ISSN: 1692-7257 - Volumen 2 – Número 44 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v2i44.3010

Exploración del sesgo de género en la clasificación de ocupaciones de Colombia utilizando aprendizaje automático

Exploring gender bias in Colombian occupation classification using machine learning

MSc. Deimer de Jesús Ramos Cuello[1], MSc. Alveiro Rosado Gómez[2]
PhD. Maritza Liliana Calderón Benavides[1]

[1] ***Universidad Autónoma de Bucaramanga,*** *Facultad de Ingeniería, Maestría en Gestión Aplicación y Desarrollo de Software, Bucaramanga, Santander, Colombia.*
[2] ***Universidad Francisco de Paula Santander,*** *Facultad de Ingeniería, Grupo de Investigación en Desarrollo Tecnológico en Ingeniería (GITYD), Ocaña, Norte de Santander, Colombia.*

Correspondencia: *aarosadog@ufpso.edu.co*

Recibido: *21 enero 2024.* ***Aceptado:*** *11 junio 2024.* ***Publicado:*** *19 julio 2024.*

Cómo citar: *D. de J. Ramos Cuello, A. A. Rosado Gomez, y M. L. Calderón Benavides, «Exploración del sesgo de género en la clasificación de ocupaciones de Colombia utilizando aprendizaje automático», RCTA, vol. 2, n.º 44, pp. 83–88, jul. 2024. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3010*

Resumen: El artículo explora el uso de Word2Vec y FastText para convertir nombres de ocupaciones en representaciones vectoriales y analizar su polaridad de género. Se emplearon dos bases de datos colombianas para preparar y limpiar los datos. Mediante clasificadores, se evaluó cómo la polaridad de género afecta la clasificación de ocupaciones y salarios. Se utilizó ANOVA y pruebas de Tukey para el análisis estadístico. Se descubrió que modelos como ExtraTreesClassifier y XGBClassifier presentaron menores diferencias de precisión entre géneros, sugiriendo que tienden a clasificar con mayor exactitud a los hombres. Sin embargo, no se evidenció una preferencia clara en las predicciones de los modelos hacia un género específico tras manipular las variables relacionadas con denominaciones profesionales. El estudio destaca la importancia de abordar los sesgos sistémicos en representaciones semánticas que pueden perpetuar prejuicios existentes.

Palabras clave: aprendizaje automático, aprendizaje supervisado, equidad en inteligencia artificial, incrustaciones de palabras, procesamiento del lenguaje natural.

Abstract: The paper explores using Word2Vec and FastText to convert occupational names into vector representations and analyze their gender polarity. Two Colombian databases were used to prepare and clean the data. Using classifiers, we evaluated how gender polarity affects the classification of occupations and salaries. ANOVA and Tukey tests were used for statistical analysis. It was discovered that models such as ExtraTreesClassifier and XGBClassifier presented more minor differences in accuracy between genders, suggesting that they tend to classify men more accurately. However, no clear preference was evident in the models' predictions toward a specific gender after manipulating the variables related to professional denominations. The study highlights the

importance of addressing systemic biases in semantic representations that can perpetuate existing prejudices.

Keywords: Machine learning, supervised learning, equity in artificial intelligence, word embeddings, natural language processing.

1. INTRODUCCIÓN

La inteligencia artificial (IA) es una herramienta poderosa para automatizar tareas complejas y mejorar la eficiencia en diferentes campos, pero su crecimiento rápido plantea desafíos éticos y riesgos, como la posible discriminación al reflejar sesgos humanos en datos y decisiones [1] [2] [3] [4].

El procesamiento del lenguaje natural (PLN) es un campo de la informática y la IA que se enfoca en analizar y representar lenguajes naturales como el inglés, español, francés, etc. en la era actual, que está llena de datos desestructurados, ya sean textuales o auditivos [5]. PLN permite el análisis, pronóstico e interacción con sistemas inteligentes como asistentes virtuales o vehículos autónomos y aborda desafíos como la inferencia del lenguaje, la comprensión, el reconocimiento de voz y la clasificación de texto [6].

La arquitectura de un sistema PLN se basa en los niveles que componen un lenguaje humano, como el fonológico, morfológico, sintáctico, semántico y pragmático. Estos niveles proporcionan las herramientas necesarias para que los usuarios finales se comuniquen con la máquina y esta pueda interpretar sus intenciones [7]. Además, el PLN implica diversas áreas, como la extracción de información, el análisis de sentimientos, la investigación y consultas, la síntesis automática y la minería de datos [5].

Una de las evoluciones que ha tenido el PLN, son las incrustaciones de palabras, las cuales consisten en representar palabras como vectores de valor real en un espacio multidimensional, donde métodos de incrustación semántica codifican el significado de las palabras de modo que las similares estén cercanas en el espacio vectorial [8]. A medida que los algoritmos de aprendizaje automático se vuelven los responsables de la toma de decisiones, se aumenta el riesgo de sesgos sistemáticos en estos algoritmos. No solo pueden reflejar los sesgos humanos en los datos de entrenamiento, sino que también pueden amplificar esos sesgos en su aplicación práctica [4].

Sí un conjunto de datos de entrenamiento para incrustaciones de palabras contiene una alta frecuencia de asociaciones negativas o prejuicios hacia un grupo étnico en particular, las palabras relacionadas con ese grupo podrían estar más cerca de términos negativos en el espacio vectorial, reflejando así un sesgo. De manera similar, palabras relacionadas con género, como "enfermera" o "ingeniero", podrían estar ubicadas en regiones del espacio vectorial que reflejen estereotipos de género, lo que podría influir en las decisiones de algoritmos de aprendizaje automático [9].

2. METODOLOGÍA

El desarrollo de la investigación inicio con el acceso a la base de datos de Clasificación Única de Ocupaciones para Colombia (CUOC), que contiene información detallada sobre nombres de grupos primarios, descripciones de ocupaciones, niveles de competencia, nombres de denominaciones y redacción de funciones.

Posteriormente, se eligieron representaciones vectoriales de incrustaciones de palabras en español, Word2Vec y FastText, debido a su entrenamiento específico en español y su capacidad para capturar relaciones semánticas entre palabras [10] [11]. Las incrustaciones de palabras fueron utilizadas para transformar las denominaciones en representaciones vectoriales y luego se determinó la polaridad de género utilizando la biblioteca responsibly [12].

Para determinar si las ocupaciones que tenían una mayor polaridad por un género determinado influían en la clasificación, fue necesario utilizar los conjuntos de datos de Características generales, seguridad social en salud y educación (información personal de cada persona) y ocupados (información laboral de las personas que tienen un empleo) que hacen parte de la Encuesta Nacional de Hogares suministrada por el Departamento Administrativo Nacional de Estadística (DANE).

Todos los datos fueron integrados y procesados. Los registros que fueron utilizados con la información ofrecida por el DANE, fueron aquellos que estaban asociados a ocupaciones con polaridad de genero

sesgada. Con esos datos fueron entrenados clasificadores para determinar si la transformación vectorial de los textos, puede influir en las salidas de los modelos entrenados con datos tabulares.

En el proceso de preparación de datos, se realizaron varias etapas. Primero, se generó una llave única concatenando variables como Directorio, SECUENCIA_P y ORDEN para identificar a las personas encuestadas en diferentes conjuntos de datos. Luego, esta llave se asignó a los datasets "ocupados" y "caracteristicas_generales", y se mapearon nombres legibles a las variables/columnas en estos conjuntos. Posteriormente, se unieron los dos conjuntos de datos basados en la llave única, combinando así la información de características generales y ocupaciones. Se extrajo y generó un nuevo conjunto de datos que incluye el código CIUO (Clasificación Internacional Uniforme de Ocupaciones) de ocupaciones sesgadas. Luego, se unieron los conjuntos de datos nuevamente usando identificadores de ocupación. Se corrigieron las edades, calculando las edades faltantes y mapeando valores numéricos de género a etiquetas literales. También se depuraron registros con edades menores de cinco años y se corrigieron salarios en cero utilizando el algoritmo KNNImputer en registros similares de la misma ocupación para imputar valores de remuneración [13].

Se efectuaron experimentos con clasificadores para evaluar la denominación de las ocupaciones, el género y las categorías de ingresos, utilizando las denominaciones como atributos binarios y midiendo la precisión de cada modelo. Finalmente se realizó un análisis estadístico con ANOVA y pruebas de Tukey, para determinar la interacción entre el género y las estrategias de procesamiento [14].

3. RESULTADOS

Se utilizó la biblioteca "responsibly" para evaluar el sesgo de género en ocupaciones laborales. Para evaluar el sesgo de género en español, se identificaron palabras neutras en cuanto al género, ya que muchas palabras en español tienen marcas de género morfológicas. La tabla 1 resultante muestra las ocupaciones laborales junto con su dirección de género, que indica si tienden hacia el sexo masculino (M) o femenino (F) en función del análisis de las palabras utilizadas para describir la ocupación, y su género asociado. En la tabla, se puede observar, como para la mayoría a la mayoría de las profesiones se le asigno el sexo masculino como la polaridad más cercana.

***Tabla 1**: Modelos pre-entrenados vs CUOC*

Denominación	FastText	Word2Vec
Almacenista	M	M
Apuntador	M	M
Arquitecto	M	M
Chef	M	M
Cocinero	M	F
Electricista	M	M
Secretario	M	F

***Fuente:** elaboración propia*

Las ocupaciones se muestran en la tabla 1, corresponde a las finales, restantes luego de eliminar registros con atributos con un 75% de sus valores nulos. Adicionalmente, solo fueron tomadas las ocupaciones que estuvieran conformadas por una única palabra, dado que la librería utilizada para determinar la polaridad, solo acepta esta clase de textos. Luego de eliminar los registros, el conjunto de datos quedo conformados por 3630 instancias y 13 atributos. Al hacer la transformación de los datos los atributos aumentaron a 32.

El conjunto de datos estaba conformado por un total de 1745 hombres y 1885 mujeres. En la tabla 2, muestra la distribución de los valores del conjunto de datos que se utilizó para la investigación, las columnas F y M, corresponden a la cantidad de registros por cada género. Las columnas ING_F y ING_M corresponden a los ingresos promedio en cientos de miles. Se puede observar cómo en este conjunto de datos existe más cantidad de mujeres que hombres por denominación y que en los ingresos no existe un comportamiento sesgado hacia uno de los géneros.

***Tabla 2**: Cantidades e ingresos por genero*

Denominación	F	M	ING_F	ING_M
Almacenista	3	1	12	15
Apuntador	466	340	14	15
Arquitecto	80	2	6	6
Chef	245	157	20	27
Cocinero	160	151	17	17
Conserje	344	518	12	9
Electricista	1	3	12	9
Secretario	274	268	13	13

***Fuente:** elaboración propia*

Para estudiar el comportamiento de las denominaciones, se realizaron varios experimentos. En el primero, se usó la denominación (Dem) como etiqueta de salida; en el segundo, se empleó el atributo "sexo". En un tercer experimento, se

clasificaron los ingresos laborales en tres grupos (Ingr) y se usaron como etiquetas de salida. En este caso, el valor "cero" indica aquellos con ingresos menores a un salario mínimo, el valor "uno" a aquellos con ingresos entre uno y dos salarios mínimos, y el valor "tres" a quienes superan los dos salarios mínimos. En los experimentos primero y tercero, las denominaciones se consideraron como atributos binarios independientes. La Tabla 3 muestra la precisión (Accuracy) de cada modelo. La mejor precisión se obtiene cuando los ingresos son clasificados y usados como categoría. Además, los resultados del clasificador RidgeClassifierCV se establecen como referencia o línea base.

***Tabla 3**: Exactitud por clase*

Clasificador	Dem	Sexo	Ingr
Logistic Regression	0.22	0.51	0.57
KNN	0.50	0.63	0.62
Decision Tree	**0.75**	**0.82**	**0.81**
Random Forest	**0.77**	**0.81**	**0.82**
SVM	0.28	0.51	0.40
Gradient Boosting	0.59	0.52	0.68
XGBClassifier	**0.80**	**0.76**	**0.82**
AdaBoostClassifier	0.34	0.46	0.58
BaggingClassifier	**0.74**	**0.81**	**0.80**
ExtraTreesClassifier	**0.75**	**0.82**	**0.82**
LogisticRegressionCV	0.23	0.51	0.59
PassiveAggressiveClassifier	0.20	0.50	0.42
RidgeClassifierCV	0.20	0.52	0.59
SGDClassifier	0.11	0.49	0.45
Perceptron	0.09	0.49	0.46

***Fuente:** elaboración propia*

Los algoritmos ExtraTreesClassifier (ET) y XGBClassifier (XG) fueron elegidos tras constatar que la diferencia en la exactitud entre hombres y mujeres era mínima. Este comportamiento continuo al utilizar el sexo como etiqueta de clasificación. La intervención siguiente se centró en los atributos relacionados con las profesiones, unificándolos en un solo atributo representado por un vector de dimensión diez, generado mediante Word2Vec (W2V) y FastText (FAS). Según se expone en la Tabla 4, la métrica F1 se estableció como el criterio de referencia para evaluar la eficacia del modelo en la clasificación entre el sexo masculino (F1_M) y femenino (F1_F). De acuerdo con estos datos, no se detectaron diferencias significativas entre las distintas intervenciones. Además, los pequeños incrementos o decrementos en los valores obtenidos con ambos clasificadores no son suficientes para concluir que una estrategia sea superior a la otra, o que un clasificador específico mejore los resultados de una intervención determinada. Un patrón observado en los experimentos es que los modelos tienden a clasificar con mayor precisión a los hombres que a las mujeres.

***Tabla 4**: Rendimiento por intervención*

Intervención	F1_F	F1_M
NOR_ET	0.85	0.92
W2V_ET	0.84	0.92
FAS_ET	0.85	0.92
NOR_XG	0.86	0.91
W2V_XG	0.86	0.93
FAS_XG	0.86	0.92

***Fuente:** elaboración propia*

Para investigar la influencia de la profesión o su codificación vectorial en los resultados de los modelos, se examinó la interacción entre el género y las estrategias de procesamiento de datos, comparando la precisión entre hombres y mujeres. A través de un análisis de ANOVA y las pruebas post-hoc de Tukey, se exploró cómo estas interacciones afectan la exactitud de los modelos predictivos. En términos generales, se descubrió que la manipulación de las variables relacionadas con las denominaciones profesionales era estadísticamente significativa ($p = 0.001$), lo que indica que la estrategia de procesamiento de datos ejerce una influencia en la precisión de los modelos. Sin embargo, al analizar la interacción entre género y la estrategia, los resultados no alcanzaron el nivel convencional de significancia estadística ($p = 0.05$), aunque un valor cercano ($p = 0.051$) sugiere una posible tendencia que, por sí sola, no es suficiente para establecer una conclusión definitiva sobre la existencia de un efecto significativo.

Como se detalla en la Tabla 5, las estrategias de procesamiento de datos parecen afectar la precisión con respecto al género, la evidencia no es concluyente para afirmar una interacción significativa de manera categórica. No obstante, en dos instancias específicas se observaron diferencias significativas en la precisión asociadas a combinaciones particulares de género y estrategia de procesamiento de datos: las mujeres (G1) que utilizaron la estrategia 'NOR' superaron en precisión a los hombres (G2) que emplearon 'FAS' o W2V. Pero, al tratarse de ‘NOR’ que no hace ninguna intervención y utiliza las denominaciones como atributos nominales binarios (1 para la presencia y 0 para la ausencia), indica que esta diferencia existe desde la línea base y no son generados por los nuevos experimentos.

***Tabla 5**: Diferencias sexo e intervención*

G 1	G 2	M diff	p-adj	lower	upper
FAS	FAS	-0.0241	0.9868	-0.13	0.0818
FAS	NOR	-0.04	0.8881	-0.1458	0.0659
FAS	W2V	-0.0273	0.9808	-0.1377	0.0832
NOR	FAS	0.1174	**0.020**	0.0115	0.2233
NOR	NOR	0.1015	0.0688	-0.0044	0.2074
NOR	W2V	0.1142	**0.0381**	0.0037	0.2246
W2V	FAS	0.0207	0.9942	-0.0881	0.1294
W2V	NOR	0.0048	1.0	-0.1039	0.1136
W2V	W2V	0.0175	0.9978	-0.0957	0.1307

***Fuente:** elaboración propia*

4. DISCUSIÓN

La utilización de modelos pre-entrenados de incrustaciones de palabras en español para la investigación, se buscaba acceder a un vocabulario en ocupaciones sesgadas. Se planteó que los word embeddings podrían asociar ocupaciones con un género basándose en la terminación de las palabras, como aquellas que terminan en "-a" con el femenino, por lo cual se optó por incluir todas las ocupaciones independientemente de su "dirección sexual", dado que los datos obtenidos incluían hombres y mujeres en ocupaciones tradicionalmente asociadas al género masculino [15].

El análisis de los datos mostró un desequilibrio en la representación de género, con una inclinación de los modelos de inteligencia artificial a predecir con mayor precisión para el género masculino. A pesar de que la distribución de los datos podría predisponer a un sesgo hacia los hombres, los modelos no evidenciaron una preferencia concluyente en sus predicciones [16] [17]. Los resultados indican que la aplicación de estos modelos en el mercado laboral de Colombia no necesariamente conduce a una discriminación por género de manera generalizada. Por otro lado, el rendimiento del conjunto de datos tabular inicial, sin la aplicación de representaciones vectoriales, fue superior en la clasificación de perfiles masculinos en comparación con los femeninos [18]. Los modelos que implementaron transformaciones en las denominaciones de las ocupaciones no modificaron significativamente este comportamiento, lo que sugiere que los sesgos presentes en las representaciones vectoriales no se trasladaron de manera directa a los conjuntos de datos tabulares, contrariando la hipótesis de que la discriminación preexistente podría perpetuarse mediante la integración de estas representaciones sesgadas [19] [20].

5. CONCLUSIONES

El uso de modelos pre-entrenados de incrustaciones de palabras en español para analizar el sesgo de género en ocupaciones laborales reveló que, aunque ciertos patrones de sesgo como la asociación de género basada en la terminación de las palabras en español fueron considerados, la inclusión de todas las ocupaciones independientemente de su "dirección sexual" no afecto las salidas del modelo. Esto se debe a que la realidad laboral incluye tanto a hombres como a mujeres en roles tradicionalmente asociados a un género específico, lo que sugiere que cualquier automatización o decisión impulsada por IA debe ser cuidadosamente examinada para evitar la perpetuación de estereotipos de género.

A pesar de que los datos mostraron un desequilibrio en la representación de género, con una ligera tendencia de los modelos de IA a predecir con mayor precisión para el género masculino, no se encontró un sesgo absoluto en las predicciones. Esto sugiere que la discriminación de género en la IA no es un fenómeno omnipresente y puede variar según el conjunto de datos específico y la forma en que se procesan y utilizan estas incrustaciones de palabras en modelos de aprendizaje automático.

Los experimentos realizados con diferentes clasificadores y estrategias de procesamiento de datos revelaron que los sesgos presentes en las representaciones vectoriales no se trasladaron de manera directa a los conjuntos de datos tabulares. Además, el análisis estadístico con ANOVA y pruebas de Tukey mostró que, aunque la manipulación de las variables relacionadas con las denominaciones profesionales tenía un efecto estadísticamente significativo en los resultados, no se corroboró la hipótesis inicial de que la discriminación preexistente se heredaría automáticamente en la IA al integrar estas representaciones sesgadas.

REFERENCIAS

[1] N. Bantilan, «Themis-ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation,» *Journal of Technology in Human Services*, pp. 15-30, 2018.

[2] J. Borana, «Applications of Artificial Intelligence & Associated Technologies,» de *International Conference on Emerging Technologies in Engineering, Biomedical, Management and Science*, Jodhpur, 2016.

[3] R. Burke, «Multisided Fairness for Recommendation,» de *Fairness, Accountability, and Transparency in Machine Learning*, Halifax, 2017.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman y A. Galstyan, «A Survey on Bias and Fairness in Machine Learning,» *arXiv*, pp. 1-31, 2019.

[5] S. Chowdhury y A. Nath, «Trends In Natural Language Processing : Scope And Challenges,» *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2021.

[6] B. Dev, A. Singh, N. Uppal, A. Rizwan, V. Sri y S. Suman, «Survey Paper: Study of Natural Language Processing and its Recent Applications,» *International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pp. 1-5, 2022.

[7] A. Nohria y H. Kaur, «Evaluation of Parsing Techniques in Natural Language Processing,» *International Journal of Computer Trends and Technology*, 2018.

[8] A. Gerek, M. C. Yüney, E. Erkaya y M. C. Ganiz, «Effects of Positivization on the Paragraph Vector Model,» *IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1-5, 2019.

[9] N. Swinger, M. De-Arteaga, N. T. Heffernan IV, M. Leiserson y A. T. Kalai, «What Are the Biases in My Word Embedding?,» de *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, 2019.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado y J. Dean, «Distributed Representations of Words and Phrases and their Compositionality,» *arXiv*, pp. 1-9, 2013.

[11] P. Bojanowski, E. Grave, A. Joulin y T. Mikolov, «Enriching Word Vectors with Subword Information,» *arXiv*, 2016.

[12] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama y A. Kalai, «Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,» *arXiv*, 2016.

[13] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, Sebastopol: O'Reilly, 2019.

[14] C. Lopez, A. Gazgalis, V. Boddapati, R. Shah, J. Cooper y J. Geller, «Artificial Learning and Machine Learning Decision Guidance Applications in Total Hip and Knee Arthroplasty: A Systematic Review,» *Arthroplasty Today*, pp. 103-112, 2021.

[15] A. Caliskan, P. P. Ajay, T. Charlesworth, R. Wolfe y M. Banaji, «Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics,» *arXiv*, pp. 1-15, 2022.

[16] Y. Shrestha y Y. Yang, «Fairness in Algorithmic Decision-Making: Applications in Multi-Winner Voting, Machine Learning, and Recommender Systems,» *Algorithms*, vol. 12, pp. 1-28, 2019.

[17] H. Chung, C. Park, W. S. Kang y J. Lee, «Gender Bias in Artificial Intelligence: Severity Prediction at an Early Stage of COVID-19,» *Front Physio*, 2021.

[18] U. Mahadeo y R. Dhanalakshmi, «Stability of feature selection algorithm: A review,» *Journal of King Saud University – Computer and Information Sciences*, p. 1060 –1073, 2022.

[19] P. S. Varsha, «How can we manage biases in artificial intelligence systems – A systematic literature review,» *International Journal of Information Management Data Insights*, pp. 1-9, 2023.

[20] A. Bhattacharya, Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more, Birmingham: Packt, 2022.