arXiv:1810.04528v1 [cs.CL] 10 Oct 2018

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,