# Detecting Bias and Intersectional Bias in Italian Word Embeddings and Language Models

**Alexandre Puttick and Mascha Kurpicz-Briki**
Berner Fachhochschule, Technik und Informatik
Quellgasse 21, 2501, Biel, Switzerland
alexandre.puttick@bfh.ch, mascha.kurpicz@bfh.ch

## Abstract

Bias in Natural Language Processing (NLP) applications has become a critical issue, with many methods developed to measure and mitigate bias in word embeddings and language models. However, most approaches focus on single categories such as gender or ethnicity, neglecting the intersectionality of biases, particularly in non-English languages. This paper addresses these gaps by studying both single-category and intersectional biases in Italian word embeddings and language models. We extend existing bias metrics to Italian, introducing GG-FISE, a novel method for detecting intersectional bias while accounting for grammatical gender. We also adapt the CrowS-Pairs dataset and bias metric to Italian. Through a series of experiments using WEAT, SEAT, and LPBS tests, we identify significant biases along gender and ethnic lines, with particular attention to biases against Romanian and South Asian populations. Our results highlight the need for culturally adapted methods to detect and address biases in multilingual and intersectional contexts.

## 1 Introduction

Bias in Natural Language Processing (NLP) applications has become a widespread problem. Various methods have been developed to measure and partially mitigate bias in word embeddings, e.g., Caliskan et al. (2017); Bolukbasi et al. (2016), and language models, e.g., Ahn and Oh (2021); Guo and Caliskan (2021). However, bias appears across many dimensions and contexts. Therefore, the majority of existing approaches address only one type of bias at a time (e.g., gender or ethnicity). Only and handful of studies, e.g., Guo and Caliskan (2021); Charlesworth et al. (2024), explore intersectional bias, especially in languages other than English. Additional challenges arise when adapting existing bias metrics to gendered languages (Zhou et al., 2019; Omrani Sabbaghi and Caliskan, 2022).

In this paper, we extend the state-of-the-art by providing insights into both single category and intersectional biases in Italian word embeddings and language models. We leverage known metrics and culturally adapt them to the Italian context, in close collaboration with an interdisciplinary team and native speakers. In particular, we introduce GG-FISE, a method for studying intersectional bias based on Charlesworth et al. (2024) that partially corrects for measurement errors resulting from grammatical gender.

**Bias Statement.** This work focuses on *diversity bias*, defined as the unfair positive or negative treatment of individuals based on protected grounds[1], with particular attention to intersectional categories. The technical methods presented here aim to quantify the extent to which potentially harmful social stereotypes are intrinsically encoded within word embeddings and language models. In other words, *a model is understood to be biased if it encodes harmful social stereotypes*. This connects to diversity bias because the use of biased models could lead to harmful outcomes in downstream tasks, where our main research concerns are *gender, ethnic* and/or *intersectional* bias in *AI-assisted hiring decisions*. For example, an NLP hiring system that computes the similarity between job ads and candidate applications using a model encoding stereotypical occupational associations could lead to unfair outcomes.

**Research Questions** This work seeks to explore the following research questions:

**RQ1** To what degree are culture-specific biases based on sensitive attributes (gender, race, etc.) reproduced in Italian-language (contextual) word embeddings?

[1]These grounds include sex, race, color, ethnic or social origin, genetics, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, or sexual orientation.



*Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 33–51

**RQ2** What adaptations to bias detection methods developed in the English-language context are required in order to apply such methods in the Italian context?

**RQ3** How does grammatical gender interact with semantic gender when measuring bias in word embeddings in language models and how can the two concepts be decoupled?

## 2 Methods

### 2.1 WEAT (Caliskan et al., 2017)

The Word-Embedding Association Test (WEAT) requires two categories of wordlists: *attributes* and *targets*. The attributes consist of wordlists $A$ and $B$ representing opposing concepts relating to an aspect of social bias. For example, $A$ = {*executive, management,...*} and $B$ = {*home, parents,...*} are attribute lists representing the concepts of *career* and *family* respectively. The targets are also wordlists $X$ and $Y$; in the case of gender, e.g., $X$ = {*male, man,...*} and $Y$ = {*female, woman,...*}. Using these wordlists, the WEAT test offers a quantitative measure of the degree of bias present in the word embeddings being studied. A detailed explanation of how WEAT metrics are computed is provided in the appendices.

In addition to WEAT tests 6-8 from Caliskan et al. (2017), we include two additional tests, GER1 and GER2, originally conducted in German in Kurpicz-Briki (2020) and in the Swiss context (where Italian is also an official language). All translations to Italian were carried out by native speakers. Five new tests are introduced in this work based on co-creation workshops concerning bias, AI and job recruitment held in Italy, with particular attention to region-specific biases. The tests IT1 and IT2 concern known biases against the Romanian population, while IT3 and IT4 concern biases against individuals with roots in South Asia. The final test, IT5, is meant to detect bias against individuals/communities identifying as queer or trans. A full list of the WEAT experiments carried out in this work is contained in Table 16.

### 2.2 SEAT (May et al., 2019)

The Sentence Embedding Association Test (SEAT) was introduced to extend WEAT to contextual word embeddings. Target and attribute words are inserted into semantically bleached templates such as 'This is WORD' or 'WORD is here.' The word embeddings from WEAT are then replaced with sentence embeddings (our templates are provided in the appendices).

SEAT is intended to work with both static and contextual word embeddings, but the manner in which embeddings are obtained depends on the model being used. For example, for fasttext static embeddings, sentence representations are simply the average of the word vectors over all words in the sentence. For BERT and GPT-2 models, we have implemented multiple methods for obtaining the contextual word embedding associated with a given sentence: *sentence-level* and *token-level*. Details can be found in the appendices.

### 2.3 LPBS (Kurita et al., 2019)

The Log Probability Bias Score (LPBS) is a WEAT-based bias metric specifically designed for masked language models (MLMs) such as BERT. Instead of using cosine-similarity as a measure of the level of association between a target (e.g., *man*) and an attribute (e.g., *programmer*), LPBS uses templates such as 'TARGET is ATTRIBUTE' and computes a similarity score for any target-attribute pair by inserting each into the template and using the corresponding probability scores outputted by the model. Details can be found in the appendices.

The requirement of grammatical gender agreement between targets and attributes in Italian sentences makes the creation of grammatically correct sentences from templates and arbitrary target/attribute lists very difficult[2]. We therefore elect to use the simplified template 'TARGET ATTRIBUTE' for all LPBS tests.

### 2.4 CrowS-Pairs (Nangia et al., 2020)

The use of templates such as those in Kurita et al. (2019) has been criticized for the limited scope and contrived nature of the resulting sentences. Nangia et al. (2020) address this by compiling the *Crowd-sourced Stereotype Pairs (CrowS-Pairs)* dataset, which consists of 1508 sentence pairs dealing with nine types of social bias: race, gender, sexual orientation, religion, age, nationality, disability, physical appearance and socioeconomic status/occupation. As opposed to template-based methods, it is asserted that the crowd-sourced nature of the dataset results in greater diversity and realism in both sentence structure and the stereotypes expressed. Bias is then measured as the percentage of sentence pairs for which the model assigns a higher probability to

[2]E.g., 'Lui è un programmatore.' and 'Lei è una programmatrice.' ('He/She is a programmer').

the stereotypical sentence. Details can be found in the appendices.

The original CrowS-Pairs dataset address bias in a U.S. context. Névéol et al. (2022) adapt CrowS-Pairs to French by first removing all sentences pertaining to stereotypes that do not apply in the French sociocultural context, and then translating and adapting the remaining sentence pairs. They use crowd-sourcing to add additional pairs unique to the French context. We use this French dataset as the basis for the Italian version under the assumption that, as neighboring countries, the regional stereotypes would be more transferable. Given our research interests and time constraints, we extracted only the sentences concerning *gender, nationality* and *race*. The sentences were divided amongst four Italian colleagues, who were instructed to remove sentences that did not apply in Italy and adapt the remaining sentences to the Italian social context. This resulted in 959 sentence pairs: 306 pertaining to gender and 653 to race/nationality.

### 2.5 FISE (Charlesworth et al., 2024)

Flexible Intersectional Stereotype Extraction (FISE) is a novel method for studying intersectional bias in word embeddings. The original work studies bias along three dimensions: *race, gender* and *class*. Similar to the WEAT test, each dimension is represented by a pair of attribute word lists $A$ and $B$ (*white/black, men/women, rich/poor*). A bias score is then computed along each dimension for each word in an additional list of target words, representing the context in which bias is being tested. The authors use two target lists for their analyses. The first consists of 627 *character traits* and the second consists of 130 *occupations*. The computed bias scores yield a scattering of points in the $xy$-plane, with each of the four quadrants representing a single intersectional category (Fig. 2). Once the target words have been divided across quadrants, intersectional bias is measured as two metrics: 1) *word distribution*, 2) *percentage of positive affect*.

**Word distribution.** The proportion of words falling into each quadrant gives an indicator of the degree to which the model associates the concept represented by the target list (*character traits, occupations*) to the corresponding demographic group. For example, if the majority of occupation words fall into the *white male* quadrant, this indicates that the model associates occupations more to white men in general compared to the other intersectional categories.[3]

**Percentage of positive affect.** Charlesworth et al. (2024) also use the percentage of positive vs. negative affect words in each quadrant as bias metrics. Five types of affect are measured: Valence, warmth, competence, arousal and dominance.

#### 2.5.1 Additions and Adaptations

**Identifying intersectional traits.** We define the traits most strongly associated with each intersectional category as those with the largest projection onto the main diagonal of the corresponding quadrant. This implies, for example, that if the occupation *physicist* demonstrates both strong *male* bias and strong *white* bias, it would be strongly associated with the *white male* category.

**Measuring Affect: Valence and Ingressivity.** We measure the percentage of positive vs. negative affect words in each quadrant according to two qualities: *valence* and *ingressivity*. The concepts of ingressivity and *congressivity* are introduced in Cheng (2020) as a means to decouple character traits from the gender identity they are stereotypically associated with. Ingressive traits include being assertive, driven, dominant, competitive and analytical, traits Cheng asserts are both stereotypically masculine and valued/rewarded in a patriarchal (ingressive) society, particularly in the workplace context. In contrast, congressive traits include being empathetic, collaborative, supportive, and open-minded, which are stereotypically feminine and undervalued in society.

We measure valence following Charlesworth et al. (2024), but elect to replace the other affect qualities with ingressivity for three reasons: 1) Unavailability of Italian affect dictionaries analogous to those used in Charlesworth et al. (2024) to measure affect, 2) Warmth, dominance and competence being closely related to ingressivity/congressivity, and 3) Research interest in structural inequalities in the labor market related to social bias concerning gender/ethnic identity and 'desirable/undesirable' character traits in employment.

The affect of a given word is measured using Eq. 1, with attribute lists corresponding

[3]To mitigate the contribution of the choice of occupations, the list was chosen so that jobs associated with different demographic groups and across employment sectors, according to the 2022 U.S. Bureau of Labor Statistics, would be represented equally.

to *pleasant/unpleasant* for valence and *ingressive/congressive* for ingressivity. We use the same valence stimuli as Charlesworth et al. (2024), while the ingressivity stimuli were manually created for this work.

**Translating Wordlists.** As an initial step, machine translation was applied to all word lists from Charlesworth et al. (2024). Then, word lists representing *race* were adapted to the Italian context (e.g. *americano→italiano*). To mitigate the contribution of grammatical gender, both the masculine and feminine forms of all adjectives were included in the *class* and *race* categories, resulting in the `FISE_IT1` test. Finally, the category *black* in the *race* dimension was replaced by an analogous list representing *Romanian*[4], resulting in `FISE_IT2`. Character traits, occupations, and valence and ingressivity stimuli were also machine translated.

**Grammatical Gender and GG-FISE.** Omrani Sabbaghi and Caliskan (2022) provide evidence that grammatical gender has a significant effect on WEAT measurements. Similar effects are therefore expected in attempting to translate FISE to gendered languages.[5] A variant of the FISE method is carried out in this work by replacing the embeddings of character traits and occupations with the *average of the embeddings corresponding to the masculine and feminine forms* (in cases where the two differ). We call this new method *Grammatical Gender FISE* (GG-FISE).

## 3 Experiments

**Models:** The following models were used in this work: 1) Fasttext: `cc.it.300`[6], 2) BERT: `dbmdz/bert-base-italian-uncased`[7], 3) GPT-2: `GroNLP/gpt2-small-italian`[8] Fasttext embeddings were obtained using the *fasttext* Python library, while the BERT and GPT-2 models were implemented using the Huggingface `transformers` library. Additional model details can be found in the appendices.

[4]Romanians make up the largest immigrant demographic group in Italy and face many harmful stereotypes there.

[5]Charlesworth et al. (2024) provide preliminary tests in French in supplementary material. However, grammatical gender is not addressed.

[6]https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.it.300.bin.gz

[7]https://huggingface.co/dbmdz/bert-base-italian-uncased

[8]https://huggingface.co/GroNLP/gpt2-small-italian

**Single Category Bias Detection:** WEAT and SEAT tests were computed for all three models. For transformer models, SEAT was conducted using both token- and sentence-level embeddings. For the BERT model, LPBS effect sizes and CrowS-Pairs scores were also computed.

**FISE - Intersectional Bias Detection** This work focuses on *gender+race* intersectional categories. The following variables are explored in FISE experiments: (a) Model choice: `fasttext`, `bert`, `gpt-2`, (b) Test type: `FISE_IT1`, `FISE_IT2` and (c) Grammatical Gender: unbalanced vs. balanced vs. GG-FISE Token-level embeddings were used for BERT and GPT-2 models.

**Study 1: Analyzing Intersectional Bias**

In order to minimize the effect of grammatical gender, experiments were performed using grammatically gender-balanced affect stimuli and the GG-FISE method. In addition to studying word distribution and the proportion of positive affect words in each quadrant, a qualitative analysis was performed on the words most strongly associated with each intersectional category (see Appendix A.4). For consistency, affect was always measured using fasttext embeddings, independent of the model being tested. In a first experiment, we carried out all FISE tests using both occupation and character trait lists and fasttext embeddings. Following this, we restricted our attention to occupations and compared across model types.

**Study 2: Grammatical Gender**

To study the effect of grammatical gender on FISE, attention was restricted to occupations and FISE-IT1. The occupation lists were varied across gg-unbalanced, gg-balanced and GG-FISE, and tests were carried out on both fasttext and BERT embeddings.

## 4 Results

### 4.1 WEAT

**Fasttext:** Table 1 shows the results of all WEAT tests on fasttext static embeddings. Bias was detected for WEAT 6, GER1, IT1, IT2 and IT4, each with relatively large effect sizes. This indicates that Italian fasttext embeddings demonstrate significant gender bias with respect to societal roles and areas of study. In terms of ethnicity, the embeddings encode noticeable bias against Romanians, linking

them to unpleasantness and low-skilled jobs. Indian ethnicity was also associated with low-skilled work, but did not demonstrate bias with respect to pleasantness. While our threshold $p$-value was set to $0.05$, it is worth noting that relatively small $p$-values were measured for all tests aside from WEAT 7 and WEAT 8, indicating an elevated possibility that the embeddings encode the associated biases.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| WEAT_8 | 0.3047 | 0.27896145 | ✗ |
| WEAT_7 | 0.1456 | 0.5567624 | ✗ |
| WEAT_6 | 0.0001 | 1.7019253 | ✓ |
| GER1 | 0.0203 | 1.2168527 | ✓ |
| GER2 | 0.0982 | 0.60425156 | ✗ |
| IT_1 | 0.0009 | 1.364754 | ✓ |
| IT_2 | 0.0001 | 1.4764705 | ✓ |
| IT_3 | 0.0678 | 0.69634694 | ✗ |
| IT_4 | 0.0001 | 1.7231854 | ✓ |
| IT_5 | 0.0695 | 1.1506343 | ✗ |

Table 1: WEAT effect sizes and p-values for Italian fasttext.

**BERT:** The results for WEAT testing on the Italian BERT model are displayed in Table 2. The model demonstrated significant bias in WEAT 6, IT2 and IT4, i.e., stereotypical gender associations regarding career vs. family, as well as linking Romanian and Indian ethnicities to low-skilled labor.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| WEAT_8 | 0.1178 | 0.6117299 | ✗ |
| WEAT_7 | 0.1649 | 0.5167636 | ✗ |
| WEAT_6 | 0.0159 | 0.9318852 | ✓ |
| GER1 | 0.4518 | 0.08670033 | ✗ |
| GER2 | 0.9785 | -1.2163782 | ✗ |
| IT_1 | 0.7694 | -0.37623438 | ✗ |
| IT_2 | 0.0113 | 0.98730487 | ✓ |
| IT_3 | 0.9649 | -0.8377872 | ✗ |
| IT_4 | 0.0002 | 1.3696517 | ✓ |
| IT_5 | 0.5721 | -0.117931664 | ✗ |

Table 2: WEAT effect sizes and p-values for Italian BERT.

**GPT-2:** Of the three models, Italian GPT-2 demonstrated bias in the fewest categories, GER2 and IT4; the model appears to associate rationality to men and emotion to women, as well as Indians to low-skilled work. The $p$-value for IT2 is also relatively low, providing grounds to further study model bias with respect to Romanians and low-skilled work. See Table 3 for details.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| WEAT_8 | 0.2163 | 0.42313254 | ✗ |
| WEAT_7 | 0.295 | 0.27914122 | ✗ |
| WEAT_6 | 0.6445 | -0.17659171 | ✗ |
| GER1 | 0.9777 | -1.2074564 | ✗ |
| GER2 | 0.0217 | 1.1980162 | ✓ |
| IT_1 | 0.2552 | 0.31297994 | ✗ |
| IT_2 | 0.0854 | 0.62772095 | ✗ |
| IT_3 | 0.1759 | 0.44588587 | ✗ |
| IT_4 | 0.0045 | 1.1476842 | ✓ |
| IT_5 | 0.6283 | -0.5059762 | ✗ |

Table 3: WEAT effect sizes and p-values for Italian GPT-2.

### 4.2 SEAT

The results for SEAT tests are detailed in Appendix A.1. In general, SEAT tests identified less bias than WEAT tests. The bias detected for fasttext in SEAT-IT2 (Table 7), corroborates the bias detected in WEAT experiments. Similary, the BERT results for SEAT-IT4 corroborate the bias detected in the WEAT IT4 test, see Tables 8 and 9. SEAT-WEAT-7 also yielded a relatively small $p$-value, indicating the need for further investigation with respect to stereotypical gender bias in math vs. art. The use of token-level vs. sentence-level embeddings did not yield significant differences. GPT-2 SEAT results in Tables 10 and 11 demonstrate the same biases as WEAT tests. Interestingly, both sentence-level and token-level embedding SEAT were required to redetect the two biases detected in GPT-2 WEAT experiments.

### 4.3 LPBS

Table 4 contains the LPBS results on the BERT model. Although the tests only detected bias in GER2, this particular bias was not detected by WEAT or SEAT.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| LPBS_WEAT_8 | 0.8217 | -0.4967 | ✗ |
| LPBS_WEAT_7 | 0.6597 | -0.2216 | ✗ |
| LPBS_WEAT_6 | 0.6111 | -0.1511 | ✗ |
| LPBS_GER1 | 0.5733 | -0.09881 | ✗ |
| LPBS_GER2 | 0.0257 | 0.9904 | ✓ |
| LPBS_IT_1 | 0.4259 | 0.1052 | ✗ |
| LPBS_IT_2 | 0.7803 | -0.3865 | ✗ |
| LPBS_IT_3 | 0.5181 | -0.0189 | ✗ |
| LPBS_IT_4 | 0.7964 | -0.4417 | ✗ |
| LPBS_IT_5 | 0.5511 | -0.0595 | ✗ |

Table 4: Effect sizes and $p$-values for LPBS on BERT.

### 4.4 CrowS-Pairs

Table 5 contains the CrowS-Pairs bias score on the BERT model. The results indicate that the model demonstrates some gender and race/nationality

bias, which (Nangia et al., 2020) define as any score above 50. For reference, the gender, race/color and nationality bias scores measured for English BERT in (Nangia et al., 2020) are 58.0, 58.1 and 62.9 respectively. The adaptation of the dataset from the U.S. to Italian context via French may explain the lower bias measurements, as many common stereotypes relevant to Italian may not appear. The bias was more pronounced concerning positive stereotypes about privileged groups (i.e. *men* and *Italians*).

| Test | Bias | Bias$^-$ | Bias$^+$ | % Neutr. |
|---|---|---|---|---|
| All | 51.3 | 51.44 | 55.86 | 1.56 |
| Gender | 53.27 | 52.43 | 55.56 | 0.33 |
| Race/Nationality | 50.38 | 51.1 | 56.52 | 2.14 |

Table 5: CrowS-Pairs bias scores for BERT. **Bias**$^+$ is the score when restricted to sentence pairs concerning negative stereotypes about underprivileged groups, while **Bias**$^-$ corresponds to positive stereotypes about privileged groups. The last column shows the percentage of total sentence pairs for which the model displayed no preference.

## 4.5 FISE

### 4.5.1 Study 1: Analyzing Intersectional Bias

#### Experiment 1: Fasttext, GG-FISE, Traits and Occupations

**Word Distributions:** The first column of Table 6 contains the word distributions across both FISE tests. Surprisingly, in FISE-IT1 the word distributions skewed towards the *black* (75.6% of occupations, 54.5% of traits) and *women* (52.9% occupations, 71% traits), with most words landing in the *black women* quadrant. Figure 1 depicts a plot of the word distributions for FISE-IT1.

The word distributions for FISE-IT2 aligned with expectations given negative stereotypes in Italy against people of Romanian descent. In the Romanian quadrants, words were skewed in the female direction. In both FISE-IT1 and FISE-IT2, with ethnicity fixed, character traits all skewed towards female. In the Italian quadrants, occupations skewed male, while the non-Italian quadrants showed the reverse trend.

**Valence:** In all cases the *white/italian* quadrants contained higher percentages of words with positive valence, with the exception of character traits in men, where the Romanian quadrant contains a higher proportion of positive words. This is likely an artifact of the fact that only about 3% of the total words are contained in the Romanian men quadrant. FISE-IT1 demonstrated particularly strong bias, with a large majority of character traits and most jobs in the white quadrants being positive. The majority of jobs in the black quadrants had negative valence, with 70% of the jobs associated with black men having negative (unpleasant) associations. The occupation valence skew was even more pronounced in FISE-IT2, with a large majority of the occupations in the *Romanian* quadrants having negative associations. In terms of character traits, *black men* and *Romanian women* were the only intersectional categories with the majority of character traits being negative.

| Test | Quadrants | Word Distributions | Pos. Valence(%) | Pos. Ingressivity(%) |
|---|---|---|---|---|
| FISE_IT1_occ. | men white | 14.600 | **55.600** | 77.778 |
| | men black | 32.500 | 30.000 | 37.500 |
| | women black | **43.100** | 39.600 | 24.528 |
| | women white | 9.800 | 50.000 | **83.333** |
| FISE_IT1_traits | men black | 15.500 | 41.200 | 61.250 |
| | women black | **39.000** | 54.200 | 51.244 |
| | men white | 13.400 | **78.300** | **68.116** |
| | women white | 32.000 | 78.200 | 62.424 |
| FISE_IT2_occ. | men italian | **39.000** | 43.800 | **56.250** |
| | women italian | 35.800 | **47.700** | 38.636 |
| | men romanian | 8.100 | 10.000 | 20.000 |
| | women romanian | 17.100 | 28.600 | 28.571 |
| FISE_IT2_traits | men italian | 25.800 | 57.900 | 61.654 |
| | women italian | **65.800** | **66.700** | 54.572 |
| | women romanian | 5.200 | 44.400 | 77.778 |
| | men romanian | 3.100 | 62.500 | **87.500** |

Table 6: Results for Experiment 1. This experiment used gender balanced affect stimuli lists with each word appearing in both masculine and feminine form. The word embeddings for character traits or occupations were obtained by averaging the embeddings for the masculine and feminine forms (in cases where the two forms differ).

**Ingressivity:** In terms of ingressivity, occupations were skewed much more along the race/ethnicity axis than along the gender axis, to the extent that the white women quadrant in FISE-IT1 contained the highest proportion of ingressive jobs. However, only a small number of jobs overall landed in that quadrant. Also of note is that the majority of occupations associated with Italian men are ingressive in both tests. On the other hand, the majority of jobs associated with non-Italian quadrants were congressive. In terms of character traits, black women in FISE-IT1 and Italian women in FISE-IT2 showed the lowest ingressivity. Whereas ingressive character traits tended toward Italian when compared to black, the skewed strongly towards Romanian in FISE-2. In the Italian quadrants character traits only demonstrated a slight stereotypical ingressivity skew towards men, whereas the

gender difference was much more pronounced in non-Italian quadrants.

**Experiment 2 - All models, gender-balanced occupation wordlists** The second experiment tested fasttext, BERT and GPT-2 models. Because preliminary tests using GG-FISE yielded extremely skewed results for the transformer models, these tests used grammatically gender-balanced occupation list instead. Results for Experiment 2 can be found in Table 12.

In FISE-IT1, both BERT and GPT-2 demonstrated a dramatic skew towards the *black women* quadrant, with on the order of only ten words landing in the *white* quadrants. For that reason, valence and ingressivity measures do not have a meaningful interpretation for the corresponding categories. Occupations associated with *black men* are more positive and more ingressive compared to *black women*.

In contrast to fasttext, the BERT and GPT-2 demonstrated a similarly unexpected skew towards the *Romanian* quadrants in FISE-IT2. For BERT, nearly half of the words landed in the *Romanian women* quadrants, while the remaining words were somewhat evenly distributed among the remaining quadrants. GPT-2 also defied expectations, with only three occupations landing in the *Italian men* quadrant. Again, the largest portion of words landed in the *Romanian women* quadrant, with 77.4% landing in the *Romanian* half-plane overall. For BERT, occupations associated with *Italian* quadrants were significantly more positive and ingressive, although ingressivity was gender-atypical for both ethnicities. Ignoring *Italian men* for GPT-2, a similar trend occurs in valence and ingressivity along the *race/ethnicity* axis, but on the *Romanian* side the proportions between genders of positive/ingressive traits were reversed relative to BERT.

#### 4.5.2 Study 2: Grammatical Gender

**Experiment 3 - Fasttext/BERT, grammatical gender** Experiment 3 tested different approaches to handling grammatical gender on fasttext and BERT models. For fasttext, the difference between GG-FISE and using a gender-balanced occupation list was not very significant. The most notable change was the reversal of the distribution imbalance between *black women* and *black men*. As expected, word distributions shifted dramatically towards the *male* quadrants when only masculine forms of occupations were used. Figure 13 depicts the top (up to) 15 intersectional words in each of the different cases for fasttext embeddings. Grammatical gender played a significant role: When exclusively male forms were used, only grammatically gender-neutral occupations appear in the *women* quadrants. When the occupation list was augmented to include feminine forms, the resulting words are clearly distributed according to grammatical gender. The contribution of grammatical gender appears to vanish if GG-FISE is used.

## 5 Discussion

**Single Category Bias** In the case of static word embeddings, WEAT tests provided ample evidence that Italian fasttext embeddings encode stereotypical biases regarding gender roles and societal expectations. Men are more associated with career, while women are more associated with family. Although the wordlists containing the academic fields of study with the highest gender imbalances were compiled in a Swiss context, indicators for the same biases were also detected, associating fields such as engineering and computer science to men, and pedagogy and psychology to women.

In 2021, The Italian National Institute of Statistics (ISTAT) reported that Romanians make up the largest immigrant group in Italy, nearly a quarter of all foreign residents. Together, Indian and Bangladeshi residents make up 6.5% of the immigrant population, making South Asia the most represented region of the Asian continent. Of the members of each minority with work experience in a foreign country, the majority of that experience was in low-skilled work.[9] Negative stereotypes against the Romanian population in Italy are documented in existing work, e.g., Popescu (2008). Prejudices are exacerbated by the association between Romanians and the Roma people, who face prejudice and marginalization across Europe (Sam Nariman et al., 2020). Our findings provide evidence that all of these biases are encoded within language models, demonstrating that fasttext embeddings associate low-skilled labor and unpleasantness to both groups. Moreover, the results provide good evidence (p=0.07) that the embeddings also encode harmful prejudices against queer and transgender identities.

[9] https://www.istat.it/it/files//2023/02/Focus_stranieri-e-naturalizzati-nel-mondo-del-lavoro.pdf

WEAT testing also indicated gender and ethnic biases in contextual models. In the case of gender, the model appear to encode stereotypes regarding career and family, as well as the gendering of rational vs. emotional character. Although our results do not indicate negative associations to Romanian and South Asian minorities, both BERT and GPT-2 transformer models appear to associate the two groups to low-skilled labor.

The biases encoded in these word embeddings and language models can have dire social consequences. In particular, our findings indicate significant gender and ethnic encoded bias in the occupational context. The use of technologies built on such models in the labor market could reinforce existing inequalities in hiring practices and prolong structural inequalities.

**Intersectional Bias** According to a national report on Romanian immigrants in Italy[10], most Romanian men work in the construction sector, whereas Romanian women are associated with domestic or care work, but are also often employed in shops, hotels and restaurants, health care, and social services. Our findings indicate that similar biases are encoded in word embeddings, most notably through the low ingressivity of occupations in Romanian quadrants (6) and the top intersection occupations identified in Table 15. Romanian men are also associated with corruption and crime (Bratu, 2014), which is reflected in both the IT1 WEAT test and the character traits most associated with Romanian men (Table 15), which are largely negative and include qualities like autocratic and bellicose. In Italy, there is also a large gender divide in the Romanian population (41.7% male, 58.3% female in 2021),[11] which may explain why nearly twice as many words landed in the Romanian women category compared to Romanian men.

Our findings also demonstrate particular intersectional biases with respect to women, most visible in the top intersectional traits corresponding to each intersectional group (Tables 14 and 15). With regard to character traits, Italian women are associated with femininity, romance, worldliness, and refinement, with proportionally more positive traits relative to women of other ethnicities. Black women are associated with more sexualized traits as well as superstition and other negative words. The traits most associated Romanian women are uniformly negative.

Not all of our results align with expectations regarding known stereotypes. For example, the fact that the majority of words landed in black quadrants (Table 6) defies intuition. In general, comparing Italian to Romanian led to results that were more aligned with the expectation that stereotypical biases are reproduced in word embeddings. This could stem from the fact that the black population in Italy is relatively small, with no predominantly black countries among the top ten countries of origin for foreign residents. This could correlate to a dearth of examples in the models' training corpus, resulting in noisy embeddings and less validity in our experiments, with the potential for additional noise pertaining to a proportionally large number of corpus occurrences of the color 'black' Future work could compensate for such noise by making use of appropriate context when testing contextual word embeddings.

There were also further unexpected observations regarding occupations. We draw particular attention to the presence of jailer, lawyer and paralegal in black quadrants. Such observations do not necessarily imply that those jobs employ more black people, but could instead reflect more frequent encounters with a discriminatory justice system.

The high level of ingressivity measured for Italian women in FISE-IT1 were also surprising. While statistical error could contribute, a portion of the measured ingressivity could also come from a general higher attribution of ingressive traits to white Italians in comparison to black Italians. The relatively close levels of ingressivity between Italian men and women in the same test could also be linked to the 'strong/fiery' stereotypes often associated with Southern European women.

**Technical Methods** Our findings suggest that grammatical gender plays a significant role in bias measurements and should be carefully accounted for. In addition. SEAT methods did not measure any biases that were not already detected by WEAT methods. This is not surprising, as using the same sentence templates for every word would be expected to make the corresponding embeddings more similar. LPBS, however, computes word similarity in an entirely different manner. Although not as many biases were detected using this method, LPBS proved to be an important complement be-

[10] https://www.participation-citoyenne.eu/sites/default/files/report-italy.pdf

[11] https://www.istat.it/it/files//2023/02/Focus_stranieri-e-naturalizzati-nel-mondo-del-lavoro.pdf

cause it detected bias in BERT that other methods overlooked. CrowS-Pairs tests also indicate that the BERT model encodes a significant number of harmful stereotypes. It would be interesting to expand the dataset to encompass a broader variety of stereotypes particular to the Italian context.

**Conclusion** With respect to RQ1, the bias measurement techniques carried out in this work demonstrate strong evidence that harmful gender and ethnic (intersectional) biases are encoded in both static and contextual Italian word embeddings. Of particular value are the tests particularly tailored to the Italian cultural context, which detect encoded biases that would not be detected by simple translation of existing methods.

In response to RQ2, we find that many elements of existing bias detection methods are particular to an English-language and American context. Careful cultural adaptation requires extensive investigation of local stereotypes and validation by native-speakers. Connecting to RQ3, linguistic adapation is also essential, particularly regarding interference between grammatical and semantic gender. However, the averaging approach we employ to mitigate the contribution of grammatical gender may not be precise enough to preserve essential semantic information. Future work could investigate more sophisticated methods for removal of the grammatical gender component of word embeddings (Omrani Sabbaghi and Caliskan, 2022; Zhou et al., 2019).

Recently, large language models (LLMs) have largely superseded the language models studied in this work. The obsolescence of the models studied here is a significant limitation. LLM bias detection is dominated by prompt-based methods, in part because many such models are proprietary and researchers do not have access to the models' inner workings. However, in cases where the necessary information is accessible, the methods described here by be adapted to state-of-the-art models as well. This could be the subject of future work.

Moreover, the datedness of the models studied in this work does not preclude their use; they may be better suited than LLMs to many NLP tasks in which social bias is relevant, even beyond saving on computational costs. For example, this work was undertaken in the context of studying fairness and bias in AI-assisted recruitment. In this context, understanding of which features about job candidates were used to render a decision is a necessity to ensure fairness. Well-developed explainable AI methods make the models studied in this work more relevant in such situations.

## Limitations

- These methods may not be as well-suited for bias detection in transformer-based contextual models. As carried out here, the FISE method did not yield convincing results for contextual embeddings and suitable adaptations in the Italian context should be further investigated. Further adaptations to address grammatical gender are also needed.
- Oversimplified LPBS templates may have adversely affected the bias detection capacity of this method.
- Machine translation and other automated adaptations for Italian may have yielded errors in FISE. Verification by native speakers would improve the reliability of our methods.
- Although the stereotypes present in the French CrowS-Pairs dataset were adapted for the Italian context, it is possible that many common stereotypes particular to Italian were omitted.
- It is also possible that grammatical gender agreement obfuscates some of the gender bias, because the pseudo-log-likelihood of the anti-stereotypical sentence would be artificially increased by gender agreement. Moreover, the binary comparison structure of CrowS-Pairs renders it difficult to extend the method to intersectional bias. Adaptations of the StereoSet dataset (Nadeem et al., 2020) may circumvent these limitations and are being explored for future work.
- Due to limitations in time and computational power, not every test was conducted on transformer-based models. These limitations also prevented testing much larger LLMs, which are rapidly replacing the models studied in this work.
- More extensive research is needed to understand how the biases detected in this work affect downstream applications.
- While affirmative detection of bias can be considered significant, failure of our methods to detect certain biases does not confirm that they are not present.

- Occupation lists were adapted from a U.S. context. Recent work provides a list of gender-imbalanced occupations in Italy, which could help validate our methods against real-world data (Ruzzetti et al., 2023). However, these occupations are not labeled according to gender. More granular demographic data by occupation would be desirable.

- The FISE method is not well-suited to the detection of *emergent bias*, i.e., stereotypes pertaining to an intersectional category that are not attributed to any of the individual component categories. For example, black women may be stereotyped as being unfeminine.

- FISE measurements do not include corresponding significance tests. This is particularly limiting in the several cases where small sample sizes within a given quadrant yielded unreliable results.

## References


Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in bert. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Roxana Bratu. 2014. Portrayals of romanian migrants in ethnic media from italy. *Journal of Comparative Research in Anthropology and Sociology*, 5(02):199–217.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tessa ES Charlesworth, Kshitish Ghate, Aylin Caliskan, and Mahzarin R Banaji. 2024. Extracting intersectional stereotypes from embeddings: Developing and validating the flexible intersectional stereotype extraction procedure. *PNAS nexus*, 3(3):pgae089.

Eugenia Cheng. 2020. *x+ y: a mathematician's manifesto for rethinking gender*. Hachette UK.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1693–1706. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. In *Proceedings of 5th SwissText and 16th KONVENS Joint Conference 2020*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531.

Shiva Omrani Sabbaghi and Aylin Caliskan. 2022. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531.

Teodora Popescu. 2008. Immigration discourses: the case of romanian immigrants in italy. *Journal of Linguistic and Intercultural Education*, 1:31–44.

Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Davide Venditti, Fabio Massimo Zanzotto, et al. 2023. Investigating gender bias in large language models for the italian language. In *CLiC-it*.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Hadi Sam Nariman, Márton Hadarics, Anna Kende, Barbara Lášticová, Xenia Daniela Poslon, Miroslav Popper, Mihaela Boza, Andreea Ernst-Vintila, Constantina Badea, Yara Mahfud, et al. 2020. Anti-roma

bias (stereotypes, prejudice, behavioral tendencies): A network approach toward attitude strength. *Frontiers in psychology*, 11:2071.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284.

# A Additional results and figures

## A.1 SEAT Results

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| SEAT_WEAT_8 | 0.683 | -0.1671093 | ✗ |
| SEAT_WEAT_7 | 0.4336 | 0.06639725 | ✗ |
| SEAT_WEAT_6 | 0.2505 | 0.22172295 | ✗ |
| SEAT_GER1 | 0.1689 | 0.36100748 | ✗ |
| SEAT_GER2 | 0.1017 | 0.46662807 | ✗ |
| SEAT_IT_1 | 0.0001 | 1.1802236 | ✓ |
| SEAT_IT_2 | 0.3168 | 0.15640602 | ✗ |
| SEAT_IT_3 | 0.4437 | 0.049905647 | ✗ |
| SEAT_IT_4 | 0.7868 | -0.25795597 | ✗ |
| SEAT_IT_5 | 0.3957 | 0.023056474 | ✗ |

Table 7: SEAT effect sizes and $p$-values for Italian fasttext.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| SEAT_WEAT_8 | 0.4602 | 0.03401969 | ✗ |
| SEAT_WEAT_7 | 0.2132 | 0.28997424 | ✗ |
| SEAT_WEAT_6 | 0.5103 | -0.016036926 | ✗ |
| SEAT_GER1 | 0.636 | -0.15082084 | ✗ |
| SEAT_GER2 | 0.4226 | 0.07441554 | ✗ |
| SEAT_IT_1 | 0.4073 | 0.08290798 | ✗ |
| SEAT_IT_2 | 0.1354 | 0.35401675 | ✗ |
| SEAT_IT_3 | 0.8568 | -0.33514008 | ✗ |
| SEAT_IT_4 | 0.0003 | 1.1244862 | ✓ |
| SEAT_IT_5 | 0.089 | 0.6980704 | ✗ |

Table 8: SEAT effect sizes and $p$-values for Italian BERT using token embeddings.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| SEAT_WEAT_8 | 0.3214 | 0.16407524 | ✗ |
| SEAT_WEAT_7 | 0.0791 | 0.51762867 | ✗ |
| SEAT_WEAT_6 | 0.8805 | -0.3841112 | ✗ |
| SEAT_GER1 | 0.8872 | -0.54121554 | ✗ |
| SEAT_GER2 | 0.4194 | 0.07997835 | ✗ |
| SEAT_IT_1 | 0.767 | -0.23425224 | ✗ |
| SEAT_IT_2 | 0.1791 | 0.29892346 | ✗ |
| SEAT_IT_3 | 0.9981 | -0.91031444 | ✗ |
| SEAT_IT_4 | 0.0062 | 0.80213153 | ✓ |
| SEAT_IT_5 | 0.3663 | 0.17249337 | ✗ |

Table 9: SEAT effect sizes and $p$-values for Italian BERT using sentence embeddings.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| SEAT_WEAT_8 | 0.1473 | 0.37586936 | ✗ |
| SEAT_WEAT_7 | 0.5336 | -0.02856302 | ✗ |
| SEAT_WEAT_6 | 0.0603 | 0.49951243 | ✗ |
| SEAT_GER1 | 0.359 | 0.1545632 | ✗ |
| SEAT_GER2 | 0.2044 | 0.3445308 | ✗ |
| SEAT_IT_1 | 0.7695 | -0.24369203 | ✗ |
| SEAT_IT_2 | 0.6974 | -0.16219279 | ✗ |
| SEAT_IT_3 | 0.9136 | -0.42532745 | ✗ |
| SEAT_IT_4 | 0.0408 | 0.5590291 | ✓ |
| SEAT_IT_5 | 0.4907 | 0.020580258 | ✗ |

Table 10: SEAT effect sizes and $p$-values for Italian GPT-2 using token embeddings.

| Name | p-value | Effect Size | Bias Detected? |
|---|---|---|---|
| SEAT_WEAT_8 | 0.9979 | -0.9598411 | ✗ |
| SEAT_WEAT_7 | 0.8857 | -0.42705518 | ✗ |
| SEAT_WEAT_6 | 0.4188 | 0.06560292 | ✗ |
| SEAT_GER1 | 0.6287 | -0.13893525 | ✗ |
| SEAT_GER2 | 0.0096 | 0.9666244 | ✓ |
| SEAT_IT_1 | 0.6403 | -0.11274278 | ✗ |
| SEAT_IT_2 | 0.8576 | -0.3475058 | ✗ |
| SEAT_IT_3 | 0.1677 | 0.30809855 | ✗ |
| SEAT_IT_4 | 0.33 | 0.14957048 | ✗ |
| SEAT_IT_5 | 0.4673 | 0.05021583 | ✗ |

Table 11: SEAT effect sizes and $p$-values for Italian GPT-2 using sentence embeddings.

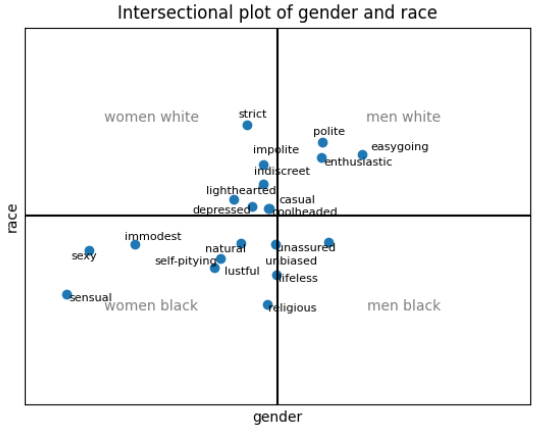


Figure 2: Example of character traits mapped into intersectional categories using word-embedding bias.

## A.2 Additional FISE results and figures

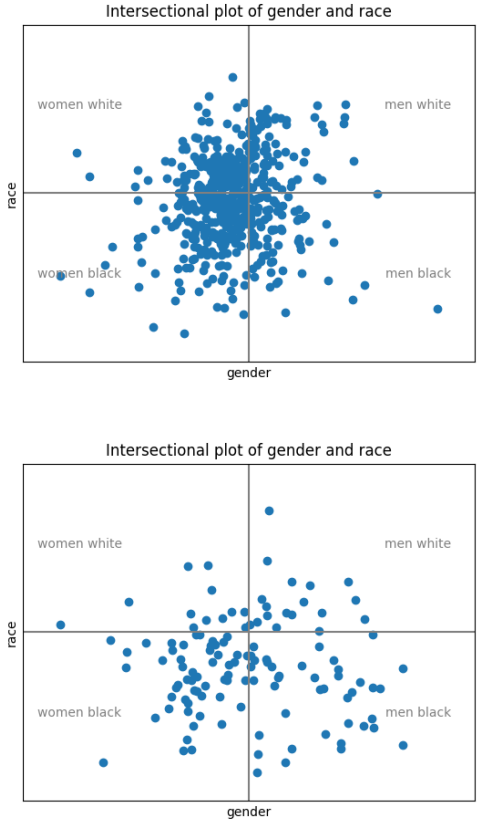


Figure 1: Distribution of character traits (top) and occupations (bottom) for FISE-IT1 in Experiment 1.

| Test | Quadrants | Word Distributions | Pos. Valence(%) | Pos. Ingressivity(%) |
|---|---|---|---|---|
| FISE_IT1_occ_fasttext | men white | 13.700 | 50.000 | 75.000 |
| | men black | **39.700** | 39.500 | 32.099 |
| | women black | 35.800 | 37.000 | 26.027 |
| | women white | 10.800 | **63.600** | **90.909** |
| FISE_IT1_occ_bert | women black | **67.200** | 41.600 | 43.796 |
| | men black | 31.900 | 44.600 | 38.462 |
| | men white | 0.500 | **100.000** | **100.000** |
| | women white | 0.500 | 0.000 | 0.000 |
| FISE_IT1_occ_gpt2 | men black | 34.800 | 46.500 | 45.070 |
| | women black | **53.400** | 38.500 | 38.532 |
| | women white | 10.800 | **54.500** | **50.000** |
| | men white | 1.000 | 0.000 | **50.000** |
| FISE_IT2_occ_fasttext | men italian | **39.200** | 45.000 | **48.750** |
| | women italian | 33.300 | **50.000** | 44.118 |
| | men romanian | 14.200 | 34.500 | 27.586 |
| | women romanian | 13.200 | 25.900 | 33.333 |
| FISE_IT2_occ_bert | women romanian | **49.000** | 39.000 | 40.000 |
| | men romanian | 15.700 | 34.400 | 28.125 |
| | men italian | 16.700 | **55.900** | 50.000 |
| | women italian | 18.600 | 47.400 | **52.632** |
| FISE_IT2_occ_gpt2 | men romanian | 34.300 | 45.700 | 44.286 |
| | women romanian | **43.100** | 33.000 | 32.955 |
| | men italian | 1.500 | 33.300 | **66.667** |
| | women italian | 21.100 | **58.100** | 55.814 |

Table 12: Results for Experiment 2. The same affect stimuli were used as in Experiment 1. Word embeddings were not averaged over grammatical gender forms, but both masculine and feminine forms of each occupation were tested.

| Test | Quadrants | Word Distributions |
|---|---|---|
| FISE_IT1_occ_gg_fasttext | men white | 14.600 |
| | men black | 32.500 |
| | women black | **42.300** |
| | women white | 10.600 |
| FISE_IT1_occ_gg_bert | women black | **85.400** |
| | men black | 14.600 |
| FISE_IT1_occ_genbal_fasttext | men white | 13.700 |
| | men black | **39.700** |
| | women black | 35.800 |
| | women white | 10.800 |
| FISE_IT1_occ_genbal_bert | women black | **67.200** |
| | men black | 31.900 |
| | men white | 0.500 |
| | women white | 0.500 |
| FISE_IT1_occ_fasttext | men white | 22.600 |
| | men black | **64.500** |
| | women black | 10.500 |
| | women white | 2.400 |
| FISE_IT1_occ_bert | women black | 46.800 |
| | men black | **52.400** |
| | men white | 0.800 |

Table 13: Results for Experiment 3. The occ_gg tests used word embeddings averaged over grammatical gender, occ_genbal refers to gender balanced wordlists and while occ used only the masculine forms.

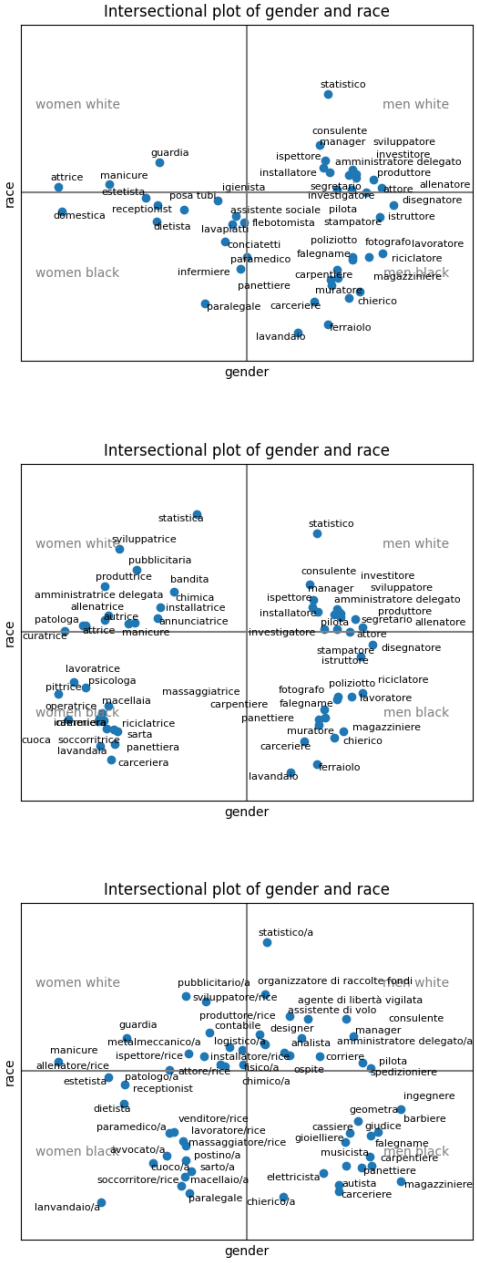


Figure 3: Effect of grammatical gender on wordlists for fasttext. The top 15 intersectional words (unnormalized) in each quadrant are displayed. (*Top*) Word embeddings averaged over grammatical gender. (*Middle*) Gender-balanced occupation list. (*Bottom*) Only masculine forms of occupations.

### A.3 Top affect traits

**Top 5 traits with highest ingressivity**: determinato/a, dominate, decisivo/a, ambizioso/a, impavido/a

**Top 5 traits with highest congressivity**: cordiale, gentile, cortese, compassionevole, premuroso/a

**Top 5 traits with highest valence**: cordiale, sereno/a, rilassato/a, piacevole, gentile

**Top 5 traits with lowest valence**: viscido/a, immorale, vigliacco/a, inefficiente, inetto/a

**Top 5 occupations with highest ingressivity**: programmatore/rice, investitore/rice, organizzatore/rice di raccolte fondi, perito/a, agente di libertà vigilata

**Top 5 occupations with lowest ingressivity**: cameriere, barista, receptionist, cuoco/a, cassiere

**Top 5 occupations with highest valence**: ospite, chef, manicure, massaggiatore/rice, disegnatore/rice

**Top 5 occupations with lowest valence**: macchinista, perito/a, bandito/a, riciclatore/rice, carceriere

### A.4 Intersectional words

Tables 14 and 15 contain the top intersectional words for each quadrant. Generally, many words appear to align with societal stereotypes, many of which relate to unfair stereotypes.

In FISE-IT1, white men are associated with words such as consultant, manager, good-humored and self-confident, while black men are associated with more blue-collar occupations and words such as dishonest and rude. Black women are associated with occupations such as masseuse and beautician and character traits like sensual, sexy and superstitious. There are a few unexpected words that seem to defy stereotypes. For example, the words paralegal and lawyer appear in the black women quadrant and the words jailer and judge appear in the black men quadrant. Rather than signifying that there is a higher representation of black people in these law and criminal-justice-related occupations, it seems more likely that the associations instead stem from harmful stereotypes connecting people of African descent to criminality.

In FISE-IT2, many character traits for non-Italian quadrants are negative. In particular, Romanian men are associated with aggressive sounding traits like autocratic, bellicose and impatient, while Romanian women are attributed with vindictive, cold and withdrawn. We also see that the occupations associated with Romanian people are almost exclusively low-skilled.

| | |
|---|---|
| men white | consulente, manager, spedizioniere, statistico, assistente di volo, corriere, agente di libertà vigilata, analista, organizzatore di raccolte fondi, paesaggista |
| men black | magazziniere, carpentiere, falegname, panettiere, autista, ingegnere, carceriere, barbiere, musicista, giudice |
| women black | lavandaia, cuoca, soccorritrice, paralegale, dietista, macellaia, estetista, avvocata, receptionist, sarta |
| women white | manicure, guardia, pubblicitaria, sviluppatrice, ispettrice, produttrice, autrice, allenatrice, attrice, metalmeccanica |

| | |
|---|---|
| men black | brontolone, rude, credulone, irascibile, disonesto, spendaccione, abile, umile, auto-denunciante, ingegnoso |
| women black | sensuale, civettuola, ficcanasa, dispettosa, soave, superstiziosa, sexy, lussuriosa, autocommiserazione, terrosa |
| men white | di alto spirito, di buon umore, privo di umorismo, di larghe vedute, privo di tatto, duro di cuore, di principio, privo di pregiudizi, sicuro di sé, troppo sicuro di sé |
| women white | sola, femminile, rassegnata, disponibile, conforme, in bilico, smemorata, particolare, tradizionale, romantica |

Table 14: Intersectional Occupations and Character Traits FISE-IT1

| | |
|---|---|
| men italian | ingegnere, professore, pilota, geometra, giudice, magazziniere, consulente, chef, manager, barbiere |
| women italian | manicure, bibliotecaria, domestica, lavandaia, conciatetti, pubblicitaria, avvocata, cuoca, paralegale, dietista |
| men romanian | falegname, spedizioniere, gioielliere, elettricista, portiere, autista, analista, dentista, estrattore, investitore |
| women romanian | guardia, estetista, receptionist, lavoratrice, venditrice, sarta, soccorritrice, paramedica, operatrice, macellaia |

| | |
|---|---|
| men italian | brontolone, duro di cuore, razionale, rude, etico, di buon umore, spendaccione, intellettuale, di alto spirito, temperante |
| women italian | sola, sensuale, femminile, civettuola, raffinata, soave, tradizionale, mondana, ficcanasa, snob |
| women romanian | rassegnata, vendicativa, trattenuta, freddolosa, ritirata, manipolatrice, guardinga, avventata, preoccupata, scortese |
| men romanian | credulone, autocratico, sicuro di sé, asistematico, consapevole di sé, troppo sicuro di sé, impotente, zestful, ricerca di sé, bellicoso |

Table 15: Intersectional Occupations and Character Traits FISE-IT2

## B Details on technical methods

### B.1 List of WEAT experiments

| Test | Bias Type | Targets | Attributes |
|---|---|---|---|
| WEAT 6 | gender | male vs. female first names | career vs. family |
| WEAT 7 | gender | math vs. arts | male vs. female terms |
| WEAT 8 | gender | science vs. arts | male vs. female terms |
| GER1 | gender | gendered study programs (CH) | male vs. female terms |
| GER2 | gender | rational vs. emotional | male vs. female terms |
| IT1 | ethnic | Italian vs. Romanian names | pleasant vs unpleasant |
| IT2 | ethnic | Italian vs. Romanian names | high- vs low-skilled jobs |
| IT3 | ethnic | Italian vs. Indian names | pleasant vs unpleasant |
| IT4 | ethnic | Italian vs. Indian names | high- vs low-skilled jobs |
| IT5 | gender/sexuality | strait/cis vs. queer/trans | pleasant vs. unpleasant |

Table 16: A list of the WEAT experiments carried out in this work.

### B.2 Computing WEAT Effect Sizes

Let $w$ be a word with corresponding word-embedding $\vec{w}$. The expression

$$s(w, A, B) = \frac{\sum_{a \in A} \cos(\vec{w}, \vec{a})}{|A|} - \frac{\sum_{b \in B} \cos(\vec{w}, \vec{b})}{|B|} \quad (1)$$

measures to what extent $w$ is more closely associated with $A$ or $B$. The sign of $s(w, A, B)$ indicates the direction of the bias, while the magnitude indicates the level of bias. For example, if $w =$ *man* and $A$ and $B$ correspond to *career* and *family* respectively, and the embedding space indeed encodes stereotypical bias, we would expect $s(w, A, B)$ to be a large positive number. The relative association between the target words $X, Y$ and the attribute words $A, B$ is then given by

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

The overall WEAT bias metric, called the *effect size*, is computed by normalizing $s(X, Y, A, B)$:

$$es(X, Y, A, B) = \frac{s(X, Y, A, B)}{\text{stddev}_{w \in X \cup Y} s(w, A, B)}. \quad (2)$$

Typically $X, Y$ and $A, B$ are chosen so that positive effect sizes reflect stereotypical bias and negative values reflect anti-stereotypical bias, as in the above examples with targets *male vs. female terms* and attributes *career vs. family*. The role of targets and attributes can be switched, and we observed several cases in the literature where wordlists originally designated as attribute sets were used as targets, particularly in the case of *male vs. female terms*. However, switching the role of targets and attributes does affect the normalization factor in the denominator of $es(X, Y, A, B)$, which should be taken into account when comparing results.

Caliskan et al. (2017) also propose a significance test, the *one-sided permutation test*, in order to ensure that random partitions of the target words $X \cup Y$ do not yield large spurious effect sizes. Let $\{X_i, Y_i\}_i$ denote the set of partitions of $X \cup Y$ into two sets of equal size. The $p$-value for the permutation test is given by

$$p := \Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)], \quad (3)$$

i.e., the fraction of partitions for which $s(X_i, Y_i, A, B) > s(X, Y, A, B)$. A common threshold for statistical significance is $p < 0.05$, meaning that the null hypothesis (that there is no significant bias present) can be rejected at a $5\%$ level of significance. To limit computational requirements, in this work all $p$-tests were conducted using 10,000 randomly sampled partitions.

### B.3 Computing LPBS

For example, given an input of the form $x =$ `[MASK]` *is a programmer*, the model will output a probability estimate $p(\texttt{[MASK]} = w|x)$, the probability that the masked token is given by the word $w$, for every word $w$ in the model's vocabulary. To compute the association between the target *he* and the attribute *programmer*, first the probability that the masked token in the sentence '`[MASK]` *is a programmer*' is given by the word *he* is computed:

$$p_{tgt} = p\big(\texttt{[MASK]} = \textit{he} \big| \texttt{[MASK]}\ \textit{is a programmer}\big) \quad (4)$$

Independent of the context, the model may be statistically more or less likely to predict the word *he* than the word *she*, for instance if the corpus the model was trained on contains many more references to male subjects. To account for this difference and isolate the contribution of the word *programmer* to the model's predictions, the probability

$$p_{prior} = p\big(\texttt{[MASK]}_1 = \text{he} \big| \texttt{[MASK]}_1\ \textit{is a}\ \texttt{[MASK]}_2\big) \quad (5)$$

is also computed and used to normalize $p_{tgt}$. In general, the association between an arbitrary target $x$ and attribute $a$ is defined as

$$asc(x, a) = \log \frac{p_{tgt}(x|a)}{p_{prior}(x)}, \quad (6)$$

where $p_{tgt}$ and $p_{prior}$ are computed exactly as in the above case with $x = he$ and $a = programmer$. Kurita et al. (2019) refer to $asc(x, a)$ as the *increased log probability score*. A positive association signifies that the likelihood of the target increases when the attribute is present, whereas a negative association indicates that the likelihood of the target decreases when combined with the attribute. The increased log probability score $asc(x, a)$ is analogous to the cosine similarity $\cos(\vec{x}, \vec{a})$, and is used to compute an *effect size* completely analogous to WEAT.

### B.4 Computing CrowS-Pairs

Each sentence pair consists of a stereotypical sentence $s_s$, such as *Fat people can never really be attractive*, and an anti-stereotypical counterpart $s_a$, e.g., *Thin people can never really be attractive.* The sentences $s_s$ and $s_a$ are as semantically similar as possible, only differing in the terms representing the demographic groups being compared. Given a sentence pair $(s_s, s_a)$, let $U$ be the set of shared words in $s_s$ and $s_a$, e.g., $U$ = {*people*, *can*, *never*, *really*, *be*, *attractive*}. Rather than using the increased log probability score (Eq. 6) to measure the likelihood of the sentence $s_s$, the metric uses the *psuedo-log-likelihood (PLL)* score (Salazar et al., 2019)

$$pll(s_s) := \sum_{u \in U} \log(p([\text{MASK}] = u | s_s \setminus u), \quad (7)$$

where $s_s \setminus u$ denotes the sentence $s_s$ with a [MASK] token in place of the word $u$, e.g., *Fat* [MASK] *can never really be attractive*. Using the above example concerning physical appearance, $pll(s_s)$ can be interpreted as the likelihood the model attributes to the remaining part of the sentence given the presence of the word *fat* in the beginning. Bias is then measured as the difference:

$$b^{p\log}_{s_s,s_a} := pll(s_s) - pll(s_a). \quad (8)$$

It measures the degree of the model's preference for the stereotypical sentence over the anti-stereotypical sentence.

The overall bias of the model is defined as the percentage of pairs $(s_s, s_a)$ in the full CrowS-pairs dataset for which the model prefers the the stereotypical sentence $s_s$ over the anti-stereotypical $s_a$, i.e.,

$$B_{CrowS} := \frac{100}{N} \sum_{(s_s,s_a)} I(pll(s_s) > pll(s_a)), \quad (9)$$

where $N$ is the total number of pairs in the dataset.

Since some of the sentences relate to harmful stereotypes about underprivileged groups and others relate to positive stereotypes about privileged groups, two further metrics are computed in the same manner by restricting to the corresponding subsets of sentence pairs. Let $S^-$ denote the sentence pairs corresponding to harmful stereotypes and $S^+$ those corresponding to positive stereotypes.

$$B^-_{CrowS} = \frac{100}{|S^-|} \sum_{(s_s,s_a) \in S^-} I(pll(s_s) > pll(s_a)), \quad (10)$$

with $B^+_{CrowS}$ defined similarly.

### B.5 Computing FISE

As a first step, given a particular target word $w$ (e.g. *friendly*), and bias dimension $d$, the word-level bias of $w$ is measured using Eq 1:

$$b_d(w) = s(w, A_d, B_d), \quad (11)$$

where $A_d$ and $B_d$ denote the target lists corresponding to bias dimension $d$.

To perform an intersectional analysis for two bias dimensions $d_1$ and $d_2$, the target $w$ is mapped to the $xy$-plane via:

$$w \to \big(b_{d_1}(w), b_{d_2}(w)\big) \in \mathbf{R}^2$$

### B.6 Embedding Methods

- *Sentence-level*: In the original SEAT implementation, May et al. (2019) use the final hidden state of the `[CLS]` token as a sentence embedding for BERT models, and the hidden state of the final token in the sentence for GPT models.

- *Token-level*: Delobelle et al. (2022) use the additional option of averaging the embedding vectors obtained from all sub-tokens of the target word and provide evidence that this method should be preferred in testing model bias.

In our experiments, token-level embeddings were used for all WEAT and FISE tests performed on transformer models. For SEAT tests, both sentence-level and token-level embeddings were used and compared.

### B.7 Model Details

- Fasttext: `cc.it.300`[12], 300-dimensional static embedding trained on Common Crawl and Wikipedia.
- BERT: `dbmdz/bert-base-italian-uncased`[13], 12 layers, 110m parameters, embedding dimension 768, trained on Wikipedia and OPUS corpora.
- GPT-2: `GroNLP/gpt2-small-italian`[14], 12 layers, 117m parameters, embedding dimension 768, trained on Wikipedia and ItWaC

[12] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.it.300.bin.gz

[13] https://huggingface.co/dbmdz/bert-base-italian-uncased

[14] https://huggingface.co/GroNLP/gpt2-small-italian

## C Wordlists and Prompts

This appendix contains all word lists and templates that were adapted for the Italian context. All remaining stimuli were translated from datasets used in previously published work.

### C.1 New WEAT lists

**IT1**

- *Italian names:* Andrea, Francesco, Alessandro, Matteo, Luca, Martina, Alessia, Giulia, Chiara, Sara
- *Romanian names:*Alexandra, Diana, Andrei, Daniel, Ionut, Adrian, Denisa, Ioana, Sorin, Alexandru
- *pleasant:* amico, gioia, more, risata, content, meraviglioso, pace, piacere
- *unpleasant:* guerra, cattivo, terrible, agonia, orribile, cattiva, male, fallimento

**IT2**

- *high-skilled:* avvocato, fondatore, CEO, dottore, ingegnere, pilota, general, artigiano
- *low-skilled:* impiegato, commesso, segretario, pulitore, assemblatore, conducente, soldato, mietitore

**IT3**

- *South Asian names:*Mohammed, Rahul, Sunil, Raju, Manoj, Puja, Anita, Priyanka, Rekha, Sunita

**IT5**

- *straight/cis:* etero, cis, cisgender, eterosessuale
- *queer/trans:* queer, omosessuale, trans, transgender

### C.2 SEAT Templates

- templates_nouns_singular=[ “WORD è qui.", “WORD è lì.", ]
- templates_nouns_plural=[ “WORD sono qui.", “WORD sono lì.", ]
- templates_adjectives = [ “È WORD.", “Era WORD.", ]
- templates_verbs = [ “Questo può WORD.", “Questa può WORD.", ]

### C.3 FISE

- White: “bianco", “bianca", “bianchi", “bianche", “caucasico", “caucasica", “caucasici", “caucasiche", “di pelle chiara", “italiano", “italiana", “italiani", “italiane", “europeo", “europea", “europei", “europee"
- Black: “nero", “nera", “neri", “nere", “negro", “negra", “negri", “negre", “di pelle scura", “africano", “africana", “africani", “africane", “etiope", “etiopi", “afro-italiano", “afro-italiana", “afro-italiani", “afro-italiane"
- Italian: “italiano", “italiana", “italiani", “italiane", “mediterraneo", “mediterranea", “mediterranei", “mediterranee", “di pelle chiara", “latino", “latina", “latini", “latine", “romano", “romana", “romani", “romane"
- Romanian: “rumeno", “rumena", “rumeni", “rumene", “daco", “daca", “daci", “dace", “di pelle chiara", “balcanico", “balcanica", “balcanici", “balcaniche", “europeo orientale", “europea orientale", “europei orientali", “europee orientali"

#### C.3.1 Affect Stimuli

| Ingressivo | Congressivo |
|---|---|
| assertivo | empatico |
| guidato | collaborativo |
| resiliente | inclusivo |
| decisivo | diplomatico |
| dominante | nutriente |
| competitivo | armonioso |
| ambizioso | solidale |
| insensibile | unificante |
| fiducioso | paziente |
| distaccato | compassionevole |
| indipendente | cooperativo |
| autosufficiente | comprensivo |
| analitico | aperto |
| orientato agli obiettivi | flessibile |
| audace | disponibile |
| sicuro di sé | gentile |
| determinato | ricettivo |
| concentrato | attento |
| impavido | gentile |
| strategico | comprensivo |
| autonomo | tollerante |