Gender Stereotypes in Parliamentary Speeches
Knowledge Extraction and Information Retrieval Project

Nicole Maria Formenti 941481 - DSE

July 2021

Gender Stereotypes in Italian Parliamentary Speeches

1 Introduction

Advancements in machine learning and automated techniques allows now to study and exploit natural language for many applications. However, natural language often embeds many stereotypes deeply rooted in our societies. Hence, it is fundamental to be able to identify and possibly remove them in order to avoid creating biased systems which carry on such stereotypes.
A large literature about the identification and removal of gender and racial stereotypes is available and several methods have been proposed. This project is focused on recognising gender stereotypes in Italian parliamentary speeches by using two methods from literature, the *WEAT* [**3**] and the *ECT* [**10**], and a novel method called *EAB*, which will be presented in details. The secondary goal is to apply the *hard debiasing* technique proposed by *Bolukbasi et.al* [**8**] to try to reduce the influence of gender stereotypes.

2 Research question and methodology

The main goal of this analysis is to study the presence of stereotypes about genders in Italian parliamentary speeches from 1948 to 2020, and the topics such stereotypes are associated with. Moreover, a debiasing method will be applied in order to understand if it is able to effectively reduce the gender stereotypes present in the corpus. In particular, the analysis will be carried out by considering historical trends, so the data have been divided in four periods of time:

- 1948 – 1968
- 1968 – 1985
- 1985 – 2000
- 2000 – 2020

In order to study the association between stereotypes and particular topics, different groups of specific words related to these topics have been created:

- Family: *famiglia, figlio, matrimonio, genitore, bambino, accudire.*
- Career: *capo, presidente, onorevole, potere, carriera, salario, lavoro, professionale, denaro, ambizione.*
- Physical appearance: *rozzo, sensuale, splendido, magro, piacevole, brutto, bello, grasso, attraente, carino, frivolo, aggraziato.*

- Rage: *intollerante, crudele, aggressivo, brutale, odioso, cattivo.*
- Kindness: *premuroso, sensibile, compiacente, delicato, buono, bravo, gentile.*
- Intelligence: *intelligente, brillante, razionale, saggio, studioso, serio.*
- Dumbness: *illogico, irrazionale, stupido, superficiale, isterico.*
- Active: *intraprendente, ambizioso, forte, assertivo, sicuro.*
- Passive: *timido, passivo, insicuro, debole, silenzioso.*
- Male stereotypes: *intraprendente, intelligente, brillante, razionale, saggio, ambizioso, forte, crudele, intollerante, assertivo*
- Female stereotypes: *bello, superficiale, frivolo, sensibile, delicato, gentile, passivo, silenzioso, insicuro, illogico, isterico, debole, irrazionale.*
- Gendered words: *sorella, fratello, femminile, maschile, nonna, nonno, ragazza, ragazzo, madre, padre.*

All the groups of words contain neutral words which are not associated with any particular gender, except for the last group which contains only gendered words by definition and it acts as a control group.

The original corpus of speeches is divided into paragraphs containing the speech of a single person, which are further subdivided into single sentences already tokenised and normalised. The single sentences are considered as the documents of the corpus.
The embedding model used to represent the text is the *word2vec model*, which is a neural language model, based on a particular neural network called autoencoder, where the input has the same size as the output. It is based on the idea of representing single words by using their context, which is able to capture their meaning. The model exists in two versions, namely the *CBOW* and the *Skip-gram* models. The former version takes the context words as input, whose size is defined by the chosen window size, and learns the central word, while the latter one does the opposite. The neural network contains a single hidden layer, whose number of neurons is arbitrarily defined in advance, and the vectors of words fed to the model are in one-hot encoding format with a length equal to the size of the vocabulary.
The true goal of the model is not that of correctly classifying the words in output, but that of providing a denser embedding of words. Hence, the weight matrix between the input and the hidden layer is retrieved after training and used to represent words. This matrix has a number of rows equal to the number of words contained in the vocabulary and a number of columns equal to the size of the hidden layer, which is usually much smaller than the input size. Each row of this matrix corresponds to the new dense embedding of the original word.
This model has proved to work well in practice since similar words having a similar context must also have a similar dense embedding in order to produce a correct classification. This results in a mapping of the original vectors onto a space where closer vectors share a similar semantic meaning.
The *word2vec model* chosen for the study is the *Skip-gram* model, since according to *J. Pennington et al.* [**4**], it performs slightly better on analogies task than the *CBOW* one. The size of the final dense vector is 300 and the window size is 5. The network performs 10 epochs for training before stopping due to time reasons. Words with a total count lower than 5 have been removed since they don't have enough statistical support, moreover they might be misspellings which only add noise. The resulting models are four, each one trained on documents belonging

to a different time period.

The methods used to study the presence of gender stereotypes are three, where two of them are coming from the existing literature. They are presented following.

2.1 WEAT

The first metric used is the renowned *WEAT* (Word Embedding Association Test) from *Caliskan et al.* [**3**], which compares two sets of target words (neutral words) and two sets of attribute words (gendered words). The null hypothesis is that there is no significant difference between the sets of target words when considering their similarity to the attribute words. The effect size of the test returns a normalised measure of the separation between the two associations distributions.

2.2 ECT

The second test found in literature is the *ECT* (Embedding Coherence Test) from *S. Dev, J.M. Phillips* [**10**]. In this case the two sets of gendered words are compared by using a single set of neutral words. A representative vector is created for each gender as the average vector between the vectors of words in each group. The method creates two vectors of similarities which contain, for each gender, the cosine similarity between each word in the neutral group and the averaged vector of the gender. Finally, the *Spearman's rank correlation coefficient* is used to assess the correlation between the two vectors. The main idea is that the higher is the correlation between the vectors of similarities, the lower is the bias embedded in the neutral group of words.

2.3 EAB

The last method has been developed in the making of this project and tested on the research question at hand. Its name is *EAB (Embedded Analogies Bias)* and, differently from the previous two methods, it exploits analogies to study the bias contained in text.
An analogy is a comparison between two things relatively to a particular aspect, which takes the form:

$$A : B \approx C : D$$

A classical example is:

$$man : king \approx woman : queen$$

Thanks to the linearity property of word embedding models which allows to perform arithmetic operations between vectors, they can be easily used to solve analogies of the type:

$$A : B \approx C : ?$$

Where D is found by considering B and C as positively influencing words and A as a negatively influencing word.

$$D \approx B - A + C$$

The idea to use analogies for studying stereotypes in the language came from the critiques moved by *M.Nissim et al.* [**1**] to some studies on gender stereotypes using analogies. One of the main problems highlighted by the authors of the paper is that the current implementations of methods to solve analogies, such as the *Gensim* library, don't allow to return the same words already present in the analogy. That's because in the original definition of analogies all the words must be different. Hence, some of the results indicating the presence of gender stereotypes are biased due to this restriction. In addition, they show that by removing this constraint the algorithm, most of the times, gives as first result a word D which is equal to the given word B. However, even if it is true that the first word returned to complete the analogy almost never contains any bias, it is also true that by looking at the other top n words returned by the algorithm some gender bias is still present.

An example is the analogy reported in the *M.Nissim et al.* paper, which is:

$$\textit{he} : \textit{computer programmer} \approx \textit{she} : ?$$

By using the Google News pretrained *word2vec* model with a vector size of 300, the top 10 words returned are: *computer programmer, homemaker, housewife, businesswoman, saleswoman, graphic designer, beautician, registered nurse, paralegal, librarian.*

Interestingly, by switching the genders of the analogy as *she* : *computer programmer* $\approx$ *he* :?, the words returned are now: *computer programmer, mechanical engineer, electrical engineer, engineer, carpenter, businessman, mechanic, salesman, programmer, machinist.*

The first word returned is the same as the word given in the analogy in both cases, however the gender bias is strikingly evident from the other words. The words positively associated to *he* and *computer programmer* refers to engineering and other jobs otfen associated with men, while the words positively associated to *she* and *computer programmer* are about caring roles or related to external appearance.

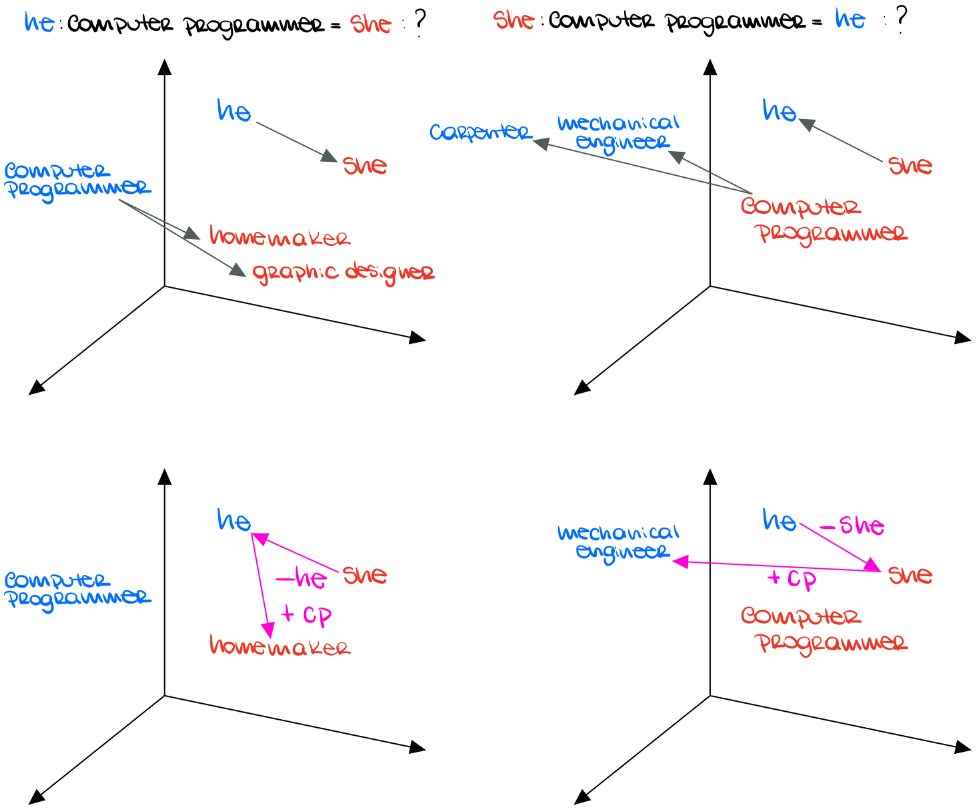


Figure 1: *Schematic representation of analogies in the embedding space.*

The main intuition is that if a neutral term doesn't really contain any gender bias, then it should return words with a similar meaning when positively associated with one or the other gender. If this is not the case, one should explore in more detail the words returned by the algorithm in order to understand where the difference comes from. In fact, there is a good chance that the difference comes from some gendered stereotypes embedded in the language.

The EAB algorithm works by first returning a set of possible words to complete an analogy of the type:

$$gender\,1\ :\ neutral\,word\ \ \approx\ gender\,2\ :?$$

Then it calculates the average cosine similarity between all the words returned by the analogy algorithm and both gendered words.

$$AvgCossim_G = \frac{\sum_{V_a \in A} cos(V_G, V_a)}{count(V_a)}$$

Where V_a are the vectors of the words returned by the analogy and V_G is the vector corresponding to the gender $G \in \{female, male\}$.
The cosine similarity between two vectors is defined as:

$$\frac{A \cdot B}{\|A\|\,\|B\|}$$

The algorithm is implemented so that a gender can either be represented by a single word, in that case V_G is the vector of the word itself, or by a group of words related to the gender. In the latter case, V_G is the average vector between all the vectors of words $g \in G$.

$$V_G = \frac{\sum_{V_g \in G} V_g}{count(V_g)}$$

The groups of words chosen for each gender are the following:

- Female: *donna, ella, femmina*
- Male: *uomo, lui, egli, maschio*

The word *lei* has not been included in the female group of words since in the Italian language it is often used as a formal 2^{nd} singular personal pronoun, regardless of the gender of the interlocutor. Since parliamentary speeches presuppose a certain degree of formality, including the term might probably have introduced some bias.

The bias of the analogy is calculated as the absolute difference between the two average cosine similarities:

$$bias(G, w_n) = |AvgCossim_{female} - AvgCossim_{male}|$$

Where w_n is the given neutral word and G is the positive gender.

The bias becomes higher as the difference in the average distances between the words returned by the analogy and the genders becomes larger.
The ratio is that if the two average similarities are quite different, the words returned by the

analogy are skewed towards the positive gender, hence it is probable that the neutral word is related to some gender stereotypes.
The bias is calculated by both considering *female* as positive and *male* as negative, and vice versa, in order to study whether there is some sort of symmetry between the biases. Finally, the total bias for a neutral word is calculated as the sum of the two biases:

$$BIAS(w_n) = bias(female, w_n) + bias(male, w_n)$$

Regarding the implementation of the *EAB* method, two types of algorithms to solve analogies have been used, namely the *3COSADD* [**11**] and the *3COSMUL* [**12**].
The *3COSADD* solves the following expression to find D:

$$\underset{D}{\operatorname{argmax}}\,(cos(D, C) - cos(D, A) + cos(D, B))$$

On the other hand, the *3COSMUL* method refines the objective function of the *3COSADD.* The main problem of the *3COSADD* is that when large terms are present in the expression, they dominate it, hence the words retrieved to complete the analogy will be more skewed towards the large term. Instead, all the terms in the objective function should contribute with a similar magnitude in order to obtain a fair analogy. To better balance all the terms in the analogy, the new expression to solve becomes:

$$\underset{D}{\operatorname{argmax}}\,\frac{cos(D, B)cos(D, C)}{cos(D, A) + \epsilon}$$

Where $\epsilon = 0.001$.
In their paper *O. Levy and Y. Goldberg* state that this is equivalent to taking the logarithm before summing the terms, hence the difference between small quantities is amplified, while that between large quantities is reduced.
Since the *3COSMUL* method is an improvement of the *3COSADD*, only the former one has been applied for the analysis of the gender stereotypes.

The algorithm has been implemented in *Python.* The class returns, for each word, the bias both when female is positive and when male is positive, along with the average cosine similarities for each gender. The results are ordered according to the total bias. It is also possible to list, for the top biased words, all the words returned by the analogy algorithm both when male or female are positive.
An additional functionality is the possibility to return two types of bar charts which provide a visual representation of the biased words. The first type of chart plots the average cosine similarity to male and to female for each neutral word, both when male is positive and when female is positive. The second plot is a stacked bar chart which shows the composition of the total bias. Below, an example:

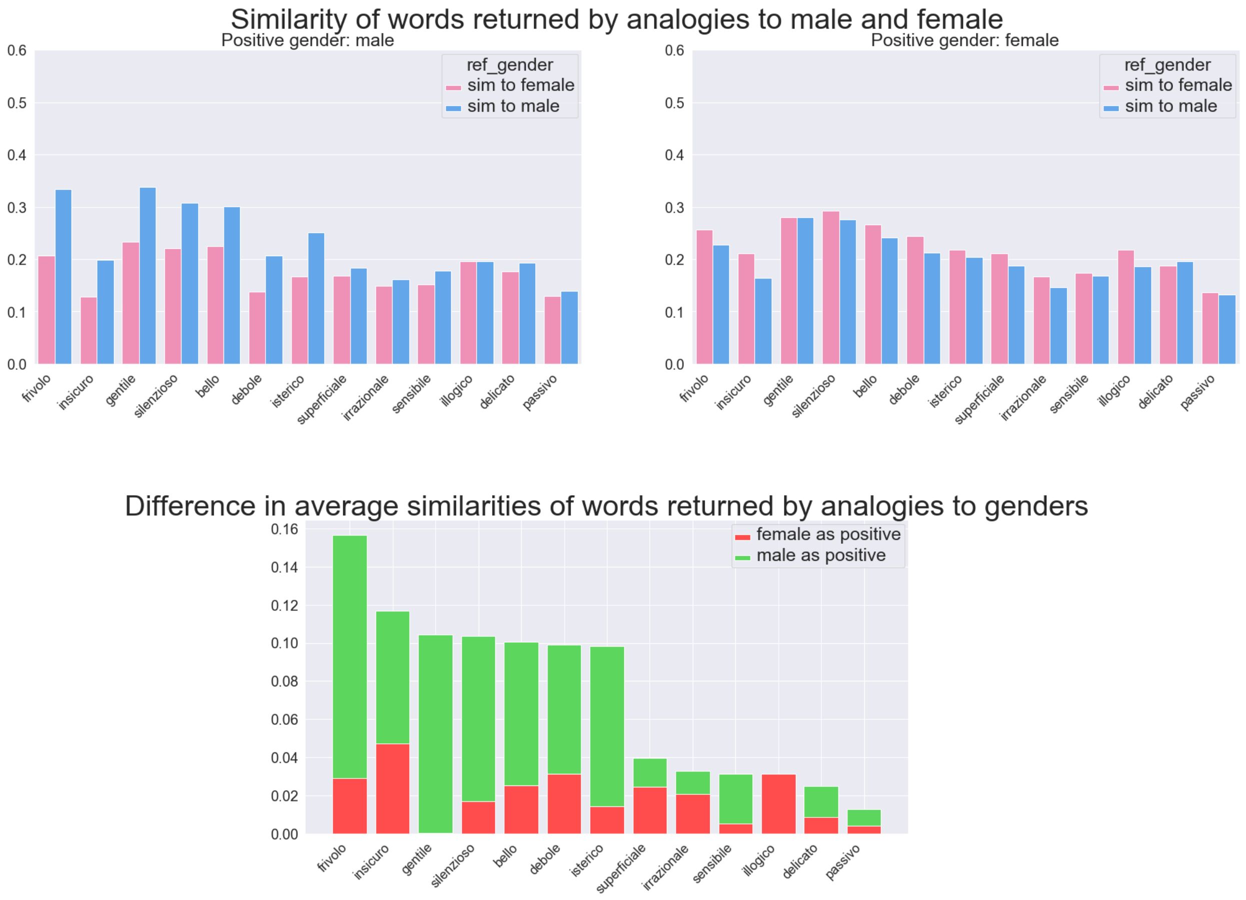


Figure 2: *Bar charts from EAB implementation for the group of words 'female stereotypes'*

As an example of the application of the *EAB* method, the group of words *family* is studied. In this case the genders are represented by the averaged vectors of words.
The word containing the highest bias is *accudire*. The words returned by the analogy when the positive gender is female are: *mangiare, bambino, andarsene, pulizia, domestico, ragazza, mamma, tranquillamente, dedicarsi, ora*. On the other hand, the words returned by the analogy when the positive gender is male are: *dedicarsi, procurarsi, sacrificare, dedicare, adempiere, sostentamento, pasto, adattare, esplicare, assolvere.*
It can be seen that when the positive gender is female, two words are specifically referred to parenting and two words are referred to the caring of the house. The same doesn't happen when the positive gender is male, where most of the words are verbs related to the compliance of some duty.
Instead, when looking at the least biased word, which is *bambino*, most of the words returned for both genders are the same or similar and they mainly refer to family.

2.4 Debiasing

Finally, the hard debiasing algorithm presented by *Bolukbasi et. al* in [**8**] is applied in order to understand whether it is able to effectively reduce gender stereotypes. The *EAB* method is used to estimate the current total bias before and after debiasing.
The hard debiasing method first identifies the direction of the gender subspace, which can be defined in three ways:

1. *Single*: normalised difference between the vectors representing the two genders, which corresponds to vectors of single words.
2. *Sum*: normalised difference between the vectors representing the two genders obtained as the sum of different vectors of gendered words.
3. *PCA*: first principal component of the difference between pairs of gendered vectors. This should explain the majority of variance in all the vectors.

For the Single method, the pair of gendered words used is *(uomo ,donna)*, while for the other two methods four different pairs of gendered words have been used: *(donna,uomo), (lui,ella), (egli,ella), (femmina,maschio)* .
Afterwards, the actual hard debias is applied through a neutralize and an equalize phase. The former consists in removing from the neutral words the projection of the neutral words themselves onto the gendered direction. This removes any influence of the gender subspace. The latter equalizes the gendered equality sets of words outside the gender subspace so that all the neutral words become equidistant from these pairs.
In order to understand the effect on the gender bias, the total bias is computed by using the *EAB* before the debiasing and after the debiasing for the single, the sum and the PCA methods. the results are first compared by using the *EAB* version using single words to represent genders and then the version using averaged vectors for genders. Only the bias of the first word returned for each group has been considered for brevity, since it corresponds to the most biased word.

2.5 Experiment

The reliability of the *EAB* will be assessed by comparing it with the other two methods, *WEAT* and *ECT*.
In particular, the intersection between groups of words selected by different methods is used as an approximation to retrieve the most and the least biased topics. Since the total topics are 12, the most biased topics are those topics selected by all the methods among the top 6 groups of words, while the least biased topics are those in the bottom 6 groups of words.
The selection is done by studying the intersection between:

1. *WEAT* and *ECT*
2. *WEAT*, *ECT* and *EAB* where the word *donna* represents the female gender and the word *uomo* represents males
3. *WEAT*, *ECT* and *EAB* where the genders are represented by the averaged vectors of gendered words.

In this way one can see whether the selected topics change when considering also the *EAB* method.

In order to make the three methods comparable, a ranking of the topics based on their bias must be obtained from all the tests.
In the case of the *EAB* and the *ECT* a ranking is naturally provided. The former ranks group of words based on the average bias of all the words belonging to the group, while for the latter the correlation is used to order them. As regards the *WEAT*, this method doesn't directly provide a ranking since it compares two groups of neutral words and returns a measure of the similarity of their distributions. In order to obtain the ranking, each group of neutral words is compared to the control group, which contains only gendered words by definition. If the p-value returned

is high, it means that the null hypothesis, stating that there is no difference between the two groups of words in terms of the gendered groups, is supported with a large confidence. Hence the neutral group of words has a strong gendered component.
Furthermore, in order to evaluate the robustness of all the methods, the control group of gendered words is ranked along with the other neutral group of words. The control group is expected to be ranked in the first positions if the tests work correctly.

A simple *TF-IDF* is also used in order to obtain a baseline for understanding whether these three methods perform well enough. The corpus of documents is further divided into males and females speakers and the average *TF-IDF* score for each group of words is calculated for the different periods of time. Then the difference between the average *TF-IDF* of the two genders is calculated and it is used to rank the group of words. The higher the difference in average *TF-IDFs* of males and females speakers, the more biased the topic, since it means that the words in the group are more specific to one gender than to the other.

Finally, the same analysis is done for single words in order to retrieve the top biased words selected by all the methods within each topic. Only the *EAB* and the *ECT* can be used in this case since the *WEAT* doesn't provide any useful metric for ranking single words. *EAB* orders single words according to their total bias, while regarding *ECT*, words have been ranked according to the difference in the cosine similarities between the neutral word and the mean vectors of genders:

$$|cos(V_n, \bar{V}_{female}) – cos(V_n, \bar{V}_{male})|$$

Where V_n is the vector of the neutral word, and $\bar{V}_{female}$, $\bar{V}_{male}$ are the mean vectors of the gendered words.
It is worth noting that the ranking of words used for the *ECT* might probably be an oversimplification, especially when compared to that used by *EAB*. That's because the ranking of single words isn't the original aim of this method.
For each topic, the words selected by both methods which are also among the top 3 biased words are taken.

3 Experimental results

3.1 Robustness

Regarding the robustness of the methods according to the ranking of the control group of words, the *EAB* seems to be the most robust one. In fact, the control group always appear in the first position in the ranking of biased groups for all the four time periods. The *WEAT* can also be considered robust since the control group always appear in the second position. On the other hand, the *ECT* method seems to be the least robust one. In this case the control group appears in the first position for the periods *1985-2000* and *2000-2020*, in the second position for the years *1948-1968* while it is ranked in position 10 for the period *1968-1985*.

3.2 Topics bias

The results of the selection of the most and least biased group of words according to the intersection of different methods is now presented. The selection has been carried out for each different time period 1.

When introducing the *EAB* for the selection some topics are filtered out, however it is still consistent with the other two methods. Since *EAB* showed to be robust, it can be assumed that by including it for the selection of the most and least biased topics the result has improved. Hence, the topics selected have a stronger support.
For the sake of brevity, only the resulting topics from the intersection between *ECT*, *WEAT* and the *EAB* version using the averaged vectors to represent the genders are shown. Possibly because this second version of *EAB* contains less noise since it relies on multiple words. The other results can be found in the appendix 2.

1. Period: *1948-1968*:
   - Top most biased topics: *physical appearance, family*
   - Top least biased topics: *active, kindness*
2. Period: *1968-1985*:
   - Top most biased topics: *family*
   - Top least biased topics: *career, kindness, male stereotypes*
3. Period: *1985-2000*:
   - Top most biased topics: *family*
   - Top least biased topics: *intelligence, male stereotypes*
4. Period: *2000-2020*:
   - Top most biased topics: *family, passive*
   - Top least biased topics: *active, female stereotypes, male stereotypes*

All the three methods always select *family* as a top biased topic for all the time periods, while for the period *1948-1968* also the topic *physical appearance* shows to be highly biased. In the period *2000-2020* words relative to *passive* contain gender stereotypes too.
On the other hand, there isn't a topic which is always among the least biased topics for all the periods. However, the topic *kindness* maintains a low bias from *1948* to *1985*. Interestingly, words related to *career* figure among the least biased ones in the years *1968-1985* which might be traced back to women's emancipation due to the feminist movement that gained popularity in that period. In the next period, from *1985* to *2000*, the topic *intelligence* figures among the least biased ones. The groups *female stereyotypes* and *male stereyotypes* often have a low bias, and that's probably because the words they contain have been arbitrarily chosen, hence they probably might not reflect true stereotypes.
When comparing it to the baseline given by the *TFIDF* score, it can be seen that the ranking of topics given by the *TF-IDF* is often in contrast with that provided by the intersection of the three methods. The *TF-IDF* gives the same result 6 times out of 7 for the top biased topics, but only 6 times out of 10 for the least biased topics 3.
In conclusion, ranking topics by using the intersection between different methods is a good starting point to further investigate what are the specific stereotypes incorporated into these topics.

3.3 Words bias

The last part of the bias analysis focuses on the retrieval of the top biased words selected by both the *EAB* and the *ECT* for each group of words in the different time periods. As before, first the version of *EAB* using single words to define genders and then the version using averaged vectors have been used. Only the top 3 most biased words have been compared. For brevity, the full results are reported in the appendix 4 .

In most of the cases there is at least one common word among the top 3 most biased words. Often, all 3 of them have been selected by both methods. This happens in the period *1948-1968* for the topics *intelligence, dumbness, active* and *male stereotypes*, and in the periods *1968-1985*, *1985-2000* and *2000-2020* only for the group *active.* On the other hand, the two methods don't have any top biased word in common in the period *1948-1968* for the topic *physical appearance*, in the period *1985-2000* for the topics *physical appearance, career* and *female and male stereotypes*, and in the years *2000-2020* for the group of words *career* and *female stereotypes.*

Thus, it can be concluded that the two tests are sufficiently consistent with each other. An additional advantage of the *EAB* is the possibility to explore more in details the stereotypes associated with the biased words. This can be done by looking at the words returned to complete the analogies when either female or male are used as positive genders.

3.4 Debiasing

To conclude with, the efficacy of the hard debiasing algorithm from *Bolukbasi et al.* will be commented following.

As regards the *EAB* version which uses averaged vectors for genders, the hard debiasing using the sum method is the only one which reduces quite consistently the bias. The PCA method tends to work well only for the period *1968-1985.* Interestingly, the hard debiasing nearly always nullify the bias for the control group of gendered words.

On the other hand, the debiasing algorithm shows a stronger efficacy for the *EAB* method with a single word for each gender. In fact, all the three methods nearly always reduce the initial bias. This is probably due to the fact that the bias is a priori higher since using a single word possibly introduces more noise when retrieving analogies. Differently from before, the bias of the control group is not entirely neutralized.

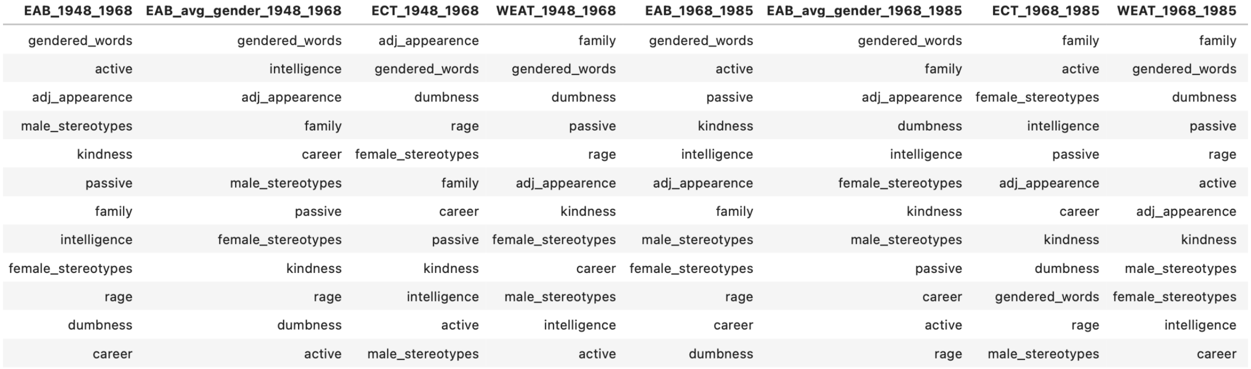
4 Concluding remarks

In conclusion, the new method *EAB* proved to be fairly robust in selecting those groups of words which contain some kind of gender stereotypes, even when compared with other methods. In addition, differently from other methods such as the *WEAT* and the *ECT*, it provides a way to study more in depth where the bias is possibly lying thanks to the use of analogies. In this way it is not only possible to study which are the stereotypes associated to genders or other groups, but also to filter out those seemingly biased words which are just false positives.

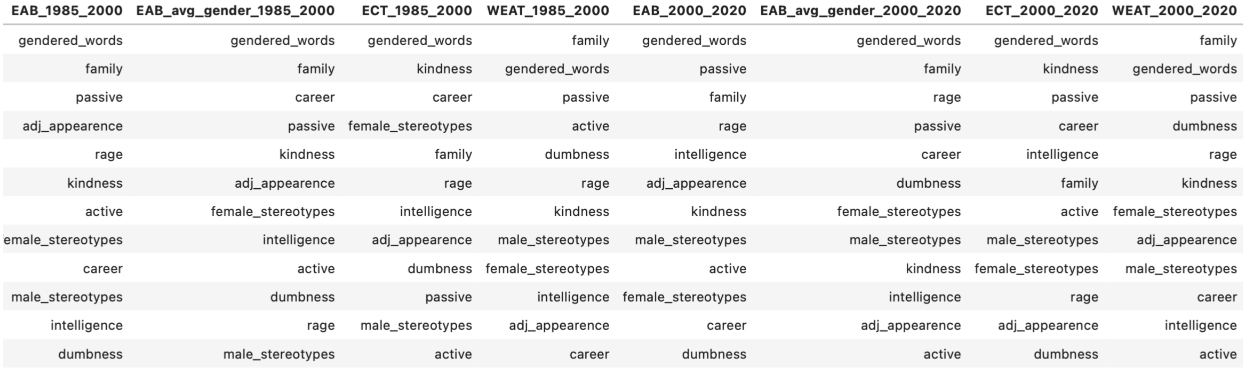
One possible drawback of the *EAB*, common to all the methods, is that the result heavily depends on the quality of the groups of words chosen. In fact, the choice of the words used for this analysis was mainly arbitrary and inspired by other works on the same topic. However, by further exploring the literature about gender stereotypes and by taking into account the specificity of the Italian language through different periods of time, it surely is possible to define better groups of words. Moreover, different methods other than the *WEAT* and *ECT* could be used to further study and enhance the robustness of the *EAB*.

Appendix

1 Ranking of topics by years for EAB, ECT and WEAT



EAB_1948_1968	EAB_avg_gender_1948_1968	ECT_1948_1968	WEAT_1948_1968	EAB_1968_1985	EAB_avg_gender_1968_1985	ECT_1968_1985	WEAT_1968_1985
gendered_words	gendered_words	adj_appearence	family	gendered_words	gendered_words	family	family
active	intelligence	gendered_words	gendered_words	active	family	active	gendered_words
adj_appearence	adj_appearence	dumbness	dumbness	passive	adj_appearence	female_stereotypes	dumbness
male_stereotypes	family	rage	passive	kindness	dumbness	intelligence	passive
kindness	career	female_stereotypes	rage	intelligence	intelligence	passive	rage
passive	male_stereotypes	family	adj_appearence	adj_appearence	female_stereotypes	adj_appearence	active
family	passive	career	kindness	family	kindness	career	adj_appearence
intelligence	female_stereotypes	passive	female_stereotypes	male_stereotypes	male_stereotypes	kindness	kindness
female_stereotypes	kindness	kindness	career	female_stereotypes	passive	dumbness	male_stereotypes
rage	rage	intelligence	male_stereotypes	rage	career	gendered_words	female_stereotypes
dumbness	dumbness	active	intelligence	career	active	rage	intelligence
career	active	male_stereotypes	active	dumbness	rage	male_stereotypes	career



EAB_1985_2000	EAB_avg_gender_1985_2000	ECT_1985_2000	WEAT_1985_2000	EAB_2000_2020	EAB_avg_gender_2000_2020	ECT_2000_2020	WEAT_2000_2020
gendered_words	gendered_words	gendered_words	family	gendered_words	gendered_words	gendered_words	family
family	family	kindness	gendered_words	passive	family	kindness	gendered_words
passive	career	career	passive	family	rage	passive	passive
adj_appearence	passive	female_stereotypes	active	rage	passive	career	dumbness
rage	kindness	family	dumbness	intelligence	career	intelligence	rage
kindness	adj_appearence	rage	rage	adj_appearence	dumbness	family	kindness
active	female_stereotypes	intelligence	kindness	kindness	female_stereotypes	active	female_stereotypes
emale_stereotypes	intelligence	adj_appearence	male_stereotypes	male_stereotypes	male_stereotypes	male_stereotypes	adj_appearence
career	active	dumbness	female_stereotypes	active	kindness	female_stereotypes	male_stereotypes
male_stereotypes	dumbness	passive	intelligence	female_stereotypes	intelligence	rage	career
intelligence	rage	male_stereotypes	adj_appearence	career	adj_appearence	adj_appearence	intelligence
dumbness	male_stereotypes	active	career	dumbness	active	dumbness	active

2 Topics selection results

The *ECT* and *WEAT* selected the following topics:

1. Period: *1948-1968*:
   - Top most biased topics: *physical appearance, dumbness, family, rage*
   - Top least biased topics: *active, career, intelligence, kindness, male stereotypes*
2. Period: *1968-1985*:
   - Top most biased topics: *active, family, passive*
   - Top least biased topics: *career, kindness, male stereotypes*

3. Period: *1985-2000*:
    - Top most biased topics: *family, rage*
    - Top least biased topics: *physical appearance, intelligence, male stereotypes*
4. Period: *2000-2020*:
    - Top most biased topics: *family, kindness, passive*
    - Top least biased topics: *active, physical appearance, female stereotypes, male stereotypes*

By considering also the *EAB* with the word *donna* representing the female gender and *uomo* representing the male gender, the selected topics are as follows:

1. Period: *1948-1968*:
    - Top most biased topics: *physical appearance*
    - Top least biased topics: *career, intelligence*
2. Period: *1968-1985*:
    - Top most biased topics: *active, passive*
    - Top least biased topics: *career, male stereotypes*
3. Period: *1985-2000*:
    - Top most biased topics: *family, rage*
    - Top least biased topics: *intelligence, male stereotypes*
4. Period: *2000-2020*:
    - Top most biased topics: *family, passive*
    - Top least biased topics: *active, female stereotypes, male stereotypes*

3 Topics selection results with TFIDF

The *ECT* and *WEAT* selected the following topics:

1. Period: *1948-1968*:
    - Top most biased topics: *family, career, gendered words, intelligence, kindness, female stereotypes*
    - Top least biased topics: *rage, passive, active, physical appearance, male stereotypes, dumbness*
2. Period: *1968-1985*:
    - Top most biased topics: *career, family, gendered words, kindness, intelligence, female stereotypes*
    - Top least biased topics: *passive, active, male stereotypes, dumbness, rage, physical appearance*

3. Period: *1985-2000*:
   - Top most biased topics: *career, family, gendered words, active, intelligence, passive*
   - Top least biased topics: *male stereotypes, female stereotypes, kindness, dumbness, rage, physical appearance*
4. Period: *2000-2020*:
   - Top most biased topics: *career, family, gendered words, rage, passive, intelligence*
   - Top least biased topics: *active, female stereotypes, kindness, male stereotypes, physical appearance, dumbness*

4 Selected biased words by EAB and ECT

Top biased words selected by *ECT* and *EAB* with single words representing genders:

1. Period: *1948-1968*:
   - Physical appearance: *carino*
   - Family: *accudire,bambino*
   - Career: *presidente*
   - Rage: *aggressivo, odioso*
   - Kindness: *bravo, premuroso*
   - Intelligence: *intelligente, saggio, studioso*
   - Dumbness: *isterico, stupido, superficiale*
   - Active: *ambizioso, forte, sicuro*
   - Passive: *debole*
   - Female stereotypes: *isterico*
   - Male stereotypes: *intelligente, saggio*
2. Period: *1968-1985*:
   - Physical appearance: *brutto*
   - Family: *bambino, matrimonio*
   - Career: *ambizione, presidente*
   - Rage: *intollerante, odioso*
   - Kindness: *bravo, premuroso*
   - Intelligence: *saggio, studioso*
   - Dumbness: *stupido*
   - Active: *ambizioso, forte, sicuro*
   - Passive: *debole, insicuro*
   - Female stereotypes: *frivolo*
   - Male stereotypes: *saggio*
3. Period: *1985-2000*:
   - Physical appearance: *frivolo*
   - Family: *bambino, famiglia*

- Career: \\
- Rage: *cattivo, intollerante*
- Kindness: *bravo, gentile*
- Intelligence: *intelligente, studioso*
- Dumbness: *irrazionale, isterico*
- Active: *ambizioso, forte, sicuro*
- Passive: *passivo, silenzioso*
- Female stereotypes: *frivolo*
- Male stereotypes: *ambizioso, intollerante*

4. Period: *2000-2020*:
   - Physical appearance: *splendido*
   - Family: *bambino, figlio*
   - Career: \\
   - Rage: *brutale*
   - Kindness: *bravo*
   - Intelligence: *intelligente, studioso*
   - Dumbness: *stupido, superficiale*
   - Active: *ambizioso, assertivo, sicuro*
   - Passive: *insicuro*
   - Female stereotypes: *insicuro*
   - Male stereotypes: *crudele*

Top biased words selected by *ECT* and *EAB* when using averaged vectors for representing genders:

1. Period: *1948-1968*:
   - Physical appearance: \\
   - Family: *accudire*
   - Career: *ambizione*
   - Rage: *crudele, odioso*
   - Kindness: *bravo, premuroso*
   - Intelligence: *intelligente, saggio, studioso*
   - Dumbness: *isterico, stupido, superficiale*
   - Active: *ambizioso, forte, sicuro*
   - Passive: *debole, timido*
   - Female stereotypes: *bello*
   - Male stereotypes: *ambizioso, intelligente, saggio*
2. Period: *1968-1985*:
   - Physical appearance: *brutto, frivolo*
   - Family: *bambino*
   - Career: *ambizione*

- Rage: *cattivo, odioso*
- Kindness: *bravo, sensibile*
- Intelligence: *saggio, studioso*
- Dumbness: *isterico, stupido*
- Active: *ambizioso, forte, sicuro*
- Passive: *debole, silenzioso*
- Female stereotypes: *sensibile*
- Male stereotypes: *intelligente, saggio*

3. Period: *1985-2000*:

    - Physical appearance: \\
    - Family: *bambino*
    - Career: \\
    - Rage: *crudele*
    - Kindness: *bravo, gentile*
    - Intelligence: *serio, studioso*
    - Dumbness: *irrazionale, isterico*
    - Active: *ambizioso, forte, sicuro*
    - Passive: *silenzioso*
    - Female stereotypes: \\
    - Male stereotypes: \\

4. Period: *2000-2020*:

    - Physical appearance: *splendido*
    - Family: *bambino*
    - Career: \\
    - Rage: *brutale, odioso*
    - Kindness: *bravo*
    - Intelligence: *intelligente, studioso*
    - Dumbness: *stupido*
    - Active: *ambizioso, assertivo, sicuro*
    - Passive: *timido*
    - Female stereotypes: \\
    - Male stereotypes: *intelligente*

Bibliography

[1] M. Nissim, R. van Noord, R. van der Goot
Fair is better than sensational: Man is to doctor as woman is to doctor. (2020)
Computational Linguistics, 46(2), 487-497.

[2] T. Mikolov, K. Chen, G. Corrado, J. Dean
Efficient estimation of word representations in vector space. (2013)
In: Journal on Educational Resources in Computing (JERIC) 4.4 (2004), p. 2.

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

[4] J.Pennington, R. Socher, C.D. Manning
GloVe: Global Vectors for Word Representation. (2014
Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[5] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou
Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. (2018)
PNAS Proceedings of the National Academy of Sciences of the United States of America, 115(16), E3635–E3644.

[6] A. Rios, R. Joshi, H. Shin
Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings. (2020)
Conference: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, 1–13.

[7] J. Zhao, Y. Zhou, Z. Li, W. Wang, K.W. Chang
Learning Gender-Neutral Word Embeddings. (2018)
Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4847–4853.
LEGGI!

[8] T. Bolukbasi, K.W Chang, J. Zou, V. Saligrama, A. Kalai
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (2016)
Proceedings of the 30th International Conference on Neural Information Processing Systems, 4356–4364.

[9] H. Gonen, Y. Goldberg
Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. (2019)
Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 609–614.

[10] S. Dev, J.M. Phillips
Attenuating Bias in Word Vectors. (2019)
AISTATS 2019.

[11] T. Mikolov, W. Yih, G. Zweig
Linguistic Regularities in Continuous Space Word Representations. (2013)
Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 746–751.

[12] O. Levy, Y. Goldberg
Linguistic Regularities in Sparse and Explicit Word Representations. (2014)
Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 171-180.