Is there Gender bias and stereotype in Portuguese Word Embeddings?

Brenda Salenave Santana$^{1[0000-0002-4853-5966]}$, Vinicius Woloszyn$^{1[0000-0003-3554-5580]}$, and Leandro Krug Wives$^{1[0000-0002-8391-446X]}$

PPGC - Instituto de Informática - UFRGS, Porto Alegre RS, Brazil
{bssantana, vwoloszyn, wives}@inf.ufrgs.br

Abstract. In this work, we propose an analysis of the presence of gender bias associated with professions in Portuguese word embeddings. The objective of this work is to study gender implications related to stereotyped professions for women and men in the context of the Portuguese language.

Keywords: Word Embeddings · Gender Bias · Portuguese Embedding.

1 Introduction

Recently, the transformative potential of machine learning (ML) has propelled ML into the forefront of mainstream media. In Brazil, the use of such technique has been widely diffused gaining more space. Thus, it is used to search for patterns, regularities or even concepts expressed in data sets [5], and can be applied as a form of aid in several areas of everyday life.

Among the different definitions, ML can be seen as the ability to improve performance in accomplishing a task through the experience [12]. Thus, [4] presents this as a method of inferences of functions or hypotheses capable of solving a problem algorithmically from data representing instances of the problem. This is an important way to solve different types of problems that permeate computer science and other areas.

One of the main uses of ML is in text processing, where the analysis of the content the entry point for various learning algorithms. However, the use of this content can represent the insertion of different types of bias in training and may vary with the context worked. This work aims to analyze and remove gender stereotypes from word embedding in Portuguese, analogous to what was done in [2] for the English language.

Hence, we propose to employ a public word2vec model pre-trained to analyze gender bias in the Portuguese language, quantifying biases present in the model so that it is possible to reduce the spreading of sexism of such models. There is also a stage of bias reducing over the results obtained in the model, where it is sought to analyze the effects of the application of gender distinction reduction techniques.

This paper is organized as follows: Section 2 discusses related works. Section 3 presents the Portuguese word2vec embeddings model used in this paper and

Section 4 proposes our method. Section 5 presents experimental results, whose purpose is to verify results of a de-bias algorithm application in Portuguese embeddings word2vec model and a short discussion about it. Section 6 brings our concluding remarks.

2 Related Work

There is a wide range of techniques that provide interesting results in the context of ML algorithms geared to the classification of data without discrimination; these techniques range from the pre-processing of data [9] to the use of bias removal techniques[8] in fact. Approaches linked to the data pre-processing step usually consist of methods based on improving the quality of the dataset after which the usual classification tools can be used to train a classifier. So, it starts from a baseline already stipulated by the execution of itself. On the other side of the spectrum, there are Unsupervised and semi-supervised learning techniques, that are attractive because they do not imply the cost of corpus annotation [16, 14, 17, 15].

The bias reduction is studied as a way to reduce discrimination through classification through different approaches [13] [3]. In [1] the authors propose to specify, implement, and evaluate the "fairness-aware" ML interface called themis-ml. In this interface, the main idea is to pick up a data set from a modified dataset. Themis-ml implements two methods for training fairness-aware models. The tool relies on two methods to make agnostic model type predictions: Reject Option Classification and Discrimination-Aware Ensemble Classification, these procedures being used to post-process predictions in a way that reduces potentially discriminatory predictions. According to the authors, it is possible to perceive the potential use of the method as a means of reducing bias in the use of ML algorithms.

In [2], the authors propose a method to hardly reduce bias in English word embeddings collected from Google News. Using word2vec, they performed a geometric analysis of gender direction of the bias contained in the data. Using this property with the generation of gender-neutral analogies, a methodology was provided for modifying an embedding to remove gender stereotypes. Some metrics were defined to quantify both direct and indirect gender biases in embeddings and to develop algorithms to reduce bias in some embedding. Hence, the authors show that embeddings can be used in applications without amplifying gender bias.

3 Portuguese Embedding

In [11], the quality of the representation of words through vectors in several models is discussed. According to the authors, the ability to train high-quality models using simplified architectures is useful in models composed of predictive methods that try to predict neighboring words with one or more context words, such as Word2Vec. Word embeddings have been used to provide meaningful representations for words in an efficient way.

In [6], several word embedding models trained in a large Portuguese corpus are evaluated. Within the Word2Vec model, two training strategies were used. In the first, namely Skip-Gram, the model is given the word and attempts to predict its neighboring words. The second, Continuous Bag-of-Words (CBOW), the model is given the sequence of words without the middle one and attempts to predict this omitted word. The latter was chosen for application in the present proposal.

The authors of [6] claim to have collected a large corpus from several sources to obtain a multi-genre corpus representative of the Portuguese language. Hence, it comprehensively covers different expressions of the language, making it possible to analyze gender bias and stereotype in Portuguese word embeddings. The dataset used was tokenized and normalized by the authors to reduce the corpus vocabulary size, under the premise that vocabulary reduction provides more representative vectors.

4 Proposed Approach

Some linguists point out that the female gender is, in Portuguese, a particularization of the masculine. In this way the only gender mark is the feminine, the others being considered without gender (including names considered masculine). In [7] the gender representation in Portuguese is associated with a set of phenomena, not only from a linguistic perspective but also from a socio-cultural perspective. Since most of the termination of words (*e.g.*, advogada and advogado) are used to indicate to whom the expression refers, stereotypes can be explained through communication. This implies the presence of biases when dealing with terms such as those referring to professions.

Figure 1 illustrates the approach proposed in this work. First, using a list of professions relating the identification of female and male who perform it as a parameter, we evaluate the accuracy of similarity generated by the embeddings. Then, getting the biased results, we apply the De-bias algorithm [2] aiming to reduce sexist analogies previous generated. Thus, all the results are analyzed by comparing the accuracies.

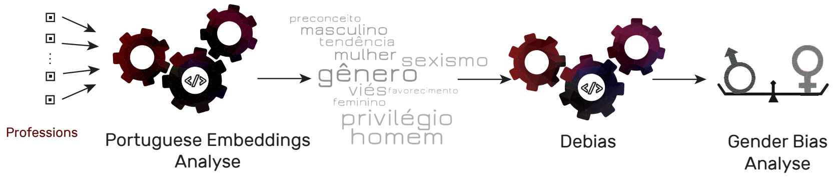

Fig. 1. Proposal

Using the word2vec model available in a public repository [6], the proposal involves the analysis of the most similar analogies generated before and after the application of the [2]. The work is focused on the analysis of gender bias associated with professions in word embeddings. So therefore into the evaluation of the accuracy of the associations generated, aiming at achieving results as good as possible without prejudicing the evaluation metrics.

Algorithm 1 describes the method performed during the evaluation of the gender bias presence. In this method we try to evaluate the accuracy of the analogies generated through the model, that is, to verify the cases of association matching generated between the words.

Algorithm 1 Model Evaluation

```
function W2V_EVALUATE($model, professions\_pair$)
    open_model($model$)
    count = 0
    for $female, male$ in $profession\_pairs$ do                ▷ read list of tuples
        x = model.most_similar(positive=[‘ela’, male], negative=[‘ele’])
        if x = female then
            count += 1
    accuracy = count/size(profession_pairs)
    return accuracy
```

5 Experiments

The purpose of this section is to perform different analysis concerning bias in word2vec models with Portuguese embeddings. The Continuous Bag-of-Words model used was provided by [6] (described in Section 3). For these experiments, we use a model containing 934966 words of dimension 300 per vector representation. To realize the experiments, a list containing fifty professions labels for female and male was used as the parameter of similarity comparison.

Using the python library gensim[1], we evaluate the extreme analogies generated when comparing vectors like: $\overrightarrow{mulher} - \overrightarrow{x} \approx \overrightarrow{y} - \overrightarrow{homem}$, where x represents the item from professions list and y the expected association. The most similarity function finds the top-N most similar entities, computing cosine similarity between a simple mean of the projection weight vectors of the given docs. Figure 2 presents the most extreme analogies results obtained from the model using these comparisons.

Applying the Algorithm 1, we check the accuracy obtained with the similarity function before and after the application of the de-bias method. Table 1 presents the corresponding results. In cases like the analogy of ‘garçonete’ to ‘stripper’ (Figure 2, line 8), it is possible to observe that the relationship stipulated between terms with sexual connotation and females is closer than between females and professions. While in the male model, even in cases of non-compliance, the closest analogy remains in the professional environment.

Using a confidence factor of 99%, when comparing the correctness levels of the model with and without the reduction of bias, the prediction of the model

[1] Available in: https://pypi.org/project/gensim/

Table 1. Accuracy Obtained in Predicting Model Analogies

Model	Accuracy
Before Debias	18.18 %
After Debias	03.03 %

with bias is significantly better. Different authors [10][18] show that the removal of bias in models produces a negative impact on the quality of the model. On the other hand, it is observed that even with a better hit rate the correctness rate in the prediction of related terms is still low.

Extreme 'he'

1. escritor → poeta
2. cantor→ músico
3. pintor → escultor
4. secretario → secretário
5. ator → actor
6. historiador → poeta
7. arquiteto→ arquitect
8. fotógrafo → cineasta
9. advogado → empresário
10. juiz → juíz

Extreme 'she'

1. atriz → actriz
2. escritora → poetisa
3. pesquisadora → bióloga
4. sindica → medicamen
5. diretora → coordenadora
6. matemática → astronomia
7. historiadora → pesquisadora
8. garçonete → stripper
9. secretaria → secretária
10. enfermeira → psicóloga

Fig. 2. Extreme Analogies

6 Final Remarks

This paper presents an analysis of the presence of gender bias in Portuguese word embeddings. Even though it is a work in progress, the proposal showed promising results in analyzing predicting models.

A possible extension of the work involves deepening the analysis of the results obtained, seeking to achieve higher accuracy rates and fairer models to be used in machine learning techniques. Thus, these studies can involve tests with different methods of pre-processing the data to the use of different models, as well as other factors that may influence the results generated. This deepening is necessary since the model's accuracy is not high.

To conclude, we believe that the presence of gender bias and stereotypes in the Portuguese language is found in different spheres of language, and it is important to study ways of mitigating different types of discrimination. As such, it can be easily applied to analyze racists bias into the language, such as different types of preconceptions.

References

1. Bantilan, N.: Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. CoRR **abs/1710.06921** (2017), http://arxiv.org/abs/1710.06921
2. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems. pp. 4349–4357 (2016)
3. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery **21**(2), 277–292 (2010)
4. Carvalho, A., FACELI, K., LORENA, A., GAMA, J.: Inteligência artificial–uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC (2011)
5. Goldschmidt, R., Passos, E.: Data Mining. Elsevier Brasil (2017)
6. Hartmann, N., Fonseca, E.R., Shulby, C., Treviso, M.V., Rodrigues, J., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. CoRR **abs/1708.06025** (2017), http://arxiv.org/abs/1708.06025
7. Hellinger, M., Motschenbacher, H.: Gender Across Languages. No. v. 4 in IMPACT: Studies in Language and Society, John Benjamins Publishing Company (2015), https://books.google.com.br/books?id=Z1uXBwAAQBAJ
8. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication. pp. 1–6 (Feb 2009). https://doi.org/10.1109/IC4.2009.4909197
9. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (Oct 2012). https://doi.org/10.1007/s10115-011-0463-8, https://doi.org/10.1007/s10115-011-0463-8
10. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 35–50. Springer (2012)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Mitchell, T.M., Michell, T.: Machine learningmcgraw-hill series in computer science. WCB/McGraw-Hill: Boston (1997)
13. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568. ACM (2008)
14. dos Santos, H.D., Ulbrich, A.H.D., Woloszyn, V., Vieira, R.: Ddc-outlier: Preventing medication errors using unsupervised learning. IEEE Journal of Biomedical and Health Informatics (2018)
15. Woloszyn, V., Machado, G.M., de Oliveira, J.P.M., Wives, L., Saggion, H.: Beatnik: an algorithm to automatic generation of educational description of movies. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). vol. 28, p. 1377 (2017)
16. Woloszyn, V., Nejdl, W.: Distrustrank: Spotting false news domains. In: Proceedings of the 10th ACM Conference on Web Science. pp. 221–228. ACM (2018)
17. Woloszyn, V., dos Santos, H.D., Wives, L.K., Becker, K.: Mrr: an unsupervised algorithm to rank reviews by relevance. In: Proceedings of the International Conference on Web Intelligence. pp. 877–883. ACM (2017)
18. Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 (2015)