Bias Score: Estimating Gender Bias in Sentence Representations

(Discussion Paper)

Fabio Azzalini[1,2], Tommaso Dolci[1] and Mara Tanelli[1]

[1]*Politecnico di Milano – Dipartimento di Elettronica, Informazione e Bioingegneria*
[2]*Human Technopole – Center for Analysis, Decisions and Society*

Abstract

The ever-increasing number of applications based on semantic text analysis is making natural language understanding a fundamental task. Language models are used for a variety of tasks, such as parsing CVs or improving web search results. At the same time, concern is growing around embedding-based language models, which often exhibit social bias and lack of transparency, despite their popularity and widespread use. Word embeddings in particular exhibit a large amount of gender bias, and they have been shown to reflect social stereotypes. Recently, sentence embeddings have been introduced as a novel and powerful technique to represent entire sentences as vectors. However, traditional methods for estimating gender bias cannot be applied to sentence representations, because gender-neutral entities cannot be easily identified and listed. We propose a new metric to estimate gender bias in sentence embeddings, named bias score. Our solution, leveraging the semantic importance of individual words and previous research on gender bias in word embeddings, is able to discern between correct and biased gender information at sentence level. Experiments on a real-world dataset demonstrates that our novel metric identifies gender stereotyped sentences.

Keywords

Gender bias, natural language processing, computer ethics

1. Introduction

Language models are used for a variety of downstream applications, such as CV parsing for a job position, or detecting sexist comments on social networks. Recently, a big step forward in the field of natural language processing (NLP) was the introduction of language models based on word embeddings, i.e. representations of words as vectors in a multi-dimensional space. These models translate the semantics of words into geometric properties, so that terms with similar meanings tend to have their vectors close to each other, and the difference between two embeddings represents the relationship between their respective words [1]. For instance, it is possible to retrieve the analogy $man : king = woman : queen$ because the difference vectors $\overrightarrow{queen} - \overrightarrow{king}$ and $\overrightarrow{woman} - \overrightarrow{man}$ share approximately the same direction.

Word embeddings boosted results in many NLP tasks, like sentiment analysis and question answering. However, despite the growing hype around them, these models have been shown

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

fabio.azzalini@polimi.it (F. Azzalini); tommaso.dolci@polimi.it (T. Dolci); mara.tanelli@polimi.it (M. Tanelli)
0000-0003-0631-2120 (F. Azzalini); 0000-0002-7172-0203 (M. Tanelli)

to reflect the stereotypes of the Western society, even when the training phase is performed over text corpora written by professionals, such as news articles. For instance, they return sexist analogies like $man : programmer = woman : homemaker$ [2]. The social bias in the geometry of the model is reflected in downstream applications like web search or CV parsing. In turn, this phenomenon favours prejudice towards social categories already frequently penalised, such as women or African Americans.

Lately, sentence embeddings – vector representations of sentences based on word embeddings – are also increasing in popularity, improving results in many language understanding tasks, such as semantic similarity or sentiment prediction [3, 4]. Therefore, it is of the utmost importance to expand the research to understand how language models perceive the semantics of natural language when computing the respective embedding. A very interesting step in this direction is to define metrics to estimate social bias in sentence embeddings.

This work expands research on social bias in embedding-based models, focusing on gender bias in sentence representations. We propose a method to estimate gender bias in sentence embeddings and perform our experiments on InferSent, a sentence encoder designed by Facebook AI [3] based on GloVe[1], a very popular word embedding model. Our solution, named *bias score*, is highly flexible and can be adapted to both different kinds of social bias and different language models. Bias score will help in researching procedures like debiasing embeddings, that require to identify biased embeddings and therefore to estimate the amount of bias they contain [2]. Similarly, techniques for improving training datasets also require to evaluate all sentences contained in them, to identify problematic entries to remove, change, or compensate for.

2. State of the Art

Although language models are successfully used in a variety of applications, bias and fairness in NLP have received relatively little consideration until recent times, running the risk of favouring prejudice and strengthening stereotypes [5].

2.1. Bias in Word Embeddings

Static word embeddings were the first to be analysed. In 2016, they have been shown to exhibit the so-called *gender bias*, defined as the cosine of the angle between the word embedding of a gender-neutral word, and a one-dimensional subspace representing gender [2]. The approach was later adapted for non-binary social biases such as racial and religious bias [6]. A *debiasing* algorithm was also proposed to mitigate gender bias in word embeddings [2], however it was also shown that it fails to entirely capture and remove it [7]. The Word Embedding Association Test (WEAT) [8] was created to measure bias in word embeddings following the pattern of the implicit-association test for humans. WEAT demonstrated the presence of harmful associations in GloVe and word2vec[2] embeddings. More recently, contextualised word embeddings like BERT [9] proved to be very accurate language models. However, despite literature suggesting

[1] https://nlp.stanford.edu/projects/glove/

[2] https://code.google.com/archive/p/word2vec/

that they are generally less biased compared to their static counterparts [10], they still display a significant amount of social bias [11].

2.2. Bias in Sentence Representations

Research is quite lacking regarding sentence-level representations. WEAT was extended to measure bias in sentence embedding encoders: the Sentence Encoder Association Test (SEAT) is again based on the evaluation of implicit associations and showed that modern sentence embeddings also exhibit social bias [11]. Attempts at debiasing sentence embeddings faced the issue of not being able to recognise neutral sentences, thus debiasing every representation regardless of the gender attributes in the original natural language sentence [12].

3. Methodology

As already mentioned, gender bias in word embeddings is estimated using the cosine similarity between word vectors and a gender direction identified in the vector space [2]. Cosine similarity is a popular metric to compute the semantic similarity of words based on the angle between their embedding vectors. Given two word vectors $\vec{u}$ and $\vec{v}$, cosine similarity is expressed as:

$$\cos(\vec{u},\vec{v}) = \cos(\theta) = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\,\|\vec{v}\|} ,$$

where θ is the angle between $\vec{u}$ and $\vec{v}$. The more $\cos(\theta)$ approaches 1, the higher is the semantic similarity between $\vec{u}$ and $\vec{v}$. In word embedding models, similarity with respect to the gender direction means that a word vector contains information about gender. Since only gender-neutral words can be biased, gendered words like *man* or *woman* are assumed to contain correct gender information.

When it comes to sentence representations, the main problem is that gender-neutral sentences cannot be easily identified and listed like words, because sentences are infinite in number. Moreover, they may contain gender bias despite their being gendered. Consider the sentence *my mother is a nurse*: the word *mother* contains correct gender semantics, but the word *nurse* is female stereotyped. Table 1 shows that the gender-neutral sentence *someone is a nurse* still contains a lot of gender information due to the bias associated with the word *nurse*.

Therefore, it is important to distinguish between the amount of encoded gender information *coming from gendered words*, and the amount *coming from biased words.* For this reason, we adopt a more dynamic approach: we keep working at the word level, using the cosine similarity between neutral word representations, and the gender direction to estimate word-level gender bias. Then, we sum the bias of all the words in the sentence, adjusted according to the length of the sentence and to the contextualised importance of each word. This decision is grounded on two observations: first, the semantics of a sentence depends largely on the semantics of the words contained in it; second, sentence embedding encoders are based on previously defined word embedding models [3, 4]. We focus our research on InferSent by Facebook AI [3], a sentence encoder that achieved great results in many different downstream tasks [13]. InferSent encodes sentence representations starting from GloVe [14] word embeddings. Therefore, we use GloVe for the first step of quantifying gender bias at the word level.

Table 1
Gender information with cosine similarity for sentence embeddings by InferSent [3] and SBERT [4].

input	*InferSent*	*SBERT*	*neutral?*	*biased?*
my mother is there	0.28877	0.46506	no	no
my mother is a nurse	0.29852	0.46018	no	yes
someone is a nurse	0.18175	0.43965	yes	yes

3.1. Gender Bias Estimation

To estimate gender bias in sentence representations, we consider four elements:

- $\cos(\vec{x}, \vec{y})$: cosine similarity between two word vectors $\vec{x}$ and $\vec{y}$,
- D: gender direction identified in the vector space,
- L: list of gendered words in English,
- I_w: a percentage estimating the semantic importance of a word in the sentence.

Our metric, named *bias score*, takes a sentence as input, and returns two indicators corresponding to the amount of female and male bias at sentence level. Respectively, they are a positive and a negative value, obtained from the sum of the gender bias of all words, estimated from cosine similarity with respect to the gender direction. Since gender bias is a characteristic of gender-neutral words, gendered terms are excluded from the computation and instead their bias is always set to zero. In detail, for each neutral word w in the sentence we compute its gender bias as the cosine similarity between its word vector emb_w and the gender direction D, and then we multiply it by the word importance I_w. In particular, for a given sentence s:

$$BiasScore_F(s) = \sum_{\substack{w \in s \\ w \notin L}} \underbrace{\cos(emb_w, D)}_{>0} \times I_w$$

$$BiasScore_M(s) = \sum_{\substack{w \in s \\ w \notin L}} \underbrace{\cos(emb_w, D)}_{<0} \times I_w$$

Notice that, for each word w that is gender-neutral, $w \notin L$. Also, word importance I_w is always a positive number, and the cosine similarity can be either positive or negative. Therefore, bias score keeps the estimations of gender bias towards the male and female directions separated. In the following sections, we go into more detail by illustrating how we derive D, L and I_w.

3.2. Gender Direction

The first step of our method is to identify in the vector space a single dimension comprising the majority of the gender semantics of the model. The resulting dimension, named *gender direction*, serves as the first term in the cosine similarity function, to establish the amount of gender semantics encoded in a vector for a given word, according to the model.

GloVe [14], the word embedding model that we use, is characterised by a vector space of 300 dimensions. Inside the vector space, the difference between two embeddings returns the direction that connects them. In the case of the embeddings $\overrightarrow{she}$ and $\overrightarrow{he}$, their difference vector $\overrightarrow{she} - \overrightarrow{he}$

represents a one-dimensional subspace that identifies gender in GloVe. However, also the difference vector $\overrightarrow{woman} - \overrightarrow{man}$ identifies gender, yet it represents a slightly different subspace compared to $\overrightarrow{she} - \overrightarrow{he}$. Therefore, following the approach in [2], we take into consideration several pairs of gendered words and perform a Principal Component Analysis (PCA) to reduce the dimensionality. We use ten pairs of gendered words: *woman–man*, *girl–boy*, *she–he*, *mother–father*, *daughter–son*, *gal–guy*, *female–male*, *her–his*, *herself–himself*, *Mary–John*.

As shown in Fig. 1, the top component resulting from the analysis is significantly more important than the other components, explaining almost 60% of the variance. We use this top component as gender direction, and we observe that embeddings of female words have a positive cosine with respect to it, whereas for male words we have a negative cosine.

3.3. Gendered Words

A list L of gendered words is fundamental to estimate gender bias, because only gender-neutral entities can be biased. Since the number of elements in the subset $\mathcal{N}$ of gender-neutral words in the vocabulary of a language is very big, while the subset $\mathcal{G}$ of gendered words is relatively small (especially in the case of the English language), we derive $\mathcal{N}$ as the difference between the complete vocabulary of the language $\mathcal{V}$ and the subset $\mathcal{G}$ of gendered words: $\mathcal{N} = \mathcal{V} \setminus \mathcal{G}$. To achieve this, we define a list L of words containing as many of the elements of the subset $\mathcal{G}$ as possible. Therefore, gender bias is estimated for all elements w_n in the subset $\mathcal{N}$ (neutral words), whereas for all elements w_g in the subset $\mathcal{G}$ (gendered words) the gender bias is always set to zero:

$$\forall\, w_n \in \mathcal{N},\ bias(w_n) \neq 0$$

$$\forall\, w_g \in \mathcal{G},\ bias(w_g) = 0$$

For this reason, all the elements from L are not considered when estimating gender bias. As a matter of fact, we consider the gender information encoded in their word embeddings to be always correct. Examples of gendered words include *he*, *she*, *sister*, *girl*, *father*, *man*.

Our list L contains a total of 6562 gendered nouns, of which 409 and 388 are respectively lower-cased and capitalised common nouns taken from [2] and [15]. Additionally, we added 5765 unique given names taken from Social Security card applications in the United States[3].

3.4. Word Importance

Following the approach in [3], word importance is estimated based on the max-pooling operation performed by the sentence encoder (in our case InferSent) using all vectors representing the words in a given sentence. The procedure counts how many times each word representation is selected by the sentence encoder during the max-pooling phase. In particular, we count the number of times that the max-pooling procedure selects the hidden state h_t, for each time step t in the neural network underlying the language model, with $t \in [0, \ldots, T]$ and T equal to the number of words in the sentence. Note that h_t can be seen as a sentence representation centred on the word w_t, i.e. the word at position t in the sentence.

[3] https://www.kaggle.com/datagov/usa-names

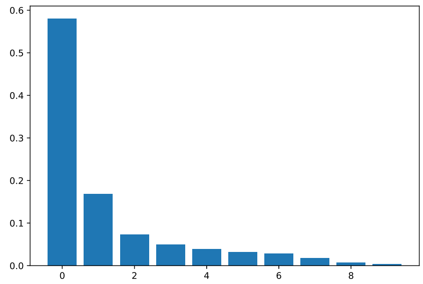

Figure 1: Top ten components in PCA.

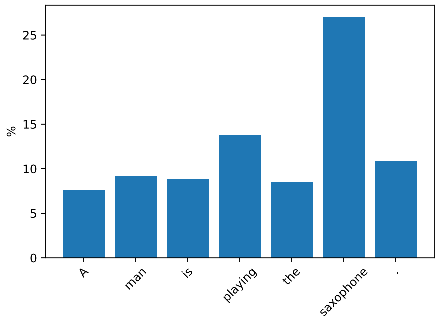

Figure 2: Example of word importance.

We consider both the absolute importance of each word, and the percentage with respect to the total absolute importance of all the words in the sentence. For instance, in the example of Fig. 2, the absolute importance of the word *saxophone* is 1106, meaning that its vector representation is selected by the max-pooling procedure for 1106 dimensions out of the total 4096 dimensions of the sentence embeddings computed by InferSent. The percentage importance is $\frac{1106}{4096} \approx 0.27$, meaning that the word counts for around 27% of the semantics of the sentence. In particular, the percentage importance is also independent of the length of the sentence, despite the fact that very long sentences generally have a more distributed semantics. For this reason, we use the percentage importance to compute bias score.

3.5. Variant

Bias score enables to discern gender bias towards the female and male directions. However, we can also take the absolute value of each word-level bias to derive a single estimation of gender bias at sentence level:

$$Abs\text{-}BiasScore(s) = \sum_{\substack{w \in s \\ w \notin L}} \mid \underbrace{\cos(emb_w, D) \times I_w}_{\textit{word-level bias}} \mid$$

4. Experimental Results

Table 2 illustrate an example of gender bias estimation via bias score, showing that gender stereotyped concepts like wearing pink dresses are heavily internalised in the final sentence representation. Additionally, we used bias score to estimate gender bias for sentences contained in the Stanford Natural Language Inference (SNLI) corpus, a large collection of human-written English sentences for classification training [16]. SNLI contains more than 570k pairs of sentences, and more than 600k unique sentences in the train set alone. According to our experiments, sentences corresponding to the highest bias score towards the male direction describe situations from popular sports like baseball and football, that are frequently associated with men and very seldom with women. Similarly, sentences corresponding to the highest bias score in the female direction illustrate female stereotypes, like participating in beauty pageants, applying make-up or working as a nurse. Table 3 displays the most-biased SNLI sentences according to our metric. Results are similar when estimating the absolute bias score. In particular, entries associated to

Table 2
Detailed bias score estimation for the sentence *She likes the pink dress.*

word	*importance*	*gender bias*	*weighted bias*
She	12.13%	0.00000	0.00000
likes	17.48%	-0.05719	-0.01000
the	8.35%	-0.10195	-0.00851
new	14.70%	-0.00051	-0.00008
pink	12.84%	0.25705	0.03301
dress	14.87%	0.28579	0.04249
overall female bias			0.07550
overall male bias			-0.01858

Table 3
Highest bias scores for sentences in SNLI train set, towards the female and male directions.

sentence	*bias score*
Beauty pageant wearing black clothing	0.134793
Middle-aged blonde hula hooping.	0.127903
A blonde child is wearing a pink bikini.	0.125145
A showgirl is applying makeup.	0.123312
Football players scoring touchdowns	-0.149844
Football players playing defense.	-0.140169
A defensive player almost intercepted the football from the quarterback.	-0.139420
Baseball players	-0.138058

the highest absolute bias score include sentences with a high bias score in either the female or male direction, like *football players scoring touchdowns* or *the bikini is pink.* Additionally, sexualised sentences like *the pregnant sexy volleyball player is hitting the ball* are also present.

5. Conclusion and Future Work

In this paper we proposed an algorithm to estimate gender bias in sentence embeddings, based on a novel metric named *bias score.* We discern between gender bias and correct gender information encoded in a sentence embedding, and weigh bias on the basis of semantic importance of each word. We tested our solution on InferSent [3], searching for gender biased representations in a corpus of natural language sentences. Since gender bias has been proven to be caused by the internalisation of gender stereotypical associations [16], our algorithm for estimating bias score allows to identify whose vector representation encapsulates stereotypes the most.

Future work will include adapting the proposed solution to different language models and different kinds of social bias. Additionally, since bias score allows to identify stereotypical entries in natural language corpora used for training language models, removing or substituting such entries may improve the fairness of the corpus. Thus, future work also includes re-training language models using text corpora made fairer with such procedure, and a comparison with the original models both from the quality and the fairness point of view.

Acknowledgments

We are grateful to our mentor Letizia Tanca for her advice in the writing phase of this work.

References

[1] T. Mikolov, W.-T. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: HLT-NAACL, 2013.

[2] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, arXiv preprint arXiv:1607.06520 (2016).

[3] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, arXiv preprint arXiv:1705.02364 (2017).

[4] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[5] K.-W. Chang, V. Prabhakaran, V. Ordonez, Bias and fairness in natural language processing, in: EMNLP-IJCNLP: Tutorial Abstracts, 2019.

[6] T. Manzini, Y. C. Lim, Y. Tsvetkov, A. W. Black, Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings, arXiv preprint arXiv:1904.04047 (2019).

[7] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, arXiv preprint arXiv:1903.03862 (2019).

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

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[10] C. Basta, M. R. Costa-Jussà, N. Casas, Evaluating the underlying gender bias in contextualized word embeddings, arXiv preprint arXiv:1904.08783 (2019).

[11] C. May, A. Wang, S. Bordia, S. R. Bowman, R. Rudinger, On measuring social biases in sentence encoders, arXiv preprint arXiv:1903.10561 (2019).

[12] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, L.-P. Morency, Towards debiasing sentence representations, arXiv preprint arXiv:2007.08100 (2020).

[13] A. Conneau, D. Kiela, Senteval: An evaluation toolkit for universal sentence representations, arXiv preprint arXiv:1803.05449 (2018).

[14] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: EMNLP, 2014, pp. 1532–1543.

[15] J. Zhao, Y. Zhou, Z. Li, W. Wang, K.-W. Chang, Learning gender-neutral word embeddings, arXiv preprint arXiv:1809.01496 (2018).

[16] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, arXiv preprint arXiv:1508.05326 (2015).