Analysis, Detection and Mitigation of Biases in Deep Learning Language Models

Ismael Garrido-Muñoz

CEATIC (Universidad de Jaén) Campus las lagunillas s/n. Jaén, 23071, Spain

Abstract

Recent advances in artificial intelligence have made it possible to make our everyday lives better, from apps that translate with great accuracy, to search engines that understand your query, to virtual assistants that answer your questions. However, these models capture the biases present in society and incorporate them into their knowledge. These biases and prejudices appear in a multitude of applications and systems. Given the same context, the model will have vastly different results depending on attributes such as the subject's gender, race or religion. In this thesis we focus on natural language bias in neural networks. However, bias is present in many areas of artificial intelligence. This behaviour is pervasive, first the models capture these associations and then replicate them as a result of their application. Bias in AI is encompassed in study areas such as Fairness or Explainability

Keywords

bias, deep learning, nlp, fairness, explicability

1. Introduction

Artificial intelligence has come a long way in recent years, and a good part of this progress is possible thanks to neural network models. These models that are trained with large amounts of data have demonstrated a great capacity to capture reality. However, capturing reality with such precision can sometimes be negative, since they also capture and replicate undesirable stereotypes

One example is the police COMPAS system in the United States. This system assigns detainees a level of risk of recidivism. From an independent analysis, it was discovered that the system failed for both white and black people[1], but the type of error was different. In the case of white people, the system would systematically provide a lower level of recidivism risk than the actual level they should have, it was failing in their favor. While in the case of black people the error was against them, the system assigned a higher level of risk. In this case, there is a social problem in which an algorithm can be disruptive in people's lives and simultaneously a resource allocation problem in which a system whose malfunctioning causes resources not to be allocated where they are really needed. A similar example can be found in a medical system called Optum[2], which would systematically allocate black patients fewer resources for their

Doctoral Symposium on Natural Language Processing from the PLN.net network 2022 (RED2018-102418-T), 21-23 September 2022, A Coruña, Spain.

igmunoz@ujaen.es (I. Garrido-Muñoz)
https://isgarrido.github.io/ (I. Garrido-Muñoz)
0000-0001-6656-9679 (I. Garrido-Muñoz)

treatment than white patients for the same level of need. This is a case of resource allocation by a biased system that can negatively influence people's health. There are several examples of automated recruitment systems such as HireVue[3] which uses artificial intelligence models to evaluate candidates. However, the system disadvantaged candidates who deviated from the model's definition of normal. This behavior is quite frequent when the model was not trained with examples that are sufficiently diverse. In this case, it is intuited that HireVue malfunctioned on candidates that were not English-native speakers, since their accents would confuse the model. The most controversial part is that in those cases where the model was not properly working, the candidate was automatically discarded and did not receive any information about the reasons. This makes us think that the application of non-explainable models may be unfair and even harmful in some situations. Amazon also discarded[4] a similar tool for recruitment, as it was found to be biased against women. Those are just some examples but those biases have been found in systems where the model is applied to tasks such as computer vision[5], audio processing[6] or linguistic corpora[7, 8].

2. Bias in NLP with deep learning

The study of bias in neural network language models can be divided into two tasks. On one hand, the evaluation task aims to discern whether a model is biased. With a biased model, the next step should be to characterize and quantify the bias. Characterization helps us understand what form it takes and thus study what causes it and where it is encoded. Quantification helps us to know to what extent its application to real-world problems can provoke harm to society. Once the evaluation has been conducted, the next step is the mitigation task, also known as correction of the bias, in this task the objective is to change the model to avoid the detrimental effect that was found in previous phases.

When quantifying or characterizing the bias present in a model, it is important to consider the task to which the model is applied and which attributes should be protected for that task. For example, in a personnel selection process, a protected attribute might be the candidate's gender, since it is a characteristic that should not influence the model's perception of the candidate. Other protected attributes could be age, race, religion, origin, etc. In the case of wanting to study more than one protected attribute in the same model, one could consider applying a method that automatically detects the bias for each protected attribute or applying an adapted version of the method developed for each given protected attribute. For example, if we want to check if there is a personnel selection bias, the system should score equally two candidates with the same qualifications in which the only change is gender, nationality or race. If this is not the case, we could study whether there is a systematic bias against each of these. In the following section, I briefly present the study of bias in data models, for more detail see a previous paper[9] in which we conducted an in-depth review of the study of bias.

2.1. Bias evaluation

All kinds of models have been evaluated, from the most basic ones such as the study by Bolukbasi et al. [10] on Word Embeddings, in which it is already found that the models associate men and

women with different professions. While some analogies like $man - woman \approx king - X$ work correctly and show us that $X = queen$, other analogies like $man - woman \approx computer\ programmer - X$, the model output is $X = homemaker$, as it does not associate computer programmer with woman, moreover it finds a set of professions strongly associated with each gender.

Later studies approach the bias issue from the point of view of Coreference Resolution, this is the approach of Zhao et al. [11] and Jieyu et al. [12] with GloVe. Also for Word Embedding models Caliskan et al. [13] did create **WEAT** as an association test between a concept and a protected attribute. Following the previous example, this test could be used to measure if any of the genders is strongly associated with the profession while the other is not. This test is used and extended later by Dev et al. [14] or the work of Manzini et al. [15] that besides gender also studies race and religion. Also, Lauscher et al. [16] has explored the problem of Word Embeddings in more languages such as German, Spanish, Italian, Russian, Croatian and Turkish.

More recent and complex models such as BERT or GPT-2 were also studied. In the work of Vig [17] he introduces a tool to visualize the attention of the model to try to understand where it is capturing unwanted biases. From the previously mentioned WEAT there are numerous adaptations such as SEAT[18], that instead of testing against a word, does test the protected attribute against a sentence for contextual models such as BERT. This work is also extended by Tan and Celis [19] that adapts SEAT to work with the full context instead of just on a sentence level. This latest evaluation method is applied to GPT-2, BERT, ELMo, among others.

2.2. Bias correction

There are several techniques for bias correction, on the one hand, the correction of the vector space of the model. In this technique the author search for the dimension or dimensions of the vector space that encodes the bias and equalizes the differences between, for example, the distance between the protected attribute (man, woman) and the biased concept (the profession). This is one of the approaches present on the work of Manzini et al. [15], Zhou et al. [20], Dev and Phillips [21]. However the work of Gonen and Goldberg [22] suggests that this technique does not eliminate bias but hides it, this work also shows how in the Word Embedding models that had the bias removed, it can still be found and is therefore the current approaches are not sufficient.

On the other hand, another line of work is helping the model acquire the ability to give an unbiased result without changing its internals. Other studies[23] train or fine-tune the model by doing some changes to its data like doing a data augmentation technique called **Counterfactual Data Augmentation** where they balance the times that appear each instance protected attribute (for example if in the data has ”he is a computer programmer”, this technique would add ”she is a computer programmer”, this would be extensible to other protected attributes with more dimensions such as race). The main advantage of fine-tuning already trained models to reach similar results since training a model from scratch is costly and this would be an impediment to its application given time or cost restrictions.

3. Relevance of the problem

Day by day, these huge models are being integrated as part of products and systems in production. On the one hand, we could speak of a purely economic disadvantage since a biased system would not work properly for all potential users. On the other hand, these models have an impact on people's lives, there are certain contexts in which it would not be appropriate to use a model that may contain some type of bias, such as systems for distributing public resources, systems related to employment, or bank credit. Therefore, bias in data models should be studied, as well as its causes, in order to either avoid using biased models or to mitigate harmful bias when necessary.

4. Hypotheses and objectives

The following hypothesis is assumed: Given a language model based on deep learning, it will be possible to discern whether it contains biases, and characterize, measure, and mitigate them.

The following objectives are established:

- Conduct an intensive study of the state of the art regarding detection, evaluation or mitigation of biases in deep learning models.
- Analyze and characterize biases present in existing models.
- Development of techniques and algorithms for unsupervised or semi-supervised detection and characterization of bias in existing models.
- Development of techniques and algorithms for the mitigation or correction of bias in existing models.
- Development of techniques that allow the creation of bias-free models in a given context.
- Development of techniques that allow the creation of robust and explainable models.

5. Methodology and the proposed experiments

To achieve these objectives, we first conducted an extensive study[9] of the state of the art regarding bias in data models. The next step was to run some experiments to look for bias in 20 deep learning models available for Spanish. We did adapt some of the existing association tests. We did adapt the gender experiments conducted in StereoSet[24] . In Stereoset they compare the result of the model by giving it a sentence that acts as a context and is followed by either a stereotyped sentence or an anti-stereotyped sentence. This way you can test the model to see if it is capturing the stereotype. We took this idea to design our experiment. We created a set of sentences where the only change is the genre. For example *Su **profesor** es muy MASK* , *Su **profesora** es muy MASK* . We ask the model to complete these sentences, expecting to receive an adjective. This way can study how the models perceive men and women. The initial results gave us some clues, you can see that the models perceive women by their physical characteristics. In the table 1 we have an example, we can see how more words referring to the physical appearance (in italics) appear in the female version of the sentence, in fact in the female version the first 3 positions are occupied by words referring to the physical appearance when we order by the probability that the model assigns to each token (score).

El chico es <mask>.			La chica es <mask>.		
Index	**token_str**	**score**	**Index**	**token_str**	**score**
0	*guapo*	0.06381441652774811	0	*preciosa*	0.1420050412416458
1	genial	0.03892676532268524	1	*guapa*	0.043913714587688446
2	buenísimo	0.03504825755953789	2	*hermosa*	0.03974246606230736
3	encantador	0.032205186784267426	3	espectacular	0.039122529327869415
4	*precioso*	0.02545679733157158	4	encantadora	0.031081615015864372
5	perfecto	0.02219405397772789	5	buenísima	0.024362897500395775
6	espectacular	0.021783752366900444	6	impresionante	0.022533416748046875
7	impresionante	0.021090665832161903	7	genial	0.019734643399715424
8	estupendo	0.01639944314956665	8	estupenda	0.017054326832294464
9	bueno	0.015286967158317566	9	maravillosa	0.01553318277001381

Table 1
Result for a pair of templates on the model BSC-TeMU\roberta-base-bne [25]

This experiment confirmed that Spanish language models capture a bias that in certain contexts can be detrimental. The proposed adjectives are categorized according to the Supersenses taxonomy[26]. The main used categories to group the adjectives are **Body, Behaviour, Social, Feel, Mind**. By applying this task to a set of template sentences and aggregating the results, we concluded that Spanish language models perceive women by their physical appearance (adjectives on the body category) while men by their behavior-related characteristics. This work is pending publication.

In order to facilitate the interpretation of the results of this study, a tool was developed that allows us to explore the results of the study in a visual way. In the tool we can:

- Visualize how the predictions of each model are distributed among the different categories, immediately comparing the weight associated with the male sentences with respect to the female sentences.
- Explore in detail the model values for each category, as well as visualize them using a heatmap to see which models are the most biased and if they are biased towards one gender or the other.
- To know the number of adjectives contributed by each model for male, female, and both.
- To explore each of the specific model responses for each of the templates.

. The tool can be found at https://ismael.codes/categoryviewer/

5.1. Future work

The next step is to find out what causes these behaviors within the model. We could pose it as a problem of causal inference in which it is a confounder that causes the biased result, by finding the confounder we could try to correct the model to eliminate its effect and thus obtain unbiased results. A concrete approach following this idea would be to study the variation of the model parameters between pairs of sentences, one biased and the other unbiased. This way we could locate what the difference is in the model and therefore be able correct the model.

References

[1] J. L. Julia Angwin, Machine bias - there's software used across the country to predict future criminals. and it's biased against blacks., 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Z. O. U. Berkeley, Z. Obermeyer, U. Berkeley, S. M. U. o. Chicago, S. Mullainathan, U. o. Chicago, O. M. A. Metrics, Dissecting racial bias in an algorithm that guides health decisions for 70 million people: Proceedings of the conference on fairness, accountability, and transparency, 2019. URL: https://dl.acm.org/doi/10.1145/3287560.3287593.

[3] D. Harwell, A face-scanning algorithm increasingly decides whether you deserve the job, 2019. URL: https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/.

[4] J. Dastin, Amazon scraps secret ai recruiting tool that showed bias against women, 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[5] A. Howard, J. Borensteion, Trust and bias in robots, 2019. URL: https://www.americanscientist.org/article/trust-and-bias-in-robots.

[6] J. Rodger, P. Pendharkar, A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application, Int. J. Hum.-Comput. Stud. 60 (2004) 529–544. doi:10.1016/j.ijhcs.2003.09.005.

[7] J. A. Bullinaria, J. P. Levy, Extracting semantic representations from word co-occurrence statistics: A computational study, Behavior Research Methods 39 (2007) 510–526. URL: https://doi.org/10.3758/BF03193020. doi:10.3758/BF03193020.

[8] M. Barlow, Michael stubbs. text and corpus analysis: Computer-assisted studies of language and culture, International Journal of Corpus Linguistics 3 (1998) 319–327.

[9] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, L. A. Ureña-López, A survey on bias in deep nlp, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/7/3184. doi:10.3390/app11073184.

[10] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, CoRR abs/1607.06520 (2016). URL: http://arxiv.org/abs/1607.06520. arXiv:1607.06520.

[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, arXiv e-prints (2018) arXiv:1804.06876. arXiv:1804.06876.

[12] Jieyu, Y. Zhou, Z. Li, W. Wang, K.-W. Chang, Learning Gender-Neutral Word Embeddings, arXiv e-prints (2018) arXiv:1809.01496. arXiv:1809.01496.

[13] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186. URL: https://www.science.org/doi/abs/10.1126/science.aal4230. doi:10.1126/science.aal4230. arXiv:https://www.science.org/doi/pdf/10.1126/science.aal4230.

[14] S. Dev, T. Li, J. M. Phillips, V. Srikumar, OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5034–5050. URL:

https://aclanthology.org/2021.emnlp-main.411. doi:10.18653/v1/2021.emnlp-main.411.
[15] T. Manzini, L. Yao Chong, A. W. Black, Y. Tsvetkov, Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 615–621. URL: https://aclanthology.org/N19-1062. doi:10.18653/v1/N19-1062.
[16] A. Lauscher, G. Glavas, S. P. Ponzetto, I. Vulic, A general framework for implicit and explicit debiasing of distributional word vector spaces, in: AAAI, 2020.
[17] J. Vig, A multiscale visualization of attention in the transformer model, 2019, pp. 37–42. doi:10.18653/v1/P19-3007.
[18] C. May, A. Wang, S. Bordia, S. R. Bowman, R. Rudinger, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 622–628. URL: https://aclanthology.org/N19-1063. doi:10.18653/v1/N19-1063.
[19] Y. C. Tan, L. E. Celis, Assessing social and intersectional biases in contextualized word representations, in: NeurIPS, 2019.
[20] P. Zhou, W. Shi, J. Zhao, K.-H. Huang, M. Chen, K.-W. Chang, Analyzing and mitigating gender bias in languages with grammatical gender and bilingual word embeddings, in: ACL 2019, 2019.
[21] S. Dev, J. M. Phillips, Attenuating bias in word vectors, CoRR abs/1901.07656 (2019). URL: http://arxiv.org/abs/1901.07656. arXiv:1901.07656.
[22] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, in: NAACL-HLT, 2019.
[23] R. H. Maudslay, H. Gonen, R. Cotterell, S. Teufel, It's all in the name: Mitigating gender bias with name-based counterfactual data substitution, CoRR abs/1909.00871 (2019). URL: http://arxiv.org/abs/1909.00871. arXiv:1909.00871.
[24] M. Nadeem, A. Bethke, S. Reddy, Stereoset: Measuring stereotypical bias in pretrained language models, CoRR abs/2004.09456 (2020). URL: https://arxiv.org/abs/2004.09456. arXiv:2004.09456.
[25] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.
[26] Y. Tsvetkov, N. Schneider, D. Hovy, A. Bhatia, M. Faruqui, C. Dyer, Augmenting English Adjective Senses with Supersenses, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4359–4365. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1096_Paper.pdf.