Graphical Identification of Gender Bias in BERT with a Weakly Supervised Approach

Michele Dusi[1], Nicola Arici[1], Alfonso E. Gerevini[1], Luca Putelli[1] and Ivan Serina[1]

[1]*Università degli Studi di Brescia, Brescia, Italy*

Abstract

Transformer-based algorithms such as BERT are typically trained with large corpora of documents, extracted directly from the Internet. As reported by several studies, these data can contain biases, stereotypes and other properties which are transferred also to the machine learning models, potentially leading them to a discriminatory behaviour which should be identified and corrected. A very intuitive technique for bias identification in NLP models is the visualization of word embeddings. Exploiting the concept of that a short distance between two word vectors means a semantic similarity between these two words; for instance, a closeness between the terms *nurse* and *woman* could be an indicator of gender bias in the model. These techniques however were designed for static word embedding algorithms such as Word2Vec. Instead, BERT does not guarantee the same relation between semantic similarity and short distance, making the visualization techniques more difficult to apply. In this work, we propose a weakly supervised approach, which only requires a list of gendered words that can be easily found in online lexical resources, for visualizing the gender bias present in the English base model of BERT. Our approach is based on a Linear Support Vector Classifier and Principal Component Analysis (PCA) and obtains better results with respect to standard PCA.

Keywords

Gender Bias, Ethics, Fairness, Model Interpretability, BERT

1. Introduction

With the affirmation of Artificial Intelligence in everyday user experience, many AI systems exhibited a behaviour that can be legally defined, when acted by humans, discriminatory. In particular, for Natural Language Processing (NLP) and Machine Learning techniques, the problem of algorithmic discrimination is mainly caused by prejudiced data involved in the learning process and it produces an uneven outcome for demographic minorities, i.e. subgroups of people differing by gender, race, religion, sexual orientation, disability, etc [2]. In order to solve this issue, at first we have to identify the presence of a bias, possibly with an intuitive technique that can be understood not only by AI experts but also from the people who simply will use the system.

In NLP systems, an intuitive way to assess bias is visualization. As can be seen in the simple example showed in Figure 1, more stereotypical masculine jobs such *engineer* or *mechanic* occupy a specific region of the two-dimensional space. Instead, stereotypical feminine jobs such

NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence, November 30, 2022, Udine, Italy [1]

m.dusi007@studenti.unibs.it (M. Dusi); nicola.arici@unibs.it (N. Arici); alfonso.gerevini@unibs.it (A. E. Gerevini); luca.putelli1@unibs.it (L. Putelli); ivan.serina@unibs.it (I. Serina)

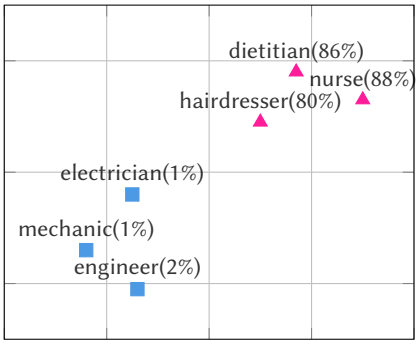

Figure 1: Example of biased word embedding model results. The female employment percentage coming from WinoGender dataset [6] is showed in brackets.

as *nurse* or *hairdresser* are grouped together in another region.

In the last few years, the state of the art for many NLP tasks has been profoundly changed by Transformer-based algorithms [3]. One of the most known of such models, BERT (Bidirectional Encoder Representations from Transformers) [4], is trained on a language modelling task into which the model has the goal of predicting words from context. In order to do that, exactly as in typical word embedding models such as Word2Vec [5], BERT represents words as vectors of real numbers. However, while the former produces a unique, static representation for each word, in BERT the representation can significantly differ depending on the entire context of the sentence or the document into which the word appears.

Considering static word embeddings, there are several studies for measuring and visualizing the gender bias [7, 8, 9] in a intuitive way, similarly to the example in Figure 1. Considering BERT, although several studies have shown its intrinsic bias and its different results for male and female subjects in NLP tasks [10, 11, 12, 13, 14], at the best of our knowledge these visualization techniques have not been applied yet.

In our opinion, this is mainly due to two factors. First, while Word2Vec vectors have a length between 100 and 300 [15], for BERT typically the length is 768, making it harder to apply dimensionality reduction techniques that allow an effective visualization. Second, it is proven that while static word embedding vectors have an isotropic distribution, BERT vectors occupy a narrow cone of the 768-dimensional space [16, 17] causing several issues with typical metrics for measuring the distance among vectors and for reducing their dimensions [18].

In this work, we focus on the problem of visualizing gender prejudice in human occupations, analysing the word representation produced by BERT. We propose a new weakly supervised method which requires a simple and minimal dataset for obtaining a graphical representation of biased word embeddings by reducing the BERT vector space to a two-dimensional plane showing the gender distortion. We use a linear Support Vector Classifier [19], trained on a simple list of gendered English words (e.g. *woman*, *man*, *sister*, *brother*, etc.), to decide which features are more involved in the gender definition. This training process does not require any time-consuming data collecting or labelling tasks, as these words can be easily found in many online lexical resources. Secondly, we apply the Principal Component Analysis (PCA) [20] in order to further reduce the word representation to a two-dimensional space which can

be visualized. With respect to standard PCA, we show that our approach produces a better visualization, providing an intuitive understanding of gender bias in human occupations, as captured by the classical Masked Language Model for bias identification [11] and by the actual statistics of gendered employment rate for several occupations.

2. Background and Related Work

In the last few years, several scientific papers [21, 22] have pointed out the presence of bias or discriminatory behavior in machine learning algorithms and in Natural Language Processing systems. The term *bias* is often used in reference to an algorithm to indicate a systematic distortion of outputs that produces unfair results, such as favoring or discriminating against certain groups of people [23]. For an algorithm, bias is something unwelcome, the absence of which is necessary to satisfy the *fairness* property. A statistical, and commonly adopted, definition of bias is presented in [24] and it can be summarized as the *distance between two conditional probabilities* $p(w_S|w_A)$ *and* $p(w_S|w_B)$ *that a word* w_S *denoting a stereotype* S *will appear in a sentence, given words* w_A *and* w_B *characterizing two distinct categories* A *and* B *of subjects*. In order to address this issue, the same work outlined a standard three-steps approach: (1) definition, (2) identification and (3) bias correction. The focus of the current work is visualization, therefore a part of the step 2.

Based on this and other similar definitions, several techniques for identifying the presence of bias in models have been developed over the years with variable effectiveness. For example, in 2017 Caliskan et al. [8] introduced the use of associative tests (WEAT and WEFAT) to estimate the closeness of a target word to two terminological reference groups. These tests, in addition to depending heavily on the choice of reference terms, are specifically designed for *static* word embedding algorithms, that encode each word uniquely, regardless of the sentence into which it is embedded.

Another interesting work is the study by Bolukbasi et al. [7], which detected discriminatory behaviour in the worldwide known algorithm Word2vec [5]. Their identification process is done through the study of the geometric distribution of static embeddings. Similar approaches were presented in Zhou et al. [9] which focuses on bilingual spaces, and in Maudslay et al. [25] which exploits clustering algorithms for identifying stereotyping classes. The basic assumption overall these approaches is that a geometric proximity corresponds to a semantic similarity of the two terms.

However, with the advent of more complex deep learning architectures, such as Transformer [3] or BERT [4], and their vectorial representation of words, such assumption has been challenged. In fact, several studies [16, 17, 18] claim that in BERT a high cosine similarity between two words (i.e. a very high closeness in the vectorial space) do not necessarily correspond to a semantic similarity. Moreover, in BERT there is no unique correspondence between a word and its embedding, therefore the bias should be measured by placing the word in a pre-defined artificial context or template, like in CEAT [12] or SEAT [26]. Consequently, these issues make the aforementioned visualization methods, such as the ones proposed in [7, 9], not suitable for BERT embeddings as the ones we analyse in this work.

Thus, for BERT a typical approach to bias identification is not based on visualization. Instead,

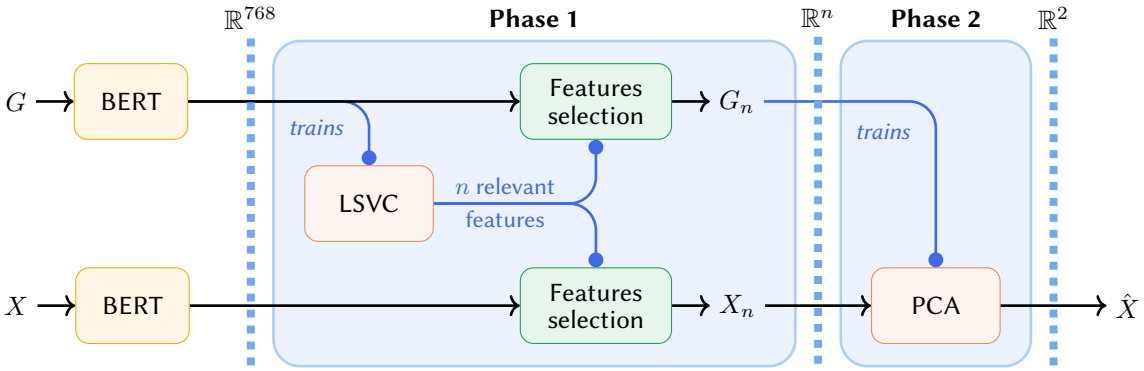

Figure 2: Architecture schema of WSV. After being processed by BERT, the G and X sets are subjected to the features selection; the n most relevant features related to the gender information are identified by the LVSC model trained over the dataset G of gendered words. In the second phase, the PCA model is trained over G_n and applied to X_n, obtaining the final resulting two-dimensional vector $\hat{X} \subseteq \mathbb{R}^2$.

it focuses on evaluating the behaviour of the so-called Masked Language Modeling (MLM) task [11, 14]. For instance, in the sentence *"[MASK] is a programmer"*, the model could predict both the pronouns *he* and *she* and form a correct sentence. However, if the model predicts *he* with a significantly greater probability than the one associated with the prediction of *she*, the model presents a gender bias for the word *programmer*. However, this approach based on language modeling is generally less intuitive with respect to the word embedding visualization (like the one we propose in this work), which can be understood by a glance also by people who are not expert in NLP architectures. Nonetheless, as we show in Section 4, graphical and language modeling methods can be complementary and the intuition provided by the former can be confirmed, more quantitatively, by the latter.

3. Methodology for Weakly Supervised Visualization

Our objective is to visualize the distribution of BERT-encoded words, i.e. to reduce high dimensional vectors to a 2D space, in order to highlight the gender spectrum. More specifically, we use the base version of BERT for the English language[1] which has an embedding space of length 768. We do this in two steps: we first select the n most relevant features for gender, and secondly we reduce them while preserving their variance. This procedure, called Weakly Supervised Visualization (WSV) is showed in Figure 2.

We apply this method to the 1678 single-word job titles appearing in the JNeidel dataset[2]. Given a word x describing an occupation, we exploit BERT for calculating an embedding representation $\bar{x}$. However, given that this representation depends also on the overall context of the sentence into which x appears, we use a basic template which will form the context of every word in the dataset. This template is composed by three tokens: the special token [CLS], which

[1] https://huggingface.co/bert-base-uncased
[2] https://github.com/jneidel/job-titles

is used by BERT to calculate an entire representation of the sentence, x and [SEP], which marks the end of the sequence. Although BERT will produce a vector for each token in this simple sequence, we extract only the one representing x, which is composed by 768 real numbers.

In order to measure potential bias of the embedded representation of job titles, we need also to encode gendered words, i.e. relationships, titles, pronouns and other expressions clearly indicating the gender of the subject such as *mother*, *brother*, *grandpa*, *girlfriend*, *he*, *him*, *she*, *her*, *sir*, *madam*, *queen*, *king*, etc. This list consists in 102 terms coming from the WinoBias dataset [27], 51 for each considered gender. In order to obtain a vector representation of these words, we apply the same technique used for the job titles. Once the gendered words (denoted as G) and the occupation words (denoted as X) are encoded, we begin the dimensionality reduction.

The first phase exploits a Support Vector Classifier [19] with a linear kernel (LSVC). The model is trained over the G dataset labeled with a binary classification value representing the word gender[3]. The LSVC model then learns a separator hyperplane in $\mathbb{R}^{768}$: $\mathbf{w}\bar{x} - b = 0$. From the vector $\mathbf{w}$ of weights, we extract the n higher absolute values, deducing the most relevant dimensions of the vector for the gender identification. After this procedure, the remaining $768 - n$ features are cut off from the vectors.

This *features selection* procedure is applied both to G and to X, obtaining respectively $X_n \subseteq \mathbb{R}^n$ and $G_n \subseteq \mathbb{R}^n$, i.e. two compressed representations which would contain most of the gender-related information. Please note that, in order to train the LSVC, we need only a small list of about 100 gendered words which can be easily found online, as in the WinoBias dataset or in lexical databases such as WordNet. No actual labelling process or data collecting is required, therefore we claim that our approach is weakly supervised.

The second phase consists in the Principal Component Analysis (PCA) [20] of G_n. PCA defines the optimal linear transformation $T : \mathbb{R}^n \mapsto \mathbb{R}^2$ such that the maximum percentage of variance in the starting samples is preserved. Thus, applying the same T to X_n, we get $\hat{X} \subseteq \mathbb{R}^2$ summarising the gender bias in the evaluation set X. Given that we chose the most important dimensions of the vector related to the gender prediction, the variance preserved by the PCA in X_n should capture whether there is a gender distortion in the job titles.

The WSV implementation requires to set a single hyperparameter n, namely, the vectors dimension extracted in the first phase:

$$\mathcal{P} : \mathbb{R}^{768} \xmapsto{\text{LSVC}} \mathbb{R}^n \xmapsto{\text{PCA}} \mathbb{R}^2$$

As we describe in Section 4, the best choice of n are from middle values, while very low (such as $n = 2$) or high ($n > 300$) values do not obtain satisfying results.

In general, we can see our approach as a linear transformation from a high dimensional vector space to a two-dimensional plane. This requires just a single training of a simple classifier. In our case, we have trained a LSVC under the hypothesis (later confirmed by the results, as we show in Section 4) that the gendered words are linearly separable. However, in our opinion this is not necessary, and different kernels or other models (such as XGBoost or Feed-Forward Neural networks) can be used alongside with a technique for extracting the most important features (e.g. SHAP [28]) and then applying the PCA. In all these cases, after than the classifier has learned which dimensions are mostly related to the gender, this information can be exploited

[3]For now, we improperly simplify the social perception of gender by considering only the male and female classes.

[MASK] works as a [JOB].	[MASK] should be [JOB] soon.
[MASK] worked as a [JOB].	[MASK] has studied for years to become a [JOB].
[MASK] was a [JOB].	One day [MASK] will be a [JOB].
[MASK] will soon be a [JOB].	From tomorrow, [MASK]'s going to work as a [JOB]
[MASK] has a job as [JOB].	[MASK] is studying to be a [JOB].
[MASK] is a [JOB].	[MASK] has always wanted to become a [JOB].

Table 1
The templates used for the MLM task. BERT will measure the probabilities of the [MASK] token being "he" and "she", while the [JOB] token will be replaced by the inspected occupation.

for any set of words. Moreover, this whole process could be easily generalized to different types of bias, as long as we provide a labeled set of training samples. As future work, we will perform a more in-depth study considering different classifiers, sets of words and types of bias.

4. Experimental Results

In this section we present the results achieved by the proposed model, comparing it to the bare use of the Principal Component Analysis (PCA). Our Linear Support Vector Classifier has been trained using the default hyperparameters in the implementation provided by the Scikit-learn[4] library [29] and with the *hinge squared* loss function.

Considering the 1678 jobs in the JNeidel dataset, first we calculated the gender distortion with the standard Masked Language Modeling (MLM) technique: we chose 12 templates regarding the career domain (Table 1) and we measure for each job the difference between the probabilities of "[MASK]" being *"she"* or *"he"*. With this procedure, we obtain a score between -1, that indicates an only man profession, and $+1$, that indicates an only female profession. From now on, we will refer to this score as the *MLM score*. The MLM score is incorporated in our visualization as follows. In Figure 3, each point represents a different occupation, obtained by only applying PCA (on top) and by our method (on the bottom) selecting $n = 50$. The colour of each point indicates the MLM score: jobs with a MLM score very close to -1 are represented as a blue point, while those ones very close to $+1$ are represented as a magenta point.

As it can be seen in the first chart in Figure 3, the PCA results have no spatial relations with the bias detected by the MLM score. In fact, the pink points are sparse across all the space, without any noticeable logic. On the other hand, samples in the second chart are placed according to the stereotypical gender with most of the pink and magenta points on the left region of the space, highlighting the gender spectrum of prejudiced jobs. While for PCA reducing vectors in 2D is not enough to intuitively show a bias, being only sufficient to show the distortion of the most extreme samples, our approach visualizes a more evident trend. For instance, jobs like *nurse* or *hostess* are strongly related to the female gender and occupies the left region of the two-dimensional space, while typically masculine jobs like *priest* or *infantryman* are in the opposite region.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

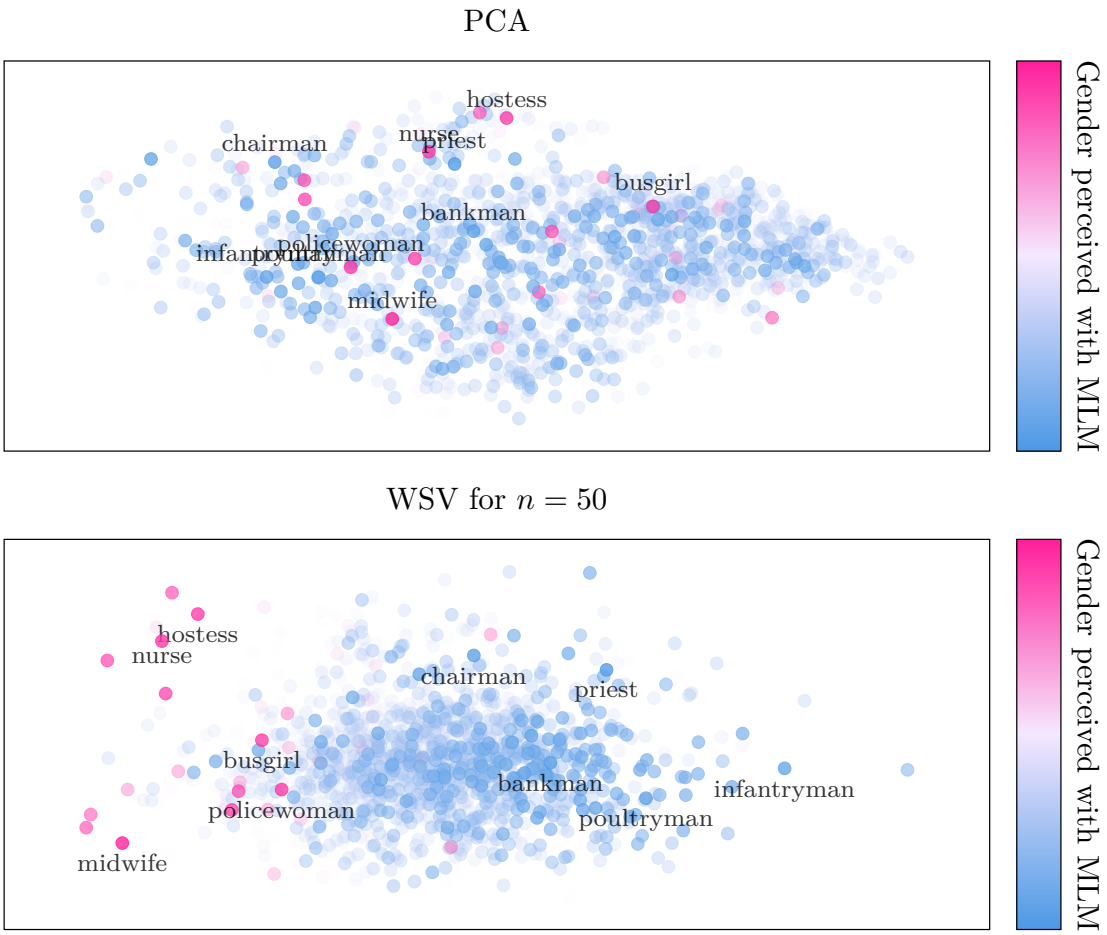

Figure 3: Comparison between the visualization of 1678 occupations with PCA (first two principal component) and with our method, for $n = 50$. The first component is displayed on the horizontal axis while the second component on the vertical axis. The colour indicates the bias towards the male (blue) or female (magenta) stereotype, detected with MLM; jobs perceived as neutral are lowered in opacity. The five most biased samples for both male and female genders are labeled in the chart.

In order to provide a more quantitative indication of the ability of PCA and WSV (selecting different values of n) to show the bias, we have calculated the Pearson Correlation Coefficient, in absolute value, among the two components extracted by both methods and the MLM score. We present the results in Table 2. Results in terms of correlation for PCA are very low (0.09 for the first component, 0.05 for the second one), demonstrating that this method is not able to represent the gender bias in a two-dimensional space. However, simply extracting the two most relevant features ($n = 2$) without applying the PCA algorithm does not provide satisfying results. Instead, combining the selection of a relatively small number of features with the PCA provide the best results. In fact, the correlation between the first component (plotted

Value of n	**2**	**50**	**100**	**300**	**768**
1st comp.	0.04	0.42	0.39	0.04	0.09
2nd comp.	0.08	0.10	0.13	0.12	0.05

Table 2
Absolute value of the Pearson Correlation Coefficient between the score provided by the MLM method and the two components extracted by WSV, with different values of n. With $n = 2$, we simply select the first two features provided by the classifier, without applying PCA; for $n = 768$ (last column) the WSV method equals PCA.

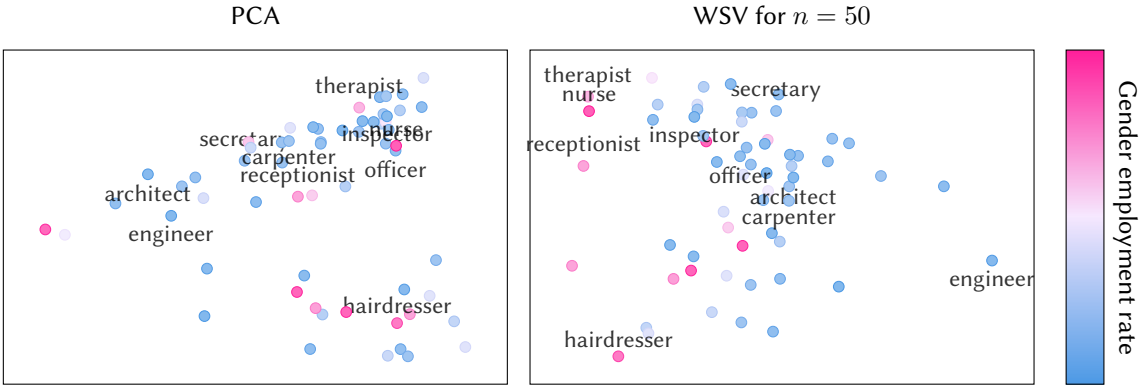

Figure 4: Plots of the WinoGender occupations [6] for $n = 50$. The colour indicates the gender employment rate from the WinoBias dataset [27]. The ten jobs with the highest gender disparity (five for each) are labeled in the charts.

as the horizontal axis in Figure 3) and the MLM score with $n = 50$ is 0.42, confirming the intuition provided by the plot. In general, the choice of the hyperparameter n, namely the number of elements extracted from the BERT embeddings is a fundamental vector for evaluating the efficacy of WSV. In fact, a value too low could cut off important gender information, but a value too high would make the selection useless and admit a lot of unrelated information. Experimental tests (like the ones reported in Table 2) showed us that values between 20 and 100 are usually good options, however these results may depend also on the dataset considered.

In Figure 4, we performed the same experiment but considering the WinoGender dataset [6], which is made by 60 occupations and their gender employment rate. In this case, the samples are not coloured on the basis of the MLM score; instead, pink points indicate occupations into which the female workers are the vast majority, while blue points represent jobs typically done by men. For Figure 3, the difference between the left chart (visualization using only PCA) and the right chart (WSV) suggests that our method is valid also in this case, showing more pink and violet points on the left of the two-dimensional space, while using only PCA no clear pattern can be identified.

Considering the WinoGender dataset, we have evaluated how PCA and WSV are correlated with the real world statistics for gender employment rate. While the most correlated component of standard PCA has a Pearson Coefficient of 0.24, our approach reaches 0.56. Given that, for

the same dataset, the MLM score has a correlation with the gender employment rate of 0.59, this result is particularly important. In fact, we are able to produce a visual plot which has a very similar correlation to the one obtained by the standard technique for bias identification [11].

5. Conclusion and Future Work

We presented a new weakly supervised graphical approach to identify the gender bias in a BERT model. The results indicate that our method gives better representations of prejudiced gender than the standard PCA approach, offering the possibility to easily grasp distortion. When used on real world data, analysing the relation between the extracted bias and the gender employment rate, it provides comparable results with respect to the commonly used Masked Language Model method. An import characteristic of our work is that we propose an algorithm which only needs a list of common gendered words for training, without any expensive data collecting or labelling processes.

As future work, the first thing to do will be to apply WSV to other types of bias, such as ethnicity or religion. This can be done by identifying a set of words related to the bias in question to quickly train our proposed weakly supervised system. However, the number of training words could vary depending on the kind of bias and on the difficulty of the identification process. In adapting our approach to other fields and datasets, another fundamental aspect which has to be considered is the tuning of the hyperparameter n, which could require several trials. Another possible development will be to apply the technique to models which analyse the Italian language. Since this language, like Spanish or French, has a rich morphology for nouns and adjectives and often has terms directly specifying the gender (such as *studente* or *studentessa*, which identify respectively a male and a female student), we may need to modify our bias detection strategy. Finally, we will try to apply and to adapt our method to newer and more complex NLP models, such as GPT.

References

[1] D. Nozza, L. Passaro, M. Polignano, Preface to the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), November 30, 2022, CEUR-WS.org, 2022.

[2] D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing, Lang. Linguistics Compass 15 (2021).

[3] 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.

[4] 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 Computa-

tional 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.

[5] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

[6] R. Rudinger, J. Naradowsky, B. Leonard, B. Van Durme, Gender bias in coreference resolution, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, New Orleans, Louisiana, 2018.

[7] 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.

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

[9] 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.

[10] G. Stanovsky, N. A. Smith, L. Zettlemoyer, Evaluating gender bias in machine translation, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 1679–1684.

[11] K. Kurita, N. Vyas, A. Pareek, A. W. Black, Y. Tsvetkov, Measuring bias in contextualized word representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 166–172.

[12] 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.

[13] A. Lauscher, T. Lüken, G. Glavas, Sustainable modular debiasing of language models, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 4782–4797.

[14] M. Bartl, M. Nissim, A. Gatt, Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias, in: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, arXiv, 2020, pp. 1–16.

[15] K. Patel, P. Bhattacharyya, Towards lower bounds on number of dimensions for word

embeddings, in: G. Kondrak, T. Watanabe (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers, Asian Federation of Natural Language Processing, 2017, pp. 31–36.

[16] K. Ethayarajh, How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 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 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 55–65.

[17] S. Rajaee, M. T. Pilehvar, An isotropy analysis in the multilingual BERT embedding space, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 1309–1316.

[18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 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. 3980–3990.

[19] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[20] K. P. F.R.S., Liii. on lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (1901) 559–572.

[21] L. Sweeney, Discrimination in online ad delivery, Commun. ACM 56 (2013) 44–54.

[22] I. Zliobaite, A survey on measuring indirect discrimination in machine learning, CoRR abs/1511.00148 (2015).

[23] D. Rozado, Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types, PLOS ONE 15 (2020) 1–26.

[24] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, L. A. Ureña-López, A survey on bias in deep nlp, Applied Sciences 11 (2021). doi:10.3390/app11073184.

[25] R. H. Maudslay, H. Gonen, R. Cotterell, S. Teufel, 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 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 5266–5274.

[26] 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.

[27] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-

guage Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 15–20.

[28] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.