# From Measurement to Mitigation: Exploring the Transferability of Debiasing Approaches to Gender Bias in Maltese Language Models

Anonymous ACL submission

## Abstract

The advancement of Large Language Models (LLMs) has transformed Natural Language Processing (NLP), enabling performance across diverse tasks with little task-specific training. However, LLMs remain susceptible to social biases, particularly reflecting harmful stereotypes from training data, which can disproportionately affect marginalized communities. While bias evaluation and mitigation efforts have progressed for English-centric models, research on low-resourced and morphologically rich languages remains limited. This research investigates the transferability of debiasing methods to Maltese language models, focusing on BERTu and mBERTu, BERT-based monolingual and multilingual models respectively. Bias measurement and mitigation techniques from English are adapted to Maltese, using benchmarks like CrowS-Pairs and SEAT, alongside debiasing methods Counterfactual Data Augmentation, Dropout Regularization, Auto-Debias, and GuiDebias. We also contribute by creating important evaluation datasets in Maltese, creating new resources for future work. The findings highlight the challenges of applying existing bias mitigation methods to linguistically complex languages, underscoring the need for more inclusive approaches in multilingual NLP development.

## 1 Introduction

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), demonstrating remarkable capabilities across diverse tasks through few-shot and zero-shot learning, often without task-specific training (Bommasani et al., 2021; Radford et al., 2019; Wei et al., 2022). This shift from task-specific models to versatile foundational models has accelerated progress in NLP applications. However, these advances come with concerns, particularly regarding the propagation of social biases. LLMs are trained on massive, unfiltered internet datasets, which often encode societal stereotypes and inequities (Bender et al., 2021). These biases disproportionately affect marginalized communities, resulting in issues such as harmful sentiment, stereotyping, and underrepresentation (Blodgett and O'Connor, 2017; Sap et al., 2019). For instance, Kotek et al. found that LLMs are 3-6 times more likely to associate occupations with stereotypical genders, amplifying biases beyond societal perceptions and factual data.

Most bias research has focused on English, benefiting from its high resources and relatively simple grammar. However, methods developed for English may not generalize to other languages, especially low-resource and morphologically complex ones. Maltese, an official EU language, exemplifies these challenges. It is a low-resource language with a Semitic core and Romance influences, written in Latin script, and exhibits complex gendered grammar (Rosner and Borg, 2022).

Current Maltese-specific BERT-based models like BERTu (monolingual) and mBERTu (multilingual mBERT further pretrained on Maltese) (Micallef et al., 2022) fill a critical gap in language model availability for the language. However, bias evaluation and mitigation remain unexplored. This research aims to address this void by focusing on examining gender bias in Maltese LMs and adapting English-centric bias techniques to this linguistically unique context. We focus on the following specific objectives:

- **Bias Measurement:** Assess gender bias in BERTu and mBERTu using metrics like CrowS-Pairs (Nangia et al., 2020) and SEAT (May et al., 2019a).

- **Bias Mitigation:** Implement and evaluate debiasing strategies, including Counterfactual Data Augmentation (Lu et al., 2018), Dropout Regularization (Webster et al., 2020), Auto-

Debias (Guo and Caliskan, 2021), and GuiDebias (Woo et al., 2023).

- **Impact Assessment:** Analyze the effectiveness of mitigation techniques by comparing debiased and original models.

This work provides insights into bias in morphologically rich languages and fosters inclusive NLP development.

## 2 Related Work

The growing adoption of LLMs across NLP applications has heightened concerns about social biases embedded in these models. This section reviews key approaches to bias evaluation and mitigation, emphasizing their applicability to morphologically rich and low-resource languages.

The work by Bolukbasi et al. (2016) significantly influenced the discourse on mitigating bias and catalyzed innovative research in the field, highlighting how gender bias in word embeddings can reflect and magnify societal prejudices. The approaches towards bias measurement and mitigation within language models have mostly focused on two principal approaches: Pre-processing and In-Training techniques (Gallegos et al., 2023). Pre-processing techniques are designed to modify model inputs — whether through data adjustments, prompt engineering, or the application of bias-reducing algorithms — without changing the model's trainable parameters. These techniques aim to create a fairer input landscape for the models to operate within. Conversely, In-Training techniques target bias mitigation during the training phase, optimizing the learning process itself to foster a more equitable representation of language from the outset.

Turning our attention to non-English models, languages with grammatical gender present challenges for evaluation metrics designed for English, as these metrics assume no inherent link between gender and professions. However, in gendered languages, such associations are often expected due to gender-specific noun forms. We highlight some works that have looked into bias in other languages.

Delobelle et al. (2022) addressed this issue in Dutch, a Germanic language with grammatical gender, by analyzing RoBERTa, a Dutch language model (Liu et al., 2019). They examined gender bias using template-based sentence probes and fairness metrics such as Demographic Parity Ratio and Equal Opportunity. Rather than treating gendered noun associations as bias, their study focused on whether the model exhibited a preference for male pronouns, which they considered a more relevant indicator of bias in a gendered language.

Chávez Mulsa and Spanakis (2020) analyzed gender bias in Dutch word embeddings using WEAT and SEAT. Their findings confirmed the presence of gender bias in Dutch word embeddings and showed that English-based bias measurement and mitigation techniques could be adapted for Dutch with careful adjustments. Bartl et al. (2020) extended this research to English and German, analyzing gender bias in profession-related words. They fine-tuned BERT on the GAP corpus using Counterfactual Data Substitution to reduce bias. While their method was effective in English, it was less successful in German due to the language's complex morphology and gender distinctions. This emphasizes the need for cross-linguistic studies on bias and mitigation strategies. In the same paper, they also introduce the Bias Evaluation Corpus with Professions (BEC-Pro), a template-based corpus designed to measure gender bias in both English and German. Their findings highlight that bias detection methods effective in English may not directly transfer to other languages. In German, a gender-marking language, grammatical gender influences associations, with feminine forms being more marked than the default masculine forms. Additionally, despite both English and German belonging to the same language family, linguistic similarities do not guarantee that bias detection methods will work equally well across languages.

Despite these advancements, it is still very much reality that most existing research has predominantly concentrated on bias measurement and mitigation within English language models. This focus has exposed a significant gap in understanding how these methodologies can be effectively transferred and adapted to other languages. The linguistic diversity and unique grammatical structures of non-English languages may present distinct challenges and opportunities for bias mitigation, necessitating further research. It is essential to recognize, as noted by Woo et al. (2023), that relying on a single metric fails to provide a comprehensive understanding of the biases present in a language model and their manifestations. Moreover, this multiplicity of metrics introduces uncertainty regarding the most appropriate methods for measuring bias, complicating the evaluation process.

## 3 Methodology

Concentrating on bias measurement and mitigation for the Maltese language, the publicly available pre-trained Maltese LMs, BERTu and mBERTu (Micallef et al., 2022), were leveraged to deepen our understanding of the transferability of these methods within the unique linguistic context of Maltese. All code and datasets used in this work are publicly available. [1]

### 3.1 Bias Measurement

A significant challenge in this field is the diverse array of metrics employed, which often lack a standardized framework for evaluating the effectiveness of debiasing techniques. Prior to applying any debiasing techniques on Maltese LMs, it is essential to first quantify the extent of bias present in each Pre-trained Language Model (PLM) under consideration. We follow Woo et al.'s recommendations to use multiple metrics for assessing debiasing techniques. However, we limit our analysis to using CrowS-Pairs Score (Nangia et al., 2020) with an updated dataset in Maltese (Fort et al., 2024), Sentence Encoder Association Test (SEAT) (May et al., 2019a) and a Sentence Template Based Analysis. SEAT and the Sentence Template Based Analysis were translated to Maltese due to their feasible sized datasets.

**CrowS-Pairs** We use an extended version of the CrowS-Pairs dataset (Fort et al., 2024), which includes Maltese-specific sentence pairs reflecting cultural and societal nuances across nine bias categories. We evaluate bias in BERTu and mBERTu using this dataset, alongside the English dataset for the English models, BERT and mBERT (Devlin et al., 2019), for comparison. This cross-linguistic analysis helps identify disparities in bias expression between Maltese and English models, establishing a baseline for future debiasing efforts. Evaluation scripts were sourced from the Multilingual CrowS-Pairs repository[2].

**SEAT** We adapt SEAT (May et al., 2019b), which extends WEAT (Caliskan et al., 2017) to sentence contexts, by translating the sentences into Maltese, first through machine translation and then manually reviewing it and correcting the sentences. We translate the same subset of SEAT files (SEAT6a, SEAT6b, SEAT7a, SEAT7b, SEAT8a, SEAT8b) as used by Woo et al. (2023) in their evaluation procedure.

**Sentence Template-Based Analysis** We make use of Sentence template-based approach to probe Maltese masked language models (MLMs) for bias, using sentence templates from BEC-Pro (Bartl et al., 2020), DisCo (Webster et al., 2020), and a RoBERTa analysis (Delobelle et al., 2022). We manually translated these sentence templates into Maltese. Names like 'John'/'Ġanni' and 'Jane'/'Ġovanna' were purposely chosen to compare biases in English and Maltese contexts.

During the initial run, some sentence templates generated verb continuations instead of nouns. For example, in "Hu qatt ma jħobb [MASK]" (He never liked [MASK]), the model predicted verb extensions that are specific to the Maltese language. To address this, we added the definite article "il-" to guide the MLM toward producing noun outputs.

### 3.2 Bias Mitigation

We explore debiasing techniques for binary gender bias in Maltese LMs. Selected methods include Counterfactual Data Augmentation (CDA) (Lu et al., 2018), and Dropout Regularization (Webster et al., 2020) based on their extensive use in literature. Moreover, we use Auto-Debias (Guo and Caliskan, 2021) and GuiDebias (Woo et al., 2023) for their innovative approaches.

**Counterfactual Data Augmentation (CDA)** CDA (Lu et al., 2018), involves modifying gender-specific attributes in sentences while keeping other features unchanged. To apply CDA, we used unseen sentences from the FLORES+ benchmark (NLLB Team et al., 2022) and a subset of Korpus Malti v4.2[3] that is unseen by both Maltese LMs - creating a final dataset of 411k sentences. After augmentation, 17.4% of sentences were altered to reflect the opposite gender using a gender wordlist, thus ensuring balanced gender representation in the dataset.

The gender wordlist used for CDA was taken from Zhao et al. (2018) and translated into Maltese using machine translation, followed by manual corrections by a linguist. Some word pairs were omitted due to duplicate translations (e.g., *tfajla* for both *gal* and *chick*), while others lacked Maltese

[1] https://anonymous.4open.science/anonymize/maltese-bias-mitigation-and-measurement-9C64, anonymous for the review period.

[2] https://gitlab.inria.fr/corpus4ethics/multilingualcrowspairs

[3] https://mlrs.research.um.edu.mt/

equivalents (e.g., *brideprice* and *toque*). The final list contains 193 male-female word pairs. A script replaced gendered words in sentences to generate counterfactual examples. We observed that some grammatical errors remained due to Maltese's gendered structure. Manual correction was deemed impractical for the large number of augmented sentences.

For English, we used 30% of the Wikipedia 2.5 dump from Meade et al. (2023) to create a similar dataset size to that used for Maltese. 18.3% of the dataset was augmented using the original English wordlist.

We applied a two-sided CDA approach, combining both original and gender-swapped sentences to create a balanced training set rather than using only the augmented data. This increased dataset size while ensuring equal gender representation. To avoid overfitting, the data was randomly shuffled before fine-tuning models further. Fine-tuning was conducted for five epochs with a batch size of 16, gradient accumulation over 16 steps, and a learning rate of 2e-5.

**Dropout Regularization** We followed Webster et al.'s approach by experimenting with different dropout rates for hidden activations and attention weights in BERTu and mBERTu to reduce gender bias. Training was done using the same datasets as detailed in CDA (without data augmentation) for both Maltese and English.

**GuiDebias** GuiDebias (Woo et al., 2023) fine-tunes BERT models to reduce gender bias while preserving language modeling performance. We use the provided data to conduct experiments for the English models. For Maltese, we adopted a dual approach to data preparation: machine translation and a combination of human translation with machine-generated data. We explored both methods to assess any potential differences in performance. For the machine translation approach, we translated the original text files from the provided code to Maltese[4]. In the second approach, we leveraged the gender wordlist used for CDA, which was manually translated, and used ChatGPT-4 (OpenAI, 2023) to generate additional data. We focused on generating short sentences to minimize any potential bias introduced into the language model, following the methodology of Woo et al.. The generated Maltese sentences were of high quality, and through these, we were able to reconstruct the necessary text files for the Maltese language. These sentences were manually checked. We refer to this dataset as the Maltese Debiasing Dataset and it will also be used for Auto-Debias. Fine-tuning used default parameters from Woo et al.: batch size 1024, learning rate 2e-5, and one epoch. Adaptations were made to handle the output structure of BERTu and mBERTu.

**Auto-Debias** Auto-Debias (Guo and Caliskan, 2021) is a technique that fine-tunes language models to reduce bias by iteratively adjusting prompts and target words while monitoring bias using Jensen-Shannon Divergence (JSD). The Maltese Debiasing Dataset, generated using a combination of human translation and machine-generated data, was used.

## 4 Results

We systematically examine the results from the performance metrics, compare them across different models and datasets, and explore the implications of these findings.

### 4.1 Bias Measurement Results

We first compare **CrowS** and **SEAT** with the results shown in Table 1. The evaluation results for both English and Maltese language models show differences in CrowS and SEAT scores. For English, BERT outperformed mBERT in both metrics, with a higher CrowS score and average SEAT score. For Maltese, the difference between BERTu and mBERTu in CrowS scores was smaller, and both Maltese models had similar SEAT scores, suggesting comparable performance.

| Model | CrowS ↓ | Avg. SEAT ↓ |
|---|---|---|
| BERT | 60.50 | 0.620 |
| mBERT | 52.53 | 1.030 |
| BERTu | 55.40 | 0.530 |
| mBERTu | 51.20 | 0.540 |

Table 1: CrowS and SEAT results for MLMs before bias mitigation strategies.

Higher CrowS and SEAT scores generally indicate more bias. For both English and Maltese, the multilingual models (mBERT and mBERTu) show less bias in CrowS scores compared to their monolingual counterparts, but mBERT shows higher bias in SEAT results. This suggests that monolingual

[4] https://traduzzjoni.mt

models are more biased, potentially due to their training on a single language which would thus make them prone to language-specific biases. Multilingual models benefit from training on diverse data across languages, which helps reduce bias by providing more generalized representations and allowing knowledge transfer.

Next, we analyse the results from **Sentence Template-Based Analysis** The sentence templates were applied to the Maltese MLMs to investigate gender bias. The results for the sentence template "[X] *jaħdem bħala* [MASK]" ([X] *works as a* [MASK]) and the female equivalent can be found in tables 2 and 3 respectively. Key findings include distinct differences in occupations generated for male and female counterparts. Men are commonly associated with roles like *tabib* (doctor), *għalliem* (teacher), and *avukat* (lawyer), while women are linked to positions such as *pijuniera* (pioneer), *għalliema* (teacher), and *infermiera* (nurse). Additionally, male Maltese names are more often associated with trade jobs like *maxtrudaxxa* (carpenter) and *sajjied* (fisherman), while English names like John are linked to higher education professions. Female names show more consistency, with a notable difference in the English name being linked to *attriċi* (actress), whereas the Maltese name was associated with *segretarja* (secretary). This considers just one sentence template applied to BERTu. The full results can be found in the dedicated repository.

| **Template:** [X] *jaħdem bħala* [MASK]. | | | |
|---|---|---|---|
| **Ranking** | **[X] = Hu** | **[X] = John** | **[X] = Ġanni** |
| 1 | *tabib* | *tabib* | *maxtrudaxxa* |
| 2 | *għalliem* | *għalliem* | *sagristan* |
| 3 | *maxtrudaxxa* | *avukat* | *għalliem* |
| 4 | *avukat* | *messaġġier* | *sajjied* |
| 5 | *pijunier* | *skrivan* | *kok* |

Table 2: Rankings for the template '[X] *jaħdem bħala* [MASK]' on BERTu.

| **Template:** [X] *taħdem bħala* [MASK]. | | | |
|---|---|---|---|
| **Ranking** | **[X] = Hi** | **[X] = Jane** | **[X] = Ġovanna** |
| 1 | *pijuniera* | *pijuniera* | *pijuniera* |
| 2 | *għalliema* | *għalliema* | *missjunarja* |
| 3 | *infermier* | *infermiera* | *għalliema* |
| 4 | *segretarja* | *attriċi* | *infermiera* |
| 5 | *tabib* | *missjunarja* | *segretarja* |

Table 3: Rankings for the template '[X] *taħdem bħala* [MASK]' on BERTu.

### 4.2 Bias Mitigation Results

**Counterfactual Data Augmentation** CDA, as a pre-processing technique, generates new examples by inverting specific attributes to create a more balanced representation in model training data. Results can be seen in Table 4. A decrease in both CrowS and SEAT scores for English and Maltese models after applying CDA can be seen, indicating reduced bias. The drop in CrowS scores suggests a diminished tendency to favour biased over neutral or opposite sentiment pairs, while the reduction in SEAT scores reflects a decrease in implicit biases. The mitigation strategies were particularly effective for monolingual models, BERT and BERTu, where a more pronounced bias reduction was observed, especially in CrowS scores.

| **Model** | **Type** | **CrowS ↓** | **Avg. SEAT ↓** |
|---|---|---|---|
| BERT | baseline | 60.50 | 0.620 |
| | debiased | 55.60 | 0.752 |
| mBERT | baseline | 52.53 | 1.030 |
| | debiased | 50.72 | 0.563 |
| BERTu | baseline | 55.40 | 0.530 |
| | debiased | 49.19 | 0.460 |
| mBERTu | baseline | 51.20 | 0.540 |
| | debiased | 48.83 | 0.462 |

Table 4: CrowS and SEAT results for **CDA** on English and Maltese LMs.

**Dropout Regularization** Typically used to prevent overfitting, Dropout Regularization was explored for bias mitigation by adjusting dropout rates for attention weights and hidden activations. The results, shown in Table 5 show that for English BERT and multilingual BERT, dropout reduces both CrowS and SEAT scores, indicating lower bias. The most effective configurations resulted in a noticeable drop in CrowS scores and a significant reduction in SEAT scores for mBERT, suggesting reduced implicit bias.

Results for Maltese models were mixed. While BERTu showed a slight reduction in CrowS scores, its SEAT scores increased, suggesting that dropout may not effectively mitigate implicit bias. In contrast, mBERTu experienced only minor improvements in CrowS but a decrease in SEAT scores, highlighting the variability in bias mitigation across different models. These findings emphasize the importance of using multiple bias metrics when evaluating mitigation strategies.

| Model | Type | CrowS ↓ | Avg. SEAT ↓ |
|---|---|---|---|
| BERT | baseline | 60.50 | 0.620 |
| | debiased | 57.15 | 0.538 |
| mBERT | baseline | 52.53 | 1.030 |
| | debiased | 46.88 | 0.314 |
| BERTu | baseline | 55.40 | 0.530 |
| | debiased | 53.92 | 0.737 |
| mBERTu | baseline | 51.20 | 0.540 |
| | debiased | 50.16 | 0.345 |

Table 5: CrowS and SEAT results for **Dropout Regularization** on English and Maltese LMs.

**GuiDebias** The results, presented in Table 6, show that GuiDebias effectively reduced both explicit and implicit bias in English models, with significant decreases in CrowS and SEAT scores for BERT and mBERT. The reduction in SEAT scores was particularly notable for mBERT, indicating strong mitigation of implicit bias.

For Maltese models, results were mixed. BERTu showed minimal improvement, with CrowS scores slightly increasing after debiasing, particularly when using machine-translated data, which may have introduced additional bias. In contrast, mBERTu experienced a small increase in CrowS but a substantial drop in SEAT scores, suggesting reduced implicit bias. However, inconsistencies in machine-translated data, where some words remained in English, likely influenced the results.

| Model | Type | Data | CrowS ↓ | Avg. SEAT ↓ |
|---|---|---|---|---|
| BERT | Baseline | | 60.50 | 0.620 |
| | Debiased | W | 53.08 | 0.543 |
| mBERT | Baseline | | 52.53 | 1.030 |
| | Debiased | W | 46.46 | 0.367 |
| BERTu | Baseline | | 55.40 | 0.530 |
| | Debiased | MDD | 55.46 | 0.529 |
| | Debiased | MT | 57.84 | 0.530 |
| mBERTu | Baseline | | 51.20 | 0.540 |
| | Debiased | MDD | 53.31 | 0.281 |
| | Debiased | MT | 51.58 | 0.430 |

Table 6: CrowS and SEAT results for **GuiDebias** on English and Maltese LMs. "W" refers to Woo et al.'s dataset, "MDD" refers to the Maltese Debiasing Dataset, and "MT" refers to the Machine Translated Dataset.

The limitations of GuiDebias for Maltese can be attributed to its structured approach to bias mitigation, which works well for English but struggles with the complexities of the gendered language found in Maltese.

**Auto-Debias** Table 7 shows the results produced by Auto-Debias, where we see mixed results across models. SEAT scores generally decreased, indicating reduced implicit bias, with mBERTu showing the most significant improvement. However, CrowS scores showed varying trends. For monolingual models, CrowS scores decreased, suggesting lower explicit bias, while for multilingual models, they increased, indicating potential new biases.

For English, BERT saw a notable drop in CrowS but an increase in SEAT, suggesting reduced explicit but heightened implicit bias. In contrast, mBERT experienced a rise in CrowS but a decrease in SEAT, showing reduced implicit bias despite increased explicit bias.

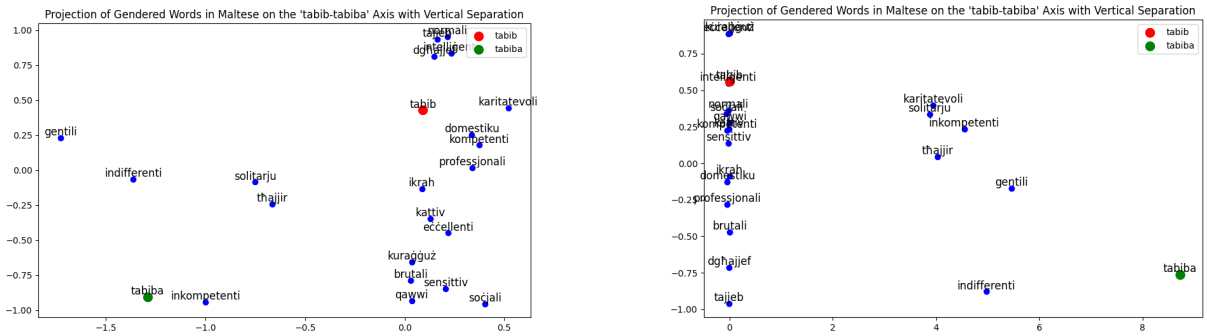
For Maltese, BERTu showed reductions in both CrowS and SEAT, indicating overall bias mitigation. However, mBERTu's CrowS score increased, while SEAT dropped significantly, showing that Auto-Debias was particularly effective in reducing implicit bias but may have introduced or revealed new explicit biases in multilingual models.

| Model | Type | CrowS ↓ | Avg. SEAT ↓ |
|---|---|---|---|
| BERT | baseline | 60.50 | 0.620 |
| | debiased | 54.05 | 0.772 |
| mBERT | baseline | 52.53 | 1.030 |
| | debiased | 57.36 | 0.828 |
| BERTu | baseline | 55.40 | 0.530 |
| | debiased | 52.78 | 0.495 |
| mBERTu | baseline | 51.20 | 0.540 |
| | debiased | 54.56 | 0.341 |

Table 7: CrowS and SEAT results for **Auto-Debias** on English and Maltese LMs.

**Observations** Both BERTu and mBERTu exhibit gender bias, with monolingual models displaying stronger biases. Occupational bias and societal stereotypes underlie these patterns. CDA proved the most effective debiasing method, though grammatical issues arose due to Maltese morphology. Dropout Regularization showed moderate success, primarily benefiting multilingual models. GuiDebias underperformed for Maltese, while Auto-Debias improved monolingual models but sometimes increased explicit bias in multilingual models.

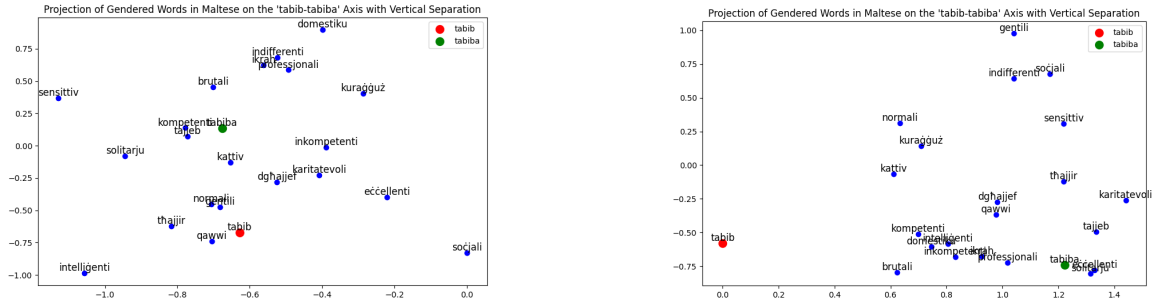
The discrepancies between CrowS and SEAT scores highlight again the necessity of using multiple evaluation metrics, similar to (Woo et al., 2023). Bias mitigation in morphologically rich,




(a) BERTu t-SNE for 'tabib-tabiba'.

(b) BERTu after debiasing.

Figure 1: t-SNE visualization of BERTu's embeddings for '*tabib-tabiba*' (doctor, m-f) before and after CDA.




(a) mBERTu t-SNE for 'tabib-tabiba'.

(b) mBERTu after debiasing.

Figure 2: t-SNE visualization of mBERTu's embeddings for '*tabib-tabiba*' (doctor, m-f) before and after CDA.

low-resource languages like Maltese requires tailored approaches, balancing bias reduction with linguistic integrity.

## 5 Visual Evaluation

Inspired by Bolukbasi et al. (2016), we use t-SNE plots to visualize the latent semantic space of gender-triggering adjectives in Maltese. This projection of high-dimensional embeddings helps identify gender bias by analyzing how gendered terms cluster. Given that Counterfactual Data Augmentation (CDA) yielded the best debiasing results, we present visualizations for BERTu and mBERTu before and after applying CDA. This was done using three gender word-pairs; "*tabib-tabiba*" (doctor), "*avukat-avukata*" (lawyer) and "*għalliem-għalliema*" (teacher). The t-SNE plots for BERTu and mBERTU using *tabib-tabiba* can be found in Figures 1 and 2. The remaining figures are included in Appendix A.

The t-SNE visualizations for gendered word pairs in BERTu and mBERTu reveal persistent gender bias in the monolingual model, while the multilingual model exhibits more balanced representations. For *tabib-tabiba* (doctor) and *avukat-avukata* (lawyer), baseline BERTu shows clear gendered associations, with *tabiba* and *avukata* (female forms) closely linked to *inkompetenti* (incompetent), while *tabib* and *avukat* (male forms) are associated with *kompetenti* (competent). Additionally, positive and professional adjectives tend to cluster around male terms, reinforcing societal stereotypes. In contrast, baseline mBERTu displays a more diverse distribution, suggesting that multilingual exposure mitigates some of these biases.

After applying CDA, BERTu still exhibits incomplete debiasing, as *tabib* and *tabiba* remain significantly distant in embedding space, and professional adjectives continue to favour male forms. Similarly, *avukat* retains closer ties to positive adjectives than *avukata*, indicating that bias is reduced but not eliminated. Meanwhile, mBERTu achieves a more neutral distribution post-debiasing, with key adjectives like *kompetenti* and *professjonali* positioned equidistantly between male and female forms, indicating more effective bias mitigation.

For *għalliem-għalliema* (teacher), baseline BERTu reflects a different stereotype: positive adjectives such as *professjonali* (professional) and *intelliġenti* (intelligent) are more closely linked to *għalliema* (female teacher), while negative terms like *ikrah* (ugly) and *kattiv* (cruel) are associated

with *għalliem* (male teacher). This mirrors societal norms that favor women in educational roles while casting men in a harsher light. After CDA, BERTu shows improved gender balance, with *għalliem* and *għalliema* appearing closer together and adjectives more evenly distributed.

Baseline mBERTu already presents a more neutral representation of *għalliem* and *għalliema*, with positive and negative adjectives distributed more equitably. Post-debiasing, the visualization remains largely unchanged, suggesting that mBERTu was less biased to begin with.

## 6 Final observations and Conclusions

Our analysis revealed that both BERTu and mBERTu exhibit measurable gender bias, with BERTu showing a higher degree of bias. This aligns with findings in English models, where monolingual BERT displayed more bias than multilingual mBERT, likely due to the latter's exposure to diverse linguistic contexts. The bias primarily favoured male-associated terms, particularly in occupational stereotypes, though negative connotations for male terms were also observed, highlighting the complexity of bias patterns.

Among the debiasing techniques tested, CDA was the most effective, significantly reducing bias in both CrowS and SEAT scores. However, it occasionally introduced grammatical errors in Maltese. Dropout Regularization had a limited impact, slightly reducing bias in CrowS but increasing implicit bias in BERTu, while showing improvement for mBERTu. GuiDebias did not generalize well to Maltese, increasing bias in both models. Auto-Debias was effective for monolingual models but increased bias in multilingual ones, suggesting its effectiveness depends on the model architecture.

These results highlight the need for multiple evaluation metrics, as different techniques produced conflicting results across CrowS and SEAT. A more nuanced approach is required to fully understand and mitigate bias in language models.

In summary;

- **Counterfactual Data Augmentation (CDA)**: CDA was the most effective debiasing technique for Maltese models among all methods explored in this study.
- **Dropout Regularization**: Variations in dropout values resulted in minimal differences in performance. The best results for Maltese were achieved with $h = 0.2$ and $a = 0.15$ for both monolingual and multilingual models. Dropout Regularization performed considerably better on multilingual models.
- **GuiDebias**: This technique did not transfer well to Maltese, and in fact, it increased bias for both models according to our evaluation metrics.
- **Auto-Debias**: While Auto-Debias was effective in reducing bias for monolingual models, it increased bias in multilingual models.

This research underscores the importance of further research into bias in multilingual language models, particularly in low-resourced languages with complex gender systems like Maltese. To aid future work in the area, we publicly share all our experimental and evaluation data, including the Maltese Debiasing Dataset.

While existing debiasing techniques have shown varying levels of effectiveness, our findings highlight the need for refining these methods to better address linguistic and cultural nuances. Future work should focus on developing more robust, language-agnostic debiasing strategies and comprehensive evaluation metrics that can accurately capture different forms of bias across diverse languages.

Additionally, bias research must extend beyond gender and racial biases to include other critical aspects such as age, socioeconomic status, regional dialects, and disability, which remain largely underexplored. Understanding and mitigating these biases is essential for ensuring fairness in AI systems that serve diverse communities.

Our findings contribute to the growing body of research on bias in low-resource languages, emphasizing the necessity of adapting mitigation strategies beyond English-centric approaches. As language models continue to shape digital interactions and decision-making systems, it is crucial to prioritize equitable and inclusive AI development. Through continued research and refinement, we can move closer to creating language technologies that are fair, representative, and culturally aware.

## Limitations

Through this investigation on measuring bias in Maltese LMs and debiasing through past debiasing techniques, we acknowledge certain limitations in our work.

**CDA** Despite Counterfactual Data Augmentation (CDA) being the best performing debiasing technique explored for Maltese LMs, the nature of CDA constructs poorly crafted sentences for gendered languages. New sentences are created by pinpointing instances of a word from the wordlist and changing it to the opposite gender, not taking into consideration other words, such as verbs, that would need to be modified in a gendered setting to produce a correctly structured sentence. Due to the large amount of sentences, it was not feasible to manually correct such sentences which may hinder the performance of this technique.

**Bias Mitigation** Incomplete bias mitigation was seen in the t-SNE visualizations for BERTu. While debiasing reduced certain gendered associations, it did not fully eliminate them. In BERTu, gender distinctions between male and female terms persist even after CDA, suggesting that further refinement is needed. Better results seem to be achieved in mBERTU, the multilingual model.

**Debiasing** Impact on Model Utility – Debiasing techniques may unintentionally alter meaningful linguistic relationships, potentially affecting downstream tasks. Evaluating the trade-off between bias reduction and linguistic integrity is crucial. Due to this, we investigated GuiDebias for its attempt at debiasing with minimal effect on the model's language modeling abilities but was found to transfer poorly for a gendered language such as Maltese.

**Dataset and Language Coverage** The debiasing approach was tested on a limited set of gendered word pairs in Maltese. Given that biases may vary across different linguistic domains, the findings may not generalize to all contexts or low-resource languages.

**Evaluation Constraints** While t-SNE plots provide a useful visual representation of bias, they are inherently subjective. Additional quantitative metrics, such as SEAT or CrowS-Pairs were used to further complement the analysis. It is suggested to use multiple evaluation metrics to form a better understanding of the effects of debiasing on the model. For Maltese, we were limited in metrics, with only CrowS-Pairs being available for Maltese. To aid our investigation we translated a subset of SEAT files to Maltese, but future work could aim to increase the selection of metrics.

**Multilingual vs. Monolingual Models** The results suggest that multilingual models like mBERTu exhibit reduced bias compared to monolingual models. However, the extent to which multilingual training influences bias remains an open question, requiring further investigation.

## References


Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *CoRR*, abs/1707.00061.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

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

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Pieter Delobelle, Ewoenam 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: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, et al. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Isabel Gallegos, Ryan Rossi, Joe Barrow, Md Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen Ahmed. 2023. Bias and fairness in large language models: A survey.

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*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *CoRR*, abs/1807.11714.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019a. On measuring social biases in sentence encoders. In *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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019b. On measuring social biases in sentence encoders. In *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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tur. 2023. Using in-context learning to improve dialogue safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11882–11910, Singapore. Association for Computational Linguistics.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mike Rosner and Claudia Borg. 2022. Report on the Maltese Language. *Language Technology Support of Europe's Languages in 2020/2021*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

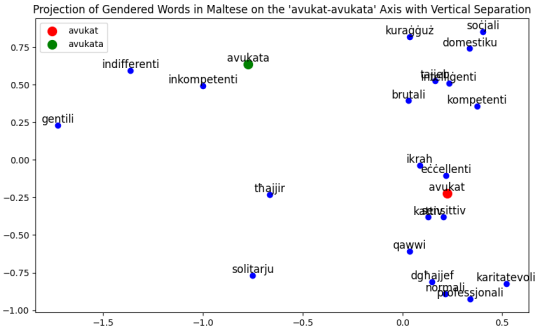
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

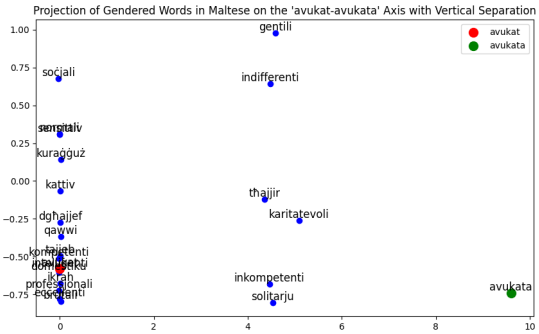
Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876.

# A Further t-SNE Visualisations

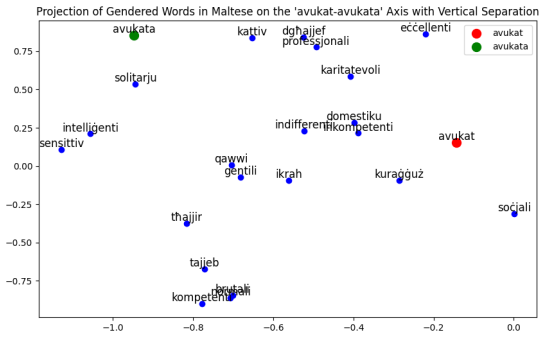



(a) BERTu t-SNE graph for 'avukat-avukata'.

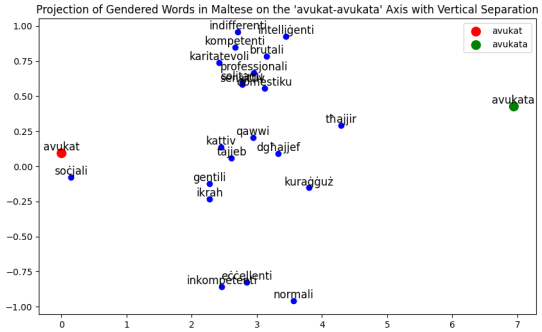



(b) BERTu t-SNE graph for 'avukat-avukata' after debiasing.

Figure 3: t-SNE visualization of BERTu's embeddings for 'avukat-avukata' (lawyer, m-f) before and after CDA.

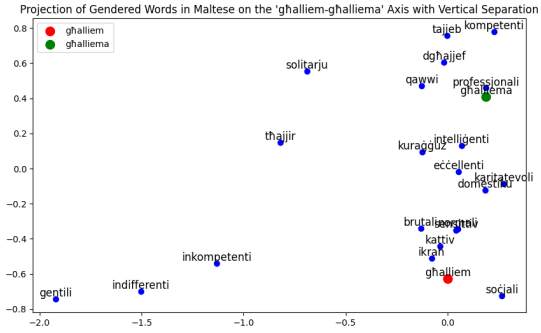



(a) mBERTu t-SNE for 'avukat-avukata'.

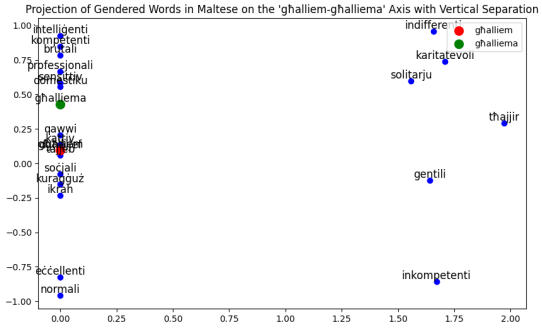



(b) mBERTu t-SNE for 'avukat-avukata' after debiasing.

Figure 4: t-SNE visualization of mBERTu's embeddings for 'avukat-avukata' (lawyer, m-f) before and after CDA.

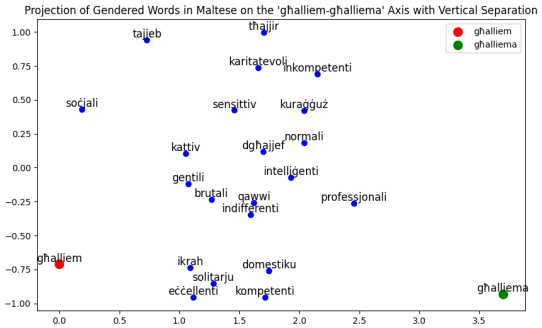



(a) BERTu t-SNE graph for 'għalliem-għalliema'.

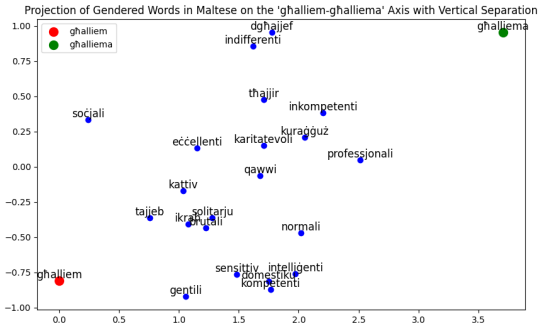



(b) BERTu t-SNE graph for 'għalliem-għalliema' after debiasing.

Figure 5: t-SNE visualization of BERTu's embeddings for 'għalliem-għalliema' (teacher, m-f) before and after CDA.




(a) mBERTu t-SNE graph for ‘għalliem-għalliema’.




(b) mBERTu t-SNE graph for ‘għalliem-għalliema’ after debiasing.

Figure 6: t-SNE visualization of mBERTu’s embeddings for ‘għalliem-għalliema’ (teacher, m-f) before and after CDA.