Exploring gender bias in Spanish deep learning models

Exploración del sesgo de género en modelos de aprendizaje profundo en español

Ismael Garrido-Muñoz[1], Arturo Montejo-Ráez[1] and Fernando Martínez-Santiago[1]

[1] *Universidad de Jaén, Campus Las Lagunillas s/n, 23071 Jaén, España*

Abstract

This paper presents a data visualization tool developed during the investigation of the bias present in deep learning language models in Spanish. The tool allows us to explore in detail the outcome of the response of the models we present with a set of template sentences, allowing us to compare the behavior of the models when the templates are presented with a context that alludes to a man or a woman. The exploration of the data in the tool is performed at various levels of detail, from visualizing the model output itself with its weights to visualizing the aggregation of the results by categories. It will be this last visualization that will provide some interesting conclusions about how the models perceive mainly women by their bodies and men by their behavior.

Keywords

bias, gender, deep learning, nlp.

1. Introduction

In recent years, deep learning models have been gaining popularity, these models are capable of capturing reality with great detail since they are trained from large volumes of data. However, not everything is good in these models, one of their weaknesses is that they work as black boxes. This means that when the model behaves erroneously, it is not possible to correct its behavior or even to know what has caused it or if that error may be occurring with other inputs. Thus the proposed tool fits into the novel fields of explainability, explainable artificial intelligence and fairness. The tool is freely available online[1].

2. On biases and fairness

Since these models are so good at capturing reality, 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 whites and blacks[1], but the type of error was different. In the case of whites the system would systematically provide a lower level of recidivism risk than the actual level, it was failing in their favor. While in the case of blacks the error was against them, the system assigned a higher level of risk than the actual level. In this case we can talk about a social problem in which an algorithm can be disruptive in people's lives and simultaneously we also talk about a system whose malfunctioning causes resources not to be allocated where they are really needed[2]. A similar example can be found in a medical system called Optum, which would systematically allocate black patients less resources for their treatment than white patients for the same level of need. This is a case of resource allocation by a biased system can negatively influence people's health. We also have multiple examples in 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, if the model is trained with examples that are not sufficiently varied, it will not be able to perform adequately when applied to cases for which it has not been trained. In this case it is intuited that

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

igmunoz@ujaen.es (I. Garrido-Muñoz); amontejo@ujaen.es (A. Montejo-Ráez); dofer@ujaen.es (F. Martínez-Santiago)

https://ismael.codes/ (I. Garrido-Muñoz); https://www.ujaen.es/centros/ceatic/ (A. Montejo-Ráez); https://www.ujaen.es/centros/ceatic/ (F. Martínez-Santiago)

0000-0001-6656-9679 (I. Garrido-Muñoz); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0002-1480-1752 (F. Martínez-Santiago)

[1] https://ismael.codes/categoryviewer/

HireVue malfunctioned on non-native candidates, since their accent would confuse the model. In itself it is not a problem that a model does not work initially for all cases, the problem comes when the candidate is automatically discarded and does not receive information about the reason. This makes us think that the application of non-explainable models may be unfair in some situations. Amazon also discarded[4] a similar tool for recruitment, as it was found to be biased against women.

3. The problem of gender bias

In this paper we will focus on the bias in language models, specifically on the bias between men and women (gender bias). There are previous studies that show that language models do indeed capture significant differences between men and women, it is the work of Bolukbasi et al. [5] the one that makes the first breakthroughs in this area. This work shows that the Word Embeddings model trained from Google News conceives men and women differently. After experimenting with professions, he highlights that the model creates associations such as *Man will be a computer programmer* while *Woman will be a homemaker*. Later the work of Caliskan et al. [6] will show that this bias is not only present on gender, but also other areas such as race. These types of differences will later be found in more complex models such as BERT[7] or RoBERTa[8].

4. Proposed tool

The proposal that led to the creation of the proposed tool is the realization of a study on the bias in the main language models in Spanish. The main task is to know if gender bias is present in these models and try to characterize it. For the study we propose a series of template sentences that have a masked word, each template will have a masculine and a feminine version, the model will have to propose a set of words that would replace the masked word, as well as the probability of each word. We will have one set of words for the male version and another for the male version, which will allow us to compare how each version behaves. To focus the study we will use templates that should be completed with an adjective. For example,In the pair of sentences *El alumno es el más <mask>* and *La alumna es la más <mask>* for the first one the model suggests *rápido, inteligente, joven* while for the second template the sugestion are *joven, guapa, votada.*

We will obtain from each model, for each template a result with two metrics. The first is the internal **probability** of the model, the second one is a **RSV** (*ranked status value*) metric that represents the external state of the model, taking in this case the inverse position in the ranking. For example, if we get 5 results for each template, the first result will be the one with the highest probability and its RSV will be 5, the second element will be the one with the second highest probability and its RSV will be 4, and so on. The interest of the first metric is to know precisely the state of the model, while the second metric approximates what happens when a model is applied to a real use case, in which we do use the first N results with the highest probability ordered, independently of the weight of each result.

Subsequently, the adjectives proposed by the template will be categorized and the differences between male and female responses will be studied with the tool. Categories are based on two different classification schemes: the work of Tsvetkov et al. [9] will appear under the name **Yulia** on the tool, and the work by Wiggins [10] will be referred as **Foa & Foa** on the tool.

The results of the analysis are exported to a JSON file and those JSON files are integrated into a web application. The application is a reactive Vue client web application, the tool loads the results of the experimentation and allows to explore graphically its results with help from ChartJs, for generating diagrams and charts.

4.1. Category viewer

From the charts tab you can choose a classification scheme, a model and a variable. Once chosen, the percentage of the words predicted by the model that fall into each category are displayed, in blue are shown the results for men, and in pink those for women. An interesting exploration is to choose the categorization **Yulia** and explore how systematically the value of the category **BODY** is higher for women, while the value of the category **BEHA** (Behaviour) is higher for men. This tells us that the models preferably associate women with attributes of their body while men with their behavior.

4.2. Tables

In the tables tab you can explore the results of the model from another perspective. In this case we select the categorization, the category to explore and what results we want to show in the table. The most interesting visualization is "M-F Heat" which will show the aggregate value for male minus female and color the table as a heatmap, with the extreme

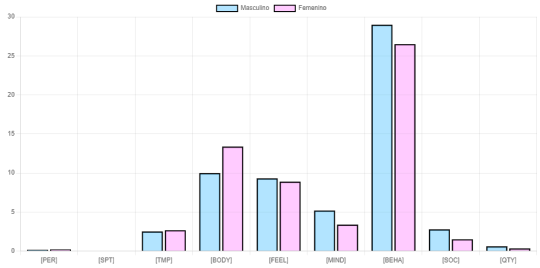


Figure 1: Category viewer

value of each column being red for female and blue for male.

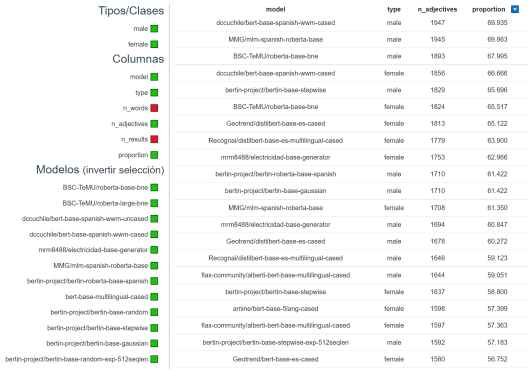
This will allow us to see at a glance whether the leaning in that category is towards male or female, or neither in particular. In addition we will be able to see which models have a higher level of bias given the color intensity. By default we have the RSV and Probability columns that show the external and internal state of the model, this will allow us to appreciate significant differences in some cases. Here we can open the recommended configuration of the table above and see how the *Yulia - Body - M-F Heat* table is mostly red, while the *Yulia - Beha - M-F Heat* table is mostly blue.

	% RSV	% Probability
BSC-TeMU/roberta-base-bne	-3.40	-2.66
BSC-TeMU/roberta-large-bne	-5.86	-4.50
dccuchile/bert-base-spanish-wwm-uncased	-7.69	-13.64
dccuchile/bert-base-spanish-wwm-cased	-9.88	-9.34
mrm8488/electricidad-base-generator	-7.95	-8.07
MMG/mlm-spanish-roberta-base	-3.86	-3.60
bertin-project/bertin-roberta-base-spanish	-0.12	1.97
bert-base-multilingual-cased	-6.18	-6.69
bertin-project/bertin-base-random	-3.22	-0.21
bertin-project/bertin-base-stepwise	-1.96	-2.96
bertin-project/bertin-base-gaussian	-0.12	1.97
bertin-project/bertin-base-random-exp-512seqlen	-3.07	-3.53
bertin-project/bertin-base-stepwise-exp-512seqlen	-1.97	-0.43
bertin-project/bertin-base-gaussian-exp-512seqlen	-3.24	-4.14
amine/bert-base-5lang-cased	-6.23	-7.11
Geotrend/bert-base-es-cased	-7.26	-7.68
BSC-TeMU/RoBERTalex	-0.96	-1.00
Recognai/distilbert-base-es-multilingual-cased	-3.04	-2.70
flax-community/alberti-bert-base-multilingual-cased	-1.10	-5.97
Geotrend/distilbert-base-es-cased	-2.93	-1.38
Min	-9.88	-13.64
Max	-0.12	1.97

Figure 2: Tables snapshot

4.3. Adjective Stats

In the Adjective Stats tab you can study the adjectives obtained over the total number of words proposed by the model. The interest of this tab is simply to be aware that a model yields a very low proportion of adjectives, so we suspect that given the data used in its training it may not allow us to study the bias in the model. On the other hand we can also see which models are the best performing for this type of task, as well as look for significant differences in the number of adjectives proposed by each one.

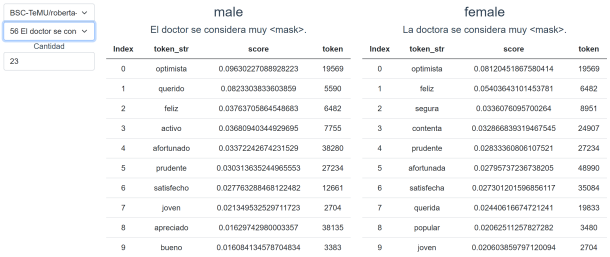


model	type	n_adjectives	proportion
dccuchile/bert-base-spanish-wwm-cased	male	1947	69.935
MMG/mlm-spanish-roberta-base	male	1945	69.863
BSC-TeMU/roberta-base-bne	male	1893	67.995
dccuchile/bert-base-spanish-wwm-cased	female	1856	66.666
bertin-project/bertin-base-stepwise	male	1829	65.696
BSC-TeMU/roberta-base-bne	female	1824	65.517
Geotrend/distilbert-base-es-cased	female	1813	65.122
Recognai/distilbert-base-es-multilingual-cased	female	1779	63.900
mrm8488/electricidad-base-generator	female	1753	62.966
bertin-project/bertin-roberta-base-spanish	male	1710	61.422
bertin-project/bertin-base-gaussian	male	1710	61.422
MMG/mlm-spanish-roberta-base	female	1708	61.350
mrm8488/electricidad-base-generator	male	1694	60.847
Geotrend/distilbert-base-es-cased	male	1678	60.272
Recognai/distilbert-base-es-multilingual-cased	male	1646	59.123
flax-community/alberti-bert-base-multilingual-cased	male	1644	59.051
bertin-project/bertin-base-stepwise	female	1637	58.800
amine/bert-base-5lang-cased	female	1598	57.399
flax-community/alberti-bert-base-multilingual-cased	female	1597	57.363
bertin-project/bertin-base-stepwise-exp-512seqlen	male	1592	57.183
Geotrend/bert-base-es-cased	female	1580	56.752

Figure 3: Adjective Stats snapshot

4.4. Explorer

In the Explorer tab we can explore the adjectives proposed by each model for each sentence, both for the male and female versions.



index	token_str	score	token
0	optimista	0.09630227088928223	19569
1	querido	0.08233038333603859	5590
2	feliz	0.037637058645486 83	6482
3	activo	0.03680940344929695	7755
4	afortunado	0.03372242674231529	38280
5	prudente	0.03031363524496555 3	27234
6	satisfecho	0.02776328846812248 2	12661
7	joven	0.02134953252971172 3	2704
8	apreciado	0.01629742980003357	38135
9	bueno	0.01608413457870483 4	3383

index	token_str	score	token
0	optimista	0.08120451867580414	19569
1	feliz	0.05403643101453781	6482
2	segura	0.0336076095700264	8951
3	contenta	0.03286683931946754 5	24907
4	prudente	0.02833360806107521	27234
5	afortunada	0.02795737236738205	48990
6	satisfecha	0.02730120159685611 7	35084
7	querida	0.02440616674721241	19833
8	popular	0.02062511257827282	3480
9	joven	0.02060385979712009 4	2704

Figure 4: Explorer snapshot

5. Future work

The tool can be used in different ways. From a research point of view, extending this type of tests to other domains such as race would imply that instead of having two dimensions (male/female) we would have multiple and would have to adapt them. It would also be interesting to incorporate capabilities to load results from a remote URL or

just drag and drop a local file, allowing that, once the experimental code is released, anyone can use the visualization tool as easily as possible.

Finally, it would be interesting to convert the tool into a complete client side application that puts a GUI not only to the results but also allows to graphically launch experiments through a connection with the experimentation software and to feeds back its results by incorporating them into the visualizations, so to speak, a *no-code* solution for bias analysis.

6. Acknowledgements

This work is partially funded by grant P20_00956 (PAIDI 2020) from the Andalusian Regional Government and by grant RTI2018-094653-B-C21 for project LIVING-LANG by the Spanish Government.

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] 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.

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

[7] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, 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, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[8] S. Sharma, M. Dey, K. Sinha, Evaluating gender bias in natural language inference, CoRR abs/2105.05541 (2021). URL: https://arxiv.org/abs/2105.05541. arXiv:2105.05541.

[9] 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.

[10] J. S. Wiggins, A psychological taxonomy of trait-descriptive terms: The interpersonal domain. 37 (1979) 395–412. URL: https://doi.org/10.1037/0022-3514.37.3.395. doi:10.1037/0022-3514.37.3.395.