ISSN: 1692-7257 - Volume 2 – Number 44 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v2i44.3010

Exploring gender bias in Colombian occupation classification using machine learning

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

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.*

Correspondence: *aarosadog@ufpso.edu.co*

Received: *january 21, 2024.* ***Accepted:*** *june 11, 2024.* ***Published:*** *july 19, 2024.*

How to cite: *D. de J. Ramos Cuello, A. A. Rosado Gomez, and M. L. Calderón Benavides, "Exploring gender bias in Colombian occupation classification using machine learning", RCTA, vol. 2, no. 44, pp. 83–88, Jul. 2024. Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3010*

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.

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.

1. INTRODUCTION

Artificial intelligence (AI) is a powerful tool for automating complex tasks and improving efficiency in different fields. However, its rapid growth poses ethical challenges and risks, such as potential discrimination by reflecting human biases in data and decisions [1] [2] [3] [4].

Natural language processing (NLP) is a field of computer science and AI that focuses on analyzing and representing natural languages such as English, Spanish, French, etc., in the current era, which is full of unstructured data, whether textual or auditory [5]. PLN enables analysis, prognostics, and interaction with intelligent systems such as virtual assistants or autonomous vehicles and addresses challenges such as language inference, comprehension, speech recognition, and text classification [6].

The architecture of a PLN system is based on the levels that compose a human language, such as phonological, morphological, syntactic, semantic, and pragmatic. These levels provide the necessary tools for end users to communicate with the machine, and the machine can interpret their intentions [7]. In addition, PLN involves several areas, such as information extraction, sentiment analysis, research and queries, automatic synthesis, and data mining [5].

One of the evolutions of PLN is word embeddings, which represent words as real-valued vectors in a multidimensional space, where semantic embedding methods encode the meaning of words so that similar words are close in the vector space [8]. As machine learning algorithms become the decision-makers, the risk of systematic biases in these algorithms increases. Not only can they reflect human biases in the training data, but they can also amplify those biases in their practical application [4].

If a training dataset for word embeddings contains a high frequency of negative associations or bias toward a particular ethnic group, words related to that group may be closer to negative terms in the vector space, reflecting a bias. Similarly, gender-related words, such as "nurse" or "engineer," could be located in regions of the vector space that reflect gender stereotypes, which could influence the decisions of machine learning algorithms [9].

2. METHODOLOGY

The research development began with access to the Single Classification of Occupations for Colombia (CUOC) database, which contains detailed information on names of primary groups, descriptions of occupations, levels of competence, names of designations, and wording of functions.

Subsequently, vector representations of Spanish word embeddings, Word2Vec and FastText, were chosen due to their specific training in Spanish and ability to capture semantic relationships between words [10] [11]. The word embeddings were used to transform the names into vector representations, and then gender polarity was determined using the responsible library [12].

To determine whether the occupations with a greater polarity for a given gender influenced the classification, it was necessary to use the data sets of General characteristics, social security in health, and education (personal information of each person). It employed (labor information of people with a job) that is part of the National Household Survey provided by the National Administrative Department of Statistics (DANE).

All data were integrated and processed. The records used with the information provided by DANE were those associated with occupations with biased gender polarity. Classifiers were trained with these data to determine whether the texts' vector transformation can influence the models' outputs trained with tabular data.

Several steps were carried out during the data preparation process. First, a unique key was generated by concatenating variables such as Directory, SEQUENCE_P, and ORDER to identify respondents in different datasets. Then, this key was assigned to the datasets "occupied" and "general_characteristics," and readable names were mapped to the variables/columns in these sets. Subsequently, the two datasets were merged based on the single key, thus combining the information on general characteristics and occupations. A new

dataset was extracted and generated, including the ISCO (International et al. of Occupations) code of biased occupations. The datasets were then merged again using occupation identifiers. Ages were corrected by calculating missing ages and mapping numerical gender values to literal labels. Records of people under five years were also cleaned, and wages were zero-corrected using the KNNImputer algorithm on similar records of the same occupation to impute pay values [13].

Experiments with classifiers were conducted to evaluate the naming of occupations, gender and income categories, using the names as binary attributes and measuring the accuracy of each model. Finally, statistical analysis was performed with ANOVA and Tukey tests to determine the interaction between gender and processing strategies [14].

3. RESULTS

The responsible library was used to assess gender bias in occupational occupations. To assess gender bias in Spanish, gender-neutral words were identified, as many Spanish words have morphological gender markers. The resulting Table 1 shows the occupational occupations and their gender direction, which indicates whether they tend toward male (M) or female (F) based on the analysis of the words used to describe the occupation and their associated gender. In the table, it can be seen how, for most occupations, most occupations were assigned male as the closest polarity.

***Table 1**: Embedded words vs CUOC*

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

***Source**: author's elaboration*

The occupations shown in Table 1 correspond to the final ones, remaining after eliminating records with attributes with 75% of their values null. Additionally, only the occupations of a single word were taken since the library used to determine polarity only accepts this type of text. After eliminating the records, the data set comprised 3630 instances and 13 attributes. When the data were transformed, the attributes increased to 32.

The data set consisted of a total of 1745 men and 1885 women. Table 2 shows the distribution of the dataset values used for the research; columns F and M correspond to the number of records for each gender. Columns ING_F and ING_M correspond to the average income in hundreds of thousands. In this data set, there are more women than men per denomination, and there is no gender bias in income.

***Table 2**: Amounts and revenues by gender*

Designation	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

***Source**: author's elaboration*

Several experiments were carried out to study the behavior of the names. In the first, the denomination (Den) was used as the output label; in the second, the attribute "sex" was used. In a third experiment, labor income was classified into three groups (Ingr) and used as output labels. In this case, the value "zero" indicates those with incomes below one minimum wage, the value "one" means those with incomes between one and two minimum wages, and the value "three" means those above two minimum wages. The denominations were considered independent binary attributes in the first and third experiments. Table 3 shows the accuracy of each model. The best Accuracy is obtained when income is classified and used as a category. In addition, the results of the RidgeClassifierCV classifier are set as a reference or baseline.

***Table 3**: Accuracy by class*

Classifier	Den	Sex	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

Source: *author's elaboration*

The ExtraTreesClassifier (ET) and XGBClassifier (XG) algorithms were chosen after finding that the difference in accuracy between males and females was minimal. This behavior continued when using sex as the classification label. The following intervention focused on the attributes related to professions, unifying them into a single attribute represented by a vector of dimension ten, generated using Word2Vec (W2V) and FastText (FAS). As shown in Table 4, the F1 metric was established as the reference criterion to evaluate the model's effectiveness in classifying males (F1_M) and females (F1_F). According to these data, no significant differences were detected between the different interventions. Furthermore, slight increases or decreases in the values obtained with both classifiers are insufficient to conclude that one strategy is superior to the other or that a specific classifier improves the results of a given intervention. One pattern observed in the experiments is that the models classify males more accurately than females.

Table 4*: Performance by intervention*

Intervention	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

Source: *author's elaboration*

To investigate the influence of profession or its vector coding on model results, the interaction between gender and data processing strategies was examined by comparing accuracy between men and women. Through ANOVA analysis and Tukey's post-hoc tests, we explored how these interactions affect the accuracy of predictive models. Overall, the manipulation of variables related to professional designations was found to be statistically significant ($p = 0.001$), indicating that the data processing strategy influences the accuracy of the models. However, when analyzing the interaction between gender and strategy, the results did not reach the conventional level of statistical significance ($p = 0.05$), although a close value ($p = 0.051$) suggests a possible trend that, by itself, is not sufficient to establish a definitive conclusion about the existence of a significant effect.

As detailed in Table 5, data processing strategies appear to affect accuracy concerning gender, but the evidence is not conclusive in stating a significant categoric interaction. However, in two specific instances, significant differences in accuracy associated with particular combinations of gender and data processing strategy were observed: females (G1) using the 'NOR' strategy outperformed males (G2) using 'FAS' or W2V in accuracy. However, since 'NOR' makes no intervention and uses the names as binary nominal attributes (1 for presence and 0 for absence), it indicates that this difference exists from baseline and is not generated by the new experiments.

Table 5*: Gender differences and intervention*

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

Source: *author's elaboration*

4. DISCUSSION

The use of pre-trained models of Spanish word embeddings for the research sought to access a vocabulary in biased occupations. It was posited that word embeddings could associate occupations with a gender based on word endings, such as those ending in "-a" with the feminine, so it was chosen to include all occupations regardless of their "sexual direction," given that the data obtained included men and women in occupations traditionally associated with the male gender [15].

Data analysis showed an imbalance in gender representation, with an inclination of the artificial intelligence models to predict more accurately for the male gender. Although the data distribution might predispose to a bias toward males, the models did not evidence a conclusive preference in their predictions [16] [17]. The results indicate that applying these models in the Colombian labor market does not necessarily lead to gender discrimination. On the other hand, the performance

of the initial tabular dataset, without the application of vector representations, was superior in classifying male profiles compared to female profiles [18]. Models that implemented transformations in the occupational designations did not significantly modify this behavior, suggesting that the biases present in the vector representations did not carry over directly to the tabular datasets, contradicting the hypothesis that pre-existing discrimination could be perpetuated by the integration of these biased representations [19] [20].

5. CONCLUSIONS

Using pre-trained models of Spanish word embeddings to analyze gender bias in occupational occupations revealed that, although specific patterns of bias such as gender association based on Spanish word endings were considered, including all occupations regardless of their "sexual direction" did not affect model outputs. This is because the reality of work includes both men and women in roles traditionally associated with a specific gender, suggesting that any automation or AI-driven decisions should be carefully examined to avoid perpetuating gender stereotypes.

Although the data showed an imbalance in gender representation, with a slight tendency for AI models to predict more accurately for the male gender, no absolute bias in predictions was found. This suggests that gender discrimination in AI is not ubiquitous and may vary depending on the specific dataset and how these word embeddings are processed and used in machine learning models.

Experiments with different classifiers and data processing strategies revealed that the biases present in the vector representations did not carry over directly to the tabular data sets. Furthermore, statistical analysis with ANOVA and Tukey tests showed that, although manipulation of variables related to professional designations had a statistically significant effect on the results, the initial hypothesis that pre-existing discrimination would be automatically inherited in the AI when integrating these biased representations was not corroborated.

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.