Center for Mind/Brain Sciences (CIMEC)
University of Trento
Corso Bettini, 31, 38068
Rovereto,Trento TN
g.melonicraveiro@studenti.unitn.it

Giovana Meloni Craveiro
(225249)

Not adopting a GCN for abusive language detection, not reviewing an ethical paper on ever-larger language models, just analyzing vector gender biases on Portuguese and English: Only the third Part of my Plural Final Project

Introduction to Machine Learning for Natural Language Processing

Professor Aurelie Georgette Geraldine Herbelot

2021

Part III - Algorithm Analysis Experiment

How biased are the datasets we use? Analyzing gender inclination in Portuguese and English word vectors

Introduction

Distributional semantic systems offer the meaning of words based on the context that they were repeatedly found. Therefore, any structural biases expressed by people will be caught by algorithms and reproduced. To mitigate possible harms, processes of data curation are often used to identify and remove those biases. A common bias is based on gender prejudice, which occurs frequently in the sport field and as women's sexualization, chosen targets for this project. As a social aspect, those biases differ according to the culture and thus, language. In this experiment, a cosine similarity is performed to identify gender biases related to the words *sport/esporte* and *prostitution/prostituição* on two word vector datasets, one for English and one for Portuguese, with the additional goal of comparing the degree of each bias on each language. This experiment is a followup to a lecture on vector biases, that provided an algorithm that calculates cosine similarity. Finally, the language choice was motivated by familiarity with the language and culture.

Data and model

The datasets chosen are a subset of 20,000 300-dimensional word vectors for English and Portuguese collections by Grave et al. (2018). The authors trained them on *Common Crawl* and *Wikipedia* using fastText. Also using CBOW with position-weights, character n-grams of length 5, a window size 5 and 10 negatives. The English subset of 20,000 embedding vectors was provided at the lecture but the Portuguese subset was composed by manually extracting the first 20,000 word embeddings of the complete set.

The model simply reads the word vectors and calculates the cosine similarity. Mathematically, the formula measures the cosine of the angle between two vectors projected in a multi-dimensional space. The smaller the angle, the higher the similarity. It then uses the cosine similarity of the word vector of a determined word to compare it with the word vectors of every other word and selects the nearest *n* word vector neighbours. To each of the listed neighbours, it calculates the cosine similarity between the vector and the reference male and female vectors. If either the female or the male vector is more similar to the referred word vector by more than 0.01, the analyzed word is considered biased. After the comparison to all the words of the list, an average of the *male* and *female* biases is calculated, by summing 1 to a variable for every considered bias and dividing by the total of words analyzed.

Experimental setup

As a consequence of the dynamicity of language, some words gain plural connotations, which also differ in different languages. To consider this influence in this experiment, three variations of gender expression were accounted for. Each of the investigated words was run for Portuguese and English. The variations respectively for English and Portuguese are: *male/female,* *macho/fêmea, masculine/feminine, masculino/feminino, man/woman, homem/mulher.* The algorithm retrieves the *n* most similar word vectors to the one of the chosen word. The value of *n* was set to 3, 7, 10 and 20. As can be imagined, using 20 neighbour words has a higher probability of also including words that are less related to the original and thus, alter the average bias in an unwanted manner. The decision for testing with those four different values was motivated specifically to investigate the different possible outcomes. The words investigated were *prostitution* and *sport.*

The code provided at the laboratory was, thus, modified to include portuguese and variation on quantity of neighbours and target words. The code is stored in file *experiment.py* with helper functions in file *utils.py.* Additionally, there is the file *prepare_dataset.py* used to preprocess the Portuguese dataset by conserving only its first 20,000 embeddings. As opposed to English, Portuguese contains accentuated words. Therefore, it is necessary to run the code in a terminal that supports accentuation when using the Portuguese dataset. All code is available at github[1]. The target words must be provided by the user at execution time.

Results

This section is dedicated to presenting the results of the similarity searches that were carried.

English dataset					
Target words	Gendered words	Number of neighbours			
		3	7	10	20
Sport (Esporte)	Male / Female	0.66/0.33	0.71/0.29	0.7/0.3	0.55/0.45
	Masculine / Feminine	0.66/0.33	0.57/0.42	0.7/0.3	0.65/0.35
	Man / Woman	1.0/0.0	0.86/0.14	0.9/0.1	0.75/0.25
Prostitution (Prostituição)	Male / Female	0.0/1.0	0.43/0.57	0.4/0.6	0.45/0.55
	Masculine / Feminine	0.0/1.0	0.29/0.71	0.4/0.6	0.4/0.6
	Man / Woman	0.0/1.0	0.14/0.86	0.3/0.7	0.45/0.55

Table 1 - Table reporting results of gender bias in the English dataset for the terms *sport*, *prostitution*, and quantities of neighbours 3, 7, 10 and 20 with varying gender expressions.

Portuguese dataset					
Target words	Gendered words	Number of neighbours			
		3	7	10	20
Esporte (Sport)	Masculino / Feminino	0.0/1.0	0.0/1.0	0.0/1.0	0.05/0.95
	Macho / Fêmea	1.0/0.0	0.86/0.14	0.9/0.1	0.8/0.2
	Homem / Mulher	1.0/0.0	0.86/0.14	0.8/0.2	0.75/0.25

Prostituição (Prostitution)	Masculino / Feminino	0.0/1.0	0.0/1.0	0.0/1.0	0.05/0.95
	Macho / Fêmea	0.33/0.66	0.29/0.71	0.2/0.8	0.2/0.8
	Homem / Mulher	0.0/1.0	0.14/0.86	0.1/0.9	0.1/0.9

Table 2 - Table reporting results of gender bias in the Portuguese dataset for the terms *esporte*, *prostituição*, and quantities of neighbours 3, 7, 10 and 20 with varying gender expressions.

Gender Bias of the whole corpus - 20000 embeddings		
Terms used	Male/Female	Difference
Male/Female	0.21/0.79	0.58 higher female bias
Masculine/Feminine	0.21/0.79	0.58 higher female bias
Man/Woman	0.635/0.365	0.27 higher male bias
Macho/Fêmea	0.495/0.505	0.01 higher female bias
Masculino/Feminino	0.265/0.735	0.47 higher female bias
Homem/Mulher	0.46/0.54	0.08 higher female bias

Table 3 - Table reporting results of gender bias in both datasets considering all of its word embeddings in each of the terms used to express gender.

Out of curiosity, the code was also run analyzing all of the words of the corpus. As expected, any word chosen with the number of neighbours approximate to 20.000 (19.990 or upper) results the same biases count. However, contrary to the hypothesis that analysing the whole corpus according to gender bias would provide a balanced count, the probabilities are high and vary significantly in each corpus, with differences ranging from 0.01 to 0.58.

Discussion

The hypothesis considered while mounting the experiment were that the prejudice faced by women in the sport field and female sexualization would be reflected in the corpus. Thus, in both languages, the data was expected to show a male bias towards the neighbour embeddings of the term *sport/esporte* and a female bias towards the neighbour embeddings of the term *prostituition/prostituição*.

However, these hypotheses were not confirmed by all of the results. In fact, the case of the terms *masculino* and *feminino* in Portuguese compared to the term *sport* showed the highest female bias possible, 100%, equivalent to a female bias in all neighbours for the configuration of 3, 7 and 10 neighbours, and only dropped by 5% with 20 neighbours.

On the other hand, all of the other cases showed the expected gender bias. In English, for the term *sport*, the distributions were identical for pairs *male,female* and *masculine, feminine* for 3 and 10 words, showing an approximate 70% male bias. For 7 and 20, the results diverged. The biases that also matched the approximate 70% were 7 neighbours at pair *male,female* and 20 neighbours at pair *masculine,feminine.* The remaining two showed smaller differences. The *man,woman* pair showed greater biases for all cases, higher than 85% for 3, 7 and 10, and decaying only to 75% in the 20-neighbours case. While the bias was still high and similar in the 10-neighbours case, it is possible to note the decay in the 20-neighbours case, which indicates that 20 as a quantity of neighbours analyzed might be getting too high, if not already. In Portuguese for the term *sport*, however, the most similar pairs were *macho,fêmea* and *homem,mulher,* which presented a male bias close to 85% for all cases, varying a maximum of 10% from that value. Finally, there is the eccentric pair *masculino, feminino* that showed 95%-100% female bias. Comparing the two languages, in general, the results for Portuguese had a more intense bias, similar to the English bias for the pair *man,woman.*

In the case of the term *prostitution* for English, the results were more stable between different gender expressions, except for the 7-neighbours case. While they all showed 100% female bias with 3 neighbours, that decayed to 55%-70% with 10 and 20 neighbours. The 7-neighbours showed 57%, 71% and 86% depending on the gender expression. This factor shows a clear difference between the words associated with each gender expression, giving light to the unwanted influence of the connotation factor. For Portuguese, the bias revolved around at least 80% for most cases. Only for the cases of 3 and 7 in pair *macho/fêmea* that results were approximately 70%. As these numbers are higher for the Portuguese dataset, it can be seen that the gender bias is stronger in this dataset for these two concept areas. This aspect can be related to a strongest gender bias in the use of Portuguese language, but might also be a result of different social aspects that contribute to the conditions of the data that forms the datasets.

Nevertheless, as the numbers differed significantly according to the specific comparison, the instability of the results draws suspicion to the reliability of the results. Indeed, the gender bias analysis of the whole corpus strengthens these suspicions. It shows a high difference between the number of words that carry a female bias and the ones that carry a male bias. While the expected values would show approximately 50% balance, in the real results, only the terms *macho,fêmea* reached a 0.01 difference between them. The next closest one is the pair *homem,mulher* with a difference of 0.08. The remaining differences are surprisingly high, varying from 0.27 (*man,woman*), the only case in which male bias was predominant, to 0.47 (*masculino,feminino)* and 0.58 (*male,female* and *masculine,feminine).*

The fact that such high biases are calculated when looking at the whole corpus suggests that the approach used in this project might have been too simplistic. A possible improvement could be a more refined calculation of the bias, maybe setting a higher limit to the difference between the cosine similarity of the vector of the terms and the gendered vectors. It also suggests that the gender bias problem is a much greater issue than anticipated in this directed research line, affecting not only specific fields but the entire set of concepts used in language. In that case, it could be productive to treat all words as biased instead of focusing on specific terms. Finally, it can be the mere reflection that more words include a higher relationship to the female gender, not necessarily representing harmful gender biases.

Appendix

This section includes all of the increments to the original code.

File prepare_dataset.py

```
f = open("data/fasttext-wiki-news-300d-20000-pt","w") #clear contents
f.close()
pt_new_dataset = open('data/fasttext-wiki-news-300d-20000-pt', 'a')
pt_original_dataset = open('data/cc.pt.300.vec', 'r+')

line = pt_original_dataset.readline()
for index in range(20000):
    line = pt_original_dataset.readline()
    pt_new_dataset.write(line)

pt_original_dataset.close()
pt_new_dataset.close()
```

File experiment.py

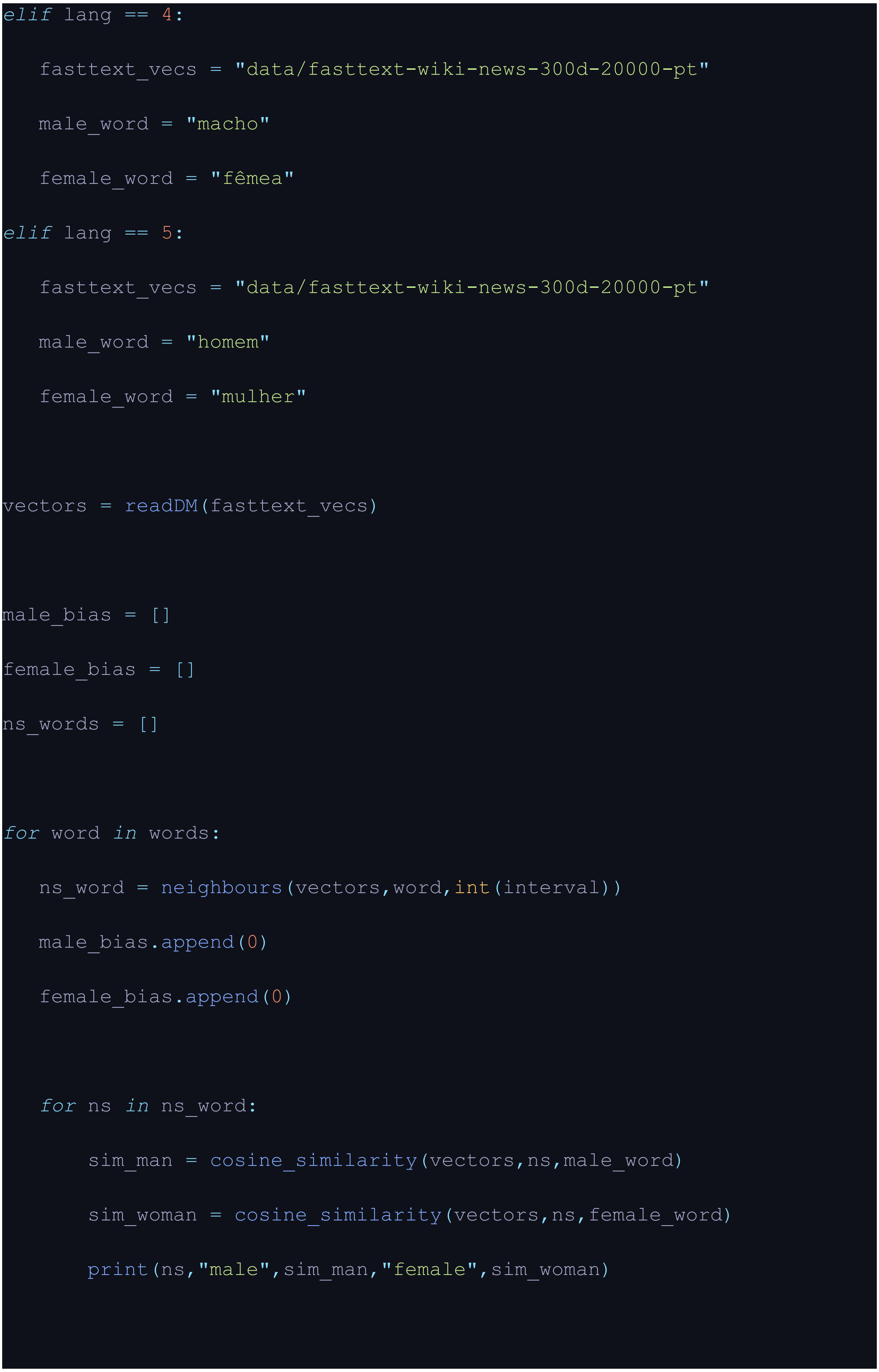
```
from utils import readDM, cosine_similarity, neighbours


print("\nQUESTION: What are the female and male biases found in those determined words?\n\n")


lang = -1
while (lang > 5 or lang < 0):
    lang = int(input("Enter\n 0 for English (male/female)\n 1 for English (man/woman) \n 2 for English (masculine/feminine) \n 3 for Portuguese (masculino/feminino)\n 4 for Portuguese (macho/fêmea)\n 5 for Portuguese (homem/mulher)\n\n"))


words = list( input("List all words separated by space").split())
interval = input("How many similar words should be analyzed?")
print(words)
i = 0


if lang == 0:
   fasttext_vecs="./data/fasttext-wiki-news-300d-20000.txt"
   male_word = "male"
   female_word = "female"
elif lang == 1:
   fasttext_vecs="./data/fasttext-wiki-news-300d-20000.txt"
   male_word = "man"
   female_word = "woman"
elif lang == 2:
   fasttext_vecs="./data/fasttext-wiki-news-300d-20000.txt"
   male_word = "masculine"
   female_word = "feminine"
elif lang == 3:
   fasttext_vecs = "data/fasttext-wiki-news-300d-20000-pt"
   male_word = "masculino"
   female_word = "feminino"
```



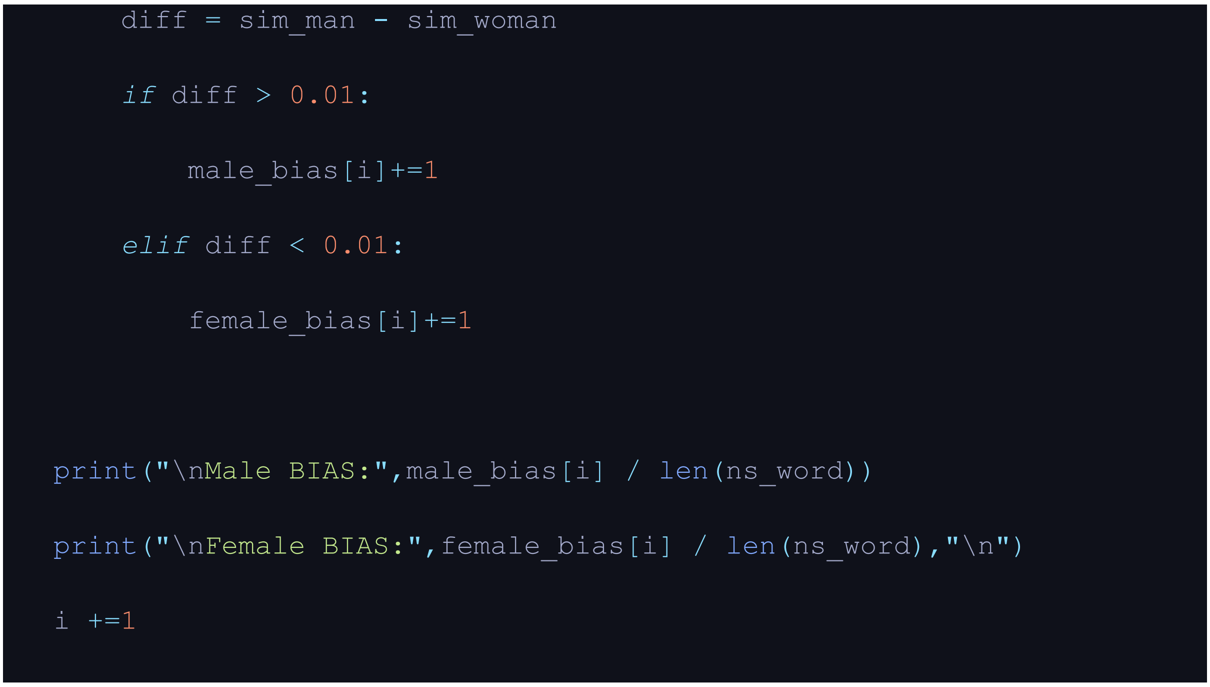
```
elif lang == 4:
    fasttext_vecs = "data/fasttext-wiki-news-300d-20000-pt"
    male_word = "macho"
    female_word = "fêmea"
elif lang == 5:
    fasttext_vecs = "data/fasttext-wiki-news-300d-20000-pt"
    male_word = "homem"
    female_word = "mulher"

vectors = readDM(fasttext_vecs)

male_bias = []
female_bias = []
ns_words = []

for word in words:
    ns_word = neighbours(vectors,word,int(interval))
    male_bias.append(0)
    female_bias.append(0)

    for ns in ns_word:
        sim_man = cosine_similarity(vectors,ns,male_word)
        sim_woman = cosine_similarity(vectors,ns,female_word)
        print(ns,"male",sim_man,"female",sim_woman)
```



```
    diff = sim_man - sim_woman
    if diff > 0.01:
        male_bias[i]+=1
    elif diff < 0.01:
        female_bias[i]+=1

print("\nMale BIAS:",male_bias[i] / len(ns_word))
print("\nFemale BIAS:",female_bias[i] / len(ns_word),"\n")
i +=1
```

References

Herbelot, Aurelie. (2021). Introduction to Machine Learning for Natural Language Understanding. Course at University of Trento.

Herbelot, Aurelie. (2021). Vector Biases. https://github.com/ml-for-nlp/vector-biases

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*. Datasets available at https://fasttext.cc/docs/en/crawl-vectors.html