Explicit Linguistic Markers of Implicit Social Biases: Gender Biases in Adjectives with Grammatical Gender

Abstract

While adjectives have received little attention in the study of gender stereotypes in word embeddings, they are of particular importance as they may carry implicit stereotypes that often go unnoticed. Most studies to date have analyzed biases with reference to nouns in English, and the limited literature on other languages, specifically those with grammatical gender, has also been confined to investigating gender biases in similar ways. Languages with binary grammatical gender systems may facilitate the study of gender stereotypes as grammatical gender may provide an explicit means for quantifying bias. In the present study, I propose a new metric to quantify bias for languages with binary grammatical gender and analyze commonalities and differences in gender stereotypes in Spanish adjectives among humans and word embeddings. Comparisons between adjectives and semantic gender and adjectives and masculine and feminine names indicate that word embeddings are indeed susceptible to learning gender biases by replicating human gender stereotypes and creating new ones.

1. Introduction

Word embeddings are a very useful tool in Natural Language Processing applications. They prove useful in a variety of downstream tasks, including sentiment analysis (Irsoy & Cardie, 2014), document ranking (Nalisnick et al., 2016), question retrieval (Lei et al., 2016), coreference resolution (Zhao et al., 2018a; Rudinger et al., 2018), and neural machine translation

(Font & Costa-jussá, 2019), among others. However, as with many machine learning systems, it has been attested that word embeddings are susceptible to learning human biases with regards to gender, race, ethnicity, age and other social variables (Bolukbasi et al., 2016; Caliskan et al., 2017, Garg et al., 2018).

The research community has been invested in creating techniques to mitigate and minimize these biases. Most analyses have focused on gender biases and have been limited to English and to certain domains (e.g., stereotypes in occupations). In languages with grammatical gender, biases have been similarly studied in regard to occupations (Zhou et al., 2019) and family and career terms (McCurdy & Serbetçi, 2017). However, analyses that focus primarily on nouns may only encompass a small portion of the bias present in language. In this paper, I take a different approach in analyzing gender biases with respect to the use of nouns and adjectives. I quantify bias compare results to human ratings from the Current Gender Stereotypes Scale (CGSS) (Mayén & Berges, 2007) to address the following two questions:

1. Are adjectives susceptible to capturing biases in word embeddings?
2. If they are, how are they similar to gender stereotypes attested in humans?

2. Gender

Throughout this paper, I make reference to two types of gender: grammatical gender and semantic gender. While a broader definition of gender is beyond our scope, I must emphasize that this study intends not to diminish gender identities outside the male-female dichotomy but to account for the linguistic properties of Spanish. Grammatical gender is a linguistic feature that differentiates noun classes with shared characteristics (Corbett, 1991; 2006). In Spanish, grammatical gender may make a reference to biological sex and it may be assigned in different ways (Lang, 1990). The most productive way to assign gender is to inflect a noun or adjective stem with -o, -e or -∅ for the masculine gender and -a for the feminine (see example 1). In some occupations, professions, and gentilic adjectives, the feminine gender is marked by augmenting the stem with one the following morphemes: -ina, -esa, or -isa (see example 2). For gender-neutral nouns or adjectives, grammatical gender is assigned by a determiner or modifier (see example 3).Additionally, grammatical gender can also be expressed lexically (see example 4). I use the term semantic gender to refer animate nouns where the masculine and femine forms of a concept are related (though not always) to biological sex or its gender is expressed lexically. For example, mother:father or woman:man are semantically gendered. Semantic gender may be intertwined with the definition of gender in a social sense. For the purpose of this study, semantic gender will be treated in its most linguistic sense.

1. niño:niña monje:monja encantador:encantadora
   boy:girl monk:nun charming.MASC:charming.FEM

2. zar:zarina conde:condesa profeta:profetisa
   tsar:tsarina count:countess prophet:prophetess

3. el atleta:la atleta
   he.MASC athlete the.FEM athlete

4. Padre:madre hombre:madre
   Father:mother man:woman

Previous work.

Caliskan et al. (2017) provide a qualitative study to investigate whether prejudicial attitudes and human behaviors are also present in English word embeddings. With that purpose, they adapted the Implicit Association Test (Greenwald et al., 1998) into the Word-Embedding Association Test (WEAT) and the Word-Embedding Factual Association Test. They use cosine similarity as analogous to response time in experiments with humans as a means of evaluation. Their results suggest that mundane associations between words in humans are replicated: pleasantness in flowers and instruments and unpleasantness in insects and weapons. In addition, a wide variety of social biases were also found: racial and ethnic biases include European-American names as having a more pleasant connotation than African American names and sexist biases portray male names as having a stronger association with career-like words and female counterpart with family-like words. Garg et al. (2018) use English word embeddings to study stereotypes and attitudes towards women and ethnic minorities in the United States through the use of adjectives. In a large-scale, longitudinal study, they analyze word embeddings diachronically by reference to historical Census data from the 20th and 21st centuries. In examining gender disparities, the researchers identified a decrease in gender biases over time, a trend that substantially changed during the women's movement in the 1960s and 1970s. In regards to immigration, it was found that before the 1960s biases towards Asians were related to themes of 'foreignness' and after the 1980s these biases were replaced by current stereotypes of Asian Americans as 'sensitive', 'passive', and 'complacent'. McCurdy and Serbetçi (2017) study gender stereotypes in German and Spanish. They use the WEAT to compute associations in two different settings: 1) comparing grammatically gendered words to male and female words and 2) comparing career and family to male and female words. They found that grammatical gender has greater bias than family-career terms even when gender bias mitigation methods have been applied. More specifically, grammatically gendered words tend to associate with male and female terms with the same grammatical gender to a greater extent than career or family terms are to associate with a certain gender. Similarly, Zhou et al. (2019) investigate gender biases in French and Spanish through occupation words. They introduce new metrics for quantifying gender biases and expand current debiasing methods to account for languages with grammatical gender. Their results suggest that gender stereotypes are also present in languages with grammatical gender and debiasing methods prove helpful in minimizing biases while keeping word embeddings useful. They also find that biases fluctuate across languages. Occupations like mathematician and nurse have male and female bias, respectively, while other professions like chef tend to associate differently to different genders across languages.

3. Quantifying bias

Gender Bias has been quantified in a variety of ways. Caliskan et al.(2017) quantify implicit biases in the English language using the Word Embedding Association Test (WEAT). The WEAT measures the difference of association between target terms (e.g., male vs. female words) and attribute terms (e.g., career vs. family words). Formally, if X and Y represent the masculine and target words (e.g., Names: man, woman) and A and B represent the attribute words (e.g., family terms, career terms), the test statistic can be computed as follows:

$$s(X,\ Y,A\ ,B)\ = \sum_{x\in X} s(x,A,B) - \sum_{y\in Y} s(y,A,B)\ \ (1)$$

$$\text{And}\ \ s(w,A,B) = \frac{1}{|A|} \sum_{a\in A} cos(w,A) - \frac{1}{|G|} \sum_{b\in B} cos(w,B)\ \ (2)$$

Where cos(*w*,*a*) denotes the cosine similarity between *w* and *a*, *s*(*w*,*A*,*B*) is the difference between average associations of *w* and attributes *A* and *B*, and *s*(*X*,*Y*,*A*,*B*) measures the difference of association between target words with respect to the attributes. Zhou et al. (2019) adapt the WEAT for languages with grammatical gender and introduce the Modified Word Embeddings

Association Test (MWEAT). The MWEAT takes the absolute value of the difference between the absolute value for each instance of Eq.2 in Eq.1, resulting in the following:

$$s(X,\ Y, A\ , B)\ = \|\sum_{x\in X} s(x,A,B)\ | - |\sum_{y\in Y} s(y,A,B)\| \quad (3)$$

Zhou et al. (2019) use the feminine and masculine alternations of occupations as target terms and gender related words as attribute words in Spanish and French. However, a linguistically grounded approach for quantifying gender biases that takes into account grammatical gender should also take into account grammatical agreement. Measuring the difference between grammatical matches and mismatches as in Eq.3 may not fully describe the grammatical properties of a language and how they are produced. For example, if we are studying gender biases in adjectives, we are interested in computing association scores for adjectives and nouns that co-occur and are directly related. For example, for the words ‘niño’ and ‘bueno’ (or ‘niña’ and ‘buena’), the adjective ‘bueno’ may describe the noun ‘niño’ as they both share the same grammatical gender whereas for the words ‘niño’ and ‘buena’ (or ‘niña’ ‘bueno’), due to grammatical agreement in Spanish, the adjective ‘buena’ cannot refer to ‘niño’ as they do not share the same grammatical gender. In addition, for adjectives where gender is not marked (e.g., ‘inteligente’) bias quantified using the MWEAT will result in a value of 0. Thus, I propose the Binary Association Difference (BAD) Test to better account for grammatical agreement and adjectives with marked and unmarked grammatical gender. Formally, if X and Y have the same dimensions and represent the masculine and feminine alternations of a list of gender unbearing target words (e.g, adjectives, occupations) and A and B have the same dimensions and represent the masculine and feminine forms of a list of gender bearing attribute words (e.g., semantic gender terms, names), the test statistic can be computed as follows:

$$s(X,\ Y,A\ ,B)\ =\sum_{x\in X} s(x,A) - \sum_{y\in Y} s(y,B)\ \ (4\ )$$

$$\text{And}\ \ s(w,G) = \frac{1}{|G|}\sum_{g\in G} cos(w,G)\ \ (5)$$

Where *cos(**a**,**b**)* denotes the cosine similarity between vectors ***a*** and ***b***, s(**w**,A) computes the average association of word vector **w** and attributes from A, and s(X,Y,A,B) measures the difference of association between the masculine and feminine forms of gender unbearing target words with respect to gender bearing attribute words.

4. Gender Biases in Humans and embeddings

Caliskan et al. (2017) compare response times in the IAT test (Greenwald et al., 1998) to WEAT scores in order to investigate whether human biases are also present in word embeddings. In the same way, I use human ratings from the CGSS scale (Mayén & Berges, 2007) and compare them to BAD Test scores to investigate human biases specifically for the Spanish language. Mayén & Berges study implicit gender stereotypes in Spanish speakers from Spain using the current gender stereotypes scale. For that purpose, they created a list of stereotypical adjectives by compiling 170 adjectives and measured bias by asking participants to rate adjectives on a scale from 1 to 7 in a way that reflects how the adjectives in question are associated to men and women according to current societal standards. After finding inconsistencies in the number of positive and negative adjectives for each gender, they refined the list to 40 adjectives that comprised 10 positive and 10 negative adjectives for each gender (38 grammatically marked, 12 grammatically unmarked). I utilize the 40-adjective CGSS list as target terms and semantic gender and masculine and feminine names as attribute terms to compute BAD Test scores for two different pre-trained word embedding datasets. The first dataset was trained on Spanish Wikipedia (Bojanowski et al., 2017) and the second dataset was trained on the Spanish Billion Word Corpus (Cardellino, 2016; Perez, 2017). Both datasets were trained on fasttext algorithm (Bojanowski et al., 2017).

Experiment 1: I investigate whether the ratings for masculine and feminine adjectives within the same sample are significantly different. The null hypothesis is that there is no difference between masculine and feminine scores (e.g., in CGSS scores or BAD Test scores). Results from two-tailed paired t-tests in Table 1 indicate that the null hypothesis is accepted for scores in the CGSS but rejected for all BAD Tests. Significantly different scores for masculine and feminine scores for adjectives in the BAD Test suggest that word embeddings are indeed capturing and exerting a gender bias.

Experiment 2: I compare bias in the CGSS and BAD Test scores to test for correlation and analyze similar and different trends. The null hypothesis is that CGSS bias and BAD Test are uncorrelated. CGSS bias is the difference between the masculine and feminine human ratings for an adjective. To provide a commensurate magnitude of comparison for the BAD Test, I scale CGSS values to have values between 0 and 1 and assume cosine similarities fall between 0 and 1. Positive values indicate a masculine bias and negative values indicate a feminie bias for both the CGSS bias and the BAD Test. Results from Pearson correlations in Table 2 indicate statistically significant moderate correlations among CGSS bias and the BAD Test with respect to semantic gender terms and and insignificant weak correlations among CGSS bias and the BAD Test with respect to names. Significantly different scores suggest similar biases in humans and biases found for adjectives with respect to semantic gender.

Metric	CGSS	SG_Wiki	SG_SBWC	N_Wiki	N_SBWC
t-statistic	0.4835	2.9510	2.5743	-6.8895	-9.7429
p-value	0.6314	0.0053	0.0139	0.0000	0.0000

Table 1. Results for two-tailed paired t-tests for CGSS, and BAD Test scores computed for CGSS adjectives relative to semantic gender and names using fasttext word embeddings trained on Spanish Wikipedia and the Spanish Billion Word Corpus (p-value > 0.05).

Dataset	BAD Test	Semantic Gender	Names
Wikipedia	t-statistic	0.3513	0.0075
	p-value	0.0262	0.9632
SBWC	t-statistic	0.3500	0.0834
	p-value	0.0268	0.6088

Table 2. Results for pearson correlations among CGSS bias and BAD Test scores computed for CGSS adjectives relative to semantic gender and names using fasttext word embeddings trained on Spanish wikipedia and the Spanish Billion Word Corpus (p-value > 0.05).

5. Discussion

The BAD Test discerns gender stereotypes without reinforcing current stereotypes, that is, by avoiding comparisons that are themselves instituting gender bias (e.g., comparing female and masculine terms to family and career terms). In this study, scores for the BAD Tests suggest that adjectives indeed carry gender biases but they differ depending on the type of attribute words used in the comparison. As seen in Figure 1, gender bias in semantic gender oscillates between the positive and negative scores for both Wikipedia and Spanish Billion Word Corpus embeddings; in contrast, as seen in Figure 2 gender bias in names remains in the negative scores, thus denoting a strong female bias. Critical adjectives, or adjectives with a strong magnitude in either direction, share some similarities across embeddings. For semantic gender, critical adjectives with a male bias include ‘free’, ‘powerful’, ‘braggart’, ‘scientific’, ‘maniac’, ‘dirty’ and ‘conceited’; and critical adjectives with a feminine bias include ‘sexual’, ‘fancy’, and ‘warm’. For names, with an overall strong female-leaning bias, critical adjectives with a male bias include ‘braggart’, ‘dirty’, ‘scientific’ and ‘clean,’ and critical adjectives for names with a female bias include ‘flirty’, ‘hysterical’, ‘fancy’, ‘bully’ and ‘vicious’. Overall, word embeddings seem to capture similar information with regards to gender stereotypes present in humans like the association of ‘braggart’, ‘dirty’, and ‘powerful’ with males and ‘flirty’, ‘fancy’, ‘warm’, and ‘hysterical’ with females. Further they seem to capture new biases, such as the association of ‘clean’ and ‘conceited’ with males and ‘sexual’, ‘bully’ and ‘vicious’ with women.

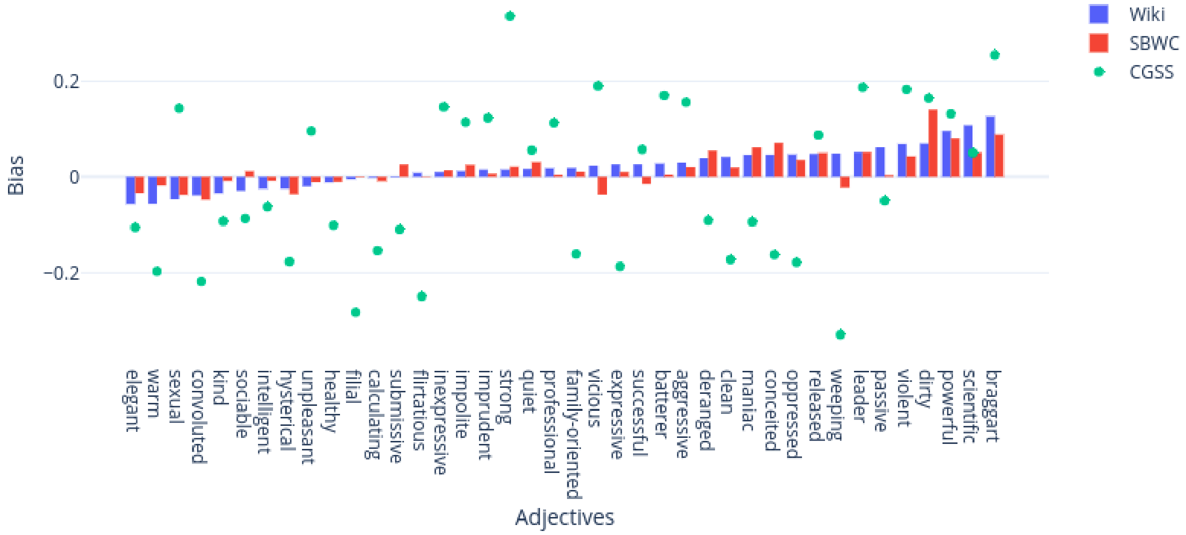

Figure 1. Bias scores plotted from lowest to highest for Wikipedia and Spanish Billion Word Corpus BAD Tests with respect to semantic gender. Negative values represent a feminine bias and positive values represent a masculine bias.

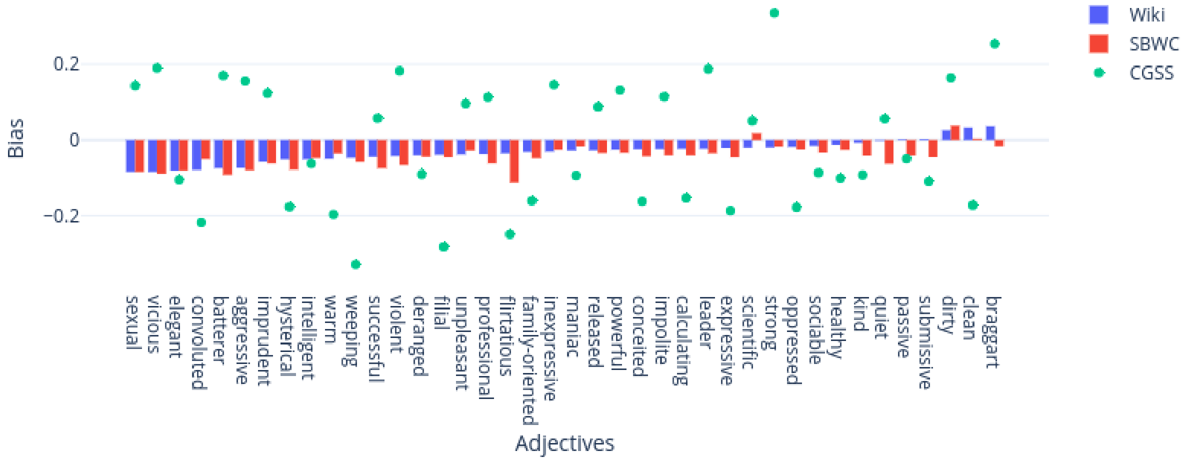

Figure 2. Bias scores plotted from lowest to highest for Wikipedia and Spanish Billion Word Corpus BAD Tests with respect to names. Negative values represent a feminine bias and positive scores represent a masculine bias.

6. Conclusion and Future Work

In this study, I introduced a new metric that measures biases in languages with binary grammatical gender. The BAD Test computes gender bias with respect to a specific term and avoids making stereotypical binary comparisons. I performed 2 experiments to test 1) whether gender stereotypes in the masculine and feminine forms of an adjective are significantly different and 2) whether gender bias in human ratings are correlated with BAD Test scores when using different datasets and different attributes as points of comparison. Results suggest that word embeddings seem to capture certain human gender stereotypes and produce new ones. Future work would involve using more extensive lists of target and attribute words and comparing BAD Test scores to empirical data acquired in similar studies carried out in other Spanish speaking countries. In addition, probing and modifying mitigation techniques like those proposed by McCurdy and Serbetçi (2017) and Zhou et al. (2019) to account for the adjectives with marked and unmarked grammatical gender would prove crucial in mitigating gender biases that are replicated or newly created.

References

Bojanowski, P., Grave, E., Joulin, A., & 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. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

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

Cardellino, Cristian.(2016) Spanish Billion Words Corpus and Embeddings. https://crscardellino.github.io/SBWCE/

Corbett, G. G. (2006). Gender, grammatical. *The Encyclopedia of language and linguistics. 2nd Edition. Oxford: Elsevier,* 749-756.

Corbett, G. (1991). 1991: Gender. Cambridge: Cambridge University Press.

Font, J. E., & Costa-Jussa, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116.*

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences,* 115(16), E3635-E3644.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74(6)*, 1464.

Irsoy, O., & Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In *Advances in neural information processing systems* (pp. 2096-2104).

Lang, M. F. (1990). Spanish Word formation. London and New York.

Lei, T., Joshi, H., Barzilay, R., Jaakkola, T., Tymoshenko, K., Moschitti, A., & Marquez, L. (2015). Semi-supervised question retrieval with gated convolutions. *arXiv preprint arXiv:1512.05726.*

Mayén, M. D. R. C., & Berges, B. M. (2007). Escala de estereotipos de género actuales. Iniciación a la Investigación, (2), 9.

McCurdy, K., & Serbetci, O. (2017). Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *Proceedings of WiNLP.*

Nalisnick, E., Mitra, B., Craswell, N., & Caruana, R. (2016, April). Improving document ranking with dual word embeddings. *In Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 83-84).

Pérez, Jorge. (2017) Spanish Billion Words Corpus and Embeddings trained on FastText. https://github.com/dccuchile/spanish-word-embeddings

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301.*

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876.*

Zhou, P., Shi, W., Zhao, J., Huang, K. H., Chen, M., Cotterell, R., & Chang, K. W. (2019). Examining Gender Bias in Languages with Grammatical Gender. *arXiv preprint arXiv:1909.02224.*