en específico, ambos son casos particulares de la principal limitante de este conjunto de vectores, que es su escasez de vocabulario. Es posible que existan más contraindicaciones vinculadas a esta limitante, por ende, es pertinente verificarla en base al uso que le daremos al conjunto de datos.

9 Metadatos

cial.

Cómo citar: este documento es principalmente un resumen de información previamente publicada. Por lo tanto, al citar este trabajo, se debe citar también el estudio original. Cuando el espacio para citar es limitado, siempre se debe dar preferencia a la(s) fuente(s) original(es).

Métricas de calidad del conjunto de datos: se realizaron dos pruebas para la calidad de este proyecto. La primera consistió en el cálculo de la correlación de Pearson entre la frecuencia de las palabras y el tiempo de reacción de los participantes (en escala logarítmica) en tareas de decisión léxica. Estas tareas son utilizadas en los campos de la psicología y psicolingüística, consisten en ir presentándole a las personas un conjunto de palabras y no-palabras, donde para cada ejemplo deben indicar si creen que se trata de una palabra perteneciente al léxico o no. Se guardan datos como la precisión de las respuestas y los tiempos de reacción de las personas, los resultados de estas pruebas se suelen usar para análisis de memoria semántica. La frecuencia para esta prueba está definida como la cantidad de veces que aparece una palabra como estímulo de otra. En la Figura 3 mostramos los resultados con diferentes conjuntos de datos de decisión léxica. Luego, la segunda prueba fue de similitud semántica. Para esto los autores utilizaron la correlación de Spearman entre los pares de palabras de distintos conjuntos de similitud con la distancia coseno de los *word embeddings* de esas palabras, es decir, imaginemos que tenemos un par de palabras, "palabra1"-"palabra2", en un conjunto con una asociación de 0.9. Se calculan los *word embeddings* de "palabra1" y de "palabra2", y se calcula la distancia coseno entre los dos *word embeddings*. Se realiza este proceso para todos los pares de palabras del conjunto de similitud, y por último se utiliza la correlación de Spearman entre las distancias coseno calculadas y el valor de asociación de cada par de palabras. En la Figura 4 mostramos los resultados.

10 Glosario

Word embeddings Son representaciones vectoriales de palabras que capturan su significado y relaciones en el lenguaje. Se suelen utilizar en tareas de procesamiento de lenguaje natural. 1

Sobre este documento

Una declaración de datos es una caracterización de un conjunto de datos que proporciona contexto para permitir que los desarrolladores y usuarios comprendan mejor cómo podrían generalizarse los resultados experimentales, cómo podría desplegarse el software de manera adecuada y qué sesgos podrían reflejarse en los sistemas construidos sobre el software.

Esta declaración de datos fue redactada según el esquema de la versión 2 de las declaraciones de datos. La plantilla fue preparada por Angelina McMillan-Major, Emily M. Bender y Batya Friedman, y puede encontrarse en `http://techpolicylab.uw.edu/data-statements`.

Realizamos los siguientes cambios al formato original del esquema:

- Traducimos los títulos de inglés a español.
- Quitamos las secciones "Documentación para conjunto de datos fuentes", "Demografía de los anotadores", "Situación de discurso y características del texto", "Divulgaciones y revisión de ética" y "Otros".

References

Cabana, Á., Zugarramurdi, C., Valle-Lisboa, J. C., & De Deyne, S. (2023). The "small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, 1–18.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, *51*, 987–1006.

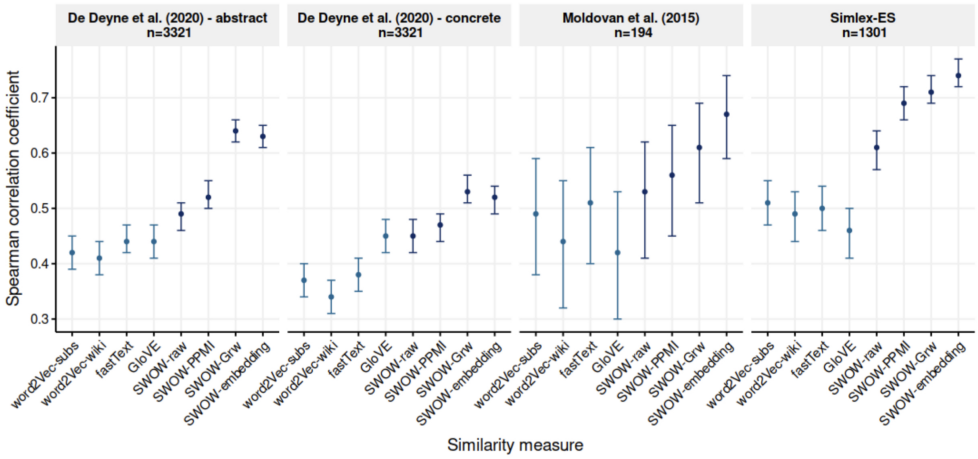

Figure 4: Correlación de Spearman. Tomado de (Cabana et al., 2023).