Detecting Stereotyped Representations of Words within Language Models Embedding Space*

Michele Dusi[1,†]

[1]Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Via Branze 38, Brescia, Italy

Abstract

Today's widespread use of Natural Language Processing techniques rises the need for control mechanisms to prevent harmful behaviors in terms of safety and ethics. Many Language Models have been shown to learn a distorted representation of words and concepts, gathering such prejudiced information from the stereotypes of the training datasets.

In this paper, a new method is presented to detect whether a language model exhibits internal bias. The proposed method is based on the Cramér's V metric [1], which measures the correlation between two categorical variables. The method operates directly on the model's internal representation by analyzing its word embeddings.

Empirical results on gender and religion biases suggest that a cardinality of 50 words (for each class) is sufficient to obtain stable values, although even a dozen words per class can provide an acceptable estimate of the measurement.

1. Introduction

Scientific literature on AI Fairness has increased in recent years, as fairness began to be considered a requirement in system development and various methodologies have been developed to ensure its presence in AI models. Fairness has been defined in several ways, but to grasp the general idea, fairness is the "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [2].

In this paper, we address the problem of fairness in the field of Natural Language Processing (NLP): our analysis focuses on words and texts fed to models [3], with the purpose of understanding whether their processing can be seen as fair or unfair. More specifically, we rely on carefully selected datasets of words that clearly designate a human attribute, such as gender or religion. Our aim is to detect whether the representation of these attributes is somewhat biased within the Language Model's inner embedding space, i.e. whether the embedding of the words suggest an unwanted correlation with other attributes, such as job salary or criminality.

Approaching the study of bias by analyzing the relationship between two attributes is a standard procedure in the literature [3]; stereotypes are often defined as an undesirable association between human properties. For example, a stereotype could suggest the association between women (gender property) and a lower salary, or the association between Muslim people (religion property) and a stronger tendency to criminal behavior.

Our computational approach diverges from those described in existing literature [4, 5]. Specifically, we utilize access to a white-box model to characterize the distribution of embeddings associated with a primary attribute, namely, the protected property, and subsequently compare this distribution with that of embeddings corresponding to a secondary attribute, namely, the stereotyped property. The association is quantified by calculating a score within the interval $[0, 1]$ using the Cramér's V metric [1].

Doctoral Consortium at the 23rd International Conference of the Italian Association for Artificial Intelligence Bolzano, Italy, November 25-28, 2024.

‡Michele Dusi was enrolled in the Italian National PhD Program in Artificial Intelligence conducted by Sapienza, University of Rome, with the University of Brescia.

michele.dusi@unibs.it,michele.dusi@uniroma1.it (M. Dusi)

The methodology and preliminary findings of this ongoing research are outlined in this short paper; additional details are available in the full paper [6]. This work draws some of its initial insights from a prior study on bias visualization [7].

2. Background and Related Works

The seminal work that first highlighted fairness issues in natural language processing (NLP) was published in 2016 [8]. This study focused on evaluating and mitigating gender bias in early word embedding models, demonstrating the significant drawbacks of training these models on large text corpora without critical oversight.

The observation of biases in language models prompted a series of studies examining the geometry of the embedding space to assess whether the embeddings exhibit any undesired distributions [4, 9]. At first, the models considered were based on static word embeddings, like Word2vec, or GloVe. Over the following few years, the same approach was applied on contextual models [10], such as Transformers-based [11] and BERT-based [12, 13] models.

Our study focuses on the same contextual transformer-based models, which are among the most widely studied open-source language models in scientific literature. Compared to other techniques, our method requires fewer words to define the analyzed properties, thereby making it more practical and easier to apply.

For the definition of bias, we refer to a survey paper that presents a structured framework: in [3], the authors outline an ontology-based approach by defining bias at the semantic level. This approach characterizes bias as the (undesirable) correlation between two human properties, serving as the foundation for our bias detection technique.

3. Methodology

In this section, we outline our methodology for measuring the bias of a language model. As mentioned briefly earlier, the procedure involves accessing the inner embedding space of the model to examine and analyze the distribution of word vectors.

Encoding the properties. We start from the two properties involved in the bias we want to detect (e.g. the gender-jobs bias, or the religion-criminality bias). These are the two properties that need to be analyzed within the language model.

The first step is to define a word list for each value of the property, with the goal of collecting the terms used in the language to describe a specific value of a given property. For instance, the male class of the gender property can be represented by terms such as “he”, “him”, “father”, and “king”, while the female class of the same property can be represented by terms such as “she”, “her”, “mother”, and “queen”.

Each term is then converted into a vector, referred to as the word embedding, by the language model. Since we work with transformer-based models, the word embedding is context-dependent, meaning it varies based on the entire sentence in which the word is used. Therefore, each term appears in multiple sentences, and the final embedding is computed by averaging the embeddings from these sentences. The result of this pre-processing step is a list of vectors within the model’s embedding space, each encoding an average representation of the corresponding terms.

Learning the protected property. In the next step, the protected embeddings - that is, the embeddings of the terms associated with the protected property – are used to train an auxiliary classifier to distinguish between the different values of the protected property.

This step aims to identify how the language model encodes the protected classes (e.g., male and female) and addresses questions such as: which vector components are most relevant to encoding gender/religion? How are the protected classes distributed within the model’s embedding space?

religion × *adjectives*		*Predicted* values (protected)		Σ
		Christian	**Muslim**	
Actual values (stereotyped)	**positive**	59.2	60.8	120
	negative	46	74	120
		105.2	134.8	240

Table 1
Example of contingency matrix that shows the average distribution (over 10 testcases) of *religion* predicted labels for actual *adjectives* words.

Evaluating the stereotyped property. The stereotyped embeddings – that is, the embeddings of the terms of the stereotyped property — are then used to test the auxiliary classifier trained at the previous step. Each stereotyped embedding corresponds to a single value of the stereotyped property, but is also classified as one of the values of the protected property. As a result, each embedding in this test set is identified by a pair of values.

The expected outcome is random classification by the auxiliary classifier: the stereotyped embeddings should not contain any encoding of the protected property, so they could be classified into any of the protected values. However, if the model exhibits bias and the word embeddings are not neutral, we should observe a statistical shift in the classifiers predictions.

For example, in the case of religion bias, the terms "criminal" and "peaceful" should ideally be labeled as either Muslim or Christian independently, as they do not inherently carry any religious connotation. However, if bias is present, we may observe that crime-related terms are classified as Muslim and good-related terms as Christian.

Measuring the bias. The final step aims to measure the distortion in the classification and express it as a quantifiable metric. The predicted labels are collected and counted by class, resulting in an aggregate measure known as the contingency table. An example of this can be seen in Table 1.

A contingency table is a matrix where each row corresponds to a stereotyped class (e.g. "positive adjectives" and "negative adjectives" for criminal behavior), and each column corresponds to a protected class (e.g. "Christian" and "Muslim" for religion). The values in each cell represent the (average) number of terms belonging to the row-associated stereotyped class and labeled as the column-associated protected class. For instance, in Table 1, the 120 negative adjectives are split and classified into "Christian" (46) and "Muslim" (74).

As stated before, an unbalanced distribution may suggest a biased representation of such concepts; in the previous example, Muslim people are more likely to be associated with negative terms, with respect to christian people and positive terms. To compute the strength of this association, we use Cramér's V metric [1], which measures the correlation between two categorical variables. In our context, these variables correspond to the protected and stereotyped properties, and their values represent the classes into which the words are grouped.

Cramér's V is a metric normalized between 0 and 1, where 0 represents a situation of no correlation (i.e. the properties are independent and unrelated), whereas 1 represents a situation of maximal association between the properties (i.e. all the words of one stereotyped class have the same protected class).

To compute the score, we calculate the Minimum Square Error (MSE) between the observed distribution and the expected distribution; the observed distribution is simply the contingency matrix we gather from the auxiliary classifier, whereas the expected distribution assumes that the variables are independent. Afterwards, the MSE value is exploited to compute the Cramér's V metric score:

$$V = \sqrt{\frac{\text{MSE}}{n \cdot \min(|S| - 1, |P| - 1)}} \tag{1}$$

which normalizes the previous result in the interval $[0; 1]$. More specifically, the MSE score is

p_{prot}	p_{ster}	Language Models	
		BERT	**RoBERTa**
gender	*profession*	33.5 %	39.2 %
religion	*adjectives*	13.9 %	2.8 %

Table 2
Values of the **Cramér's V metric** over 100 testcases. Each row represents an experiment over two properties (protected and stereotyped).

divided by the total number of samples n and by the minimum between the degrees of freedom of the rows (number of stereotyped classes $|S|$ minus 1) and the degrees of freedom of the columns (number of protected classes $|P|$ minus 1).

We consider the resulting score as a measure of bias, quantifying the prejudiced correlation between two human categories. It is important to note that, due to the mathematical properties of Cramér's V metric, the outcome is an easily interpretable value that is unaffected by the size of the initial datasets and is applicable to multi-class properties [1].

4. Experimental Results

In this section, we present the results of our experiments on measuring the bias of language models. We evaluated the following two transformer-based models:

- BERT [12] in its base implementation (bert-base-uncased) by Hugging Face[1].
- RoBERTa [13], a more robust version of BERT, by Hugging Face[2].

The considered models are trained mainly for the English language.

We also tested two different kind of social bias: the gender bias, with respect to the stereotyped professions, and the religion bias, with respect to a positive or negative behavior (expressed by adjectives like "peaceful" or "aggressive").

The results are heterogeneous across different models, indicating that the language models exhibit varying amounts of bias. Table 2 summarizes the key findings of this experimental phase. We observe the highest bias values for the gender property (BERT 33.5 %, RoBERTa 39.2 %), while the religion property shows a relatively lighter presence of bias (BERT 13.9 %, RoBERTa 2.8 %).

When comparing the results from the model perspective, RoBERTa [13] demonstrates a significant difference between the scores for the two domains. This suggests that gender is strongly encoded and recognizable within RoBERTa's word embeddings, while the model does not exhibit any notable religion bias. In contrast, BERT [12] shows less disparity in its scores, implying that while biased behavior may emerge from the internal representation, its effects on the model's functions are relatively lighter.

Finally, we observe from the series of results that a cardinality of 50 words per class is sufficient to obtain stable values. However, even as few as a dozen words per class can provide an acceptable estimate of the measurement.

5. Conclusion and Future Works

In this paper, we presented a novel automatic method for detecting and measuring social bias within language models. The method requires minimal initial data, as it only necessitates the definition of two datasets corresponding to the properties being analyzed.

[1] https://huggingface.co/bert-base-uncased
[2] https://huggingface.co/docs/transformers/model_doc/roberta

Furthermore, our method operates directly on the model's internal representation by analyzing its word embeddings. This, however, also constitutes a limitation, as it requires access to a white-box model. Such an approach would not be feasible for other large language models (LLMs), which are often black-box models.

In the future, it would be valuable to expand this type of analysis in several directions. For example, new properties and additional classes could be considered. A common limitation in the technical literature on bias detection is the reduction of gender to only two classes (male and female). However, incorporating more possibilities, as supported by current psychological studies [14], could help address this issue. Our method is already capable of handling a larger number of classes; the challenge, in this case, would be to identify a sufficient number of terms that uniquely represent the newly introduced classes.

As noted earlier, this article is written in English, which is the standard language in scientific literature. However, different languages may carry distinct stereotypes and biases. For example, in some languages, gender is embedded in grammatical structures, which can affect the interpretation of certain terms. To fully address the nuances of language-specific biases, further research is required to adapt this method to other languages in a way that accounts for their unique characteristics.

References

[1] H. Cramér, Mathematical methods of statistics, Princeton: Princeton University Press, 1946.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021).

[3] 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.

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

[5] 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.

[6] M. Dusi, N. Arici, A. Emilio Gerevini, L. Putelli, I. Serina, Discrimination bias detection through categorical association in pre-trained language models, IEEE Access 12 (2024) 162651–162667. doi:10.1109/ACCESS.2024.3482010.

[7] M. Dusi, N. Arici, A. E. Gerevini, L. Putelli, I. Serina, Graphical identification of gender bias in bert with a weakly supervised approach, in: NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence, CEUR-WS, 2022. URL: http://sag.art.uniroma2.it/NL4AI/wp-content/uploads/2022/11/paper16.pdf.

[8] 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.

[9] 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, AIES '21, Association for Computing

Machinery, New York, NY, USA, 2021, p. 122133. URL: https://doi.org/10.1145/3461702.3462536. doi:10.1145/3461702.3462536.

[10] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, K.-W. Chang, Gender bias in contextualized 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. 629–634. URL: https://aclanthology.org/N19-1064. doi:10.18653/v1/N19-1064.

[11] 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.

[12] 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.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[14] C. Richards, W. P. Bouman, L. Seal, M. J. Barker, T. O. Nieder, G. TSjoen, Non-binary or genderqueer genders, International Review of Psychiatry 28 (2016) 95–102.