Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 211-222 recibido 05-04-2025 revisado 12-05-2025 aceptado 19-05-2025

Preserving Grammatical Gender when Debiasing Word Embeddings in Spanish

Propuesta de método de conservación del género gramatical en embeddings de palabras para la mitigación de sesgos en español

Aitana Morote Martínez,[1] **Juan Pablo Consuegra-Ayala,**[1] **Elena Lloret**[2]
[1]Digital Intelligence Centre (CENID), University of Alicante, Alicante, Spain
[2]Department of Language and Computing Systems, University of Alicante, Alicante, Spain
{aitana.morote, juan.consuegra}@ua.es
elloret@dlsi.ua.es

Abstract: Word embeddings are widely used in Natural Language Processing but often encode gender biases, which can lead to discriminatory outcomes. Various debiasing techniques exist, especially focusing on English, thus failing to account for the complexities of languages with grammatical gender, such as Spanish. In this paper, we propose INLP-Gram, an algorithm designed to mitigate gender bias in Spanish word embeddings while preserving grammatical gender information. It is an adaptation of the Iterative Nullspace Projection (INLP). We evaluate INLP-Gram using the Word Embedding Association Test (WEAT) and a grammatical gender classification test. Our results demonstrate that INLP-Gram effectively reduces gender bias while maintaining grammatical gender distinctions. This work advances bias mitigation techniques for word embeddings in morphologically-rich languages.
Keywords: bias mitigation, bias evaluation, gender, word embeddings.

Resumen: Los *word embeddings* son ampliamente utilizados en el Procesamiento del Lenguaje Natural, pero a menudo codifican sesgos de género, lo que puede dar lugar a resultados discriminatorios. Existen varias técnicas de mitigación de sesgos (*debiasing*), centradas en el inglés, que no tienen en cuenta las complejidades de las lenguas con género gramatical como el español. Este artículo presenta INLP-Gram, un algoritmo diseñado para mitigar el sesgo de género en *embeddings* en español que es capaz de conservar la información de género gramatical. Nuestro algoritmo es una adaptación del algoritmo INLP (Iterative Nullspace Projection), pero teniendo en cuenta las variaciones morfológicas de idiomas con género gramatical. Evaluamos INLP-Gram mediante el Word Embedding Association Test (WEAT) y una prueba de clasificación del género gramatical. Nuestros resultados demuestran que INLP-Gram reduce efectivamente el sesgo de género a la vez que mantiene las distinciones gramaticales de género. Este trabajo supone un avance en las técnicas de mitigación de sesgos para *word embeddings* en lenguas con riqueza morfológica.
Palabras clave: mitigación de sesgos, evaluación de sesgos, género, word embeddings.

1 Introduction

Large Language Models (LLMs) like GPT are gaining widespread attention (Hadi et al., 2025; Meng et al., 2024; Zhao et al., 2024; Casheekar et al., 2024). These models have advanced the field of Natural Language Processing (NLP) by providing powerful tools for understanding and generating text. However, even as LLMs grow in popularity, *word embeddings* (Johnson, Murty, and Navakanth, 2024) continue to play an important role in NLP, particularly in terms of computational efficiency and their ability to capture semantic relationships in a straightforward manner. Their simplicity and ability to capture relationships between words make them valuable in many applications (Zulfiqar et al., 2024; Sun et al., 2024).

Despite their usefulness, machine learning models often reflect and reinforce biases that

ISSN 1135-5948 DOI 10.26342/2025-75-16

exist in the data used to train them (Brunet et al., 2019; Caliskan et al., 2022). These biases can lead to unfair or harmful outcomes when the models are applied in real-world systems (Consuegra-Ayala et al., 2025). Word embeddings, in particular, can encode biases related to gender, race, and other social factors (Angwin et al., 2016; Caliskan, Bryson, and Narayanan, 2017). This issue has motivated significant research aimed at measuring and reducing bias in word embeddings to make Artificial Intelligence (AI) systems more fair and reliable.

Most efforts to address gender bias in word embeddings have focused on languages like English, which is not a grammatical gendered language (Zhou et al., 2019). However, many languages, such as Spanish, encode gender directly in their grammar. This makes it challenging to separate grammatical gender (Omrani Sabbaghi and Caliskan, 2022) from semantic gender which is prone to societal biases, which are unfair and should be removed. Solving this problem is important to ensure that AI systems work fairly across different languages.

In this paper, we present a method to reduce gender bias in Spanish word embeddings while keeping useful grammatical information. We adapt an existing technique for bias mitigation —namely, Iterative Nullspace Projection (INLP)— to address the challenges of languages with grammatical gender. Our goal is to remove harmful associations related to gender while preserving the information needed for linguistic tasks.

The main contributions of this work are:

- A new method for reducing gender bias in Spanish word embeddings (INLP-Gram).
- The creation of a dataset of Spanish words to aid the evaluation and improvement of gender bias mitigation techniques in Spanish.
- A detailed evaluation of the proposed method, showing its effectiveness and exploring how it handles grammatical and societal gender information.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on word embeddings, bias quantification, and mitigation techniques. Section 3 outlines our proposed methodology. Sections 4, 5 and 6 detail the experimental resources, setup and results, along with a discussion of key findings. Finally, Section 7 summarizes our contributions and highlights directions for future research.

2 Related Work

This section reviews prior research relevant to word embeddings, the quantification of bias in such embeddings, and methods proposed for mitigating these biases. We also position our contribution within this context by identifying gaps and opportunities for improvement.

2.1 Word Embeddings

Word embeddings are dense, low-dimensional representations of words in a continuous vector space. These representations capture semantic and syntactic relationships between words, making them foundational in many NLP tasks. Notable examples include *word2vec* (Mikolov et al., 2013), which leverages skip-gram and continuous bag-of-words models; *GloVe* (Pennington, Socher, and Manning, 2014), which constructs embeddings based on global co-occurrence statistics; and *fastText* (Bojanowski et al., 2017), which enriches embeddings with subword information.

These techniques have significantly advanced the field of NLP but have also been found to encode and propagate societal biases present in training data, motivating research into bias quantification and mitigation.

2.2 Quantification of Bias in Word Embeddings

Bias in word embeddings has drawn considerable attention due to its potential to propagate harmful stereotypes in AI systems. A key metric for quantifying bias is the *Word Embedding Association Test* , *WEAT* (Caliskan, Bryson, and Narayanan, 2017), which evaluates the association between target and attribute word sets to measure implicit biases encoded in embeddings. This method reveals biases related to gender, race, and other social constructs embedded within word vectors.

Building upon WEAT, Lauscher et al. (2020) introduced a framework that expands bias evaluation by integrating both existing and novel metrics, enabling a more comprehensive assessment of implicit and explicit biases. These contributions highlight the need for robust methodologies to identify and quantify bias, providing a foundation for subsequent mitigation techniques.

2.3 Mitigation of Bias in Word Embeddings

One of the earliest and most well-known methods to mitigate bias in word embeddings is *Hard Debias* (Bolukbasi et al., 2016). This approach identifies a bias subspace within the embedding space and then projects biased words onto an orthogonal direction to this subspace, effectively reducing gender associations while retaining semantic coherence.

Another significant contribution is *Iterative Nullspace Projection (INLP)* (Ravfogel et al., 2020). This technique iteratively removes linearly encoded bias by nullifying its representation across multiple dimensions. While effective, INLP requires several projection steps, which can disrupt other semantic properties of the embeddings.

To address some limitations of INLP, the *Mean Projection (MP)* method (Haghighatkhah et al., 2022) simplifies the process by requiring only a single projection. MP achieves comparable results with reduced computational complexity and minimal impact on non-biased dimensions.

OSCAR (Orthogonal Subspace Correction and Rectification) (Dev et al., 2021) introduced a method that rectifies identified subspaces by orthogonalizing them while preserving the integrity of unaffected embeddings. By selectively stretching or preserving embeddings outside the subspace, OSCAR provides a nuanced approach to mitigating bias.

For languages with grammatical gender, Zhou et al. (2019) explores how linguistic structures impact embedding biases. This study proposes specialized methods for mitigating gender bias in both monolingual and bilingual word embeddings, balancing bias reduction with the preservation of embedding quality.

Lastly, Takeshita et al. (2020) examine the challenges of adapting bias mitigation techniques to languages other than English. This work emphasizes that linguistic and cultural differences, such as grammatical gender and cultural norms, need tailored approaches to debiasing embeddings, as methods designed for English often fail to generalize effectively.

2.4 Our outlook: limitations and opportunities

Our proposed approach, *INLP-Gram*, builds on existing methods to address bias in embeddings for languages with grammatical gender, particularly Spanish. It adapts INLP with modifications inspired by Zhou et al. (2019) to handle grammatical gender. INLP-Gram also incorporates principles from *MP* and *OSCAR*, preserving harmless information by ensuring the grammatical gender subspace is orthogonalized without significant impact on unrelated features.

Existing research reveals several limitations and opportunities for further exploration:

- Methods for debiasing embeddings in Spanish often rely on unadapted approaches designed for English, leading to suboptimal performance in downstream tasks such as machine translation (e.g., Escudé Font and Costa-jussà (2019)).
- Previous adaptations for Spanish precede INLP and employ less efficient techniques, leaving room for improvement.
- Although improvements in INLP aim to retain valid information, they rarely address the unique challenges posed by grammatical gender.

By addressing these gaps, INLP-Gram aims to advance the state-of-the-art in bias mitigation for morphologically-rich languages.

3 INLP-Gram Method

Our INLP-Gram method mitigates the gender bias of embeddings while preserving grammatical gender information. The mitigation of gender bias is performed on the basis of the Iterative Nullspace Projection (INLP) method.

Let us start by formally defining the INLP method. Given a set of embeddings $x_i \in X$ and a set of corresponding discrete attributes $z_i \in Z$ (e.g., gender), INLP computes a transformation g such that z_i cannot be predicted from $g(x_i)$, achieving “linear guarding” of the attribute. Internally, INLP trains at each iteration a linear classifier c, specifically a linear Support Vector Machine (SVM). Classifier c is parameterized by a matrix W, which predicts an attribute z_i. Then, a projection matrix P is constructed such that $W(Px) = 0$ for all x, making W useless in the dataset X. This process is repeated until the accuracy of c is sufficiently low. The construction of P is achieved via nullspace projection (Ravfogel et al., 2020).

The original INLP is not suitable for removing gender from languages that have grammatical gender. These languages are classified as *grammatical gender languages* in the work of Prewitt-Freilino, Caswell, and Laakso (2011), which provides a taxonomy of languages according to gender expression. INLP does not differentiate between semantic and grammatical gender in its multiple projections, and negatively affects grammatical gender. That is, when the protected attribute is gender in grammatical gender languages, it is very likely that some directions that are removed from the embeddings encode grammatical gender.

INLP-Gram compensates for this deficiency by modifying, at each iteration, the trained SVM parameters (matrix W) which represent gender information found at the iteration. W is made linear independent to a direction encoding grammatical gender, thus protecting the embedding grammatical gender information.

Figure 1 illustrates the INLP-Gram method. Steps 1 and 4 belong to the INLP algorithm and represent (1) the training of a linear SVM classifier to obtain a gender subspace corresponding to the model parameters W_i and (4) projecting the embeddings onto the nullspace of the gender subspace. Steps 2, 3 and 5 are novel steps related to the preservation of grammatical gender, that is, (2) the computation of the grammatical gender direction G_i, (3) the modification of the iteration gender subspace (W_i is modified to S_i), and (5) the projection of the set of word embeddings used to compute the next iteration grammatical gender direction, noted with $X'_{projected}$. These steps will be explained in more detail in the sections 3.1, 3.2 and 3.3.

3.1 Computation of the grammatical gender direction

Grammatical gender is an inherent property of nouns and pronouns that produces effects in agreement with determiners, quantifiers, adjectives, and other classes of words. For animate nouns, the grammatical gender generally aligns with the semantic gender of the referent (for example, *"el gato"* and *"la gata"* are masculine and feminine forms of the animate noun *"the cat"*). However, inanimate nouns are also assigned a grammatical gender, demonstrating that the grammatical gender possesses features distinct from semantic gender (for example, *"the chair"* is translated by *"la silla"*, which has feminine grammatical gender).

The grammatical gender direction, which we note with $\vec{d_g}$, is a direction in the embedding space that encodes grammatical gender (Zhou et al., 2019). The method we used to calculate $\vec{d_g}$ is *Linear Discriminant Analysis (LDA)* (Li and Wang, 2014). LDA seeks to find a linear combination of features that maximizes the separation between classes while minimizing the scatter within each class. This is achieved by solving a generalized eigenvalue problem (Gambella, Ghaddar, and Naoum-Sawaya, 2021).

The word sets for each class in the LDA training dataset are systematically chosen as a combination of singular and plural nouns and adjectives of each grammatical gender. Although the neuter gender exists in grammatical gender languages such as Spanish, to simplify, we consider two classes that define grammatical gender: the feminine class and the masculine class.

3.2 Computation of the semantic gender direction

The semantic gender direction, which we will refer to with $\vec{d_s}$, encodes the gender of the animate being to which the word refers. In works focused on English (Bolukbasi et al., 2016), this direction has been predominantly obtained by computing the principal component with *Principal Component Analysis (PCA)* (Gorban et al., 2008) of two sets of definitional gendered words.

However, this definition of semantic gender is not appropriate for grammatical gender languages, as translations into these languages of the sets of gendered words that are typically used for computing the semantic gender direction (Bolukbasi et al., 2016) are not exempt from grammatical gender, so the principal component is not purely semantic and also encodes grammatical information. In our work, we adopt the approach and word sets of Zhou et al. (2019), who already provides a proposal for Spanish.

Specifically, let us define three gender directions, the gender direction obtained with PCA applied to definitional word pairs, $\vec{d}_{PCA}$, the grammatical gender direction, $\vec{d_g}$, and the semantic gender direction, $\vec{d_s}$. As our objective is to compute $\vec{d_s}$ as a direction with minimal grammatical gender information, we make $\vec{d}_{PCA}$ orthogonal with respect to $\vec{d_g}$ by removing its projection onto the grammati-

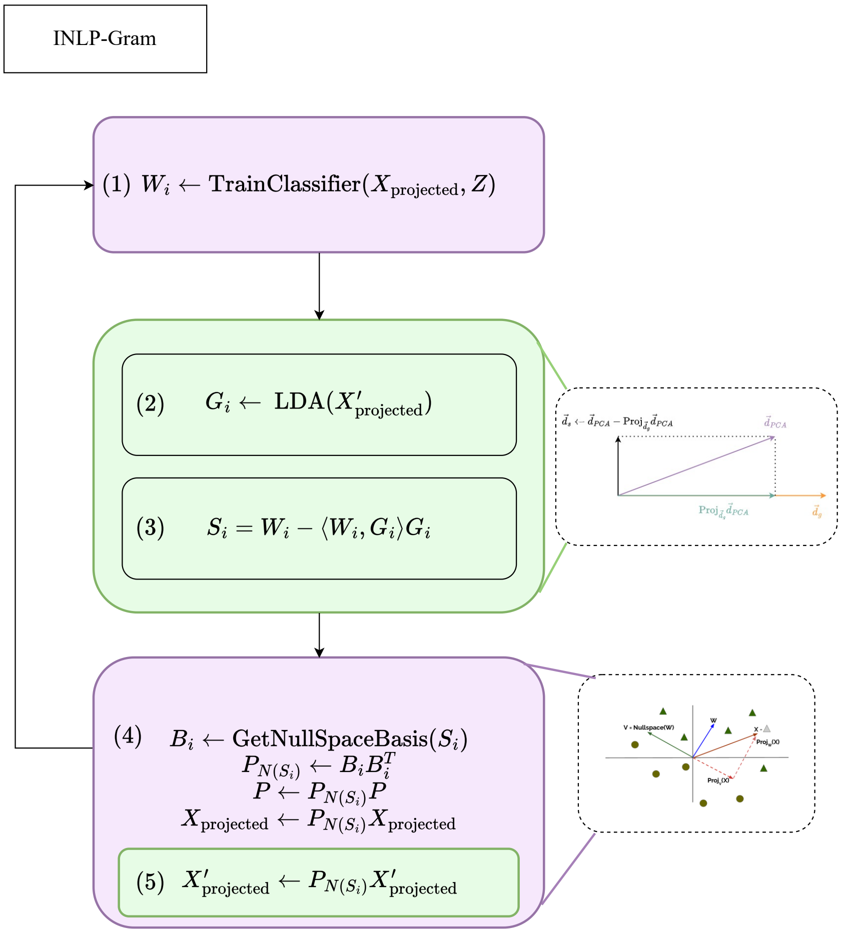
INLP-Gram
(1) $W_i \leftarrow \text{TrainClassifier}(X_{\text{projected}}, Z)$
(2) $G_i \leftarrow \text{LDA}(X'_{\text{projected}})$
(3) $S_i = W_i - \langle W_i, G_i \rangle G_i$
(4) $B_i \leftarrow \text{GetNullSpaceBasis}(S_i)$
$P_{N(S_i)} \leftarrow B_i B_i^T$
$P \leftarrow P_{N(S_i)} P$
$X_{\text{projected}} \leftarrow P_{N(S_i)} X_{\text{projected}}$
(5) $X'_{\text{projected}} \leftarrow P_{N(S_i)} X'_{\text{projected}}$
$\vec{d}_s \leftarrow \vec{d}_{PCA} - \text{Proj}_{\vec{d}_g} \vec{d}_{PCA}$
$\vec{d}_{PCA}$
$\text{Proj}_{\vec{d}_g} \vec{d}_{PCA}$
$\vec{d}_g$
V = Nullspace(W)
W
X
Proj_v(X)
Proj_w(X)

Figure 1: Steps of the INLP-Gram method. G_i and S_i represent the iteration grammatical and semantic gender directions respectively. In step 4 the initial gender subspace W_i has been replaced by the obtained semantic gender subspace S_i.

cal direction. We use the formula introduced by Zhou et al. (2019):

$$\vec{d_s} = \vec{d}_{PCA} - \left\langle \vec{d}_{PCA}, \vec{d_g} \right\rangle \vec{d_g}$$

where $\left\langle \vec{d}_{PCA}, \vec{d_g} \right\rangle$ represents the inner product between directions and $\left\langle \vec{d}_{PCA}, \vec{d_g} \right\rangle \vec{d_g}$ is equivalent to $\text{Proj}_{\vec{d_g}} \vec{d}_{PCA}$, the projection of $\vec{d}_{PCA}$ onto $\vec{d_g}$.

This calculation is illustrated in Figure 2.

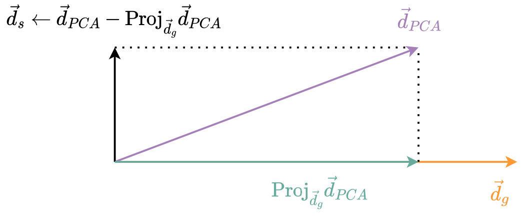
$\vec{d}_s \leftarrow \vec{d}_{PCA} - \text{Proj}_{\vec{d}_g} \vec{d}_{PCA}$
$\vec{d}_{PCA}$
$\text{Proj}_{\vec{d}_g} \vec{d}_{PCA}$
$\vec{d}_g$

Figure 2: Computation of the semantic gender direction $\vec{d_s}$. $\vec{d_s}$ is orthogonal to $\vec{d_g}$. $\text{Proj}_{\vec{d_g}} \vec{d}_{PCA}$ is the projection of $\vec{d}_{PCA}$ onto $\vec{d_g}$.

3.3 Adaptation of INLP

As explained in Section 3, the INLP algorithm performs iterative linear projections of the embedding space into the nullspace of a protected attribute, such as gender. In each iteration, the protected attribute subspace is determined by the features of a linear classifier. The issue is that, as iterations progress, the classifier may start relying on data features other than the protected attribute, thereby removing valid information from the embeddings. As mentioned earlier, the grammatical and semantic genders are concordant for words referring to animate beings, so they encode similar information. Consequently, both directions share a common subspace, and INLP may mistakenly remove grammatical gender information from the embeddings.

Our method consists of combining the processes introduced in Sections 3.1 and 3.2 with the INLP algorithm. More precisely, at each iteration of INLP, the grammatical gender direction is computed (in the way described in Section 3.1) over the embeddings of the iteration. Then, the gender subspace is obtained by removing the grammatical gender component of each of the directions composing it,[1] using the formula of Section 3.2. As we run INLP with an auto-regressive configuration, at each iteration, the embeddings are projected onto the intersection of all nullspaces computed until the current iteration. In INLP-Gram, two sets of embeddings are projected at each iteration, the embeddings used to train the SVM classifier and the embeddings used to compute the grammatical gender direction with LDA. With this process, grammatical gender information can be protected while benefiting from the effectiveness of the INLP method to mitigate gender bias.

4 Resources

This section presents the resources that allowed us to perform the experimentation for our approach, including the source code for our approach (Section 4.1), the language resources used (Section 4.2), the embeddings employed in the experiments (Section 4.3), the libraries (Section 4.4) and the hardware configuration (Section 4.5).

4.1 Source code

Our source code is available at `https://github.com/amm533/INLP_Gram.git`.

We benefited from the source code of Ravfogel et al. (2020), Zhou et al. (2019) and Caliskan, Bryson, and Narayanan (2017) to reproduce their approaches and replicate their experiments.

4.2 Language resources

For Spanish experiments regarding grammatical gender, word sets are selected from the resource CORPES XXI (Real Academia Española, n.d.). This resource allows to filter a word search by grammatical category, gender (masculine, feminine or neuter) and number (singular or plural), among other parameters, allowing to create customized datasets to encode the grammatical gender. In particular, two datasets were constructed using this resource: one for computing the grammatical gender direction with Linear Discriminant Analysis, consisting of approximately 15,000 words per gender (see Section 3.1); and another for performing the grammatical gender classification test, composed of around 100 words per gender (see Section 5.2.2).

In addition, the sets of profession words used in the visualization experiments were selected from the resource *Las profesiones de la A a la Z*.[2]

The datasets with grammatical gender word sets and profession words are released at `https://github.com/amm533/INLP_Gram-wordsets.git`.

4.3 Embeddings

For experiments, we used Fasttext embeddings from Spanish Billion Word Corpus (SBWC) (Cardellino, 2019), which were trained on a large number of resources.

4.4 Libraries

Two main libraries were used to facilitate the experimentation: `gensim` (Řehůřek and Sojka, 2010) and `scikit-learn` (Pedregosa et al., 2011). More specifically, we used the data structure `gensim.models.KeyedVectors` to store word embeddings and interact with them, and we used `scikit-learn` functions for learning and evaluating directions in the data.

4.5 Hardware

Experiments were carried out on a machine with the following details: 8-core Intel Core i7-8565U (-MT-MCP-) CPU, speed/max: 1990/4600 MHz, cache: 8000 KB, and RAM: 16 GB.

5 Experimental setup and evaluation

Sections 5.1 and 5.2 present the evaluation baselines and scenarios for the experimentation.

5.1 Evaluation baselines

We consider two baselines: (1) the initial Spanish fastText embeddings before applying any bias mitigation method and (2) the embeddings after debiasing with the original INLP method.

[1] INLP defines gender with 3 classes, masculine, feminine and neutral, so the gender subspace computed at each iteration is spanned by 3 directions.

[2] `https://www.inmujeres.gob.es`

5.2 Evaluation scenarios

Three scenarios were defined to evaluate INLP-Gram: i) the *Word Embedding Association Test (WEAT)* (Caliskan, Bryson, and Narayanan, 2017) (Section 5.2.1); ii) the grammatical gender classification test to evaluate the grammatical gender preservation in the embeddings (Section 5.2.2) and iii) a visual evaluation of the projection of occupation words onto a gender direction (Section 5.2.3).

5.2.1 WEAT

To evaluate the quality of debiasing, we use the *Word Embedding Association Test (WEAT)* (Caliskan, Bryson, and Narayanan, 2017), which measures the relative association between two sets of target words and two sets of attribute words based on cosine similarity. The objective is to learn if target words are biased with respect to attribute words. In the WEAT paper (Caliskan, Bryson, and Narayanan, 2017), attribute words are not always the words encoding gender. Experiments in Table 1 are replicated from the paper. In experiments (2) and (3) attribute words are gender terms, whereas in experiment (1) gender names are objective words whose association to family or career is tested. The null hypothesis states no difference in association between the target sets.

This test provides three metrics. Firstly, the *WEAT test statistic* measures the difference in the association of the two sets of target words with the attribute. Secondly, the *effect size* is a measure of practical significance of the test statistic. It is calculated in terms of *Cohen's d*, that is, the standardized difference between the mean association of the target sets with the attribute. According to Cohen, small, medium and large values of the effect size are 0.2, 0.5, and 0.8, respectively (Sullivan and Feinn, 2012). Finally, the p-value measures the probability that a random permutation of the attribute words would produce at least the observed test statistic. If the p-value is less than 0.01, which is the threshold used by Caliskan, Bryson, and Narayanan (2017), then the difference between targets is statistically significant.

These measures are formulated below, where X and Y represent the sets of target words, and A and B the sets of attribute concepts.

Effect size

$$\frac{\text{mean}_{x\in X} s(x,A,B) - \text{mean}_{y\in Y} s(y,A,B)}{\text{std\_dev}_{w\in X\cup Y} s(w,A,B)}$$

p-value

$$\Pr_i\left[s(X_i,Y_i,A,B) > s(X,Y,A,B)\right]$$

where

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

and

$$s(w,A,B) = \text{mean}_{a\in A} cos(\vec{w},\vec{a}) - \text{mean}_{b\in B} cos(\vec{w},\vec{b})$$

We replicate the three gender-related tests in the WEAT paper, more precisely dealing with the gender association with *career-family*, *science-arts*, and *mathematics-arts*. As the source code was available in Java, we ported it to Python.

5.2.2 Grammatical gender classification test

To evaluate the preservation of grammatical gender, we developed a test based on grammatical gender classification using a logistic regression model. More concretely, our test measures the amount of grammatical gender information in the embeddings by evaluating the effectivity of a logistic regression classifier in separating words with masculine or feminine grammatical gender. These words are balanced combinations of singular and plural adjectives and nouns. It is a different and smaller set than the one used to compute the grammatical gender direction.

To ensure statistical significance of the difference in the performance of the logistic regression model in different scenarios, we obtain the accuracies over 100 runs of training of the model, each with different training and test splits. We then perform *two sample T-tests* (Snedecor and Cochran, 1989) using the lists of accuracies in two different scenarios as the two populations. The null hypothesis of this statistical test is that the two population means are equal. We use a significance level of 0.05. We also compute the effect sizes to quantify the difference between the mean accuracies. This test indicates whether there is a change in the quality of the grammatical gender information in the embeddings between scenarios.

5.2.3 Visualization of occupation words

The previous evaluation scenarios allow us to quantify gender bias and grammatical gender preservation in the embeddings, but visualizing the gender component of the embeddings can help better understand the specific effect of the different debiasing methods. Therefore, we present three plots in Figures 3, 4 and 5 showing the projection of embeddings corresponding to occupation words onto a gender direction computer with PCA over gender definitional pairs ($\vec{d_{pca}}$). Profession words are convenient because they embody both semantic and grammatical gender, making it possible to draw inferences about the effect of the methods on both types of gender. The ten occupations displayed on each graph are chosen as the most biased in each case, but the means are computed with a larger set of occupations. For each occupation, four forms are projected and visualized, corresponding to the gender and number inflections of the profession (for example, *"actriz", "actor", "actrices", "actores"*).

The main interest of this experiment is that, as suggested in the work of Zhou et al. (2019), gender bias is reflected in the symmetry of male and female forms with respect to the origin, representing neutrality. As commented before, $\vec{d_{pca}}$ represents a direction predominantly encoding semantic gender but also containing grammatical gender in some measure, so the optimal scenario would place singular and plural gender pairs symmetric to each other (e.g. *"actriz"* symmetric to *"actor"*, *"actrices"* symmetric to *"actores"*).

6 Results and Discussion

The results of the experiments are summarized in Tables 1, 2, and 3. Table 1 shows the results for the three gender-related experiments of WEAT. Tables 2 and 3 present the results of the grammatical gender classification task. The values reported in Table 2 represent the average accuracies over 100 runs of a logistic regression model, each with different training and test splits. In Tables 1 and 3, d represents the effect sizes and p the p-values of the statistic tests. The results of projection visualization are discussed in Section 6.3.

6.1 WEAT

The WEAT measures allow us to quantify the gender bias present in the embeddings. Table 1 collects the effect sizes (d) and p-values (p) of different experiments. The higher the effect size value, the greater the bias detected, and $p - \text{value} < 0.01$ suggests that the bias is statistically significant.

The results show that, for the three experiments, the effect size achieves conventionally high (higher than 0.8) values in initial embeddings without debiasing, and p-values are below or near the threshold or 0.01, revealing a gender bias in these embeddings. INLP and INLP-Gram obtain lower effect sizes with respect to initial embeddings, and the p-values are increased to a statistically insignificant level, indicating that both methods have a positive impact in reducing gender bias. Furthermore, in experiments on associations of gender with career-family and math-arts, INLP-Gram shows better results (lower effect size and higher p-value) than INLP.

6.2 Grammatical gender classification test

Tables 2 and 3 reveal that INLP negatively affects the grammatical gender, as the mean accuracy of the logistic regression model drops from 82.2% to 74.57%. This decrease is statistically significant with a high effect size, indicating that using INLP for debiasing embeddings in grammatical gendered languages erases key grammatical gender cues in embeddings.

Fortunately, INLP-Gram recovers this loss of accuracy. In debiased embeddings with INLP-Gram the accuracy of the classification model is even improved (82.2% to 85.85%) compared to embeddings without debiasing. This increase is shown to be statistically significant with respect to the measure for both baselines, showing the effectivity of INLP-Gram in preserving grammatical gender.

6.3 Visualization of occupation words

Figures 3, 4 and 5 give us additional information about the effect of INLP-Gram with respect to the baselines.

In Figure 3, initial embeddings of masculine and feminine profession forms are asymmetrically distributed, with masculine forms closer to the origin, associating masculinity with a neutral human concept, as noted in prior studies (Costa Jussa et al., 2023).

INLP (Figure 4) reduces this bias by decreasing the mean gender difference of pro-

Target words	Attribute words	Without debiasing		INLP		INLP-Gram	
		d	p	d	p	d	p
(1) Male vs. female names	Career vs. family	1.75	<0.001	0.98	0.028	**0.82**	**0.05**
(2) Math vs. arts	Male vs. female terms	1.16	0.009	0.53	0.16	**0.25**	**0.32**
(3) Science vs. arts	Male vs. female terms	1.06	0.020	**0.16**	**0.39**	0.30	0.28

Table 1: Gender debiasing results on WEAT. d represents the effect size and p the p-value of the test statistic.

Without debiasing	INLP	INLP-Gram
82.2%	74.57%	**85.85%**

Table 2: Grammatical gender test results. Percentage of accuracy of a logistic regression model.

Scenario comparison	d	p
Without debias vs. INLP	1.80	<0.001
Without debias vs. INLP-Gram	0.73	5.94×10^{-7}
INLP vs. INLP-Gram	2.61	6.38×10^{-45}

Table 3: Grammatical gender test results significance. Effect sizes (d) and p-values (p) of the two sample T-test between scenarios.

jections. However, another side effect can be observed: before applying INLP (Figure 3), the different grammatical inflections were reflected in the sign (feminine forms have negative projections) and absolute value (singular forms have a higher absolute value) of the projections, but INLP seems to disrupt this pattern.

The plot obtained with INLP-Gram (Figure 5) appears to be a middle ground between both baselines. Although some gender asymmetry remains, it is reduced compared to the original embeddings (Figure 3), and grammatical distinctions are preserved: the feminine forms have negative projections, the masculine forms are positive and the singular forms show stronger absolute values than the plurals. This suggests that the conceptual difference between the grammatical forms that characterizes the grammatical gender languages and differentiates them from English seems to have been protected by our method.

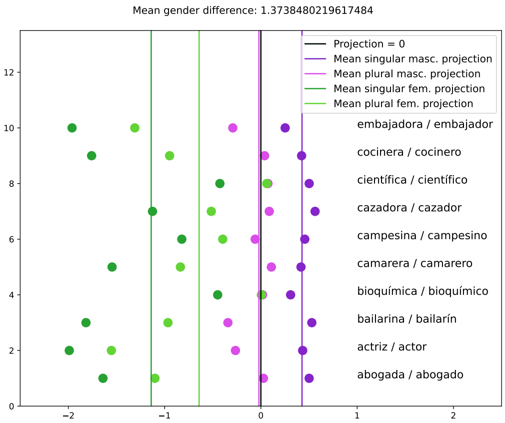

Figure 3: Projection of Spanish professions onto $\vec{d}_{pca}$. Embeddings before debiasing. Green points correspond to feminine forms and violet points to masculine forms.

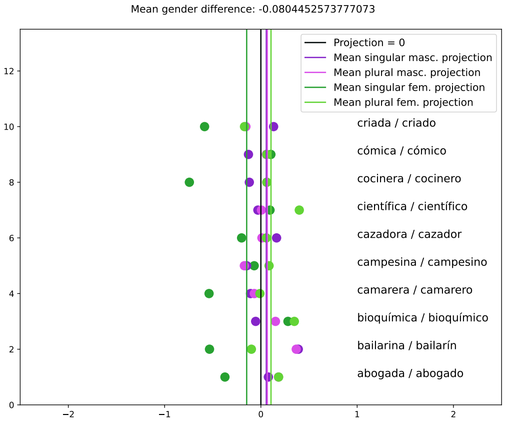

Figure 4: Projection of Spanish professions onto $\vec{d}_{pca}$. Embeddings debiased with INLP.

7 *Conclusion and Future Work*

We presented INLP-Gram as a debiasing algorithm grounded on the INLP method. As INLP is not targeted at grammatical gender languages, we propose an approach to protect grammatical gender for the case of Spanish integrated into INLP.

The quantitative evaluation of gender bias and grammatical gender shows that our method does not interfere with the mitigation of gender bias while maintaining grammatical gender information in the embeddings. Visualization experiments suggest that INLP-Gram may retain some gender asymmetry with respect to INLP, but this bias is greatly reduced

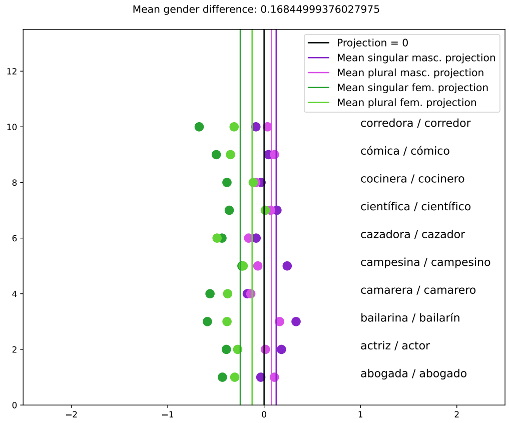

Figure 5: Projection of Spanish professions onto $\vec{d}_{pca}$. Embeddings debiased with INLP-Gram.

with respect to initial embeddings, which is supported by the WEAT quantitative tests.

This work also provides an analysis of how grammatical gender is encoded in embeddings, revealing the high correlation between semantic and grammatical gender in the embedding space, as well as the algorithms and word sets that allow us to effectively (though not completely) isolate grammatical gender from semantic gender.

In future work, we will examine the behavior and appropriate adjustments of INLP-Gram in other grammatical gender languages, particularly in French and Valencian, a low-resource Spanish dialect. Similarly, our method will be tested with other ways of computing embeddings, among which contextualized embeddings are specially rellevant. Another future direction focuses on the method proposed by Zhou et al. (2019) of shifting gendered words with respect to an anchor point, which could be combined with our method. We will also evaluate the approach in a wider framework including some extrinsic metrics to check the real behavior of debiased embeddings in downstream tasks.

Acknowledgments

The research work conducted is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; "CLEAR.TEXT:Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; "VIVES: "Pla de Tecnologies de la Llengua per al valencià" project (2022/TL22/00215334) from the Ministerio de Transformación Digital y por el Plan de Recuperación, Transformación y Resiliencia – Financiado por la Unión Europea – NextGenerationEU; and the project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021)" funded by the Generalitat Valenciana.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91.

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

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

Brunet, M.-E., C. Alkalay-Houlihan, A. Anderson, and R. Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Caliskan, A., P. P. Ajay, T. Charlesworth, R. Wolfe, and M. R. Banaji. 2022. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

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

Cardellino, C. 2019. Spanish Billion Words Corpus and Embeddings, August. Available at `https://crscardellino.github.io/SBWCE/`.

Casheekar, A., A. Lahiri, K. Rath, K. S. Prabhakar, and K. Srinivasan. 2024. A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review*, 52:100632.

Consuegra-Ayala, J. P., Y. Gutiérrez, Y. Almeida-Cruz, and M. Palomar. 2025. Bias mitigation for fair automation of classification tasks. *Expert Systems*, 42(2):e13734.

Costa Jussa, M., P. Andrews, E. Smith, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, D. Licht, and C. Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore, December. Association for Computational Linguistics.

Dev, S., T. Li, J. M. Phillips, and V. Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Escudé Font, J. and M. R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August. Association for Computational Linguistics.

Gambella, C., B. Ghaddar, and J. Naoum-Sawaya. 2021. Optimization problems for machine learning: A survey. *European Journal of Operational Research*, 290(3):807–828.

Gorban, A., B. Kégl, D. Wunsch, and A. Zinovyev. 2008. *Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE (Lecture Notes in Computational Science and Engineering) 58*. 01.

Hadi, M. U., Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, S. Z. Hassan, M. Shoman, J. Wu, S. Mirjalili, and M. Shah. 2025. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv*, February.

Haghighatkhah, P., A. Fokkens, P. Sommerauer, B. Speckmann, and K. Verbeek. 2022. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8395–8416, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Johnson, S. J., M. R. Murty, and I. Navakanth. 2024. A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, 83(13):37979–38007.

Lauscher, A., G. Glavaš, S. P. Ponzetto, and I. Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8131–8138.

Li, C. and B. Wang. 2014. Fisher linear discriminant analysis. Lecture notes, College of Computer and Information Science, Northeastern University.

Meng, X., X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang, et al. 2024. The application of large language models in medicine: A scoping review. *Iscience*, 27(5).

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Omrani Sabbaghi, S. and A. Caliskan. 2022. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pennington, J., R. Socher, and C. Manning. 2014. GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Prewitt-Freilino, J., T. A. Caswell, and E. Laakso. 2011. The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles*, 66, 02.

Ravfogel, S., Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July. Association for Computational Linguistics.

Real Academia Española. n.d. Banco de datos (CORPES XXI) [en línea]. `http://www.rae.es`.

Řehůřek, R. and P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Snedecor, G. W. and W. G. Cochran. 1989. *Statistical Methods*. Iowa State University Press, Ames, Iowa, 8th edition.

Sullivan, G. M. and R. Feinn. 2012. Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282, September.

Sun, G., Y. Cheng, Z. Zhang, X. Tong, and T. Chai. 2024. Text classification with improved word embedding and adaptive segmentation. *Expert Systems with Applications*, 238:121852.

Takeshita, M., Y. Katsumata, R. Rzepka, and K. Araki. 2020. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55.

Zhao, H., Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, et al. 2024. Revolutionizing finance with LLMs: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

Zhou, P., W. Shi, J. Zhao, K.-H. Huang, M. Chen, R. Cotterell, and K.-W. Chang. 2019. Examining gender bias in languages with grammatical gender. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China, November. Association for Computational Linguistics.

Zulfiqar, H., Z. Guo, R. M. Ahmad, Z. Ahmed, P. Cai, X. Chen, Y. Zhang, H. Lin, and Z. Shi. 2024. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. *Frontiers in Medicine*, 10:1291352.