Supervised Bias Detection in Transformers-based Language Models

Michele Dusi[1,2,*], Alfonso Emilio Gerevini[1], Luca Putelli[1] and Ivan Serina[1]

[1]University of Brescia (UniBS), Italy
[2]Sapienza University of Rome, Italy

Abstract

Training Large Language Models on biased datasets tends to teach a discriminatory behavior to the systems themselves, as it has been proven by the last years literature on fairness in AI and Machine Learning algorithms. The developed bias-detection strategies often ignores the inner body of the model, making it easy to generalize the methodology, but harder to understand the underlying motivations. In this paper, we present a general approach for detecting unwanted prejudices in Language Models, requiring only a small set of input data. Our strategy works on the embedding representation of languages, without any constraint on model architecture, but it is able to detect which parts of the representation is the most prejudice-affected.

Keywords

Natural Language Processing, Large Language Models, AI Fairness, Bias Detection, Word Embeddings.

1. Introduction

In recent years, there has been significant growth in the utilization of pre-trained models for Natural Language Processing across various sectors, including chatbots, sentiment analysis systems, and applications in fields such as medicine, marketing, and education. Naturally, the primary concern of the Machine Learning community for these applications is their performance. New, intricate architectures like BERT and other Transformer-based models have proven to deliver a notably high level of accuracy. However, it's crucial to note that these models are trained using extensive datasets directly sourced from the internet. Consequently, they may inadvertently incorporate biases, prejudices, and stereotypes related to demographic minorities, such as gender, race, religion, sexual orientation, disability, and more. These unintended characteristics can emerge in the algorithms, potentially leading to discriminatory behavior.

Numerous studies have been dedicated to this topic, revealing that both word embedding representations and pre-trained language models encompass gender bias [1], showing a deep relationship between the gender stereotypes and the use of language. Further studies have also highlighted the presence of other types of biases (ethnicity bias, religion bias, sexual orientation bias, …) in monolingual and bilingual models [2]. Works following a *geometric* approach try to

Doctoral Consortium of AIxIA 2023: 22nd International Conference of the Italian Association for Artificial Intelligence, Roma Tre University - ICITA Department, November 6–9, 2023, Rome, Italy.
[*]Corresponding author.
michele.dusi@unibs.it (M. Dusi)
0000-0002-6564-3096 (M. Dusi); 0000-0001-9008-6386 (A. E. Gerevini); 0009-0008-5055-6812 (L. Putelli); 0000-0002-7785-9492 (I. Serina)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

measure bias according to concepts representation in word embeddings; this idea follows directly the *distributional hypothesis* from linguistics [3], stating that nearby embeddings corresponds to similar words in meaning. Statistical analysis of the spatial relationships between sets of concepts has also led to bias-quantification tests [4, 5, 6, 7].

One of the key limitations in these studies lies in their treatment of the model as an opaque entity, a *black-box*. Primarily, their focus is on confirming the presence or absence of bias without delving deeper into how it is encoded within the model. Instead, this work presents an alternative methodology for studying the presence of bias in a generic model of contextual word embeddings. Furthermore, this approach could pinpoint the specific components of the word vectors that are responsible of conveying the bias. This insight aids in visualization, offering an immediate grasp of bias within a language model, and helps the development of more targeted strategies for bias reduction.

In the following, we apply our methodology to three different protected attributes: *gender*, *ethnicity* and *religion*; we thus consider different stereotypes involving perceived criminality, positive and negative terms and jobs salary. The experiments have been conducted on the BERT architecture [8] (in particular, the base English version), but our approach can be generalized considering other Transformer-based models [9].

2. Materials and Methods

Applying our method requires two different sets of word vectors: a set of **protected words**, characterising the protected attribute, and a second set of **stereotyped words**, characterising the values in which the stereotype expresses. For instance, we could have a first set of words identifying religions (*christian*, *muslim*, *church*, *mosque*, *priest*, *imam*, etc.), and a second identifying positive and negative adjectives (*good*, *faithful*, *innocent*, *bad*, *treacherous*, *guilty*, etc.). Each word is associated with a value of the corresponding attribute; in the previous example, protected words refer to values *christian* or *muslim*, whereas stereotyped words can be *positive* or *negative*. Lastly, each word has a reciprocal word vector obtained through BERT. In conclusion, items of the two sets can be seen as tuples with a **word**, a **value** for the property, and a **vector** with the size of the model embedding space.

The words datasets have been crafted from scratch to precisely identify the chosen attributes, taking word samples from internet services like WordReference[1]. The construction of the word embeddings by BERT, instead, required the usage of contextual sentences in which the words appeared. Each sentence was processed by the Language Model and converted into a series of embeddings, one for each *token*[2]. Next, only the embedding referring to the inquired word was retained, obtaining a single word vector.

The core procedure of bias detection is composed by two steps: in a first phase, the protected attribute is characterised in terms of relevant embedding features, meaning that we identify which components of the word vectors encode the analysed attribute. The idea of this preliminary step is to understand how the protected property is represented within the model

[1] https://www.wordreference.com/it/

[2] A token is a part of text in which Large Language Models like BERT split the input sentences. For the sake of simplicity, we could think of a token as a single word.

$\text{N} = 235$		*Predicted* values (protected)		$\sum$
		christian	**muslim**	
Actual values (stereotyped)	**positive**	80	40	120
	negative	39	81	120
		119	121	240

Table 1
Example of classification frequencies for adjective terms according to the *religion* attribute. The values are obtained by taking $\text{N} = 235$.

embedding space and which features best describe it. The second phase compares the stereotyped words to the previously-described protected features, effectively performing operations of bias detection and quantification. The second step provides the desired outcome of quantifying a prejudice; however, it builds upon the previous procedure and cannot be performed by itself.

More specifically, in the **first phase** we train a classifier on the protected words, teaching it to discern different values of the protected attribute by their embeddings. This way, the classifier learns which features are relevant for encoding the inquired property (gender, ethnicity or religion). After experimenting with multiple classifiers, we opted for the one that gave the best results, a Linear Support Vector Machine (LSVM). This choice can be explained by looking at the dimension of the learning dataset, around 100 samples, which is not enough for training a neural network architecture.

The LSVM learns one scalar weight for each one of the embedding dimensions; in our case, BERT works in a 768-dimensional space. Among these 768, we selected the N highest weights by their absolute value, where N is a hyperparameter chosen based on the protected and stereotyped attributes; its exact value, typically around 100, will be discussed later. The N highest weights identify the N most relevant features of the embedding space; for this, we project all the word vectors to those N dimensions, i.e. we shorten the vectors by discarding the $768 - \text{N}$ non-relevant features. This operation is applied both to protected and stereotyped word embeddings.

The outcome at the end of the first phase consists in two sets of **reduced word vectors** with size N, corresponding to the original word vectors whose only the most relevant dimensions have been retained. This enhance the encoding of the protected attribute within the embeddings, if present, and removes the unwanted noise in the vector.

The **second phase** works on the enhanced sets to detect whether a bias is present. A bias in word embedding is a **distortion** of how concepts are represented; in simpler words, if the *terrorist* word vector is more similar to vectors of value *muslim* than vectors of value *christian*, the representation is not balanced and, thus, not fair.

To measure similarity of stereotyped vectors with respect to protected ones, we train another classifier on the reduced N-dimensional protected word vector, which learns to discern among genders, religions or ethnicities. Thence, we provide to it the stereotyped embeddings as a test set and look at the predicted classes. In the ideal situation, the classifier shouldn't be able to predict anything, because the test samples (the stereotyped embeddings) are not related to the training samples (the protected embeddings). As a consequence, the outcomes for the test set should be random, with equal probabilities for each predicted class; this is our **null hypothesis**.

On the contrary, if the predictions are unbalanced and the protected values show some sort of

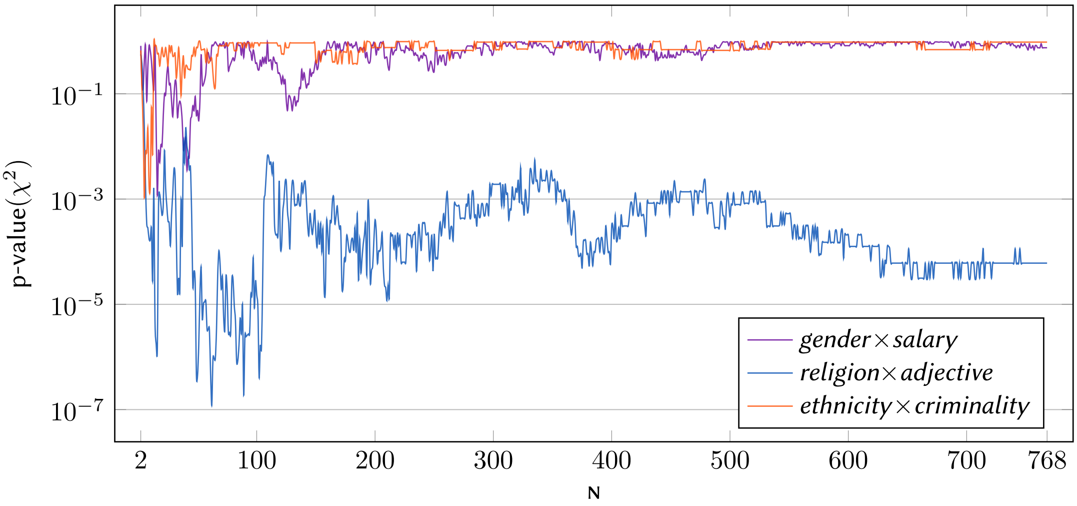

Figure 1: Each line shows the p-values for a given protected+stereotyped attributes pair, in relation with the value of the hyperparameter N. Lower pikes represent higher distortion in words, and thus a more prejudiced concepts representation.

correlation with the stereotyped values (for example, if the *low-salary jobs* vectors are classified as *female*, or if the *negative* adjectives are classifies as *muslim*, etc.), the null hypothesis should be rejected and we assess the presence of a **prejudice**.

3. Results and Discussion

In Table 1, we report the observed frequency for the classification of *positive* and *negative* adjectives in the *religion* classes. As it can be observed, the distribution is not balanced for the protected attribute: the tendency to classify the positive adjectives as *christian* is contrasted by the opposite tendency to classify negative terms as *muslim*.

We can measure how the distribution shifts from a random classification via the χ^2 test. The resulting p-value for the example in Table 1 is $1.1 \cdot 10^{-7}$, meaning that such distribution is not random, and thus the protected and stereotyped attributes are perceived as correlated with probability higher than 0.999999.

Other domains give similar outcomes: the p-value for jobs word vectors grouped by *salary* (stereotyped attribute) and classified by *gender* (protected attribute) is $6.4 \cdot 10^{-3}$ for $\text{N} = 42$; the p-value for the *ethnicity* protected attribute compared with *criminality* words is $1.8 \cdot 10^{-3}$ for $\text{N} = 10$.

Finally, in Figure 1 we address the problem of choosing the right value for the hyperparameter N. The plots show the trend of the p-value for different N going from 2 to 768, representing different percentages of retained dimensions of the original embedding space. As we can see, higher values of N includes noise in the representation, obfuscating the underlying bias which would not be detected by our method. Lower values of N, instead, give a clean representation of the protected attribute in the stereotyped words, enhancing their unwanted distortion.

4. Conclusion

In this document we briefly presented a geometric approach to assess bias detection in Large Language Models. We considered BERT in the English base implementation, evaluating common prejudices for gender, ethnic and religious groups of people. The method requires very few input resources to grasp and assess the stereotypes, making it easy to implement and use.

The hyperparameter N controls the amount of information retained from the original word embeddings; choosing the right value is crucial to detect the distortions. The plots in Figure 1 indicate that different attributes are characterised by and encoded with different amount of dimensions. Further analysis are required to understand how the best value of N can be chosen in advance, without testing all the possible values.

In the future, many other domains could be tested with this methodology, discovering connotations not only related to biases and prejudices, but also to the conception of the word embedding space. This could improve our comprehension of Large Language Models, allowing to operate on word vectors and language representation with increased precision and knowledge.

Acknowledgments

This work was carried out while the author, Michele Dusi, was enrolled in the *Italian National Doctorate on Artificial Intelligence* run by Sapienza University of Rome in collaboration with the University of Brescia.

References

[1] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 4349–4357.

[2] P. Zhou, W. Shi, J. Zhao, K. Huang, M. Chen, R. Cotterell, K. Chang, Examining gender bias in languages with grammatical gender, 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 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 5275–5283.

[3] Z. S. Harris, Distributional structure, WORD 10 (1954) 146–162. URL: https://doi.org/10.1080/00437956.1954.11659520. doi:10.1080/00437956.1954.11659520. arXiv:https://doi.org/10.1080/00437956.1954.11659520.

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

[5] A. Lauscher, G. Glavas, Are we consistently biased? multidimensional analysis of biases in distributional word vectors, in: R. Mihalcea, E. Shutova, L. Ku, K. Evang, S. Poria (Eds.), Proceedings of the Eighth Joint Conference on Lexical and Computational Seman-

tics, *SEM@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, Association for Computational Linguistics, 2019, pp. 85–91.

[6] C. May, A. Wang, S. Bordia, S. R. Bowman, R. Rudinger, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 622–628.

[7] W. Guo, A. Caliskan, 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, ACM, 2021.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.