Measuring and Mitigating Bias for Maltese as a Low-Resource Language

Melanie Galea
Department of Artificial Intelligence, University of Malta, Msida, Malta
melanie.galea.20@um.edu.mt

Keywords

bias, language models, bias mitigation, bias measurement

1 Introduction

In an era of rapid technological advancements with an ever-expanding pool of information, the presence of bias in our systems has proven to be a significant challenge across various domains. Instead of looking at Artificial Intelligence (AI) from solely a mathematical and/or computational perspective as has been done in the past, there needs to be certain insight into the human dimension where we look at AI from an ethical perspective that tries to understand the impact (whether positive or negative) that an AI system can have.

Language Models (LMs) have emerged as a powerful tool able to understand, generate and manipulate natural language. LMs, such as the widely used transformer-based architectures, learn patterns from diverse textual sources. However, the training data often reflects historical biases, stereotypes, and prejudices present in society. Consequently, LMs can inadvertently perpetuate and even amplify these biases in their responses, affecting downstream applications. Unravelling and addressing these biases is crucial for ensuring the ethical and fair deployment of LMs to prevent the reinforcement of existing social inequities.

The majority of work that has focused on producing ways to reduce the amount of bias present in these language models has been focused predominantly on the English language. There still seems to be ambiguity on the transferability of these debiasing techniques to other languages that may be more linguistically rich than English and much lower-resourced, which adds complexity to the identification and reduction of bias in such models. This can be partly seen in works such as those by Bartl et al. (2020) and Chávez Mulsa and Spanakis (2020).

Bartl et al. (2020) measure and mitigate gender bias on BERT (Devlin et al., 2019) for English and German. Gender bias was measured through an examination of the correlations between words indicating gender and the titles of occupations in both languages, comparing the findings with real-world workforce statistics. The LM was then fine-tuned on the GAP corpus (Webster et al., 2018), after applying Counterfactual Data Substitution (CDS) (Maudslay et al., 2019). Their findings show that while their method of measuring bias is appropriate for English, it was not appropriate for German - attributing this to the rich morphology and gender-marking of the language.

On the other hand, work by Chávez Mulsa and Spanakis (2020) focused on measuring and mitigating bias for Dutch and showed positive results - concluding that the techniques made use of in their paper - Word Embeddings Association Test (WEAT) (Caliskan et al., 2017) and Clustering and Sentence Embeddings Association Test (SEAT) (May et al., 2019) to quantify gender bias, and Hard-Debias (Bolukbasi et al., 2016) and Sent-Debias (Liang et al., 2020) to mitigate the bias - can be transferred to Dutch by adequately translating the data and taking into account the unique characteristics of the Dutch language.

Although there have been a few works that explore the transferability of these measuring and mitigation techniques on other languages, there has been no work found from the research that has been conducted so far that explores this scope for low-resource languages.

2 Current Work

This ongoing project delves into the topic of measuring and mitigating bias in LMs, specifically for Maltese, a low-resource language of Semitic origin written in Latin script. For this study, we will be mainly focusing on debiasing the Maltese monolingual BERTu and multilingual mBERTu (Micallef et al., 2022) using seven debiasing techniques; Auto-Debias (Guo et al., 2022), Causal-Debias (Zhou et al., 2023), Counterfactual Data Augmentation (CDA) (Lu et al., 2018), Dropout Regularization (Webster et al., 2020), Sent-Debias (Liang et al., 2020), Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020), and DensRay (Dufter and Schütze, 2019). These debiasing techniques were chosen to ensure that a variety of techniques are explored in Maltese. Initial work will be focused on gender bias before moving on to applying the same techniques to target racial bias. Further planned experimentation includes looking into the effect of combining different techniques.

Before mitigating the bias in these LMs, bias has been measured using the Crowd-sourced Stereotype Pairs benchmark (CrowS-Pairs) (Nangia et al., 2020) - a dataset covering nine types of bias spanning from gender to religion to age. In the CrowS-Pairs dataset, a sentence is presented twice; one as a stereotyping sentence and another as a less stereotyping sentence. The version of CrowS-Pairs being used has been released recently, through which the English phrases were translated into Maltese and localised according to the social context of Malta. This work was part of a larger, multilingual effort described in Fort et al., 2024. As expected, we found that the monolingual BERTu had a higher model bias compared to the multilingual mBERTu. This was also true when the dataset was applied to the monolingual BERT and multilingual mBERT (Devlin et al., 2019) for English. Additional analysis on the individual models is currently being performed through templating metrics such as DisCo (Webster et al., 2020) and BEC-Pro (Bartl et al., 2020).

Currently, we have been working on translating corpora related to the debiasing techniques chosen. As expected, apart from the literal translation, more care is needed to ensure that the context of these corpora makes sense culturally and linguistically for the Maltese language.

3 Relevance to UniDive

Participation in UniDive would be particularly helpful to better understand how biased perceptions are being formed in the LM. Multiword expressions (MWEs) often contain idiomatic or non-compositional meanings which can be difficult for LMs to interpret correctly. Potentially further understanding the characteristics of MWEs can help me to identify biases and inaccuracies in the LM's treatment of such expressions. Rather than just basing numerical results on whether a model is considered biased or not, understanding the dependency syntax, the structure of sentences, and the relationships between words in a language would provide a foundation to interpret the LM and generate syntactic structures.

It would also be valuable to seek input from fellow academics regarding the challenge of effectively measuring bias and establishing clear definitions, given the inherent ambiguity and lack of agreement surrounding this concept in the field. Additionally, the question of how pretrained language models can be evaluated and debiased to address intersectional biases, where multiple demographic attributes intersect and contribute to complex forms of discrimination, as well as the challenge of creating related datasets for a low-resource language would be interesting questions to discuss during the Brainstorming Hackathon.

References

Bartl, M., Nissim, M., & Gatt, A. (2020, December). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the second workshop on gender bias in natural language processing* (pp. 1–16). Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.1

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chávez Mulsa, R. A., & Spanakis, G. (2020, December). Evaluating bias in Dutch word embeddings. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the second workshop on gender bias in natural language processing* (pp. 56–71). Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.6

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dufter, P., & Schütze, H. (2019, November). Analytical methods for interpretable ultradense word embeddings. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1185–1191). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1111

Fort, K., Alemany, L. A., Benotti, L., Bezançon, J., Borg, C., Borg, M., Chen, Y., Ducel, F., Dupont, Y., Ivetta, G., Li, Z., Mieskes, M., Naguib, M., Qian, Y., Radaelli, M., Schmeisser-Nieto, W. S., Raimundo Schulz, E., Saci, T., Saidi, S., … Névéol, A. (2024). Your Stereotypical Mileage may Vary: Practical Challenges of Evaluating Biases in Multiple Languages and Cultural Contexts. *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. https://inria.hal.science/hal-04537096

Guo, Y., Yang, Y., & Abbasi, A. (2022, May). Auto-debias: Debiasing masked language models with automated biased prompts. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1012–1023). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.72

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L.-P. (2020, July). Towards debiasing sentence representations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5502–5515). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.488

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2018). Gender bias in neural natural language processing. *CoRR*, *abs/1807.11714*. http://arxiv.org/abs/1807.11714

Maudslay, R. H., Gonen, H., Cotterell, R., & Teufel, S. (2019, November). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 5267–5275). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1530

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019, June). On measuring social biases in sentence encoders. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 622–628). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1063

Micallef, K., Gatt, A., Tanti, M., van der Plas, L., & Borg, C. (2022). Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 90–101. https://doi.org/10.18653/v1/2022.deeplo-1.10

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020, November). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 1953–1967). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.154

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020, July). Null it out: Guarding protected attributes by iterative nullspace projection. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7237–7256). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.647

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns (L. Lee, M. Johnson, K. Toutanova, & B. Roark, Eds.). *Transactions of the Association for Computational Linguistics*, *6*, 605–617. https://doi.org/10.1162/tacl_a_00240

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., & Petrov, S. (2020). Measuring and reducing gendered correlations in pre-trained models. *CoRR*, *abs/2010.06032*. https://arxiv.org/abs/2010.06032

Zhou, F., Mao, Y., Yu, L., Yang, Y., & Zhong, T. (2023, July). Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4227–4241). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.232