Data Science and Engineering (2023) 8:177–195
https://doi.org/10.1007/s41019-023-00211-0

RESEARCH PAPER

Improving Gender-Related Fairness in Sentence Encoders: A Semantics-Based Approach

Tommaso Dolci[1] · **Fabio Azzalini**[1] · **Mara Tanelli**[1]

Received: 29 December 2022 / Revised: 20 March 2023 / Accepted: 22 March 2023 / Published online: 15 April 2023

Abstract
The ever-increasing number of systems based on semantic text analysis is making natural language understanding a fundamental task: embedding-based language models are used for a variety of applications, such as resume parsing or improving web search results. At the same time, despite their popularity and widespread use, concern is rapidly growing due to their display of social bias and lack of transparency. In particular, they exhibit a large amount of gender bias, favouring the consolidation of social stereotypes. Recently, sentence embeddings have been introduced as a novel and powerful technique to represent entire sentences as vectors. We propose a new metric to estimate gender bias in sentence embeddings, named *bias score*. Our solution leverages semantic importance of words and previous research on bias in word embeddings, and it is able to discern between neutral and biased gender information at sentence level. Experiments on a real-world dataset demonstrate that our novel metric can identify gender stereotyped sentences. Furthermore, we employ *bias score* to detect and then remove or compensate for the more stereotyped entries in text corpora used to train sentence encoders, improving their degree of fairness. Finally, we prove that models retrained on fairer corpora are less prone to make stereotypical associations compared to their original counterpart, while preserving accuracy in natural language understanding tasks. Additionally, we compare our experiments with traditional methods for reducing bias in embedding-based language models.

Keywords Gender bias · Word embeddings · Sentence encoders · Ethics of NLP · Data augmentation

1 Introduction

Language models are used for a variety of applications, such as CV parsing for a job position or document ranking for web search [5, 33]. 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 [40]. 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 to reflect the stereotypes of our 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 $man : programmer = woman : homemaker$ [7]. The social bias in the geometry of the model is then of course reflected in downstream applications like web search, CV parsing or hate speech detection [3, 7, 43]. In turn, this phenomenon favours the spread of prejudice towards social categories

Tommaso Dolci, Fabio Azzalini and Mara Tanelli have contributed equally to this work.

✉ Tommaso Dolci
tommaso.dolci@polimi.it

Fabio Azzalini
fabio.azzalini@polimi.it

Mara Tanelli
mara.tanelli@polimi.it

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy

already frequently penalised, such as women or African Americans.

Lately, sentence embeddings—vector representations of sentences based on word embeddings—are increasing in popularity, gaining exceptional results in many language understanding tasks, such as semantic similarity or sentiment prediction [17, 49]. Training language models on large corpora, that often encapsulate historical bias in the form of social stereotypes, leads to the risk of enforcing the bias originally present in our society; as a result, training datasets should be adjusted to remove bias [54]. Therefore, it is of the utmost importance to expand research on how sentence embedding encoders internalise the semantics of natural languages. An important step towards this direction is to define metrics that are able to reflect and quantify social bias in sentence encoders. Furthermore, studying and limiting the causes and consequences of bias in language models is an extremely important task [4, 6].

This work expands research on social bias in embedding-based models, focusing specifically on gender bias in sentence representations. First, we propose a method to estimate gender bias in sentence embeddings, highlighting the correlation between bias and stereotypical concepts in the sentence. Our solution, named *bias score*, is highly flexible and designed to be easily adapted to both different kinds of social biases (e.g. ethnic, religious) and various sentence encoders. Moreover, since gender bias is determined by the internalisation of stereotypical associations in language models, *bias score* allows to identify stereotyped sentences that are responsible for increasing gender bias in the output embeddings encoded by the model. Therefore, in the second part of the paper, we leverage *bias score* to retrieve the more stereotyped sentences from the Stanford Natural Language Inference corpus (SNLI) [9], a large text corpus suitable for training general-purpose sentence encoders, such as those proposed by [17] and [13]. We then outline two approaches to make SNLI fairer: removing entries associated to the highest *bias score*, and performing data augmentation by compensating stereotyped sentences with their gender-swapped counterparts. Finally, we retrain a BiLSTM sentence encoder [17] on different fairer versions of SNLI, testing and comparing it with its original counterpart from both fairness and accuracy viewpoint in downstream tasks.

Our contributions in this work include: *a*a novel metric to estimate gender bias in sentence embeddings leveraging the semantic importance of words and previous research on bias in word embeddings; *b*two methods to mitigate gender bias in sentence encoders by improving training data, performing data subtraction and data augmentation, respectively; *c*an analysis of the effect of such mitigation actions when retraining a BiLSTM sentence embedding encoder, with a comparison with traditional methods for gender bias mitigation; *d*a demonstration of the flexibility of our approach to be adapted to other language models, such as those based on transformer architectures.

The rest of the paper is structured as follows. Section 2 explores the state of the art on bias identification and reduction in language models, focusing on word and sentence embeddings. Section 3 introduces and defines *bias score*, our new metric for estimating gender bias in sentence representations. At the end of the section, we provide some examples of gender bias estimation via *bias score*. Section 4 first describes how to leverage *bias score* to make text corpora fairer, then explores a new approach to reduce bias in sentence encoders by retraining them on improved versions of their training data. Section 5 shows the results of our bias reduction methodology, discussing the benefits of the procedure from both the perspectives of quality and fairness. Section 6 describes how to extend our solution to transformer-based sentence encoders. Finally, Sect. 7 concludes the paper and outlines future work.

2 Related Work

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 [14].

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 [7]. This approach was later adapted for non-binary social biases such as racial and religious bias [34]. A *debiasing* algorithm was also proposed to mitigate gender bias in word embeddings [7]; however, it was also shown that it fails to entirely capture and remove bias [24]. The Word Embedding Association Test (WEAT) [11] 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 [46] and word2vec [38, 39] embeddings.

Recently, a number of different approaches extended the research field. A new debiasing procedure was proposed to reduce gender bias by introducing a term to the loss function used during the training phase of the model [48]. Additionally, [8] presented a regularisation procedure that aims at debiasing a language model by minimising the projection of encoder-trained embeddings onto a subspace that encodes gender. Similarly, [59] used model compression techniques, a type of regularisation techniques, to reduce toxicity and bias originally present in generative language models. The system proposed by [32] mitigates bias by employing counterfactual data augmentation, proving that

modifying the training data works better than changing the actual geometry of the embeddings. On a similar note, the approach described by [10] performs perturbations of the original embeddings training data to reduce the overall bias present in them. [27] presented a method to preserve gender-related information in feminine and masculine words while removing bias from stereotypical words. Still using GloVe as language model, [56] described an innovative procedure called Double-Hard Debias, to cope with changes in word frequency statistics that commonly have an undesirable impact on standard debasing methods. [60] describes a novel method exploiting causal inference, to reduce not only gender bias associated with a gender direction, but also gender bias from word embedding relations.

More recently, contextualised word embeddings like BERT [18] proved to be very accurate language models. However, despite the literature suggesting that they are generally less biased compared to their static counterparts [2], they still display a significant amount of social bias [36]. 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 shows that modern sentence embeddings also exhibit social bias [36]. Meanwhile, 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, leading to a loss of correct semantics [30].

Recently, [61] suggested the generation of implicit gender bias samples at sentence-level, which, along with a novel metric, can be used to accurately measure gender bias on contextualised embeddings. [28] proposed a fine-tuning method for debiasing word embeddings that can be applied to any pre-trained language model. Additionally, researchers have started working on generative transformer models. For instance, [25] proposed to mitigate gender disparity in text generation by learning a fair model with knowledge distillation. Last but not least, two comprehensive survey papers highlighted the latest advances on this front: [45] presents an overview of the most common debiasing methods in the context of vision and language research, while [37] proposes a deep empirical analysis of several bias mitigation techniques with different language models.

3 Gender Bias Estimation

Gender bias in word embeddings is typically estimated by computing the cosine similarity between word vectors and a gender direction identified in the vector space [7]. Cosine similarity is a popular metric to establish the semantic similarity of words based on the angle θ between their embedding vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$: $\cos(\theta) = \frac{u \cdot \vec{v}}{\|\vec{u}\| \, \|\vec{v}\|}$. The more $\cos(\theta)$ approaches 1, the higher is the similarity between $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$. In word embedding models, semantic similarity with respect to a gender direction (typically computed with PCA from multiple gender words) means that a word vector contains information about gender. Since only gender-neutral words can be biased, gender words like *man* or *woman* are assumed to contain correct gender information.

Table 1 Gender information from cosine similarity for sentence embeddings encoded by InferSent [17] and SBERT [49]

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

Recently, sentence representations are becoming increasingly popular, but the same approach used for measuring gender bias in word-level representations cannot be easily adopted, demanding a new methodology. In fact, the main problem is that gender-neutral sentences cannot be identified and listed. Unlike words, sentences are infinite in number. Moreover, sentences may contain gender bias despite containing explicit gender information. 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 representations of gender-neutral sentences like *someone is a nurse* still contain a lot of "false" 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 gender words, and the amount coming from biased words. For this reason, we adopt a more dynamic approach: we start 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, normalising it with respect to the length of the sentence and to the contextualised semantic importance of each word. This decision is grounded on two observations: first, the semantics of a sentence largely depends on the semantics of the words contained in it; second, sentence embedding encoders are generally based on predefined word embedding models [17, 49]. In Sect. 3.5, we test our metric on sentence representations produced by InferSent [17], a sentence encoder by Facebook AI that achieved great results in a variety of natural language understanding tasks [16]. Since InferSent is based on GloVe [46] word vectors, we adopt GloVe representations to first quantify gender bias at word level.

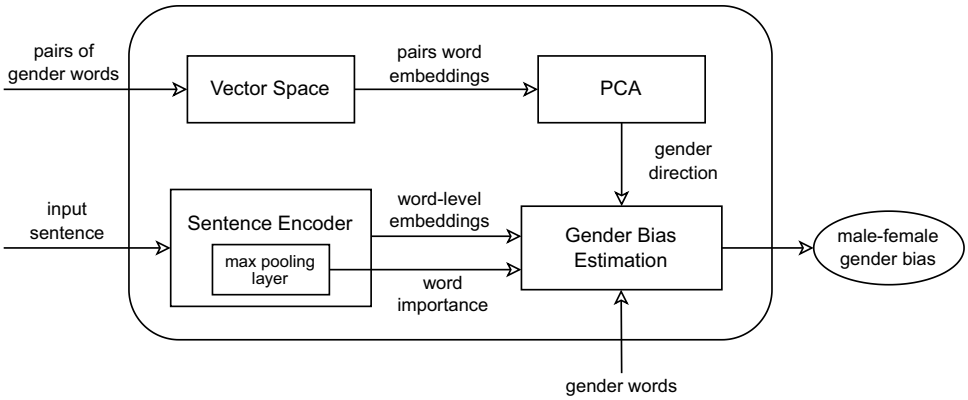


Fig. 1 Overview of the framework to compute *bias score*

3.1 Bias Score

An overview of the approach adopted to calculate *bias score* is illustrated in Fig. 1. We consider as inputs a sentence in natural language, n pairs of gender words, and a list of words with explicit gender connotation. The output are two estimations corresponding to the amount of female-related and male-related gender bias at sentence level. In particular, we consider four fundamental elements for gender bias estimation, representing the inputs to the bottom-right block in Fig. 1: *a*the gender direction $\vec{g}$ previously identified in the vector space; *b*a list L of gender words in the same language as the encoder (here is English); *c*the semantic importance I_w of each word in the sentence according to the encoder; *d*the word-level embeddings of the input sentence.

The amount of female-related and male-related gender bias is represented by a positive and a negative value, respectively, obtained from the sum of the gender bias associated to each word, 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 their bias is always set to zero. In fact, should the *bias score* of gender words be considered, it would create an additive term, i.e. an offset in the final score, which might hide the real bias in the geometry of the sentence representation. In detail, for each neutral word w in the sentence, we compute its gender bias as the cosine similarity between its word vector vec_w and the gender direction $\vec{g}$, and then we multiply it by the word importance I_w. In particular, for a given sentence s:

$$BiasScore_F(s) = \sum_{\substack{w \in s \\ w \notin L}} \underbrace{\cos(vec_w, \vec{g})}_{>0} \times I_w \tag{1}$$

$$BiasScore_M(s) = \sum_{\substack{w \in s \\ w \notin L}} \underbrace{\cos(vec_w, \vec{g})}_{<0} \times I_w \tag{2}$$

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. A slightly different approach allows us to derive a single value estimation of gender bias at sentence level, by computing the absolute value of each word-level bias:

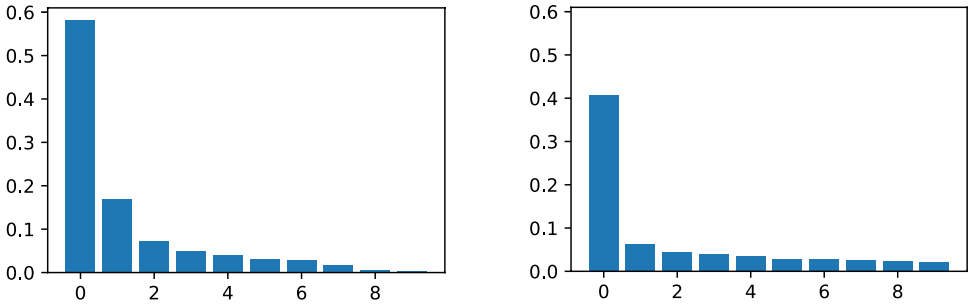
$$Abs\text{-}BiasScore(s) = \sum_{\substack{w \in s \\ w \notin L}} \Big| \underbrace{\cos(vec_w, \vec{g}) \times I_w}_{\text{word-level bias}} \Big| \tag{3}$$

This proves useful in certain situations, such as when sorting multiple sentences according to the total amount of associated *bias score*. In the following sections, we go into more detail by describing how we derive $\vec{g}$, 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 $\vec{g}$, named *gender direction*, serves as the second 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 being analysed. We mainly test *bias score* on the sentence encoder InferSent [17], which is based on GloVe [46], a word embedding model with a vector space of 300 dimensions. In general, it is important to adopt word embedding models matched to the encoder under analysis. For instance, SBERT produces sentence representations based on word-level BERT embeddings [49]. See Sect. 6 for an example of *bias score* with SBERT.

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



(a) 10 pairs (b) 50 pairs

Fig. 2 Top ten components in PCA from using 10 and 50 pairs of gender words. The top component explains 58% and 41% of the variance, respectively

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, to avoid inconsistency, we take into consideration several pairs of gender words and perform a Principal Component Analysis (PCA) to reduce their dimensionality to one. In particular, we select the following ten pairs of gender 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. 2a, the top component resulting from the analysis is significantly more important than the other components, explaining 58% 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.

Following the advise from [21], and to assess the quality of the gender direction obtained, we further perform PCA starting from an extended list of 50 pairs of gender words, taken from [7], and compare the result with $\vec{g}$. From the full list of pairs available on the author's repository,[1] we select only those consisting of words present in GloVe. Figure 2b shows that again the top component is clearly the most important, explaining a variance of 41%. Moreover, its cosine similarity with respect to $\vec{g}$ is above 93%, demonstrating the quality of the gender direction selected.

3.3 Gender Words

A list L of gender 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 gender 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 gender 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 $\mathcal{G}$ (gender 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 appropriately expressed. Examples of gender words include he, she, sister, father, councilman, heroine, princess.

Our list L contains a total of 6562 gendered nouns, of which 409 and 388 are common nouns, respectively, selected starting from the work of [7] and [63]. All words are listed in their lower-cased and capitalised version, in both singular and plural forms. Additionally, we added 5765 unique given names taken from Social Security card applications in the USA.[2]

3.4 Word Importance

Following the approach described by [17], word importance is estimated based on the max-pooling operation commonly performed by sentence encoders to reduce the dimensionality of the final sentence embedding to a fixed amount. Our approach consists in counting how many times in the max-pooling phase a word representation is selected to be part of the sentence embedding. In the case of InferSent, this approach is equivalent to counting 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

[1] https://github.com/tolga-b/debiaswe/blob/master/data/equalize_pairs.json.

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

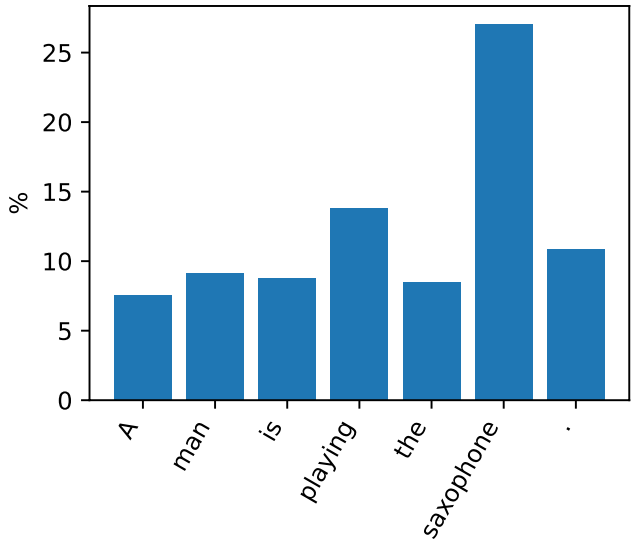


Fig. 3 Word importance for the input sentence *A man is playing the saxophone*

sentence representation centred on the word w_t, i.e. the word at position t in the sentence.

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. 3, 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 choose to adopt the percentage importance for computing *bias score*.

3.5 Bias Score Examples

Table 2 illustrates a detailed example of gender bias estimation via *bias score*, regarding the sentence *She likes the pink dress*. The example shows how gender stereotypes like this are heavily internalised in the final sentence representation: in fact, they account for the majority of the gender bias in the embedding according to *bias score*.

Additionally, we use *bias score* to estimate gender bias for sentences contained in SNLI, a large text corpus for training language models [9]. According to the experiments, sentences corresponding to the highest *bias score* towards the male direction, estimated applying Eq. 2, describe situations from popular sports like baseball and football, that are frequently associated with men and rarely with women. Similarly, sentences corresponding to the highest *bias score* in the female direction, estimated via Eq. 1, illustrate female stereotypes, like participating in beauty pageants, applying make-up or working as a nurse. Table 3 displays the most-biased sentences in the SNLI corpus according to our metric, in both male and female directions.

Results are similar when estimating the absolute *bias score* using the more general Eq. 3: top entries include sentences with a high 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.

Table 2 Bias score example for the sentence *She likes the pink dress*

Word	Importance	Gender bias	Weighted bias
She	12.13%	0	0
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, 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
The bikini is pink	0.121159
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
Dodgers player playing baseball	− 0.136282

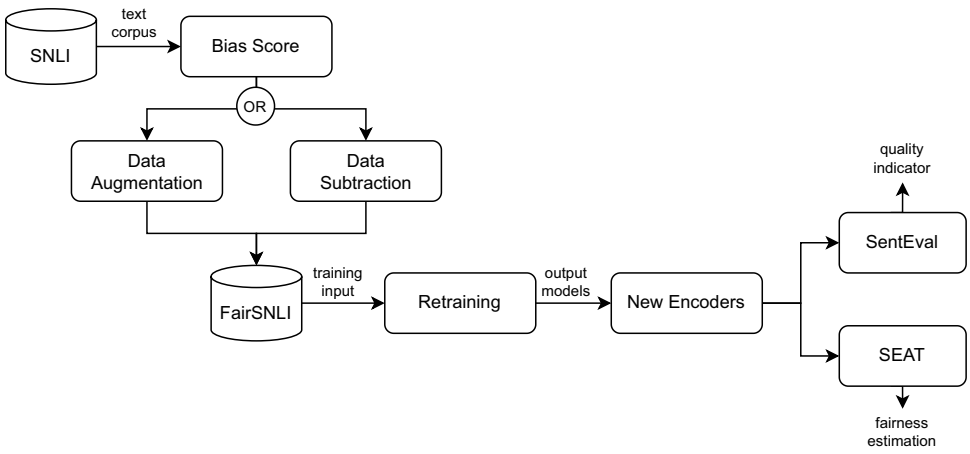


Fig. 4 Overview of the proposed methodology for gender bias reduction in sentence encoders

Table 4 Sample entries from the SNLI corpus. Each entry contains a premise sentence, an hypothesis sentence and a gold label describing their relationship. C = contradiction, N = neutral, E = entailment

Premise	Hypothesis	Label
A man inspects the uniform of a figure in some East Asian country	The man is sleeping	C
Two people are in a pond pulling a life raft	People are having fun	N
A black race car starts up in front of a crowd of people	A man is driving down a lonely road	C
A soccer game with multiple males playing	Some men are playing a sport	E
A smiling costumed woman is holding an umbrella	A happy woman in a fairy costume holds an umbrella	N

4 Gender Bias Reduction

Embedding-based language models learn stereotypical associations during the training phase, even if data are seemingly verified and safe [7]. Therefore, to mitigate gender bias and limit the internalisation of stereotypical conceptions, our approach aims to detect stereotyped entries in text corpora used for training language models, namely SNLI [9]. We explore two directions: removing stereotyped entries from the corpus, or compensating by adding counterfactual entries regarding gender. In this section, we describe in detail how to improve the fairness of a training corpus, and then test our intuition by retraining a sentence encoder on the new corpus obtained. Our goal is to improve the degree of fairness in the encoder, without losing accuracy in downstream tasks, by retraining it on a fairer and less stereotyped corpus of text. To evaluate both properties (quality and fairness), we test the newly retrained models with SentEval [16] and SEAT [36], respectively, as described in Sect. 5. An overview of the adopted methodology is illustrated in Fig. 4.

4.1 Training Corpus

The Stanford Natural Language Inference corpus (SNLI) is a large collection of English sentence pairs written by humans for textual inference tasks [9]. It is one of the largest resources of its kind, listing more than 570K pairs of sentences, and more than 600k unique sentences in the train set alone. Each entry is composed of a premise sentence, a hypothesis sentence, and a gold label with one of three possible values: entailment, contradiction, neutral. The general goal of inference tasks is to predict whether the hypothesis sentence logically follows the premise sentence (entailment), contradicts it (contradiction), or they do not share any correlation (neutral). According to the authors, the size and diversity of the dataset allow to train language models for sentence meaning representation. Table 4 illustrates some examples of SNLI entries.

Despite its large use, the literature showed that SNLI contains gender and ethnic stereotypes, alongside harmful or pejorative language associated with social categories like women, Muslims, African Americans [51]. For this reason, our intuition is that improving the fairness of SNLI by getting rid of stereotypical concepts, we effectively prevent natural language models trained on this corpus from internalising them.

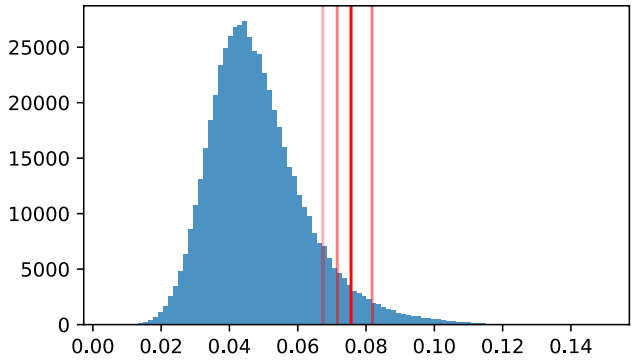


Fig. 5 *Bias score* distribution for entries in SNLI. The red lines correspond to the 90th, 93th, 95th and 97th percentiles

4.2 Improving SNLI Fairness

To improve the fairness of SNLI, we follow two approaches: data subtraction and data augmentation. The first is aimed at reducing the number of entries in the corpus, by removing stereotyped pairs of sentences, identified from the *bias score* associated to their embeddings. The second is directed to the addition of more entries to the corpus, in order to balance stereotyped sentence pairs and reach a higher degree of gender equality.

In both cases, the first step is to apply our metric to all the sentences contained in SNLI in order to find the entries associated with the highest *bias score*. This methodology is grounded on the observation that a high *bias score* is correlated with stereotyped or sexist sentences, as illustrated by the examples in Sect. 3.5. Moreover, SNLI has been shown to contain many stereotypical associations [51], proving to be a good candidate for our methodology.

4.2.1 Data Subtraction

Following this approach, we want to exclude sentences whose embedding exhibits a large amount of gender bias, without diminishing too drastically the size of the corpus. First, we compute *bias score* for both the premise and the hypothesis sentence in each sentence pair in the corpus, using the absolute *bias score* formula described in Eq. 3. Then, we only consider for each pair the highest *bias score* between the one associated to the embedding of the premise and the one related to the hypothesis, in order to discard entries for which at least one of the two scores is too large.

We define four additional corpora derived from SNLI, respectively, by subtraction of the 3%, 5%, 7% and 10% entries with the highest *bias score* associated. Therefore, we set a threshold at the 97th, 95th, 93th and 90th percentiles and discard entries with a *bias score* above the threshold. The distribution of *bias score* for all entries in SNLI and the four selected thresholds is illustrated in Fig. 5. It is evident from the chart that discarding 5% of the entries already allows to halve the highest absolute *bias score* in the corpus, from 0.1498 to 0.0756. After removing the selected entries, we randomise and split each of the four resulting corpora (*Sub90*, *Sub93*, *Sub95* and *Sub97*) into training, development and testing sets. Following the split in the original version of the corpus, we place 10k pairs each in the development and testing sets, and use the remaining pairs for training.

4.2.2 Data Augmentation

To increase the number of entries in SNLI, we adopt an approach of counterfactual data augmentation based on duplicating sentence pairs by converting all female words to their corresponding male words, and vice versa. A similar approach for swapping gender entities is described by [62] and [32], but neither of the two works consider given names in the procedure. First, similarly to the approach used for data subtraction, we set a threshold at the 90th, 93th, 95th and 97th percentiles, then perform the duplication for all sentence pairs in the corpus associated to a *bias score* higher than the threshold. In case neither the premise or the hypothesis contain gender words, the entry is not duplicated. On average, duplication affects around 60% of the considered entries. To avoid overfitting, each entry is only duplicated once.

To perform the duplication, we first tokenise the sentence with NLTK, a text processing library.[3] Each token represents either a word or a punctuation mark. Then, we consider only gender words, and specifically those for which there exists a female/male counterpart. After swapping them with their gender counterpart, we obtain a sentence with subjects of the opposite gender. For instance, *he is a young boy* is converted to *she is a young girl*, and *my father is a singer* becomes *my mother is a singer*. A total of 122 gender words are considered, mostly nouns regarding family members or occupations (e.g. uncle, aunt, chairman, chairwoman). Appendix A contains the full list of gender words involved in this procedure.

Moreover, we consider the 2500 most popular female and male given names in the USA, according to the Social Security Administration[4]; they are used to convert female names to equally popular male names, and vice versa. For instance, the sentence *Patrick is going to the supermarket* becomes *Rachel is going to the supermarket*. Some examples of sentence pair duplication by gender-swapping are provided in Table 5. Finally, we randomise the four resulting corpora (*Aug90*, *Aug93*, *Aug95* and *Aug97*) and split them into training, development and testing sets, again with 10k

[3] https://www.nltk.org.

[4] https://www.ssa.gov/OACT/babynames/limits.html.

Table 5 Sentence pair duplication by gender-swapping in SNLI

	Premise	Hypothesis
Original entries	A *girl* in pink twirls a ribbon	A ribbon is being twirled
	Two *men* are outside and talking to each other	The *men* are discussing football
	Two *men* wearing padding are fighting	Two *men* watch a fight on ESPN
Duplicate entries	A *boy* in pink twirls a ribbon	A ribbon is being twirled
	Two *women* are outside and talking to each other	The *women* are discussing football
	Two *women* wearing padding are fighting	Two *women* watch a fight on ESPN

pairs in development and testing sets, and the rest reserved for training.

4.3 Language Model and Parameters

We retrain a sentence encoder language model based on a bidirectional LSTM architecture (BiLSTM) developed by Facebook AI [17]. The network features a forward and a backward LSTM that read the input sentence in two opposite directions. The final output is a 4096 dimensions vector representation of a natural language sentence, obtained with a max-pooling discretisation technique. Authors train the network on SNLI with a supervised approach, hence our choice to focus on this specific corpus in our experiments.

The parameters used for retraining the BiLSTM network are those suggested by the original authors: batch size 64, SGD optimiser with a learning rate of 0.1 and weight decay of 0.99. Training is stopped when learning rate goes under the threshold of 10^{-5}: since the network converges fairly quickly, as pointed out by [17], we set the maximum number of training epochs to 8. Finally, as underlying word-level representations, we select the biggest and more powerful GloVe model available, trained on Common Crawl 840B.[5] For the retraining, we used a Linux machine with Ubuntu 18.04, 78GB of RAM and a GeForce GTX 1060 GPU.

5 Experimental Results

The goal of our experimental setting is to confirm a reduction in the gender bias exhibited by the language model retrained by either data augmentation or subtraction. At the same time, it is fundamental to avoid any degradation in the semantic power of the resulting sentence embeddings.

For this reason, after separately retraining the model on the eight SNLI-derived corpora obtained with data subtraction and augmentation, we test them on both fairness and accuracy, comparing them with the original encoder trained on the full unedited SNLI corpus.

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

5.1 Fairness and Accuracy Metrics

Since our goal is to improve the fairness of sentence encoders, without losing accuracy in downstream tasks, we need to check both qualities. Therefore, the retrained models are evaluated using SEAT [36], a fairness test for sentence encoders, and SentEval [16], a toolkit for assessing the accuracy of sentence embedding models in a variety of natural language tasks.

5.1.1 SEAT

The Sentence Encoder Association Test, or in short SEAT [36], is a fairness test that adapts to sentence encoders the well-known Word Embedding Association Test (WEAT) [11]. Like WEAT, SEAT measures stereotypical association between two sets of target concepts X, Y (e.g. sentences with male/female subjects) and two sets of attributes A, B (e.g. sentences related to career/family). Targets and attributes are built from simple semantically bleached templates like *This is* $<$ *term* $>$ or $<$ *term* $>$ *is here*, e.g. *This is John*. The magnitude of the association is measured by the so-called effect size, defined as:

$$d = \frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std}_{z \in X \cup Y}\, s(z, A, B)},$$

where $s(w, A, B)$ is the difference of mean cosine similarities between the embedding of the target concept w and the embedding of each attribute in the sets A and B. A higher effect size reflects a stronger correlation between target concepts and attributes, thus a more severe association bias.

SEAT contains ten tests adapted from the work of [11]. We focus only on those tests that contain gender target concepts or attributes, namely C6, C7, C8, and their alternative versions C6b, C7b and C8b, generated by replacing given names (e.g. John, Sarah) with group terms (e.g. male, women). In SEAT, positive scores describe stereotypical associations, like *male* $\rightarrow$ *career* and *female* $\rightarrow$ *family*. Thus, the lower the score, the higher the fairness of the model. Additionally, we consider tests C1 and C2, to assess the retention of correct and ethical associations, such as *instruments* $\rightarrow$ *pleasant* and *weapons* $\rightarrow$ *unpleasant*. In

Table 6 SEAT results on subtraction models for gender-related tests C6–C8b and retention tests C1–C2 for the data subtraction approach, average effect size is computed on C6–C8b

Test	Original	Sub97	Sub95	Sub93	Sub90
C1: Flowers/Insects	1.523	1.583	1.572	1.586	1.545
C2: Instruments/Weapons	0.894	1.136	0.974	1.160	0.924
C6: Career/Family	1.674	**1.644**	**1.585**	1.689	**1.591**
C6b: Career/Family	0.286	0.350	0.358	0.445	**0.258**
C7: Math/Arts	1.124	**0.796**	**0.512**	**0.819**	**0.697**
C7b: Math/Arts	1.543	**1.454**	**1.439**	**1.343**	**1.408**
C8: Science/Arts	1.269	**0.954**	**0.730**	**1.168**	**0.971**
C8b: Science/Arts	1.629	**1.628**	**1.450**	**1.615**	**1.494**
Average Effect Size	1.254	**1.138**	**1.012**	**1.180**	**1.070**

Lowest score for each test underlined; association scores lower than the original model in bold

Table 7 SEAT results on augmentation models for gender-related tests C6–C8b and retention tests C1–C2 for the data augmentation approach, average effect size is computed on C6–C8b

Test	Original	Aug97	Aug95	Aug93	Aug90
C1: Flowers/Insects	1.523	**1.498**	1.590	1.504	1.622
C2: Instruments/Weapons	0.894	0.960	1.037	0.971	1.093
C6: Career/Family	1.674	**1.579**	**1.660**	**1.660**	1.676
C6b: Career/Family	0.286	0.301	0.299	0.306	0.292
C7: Math/Arts	1.124	**0.928**	**0.952**	**0.800**	**0.489**
C7b: Math/Arts	1.543	1.651	**1.535**	1.551	1.586
C8: Science/Arts	1.269	**1.080**	**1.129**	**0.892**	**0.566**
C8b: Science/Arts	1.629	1.713	**1.605**	**1.521**	**1.560**
Average Effect Size	1.254	**1.209**	**1.197**	**1.122**	**1.028**

Lowest score for each test underlined; association scores lower than the original model in bold

this case, a reduction in the score indicates a loss of correct associations, therefore it is desirable to maintain fairly high positive scores.

5.1.2 SentEval

We employ SentEval [16] to assess the quality of our models in terms of sentence representations. SentEval is a toolkit featuring a variety of natural language downstream tasks; to test our models we select twelve of them, namely MR [42] and SST [53] for sentiment analysis, CR for text summarisation [26], SUBJ on subjectivity/objectivity [41], MPQA on opinion polarity [57], TREC for question answering [55], STS-Benchmark for semantic relatedness [12], SICK-E and SICK-R for semantic entailment and relatedness [35], STS14

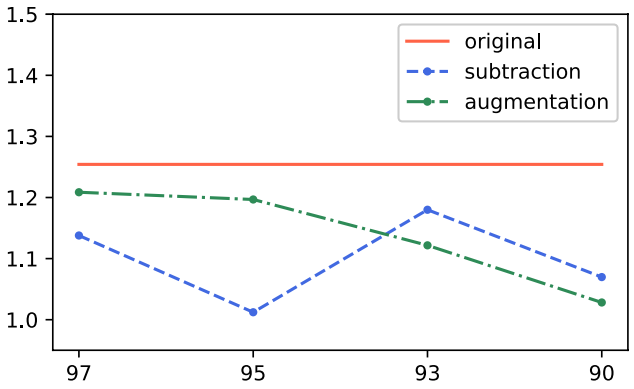


Fig. 6 Average SEAT score across models for tests C6-C8b. The lower the score, the higher the degree of fairness

for semantic similarity [1] and MRPC for paraphrase detection [19]. All tasks except for STS14 requires the model to be further trained on their respective corpus. For all the training phases, we use the default parameters suggested by the authors: 10-fold cross-validation, Adam optimisation, batch size of 64, tenacity of 5, and epoch size of 4.

5.2 Retrained Models Evaluation

Tables 6 and 7 show the effect size for all gender-related tests in SEAT and for the two association retention tests C1 and C2, for both subtraction and augmentation approaches, respectively.

Concerning gender-related tests from C6 to C8b, there is a marked improvement for all our models in almost every test. C7 and C8 in particular sees an improvement in the association score for every single model, with up to a 0.635 and a 0.703 reduction, respectively, both performed by *Aug90*, i.e. the most augmented model tested. The same trend is evident in tests C7, C7b and C8, with only C6b not showing significant improvements, probably due to the already very low association score in the original model. Furthermore, the average score from tests C6 to C8b shows again an improvement for all our models compared to the original, with best scores by *Sub95*, *Sub90* and *Aug90*, suggesting that addressing a larger percentage of sentences in the training data reduces the amount of bias. Figure 6 depicts the trend of the average SEAT scores for both data subtraction and augmentation approaches: concerning augmentation, the trend clearly decreases as the percentage of sentence pairs subject to gender-swapping augmentation increases; on the other hand, models obtained from data subtraction fluctuate in the results, yet still performing better than the original model. Finally, C1 and C2 both highlight a retention of correct

Table 8 SentEval results for all encoders trained on SNLI corpus adjusted by data subtraction and augmentation (above), and a comparison with encoders trained on hard-debiased (HD) embeddings (below)

Model	MR	CR	SUBJ	MPQA	TREC	SST-2	SST-5
Original	80.7	85.2	92.7	90.2	86.2	82.0	44.1
Sub97	80.6	85.3	92.6	90.2	89.0	82.0	45.1
Sub95	80.7	84.8	92.6	89.8	87.8	82.5	44.2
Sub93	80.9	84.5	92.4	90.3	87.0	82.3	43.9
Sub90	80.2	85.1	92.5	90.2	88.0	82.0	43.5
Aug97	80.5	85.4	92.7	89.9	87.8	80.1	43.3
Aug95	80.6	84.9	92.5	90.0	87.8	82.2	44.7
Aug93	80.5	85.4	92.5	90.2	86.8	83.5	43.3
Aug90	80.3	84.7	92.7	90.1	87.0	82.9	44.0
GloVe-HD	79.6	81.8	90.7	89.1	79.4	82.0	43.1
w2v	78.2	81.4	90.6	89.3	86.0	82.9	45.4
w2v-HD	77.9	80.0	90.2	89.1	85.4	82.6	43.1

Best score for each task underlined

Table 9 SentEval results for all encoders trained on SNLI corpus adjusted by data subtraction and augmentation (above), and a comparison with encoders trained on hard-debiased (HD) embeddings (below)

Model	STS-B	SICK-R	SICK-E	STS14	MRPC
Original	74.7/74.4	0.887	85.7	0.68/.65	75.6/82.5
Sub97	75.9/75.6	0.886	86.0	0.68/.65	75.4/82.8
Sub95	76.6/76.3	0.886	85.5	0.68/.65	76.6/83.3
Sub93	74.9/74.4	0.883	85.9	0.68/.65	73.9/81.7
Sub90	75.8/75.6	0.886	85.7	0.68/.65	75.0/82.4
Aug97	75.5/75.0	0.885	86.1	0.67/.65	75.7/82.5
Aug95	76.3/75.8	0.886	85.8	0.68/.66	75.3/82.4
Aug93	74.7/74.6	0.886	85.4	0.68/.66	75.0/82.2
Aug90	75.7/75.3	0.886	85.8	0.68/.66	75.6/82.7
GloVe-HD	76.6/75.8	0.890	84.8	0.63/.61	73.4/82.1
w2v	74.5/73.6	0.885	84.8	0.68/.66	73.6/81.6
w2v-HD	74.9/73.9	0.886	85.6	0.67/.66	72.5/82.1

Best score for each task underlined

Table 10 Overall SentEval score across all test, and difference compared to the original model

Model	Overall Score	Difference
Original	79.45	–
Sub97	*79.89*	**+0.44**
Sub95	**79.81**	**+0.36**
Sub93	79.35	−0.10
Sub90	*79.52*	**+0.07**
Aug97	79.41	−0.04
Aug95	*79.75*	**+0.30**
Aug93	*79.52*	**+0.07**
Aug90	*79.64*	**+0.19**
GloVe-HD	77.72	−1.73
w2v	78.60	−0.85
w2v-HD	78.08	−1.37

Best score in bold, scores slightly higher than the original model in italics

associations in all our models, with C2 also showing a reinforcement of the ethical correlations *instruments* → *pleasant* and *weapons* → *unpleasant*.

Tables 8 and 9 show the results from testing our models on the twelve downstream tasks provided by SentEval. In all tasks, the performance is very much similar compared to the original model, with only slight deviations. These results are confirmed by the overall mean score across all tasks, illustrated in Table 10.

5.3 Analysis and Discussion

The results of our experiments show the possibility to find stereotyped sentences by means of *bias score*, allowing to identify stereotypes in text corpora used for training language models. Additionally, retraining encoders on fairer corpora of sentences, such as an augmented or size-reduced version of SNLI, proves to be an effective way to achieve more ethical and equally powerful language models. More specifically, results from SEAT suggests that both models obtained from data augmentation and data subtraction can override unethical and gender stereotypical associations, leading to better association scores. At the same time, correct association are maintained if not even reinforced, meaning that the basic semantics of the language is retained. In fact, this is confirmed by the tests performed on SentEval, showing that all our models achieve comparable

results to the original one, with no augmentation or subtraction on the training data. Additionally, the trend from average SEAT scores depicted in Fig. 6 suggests that considering a higher percentage of sentence pairs from SNLI increasingly improves results, at least for the augmentation approach. Despite results from SentEval suggesting that accuracy remain steady regardless of the percentage of entries removed from the corpus or added to it, we expect a limit to exist, beyond which the accuracy of the model starts decreasing, despite a continuous improvement from the fairness point of view.

In general, results confirm that gender bias in sentence encoders can be ascribed to the internalisation of stereotypical concepts encountered during the training phase. Therefore, removing or compensating for stereotyped entries in the training data improves the fairness of the final model.

5.4 Comparison with Hard-Debiasing

While our approach proposes the optimisation of the training procedure of language models focusing on the training data, an alternative option is represented by the so-called vector-space manipulation [22, 45].

In particular, we compare our approach with *hard-debiasing*, a vector-space manipulation technique to reduce bias in word embeddings by forcibly removing gender semantics for all vectors associated with gender-neutral words [7]. We first hard-debias GloVe embeddings, then re-train the sentence encoder starting from debiased embeddings and the full original SNLI corpus. We tested the resulting model on both SentEval and SEAT. Results are presented in Table 8, 9 and 10, and Table 11, respectively. While we witness a sensible improvement in SEAT tests, the overall score from SentEval sees an average degradation of 1.73 points. Moreover, performance in specific tasks like CR and TREC drop drastically by up to 3.4 and 6.8, respectively. Similar results are obtained when adopting word2vec [38, 39] and hard-debiased word2vec from [7] in place of GloVe: SEAT tests improves considerably, at the price of a major reduction in many SentEval tasks, particularly in classification tasks such as CR, TREC and SST-5.

In summary, *hard-debiasing* can effectively improve the fairness of sentence encoders, but at the cost of largely losing accuracy in downstream tasks. On the contrary, our approach allows to maintain the accuracy and quality of the resulting sentence embeddings in all tasks, while still considerably improving the average SEAT scores for gender-related tests by up to 19%. Since the drop in the accuracy of the model is significant when using hard-debiased GloVe, we think it is not advisable to combine our approach with hard-debiasing, since the quality of the final sentence representations would still be lower.

Table 11 SEAT results comparison between encoders trained on original GloVe or word2vec (w2v) and their hard-debiased (HD) version

Test	GloVe	GloVe-HD	w2v	w2v-HD
C1: Flowers/Insects	1.523	1.506	1.637	1.073
C2: Instruments/Weapons	0.894	1.037	1.569	0.677
C6: Career/Family	1.674	0.889	1.894	1.584
C6b: Career/Family	0.286	0.188	0.266	0.250
C7: Math/Arts	1.124	0.737	1.088	1.018
C7b: Math/Arts	1.543	1.552	1.561	1.004
C8: Science/Arts	1.269	0.008	1.127	0.875
C8b: Science/Arts	1.629	0.702	1.664	0.981
Average Effect Size	1.254	0.679	1.351	0.933

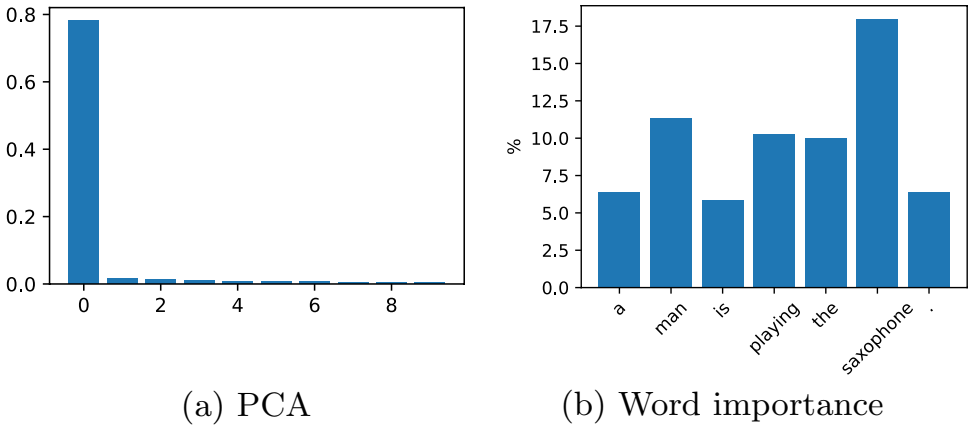
Lowest scores underlined

6 Extension to Transformer-Based Models

In this section, we briefly describe how to adapt *bias score* to transformer-based sentence encoders. Additionally, we present some examples focused on the widespread BERT family of language models, particularly on the sentence encoder SBERT [49] based on BERT-Base for NLI with max pooling discretisation. We adopt SentenceTransformer,[6] a Python framework that allows to easily switch from one language model to another, without installing additional tools.

The main difference with the methodology described in Sect. 3 is how to compute the gender direction $\vec{g}$. In fact, contextualised encoders need pairs of sentences instead of words, to fully capture the semantic of gender. To do so, we take more than 100 sentences randomly extracted from the following three datasets: POM [44], MELD [47], SST [53]. Then, we swap all female words to male words, and vice versa. Following the approach described by [30], the difference vector, between the embedding of the original sentence and the embedding of the gender-swapped counterpart, represents the gender. Again, to solidify the methodology, we perform a PCA of the resulting difference vectors to find a single direction $\vec{g}$. This approach proves to be extremely effective, resulting in a top component explaining 78% of the variance, as shown in Fig. 7a.

[6] https://www.sbert.net.



(a) PCA (b) Word importance

Fig. 7 On the left, top ten components in PCA to retrieve the gender direction for SBERT with BERT-Base vectors. On the right, token-level importance for the input sentence *A man is playing the saxophone*

Table 12 Detailed bias score estimation for SBERT encoder with BERT-Base vectors for the sentence *She likes the new pink dress*

Word	Importance	Gender bias	Weighted bias
she	11.33%	0	0
likes	10.68%	0.37394	0.03993
the	8.33%	0.39881	0.03323
new	8.46%	0.37113	0.03141
pink	11.33%	0.35007	0.03966
dress	14.06%	0.37273	0.05242
Overall female bias			0.19664
Overall male bias			0

Table 13 Examples of male and female bias score with BERT-based sentence encoders, considering different gender stereotyped sentences

Set	Sentence	F-Bias	M-Bias
1	She is a homemaker	0.16669	0
	He is a homemaker	0	–0.02642
	This is a homemaker	0.08124	0
2	She likes wearing pink dresses and playing football	0.21236	0
	He likes wearing pink dresses and playing football	0.05823	–0.00366
	Everyone likes wearing pink dresses and playing football	0.16925	0
	Everyone likes playing football and wearing pink dresses	0.14301	0

Again, repeating the procedure with different pairs of gender sentences does not produce much difference in the resulting direction, with a similarity higher than 99%.

Concerning the word importance, since our approach is based on max pooling, it is best to choose this discretisation option. Moreover, it is worth noting that BERT-based models split sentences into *tokens*, that either represent entire words or sub-words. For this reason, the importance is estimated at token-level. An example is provided in Fig. 7b

Table 12 provides a detailed example of *bias score*, showing that, differently from sentence encoders based on static word vectors like GloVe, in contextualised representations every word inherits the gender information from the context of the sentence, in this case from the female subject. This feature makes it more difficult to separate the amount of correct gender information from the gender biased information. However, results from Table 13 show that BERT-based encoders still encapsulate a lot of gender bias, which is especially noticeable in gender-neutral sentences. In fact,

in the first set of sentences, the amount of female bias with a female subject is a lot higher than the male bias with a male subject. Accordingly, with a neutral subject the same sentence still exhibits female bias due to the stereotypical profession described. The second set of sentences again shows that female stereotypical associations are internalised more than the male ones; when the context of the sentence contains both a female and male stereotype (pink dresses and playing football), the amount of bias in the female direction is higher, even when the subject is explicitly male. Moreover, with a neutral subject, the female bias is still very large regardless of the order in which the two concepts are presented. We could not find any explanation of this behaviour in the literature and we believe that it is due to the higher presence of female stereotypes in training corpora.

Finally, to adapt the methodology described in Sect. 4 to BERT-based encoders, we first consider that their training is extremely time and resource consuming. However, they offer the opportunity to be fine-tuned and adjusted with a lot less effort. To adapt our methodology to this scenario, we first identify text corpora commonly used to semantically fine-tune pre-trained language models, such as STS-Benchmark [12] and MultiNLI [58]. Similarly to the methodology described in this work, the first step is to identify the more stereotypical entries in these corpora, improve their degree of fairness, and then used them to fine-tune a pre-trained transformer-based model. Appendix B illustrates preliminary results from the identification of stereotypical entries in the two aforementioned training corpora. Similar approaches based on fine-tuning for bias mitigation showed promising results in the recent literature [15, 20, 23]

7 Conclusions and Future Work

In this paper, we proposed both an algorithm to estimate gender bias in sentence embeddings, based on a novel metric named *bias score*, and a method to mitigate gender bias in sentence encoders by retraining them on training data improved by performing either data subtraction or gender-swapping data augmentation.

Bias score discerns between gender bias and gender neutral information encoded in a sentence embedding and quantifies the presence of bias on the basis of the semantic importance of each word. We tested our solution on InferSent [17], searching for the most gender biased representations from a corpus of natural language sentences. Moreover, bias estimation is also crucial for adapting procedures like *debiasing* to sentence embeddings, since it requires to effectively identify biased sentence representations [7]. Since *bias score* is proportional to the amount of stereotypical conceptions encapsulated in sentence representations, it allows to retrieve stereotyped entries from text corpora used for training language models. In the second part of this work, we define fairer versions of the SNLI corpus by data subtraction and data augmentation of its more stereotyped entries: sentence encoders retrained on them proves to be less prone to make stereotypical associations compared to their original counterpart, while maintaining the same accuracy in a variety of natural language understanding tasks. This is crucial to maintain quality, yet reducing discrimination in a variety of web-related tasks, such as document search and ranking or hate speech detection.

Future work includes adapting *bias score* to different kinds of social bias (e.g. ethnic, religious) and further testing it on other sentence encoders such as SBERT [49]. Additionally, considering a higher percentage of SNLI for data subtraction and data augmentation may result in additional improvement in SEAT scores, either maintaining the same accuracy for the retrained encoder, or confirming the hypothesis that after a certain threshold the performance of the model starts decreasing. Moreover, combining the two approaches of subtraction and augmentation may prove even more effective on reducing gender bias. This means removing sentence pairs associated with high *bias score* and substituting them with the equivalent gender-swapped sentence. Finally, additional comparisons with debiasing techniques such as the one proposed by [63] can be useful to highlight strength and weaknesses of both approaches.

Appendix A Gender Words

Here, we provide the list of gender words involved in the process of gender-swapping sentence duplication. All words are also present in their capitalised version, here omitted. Additionally, 5000 of the most popular female and male names in the USA are also considered (2500 each), to be converted to an equally popular male/female name. Examples of given names are also provided.

boy $\Longleftrightarrow$ *girl*
boyfriend $\Longleftrightarrow$ girlfriend
boyfriends $\Longleftrightarrow$ *girlfriends*
boys $\Longleftrightarrow$ *girls*
brother$\Longleftrightarrow$ *sister*
brothers $\Longleftrightarrow$ *sisters*
businessman $\Longleftrightarrow$ *businesswoman*
businessmen $\Longleftrightarrow$ *businesswomen*
chairman $\Longleftrightarrow$ *chairwoman*
chairmen $\Longleftrightarrow$ *chairwomen*

congressman ⟺ *congresswoman*
congressmen ⟺ *congresswomen*
councilman ⟺ *councilwoman*
councilmen ⟺ *councilwomen*
dad ⟺ *mom*
daddy ⟺ *mommy*
dads ⟺ *moms*
father ⟺ mother
fatherhood ⟺ *motherhood*
fathers ⟺ *mothers*
fraternity ⟺ *sorority*
gentleman ⟺ *lady*
gentlemen ⟺ *ladies*
grandfather ⟺ *grandmother*
grandfathers⟺ *grandmothers*
grandpa ⟺ *grandma*
grandson⟺ *granddaughter*
grandsons ⟺ *granddaughters*
guy⟺ *gal*
guys ⟺ *gals*
he ⟺ *she*
him ⟺ *her*
himself ⟺ *herself*
his ⇐ *hers*
his ⇒ *her*
husband ⟺ *wife*
husbands ⟺ *wives*
king ⟺ *queen*
kings ⟺ *queens*
male ⟺ *female*
males ⟺ *females*
man ⟺ *woman*
men ⟺ *women*
mr ⟺ *mrs*
nephew ⟺ *niece*
nephews⟺ *nieces*
pa⟺ *ma*
paternity ⟺ *maternity*
prince⟺ *princess*
princes ⟺ *princesses*
schoolboy⟺ *schoolgirl*
schoolboys ⟺ *schoolgirls*
son ⟺ *daughter*
sons ⟺ *daughters*
spokesman ⟺ *spokeswoman*
spokesmen⟺ *spokeswomen*
stepfather⟺ *stepmother*
stepfathers⟺ *stepmothers*
stepson⟺ *stepdaughter*
stepsons⟺ *stepdaughters*
uncle⟺ *aunt*
uncles⟺ *aunts*
James ⟺ *Mary*
John ⟺ *Patricia*
Robert ⟺ *Elizabeth*
Michael ⟺ *Jennifer*
William ⟺ *Linda*
David ⟺ *Barbara*
Richard ⟺ *Margaret*
Joseph ⟺ *Susan*
Charles ⟺ *Dorothy*
Thomas ⟺ *Jessica*
Christopher ⟺ *Sarah*
Daniel ⟺ *Nancy*
Matthew ⟺ *Betty*
Anthony ⟺ *Karen*
Donald ⟺ *Lisa*
Paul ⟺ *Helen*
Mark ⟺ *Sandra*
George ⟺ *Ashley*
Steven ⟺ *Emily*
Andrew ⟺ *Kimberly*
Kenneth ⟺ *Donna*
Edward ⟺ *Carol*
Joshua ⟺ *Michelle*
Kevin ⟺ *Amanda*
Brian ⟺ *Melissa*
Ronald ⟺ *Laura*
Timothy ⟺ *Anna*
Jason ⟺ *Stephanie*
Jeffrey ⟺ *Rebecca*
Ryan ⟺ *Deborah*

Appendix B STS and MultiNLI Bias Score

Here, we provide preliminary results from the analysis of STS-Benchmark [12] and MultiNLI [58], two text corpora commonly adopted for fine-tuning transformer-based models on semantic textual similarity (STS) and natural language inference (NLI) tasks [18, 29, 31, 49, 50, 52]. For each entry, we estimate the amount of female and male bias with *bias score*. Tables 14 and 15 report the results for STS-Benchmark and MultiNLI, respectively. STS-Benchmark consists of 5749 entries containing two sentences each. For MultiNLI, we limited our analysis to the first 10000 entries.

Table 14 Top ten most biased sentences in both female and male directions in STS-benchmark according to bias score

Sentence	Bias Score
France to ban child beauty pageants	0.08772
Couple with newborn baby	0.08156
A dancer posing for the camera in a red and white dress	0.07757
A baby is playing with a doll	0.07533
A kitten is drinking milk	0.07189
A girl in a purple shirt and pink headband posing	0.06948
A baby in a red hat sitting in a stroller is holding a doll	0.06924
A tiny blonde child in a blue dress sits on a table near her mother	0.06922
A cat is licking from a saucer of milk	0.06875
A girl is wearing a purple sash and matching skirt	0.06301
Capello resigns as England manager	−0.12177
A baseball player hitting the ball	−0.11509
Capello quits as England manager	−0.10839
West Ham beats Newcastle 2-0 in Premier League	−0.10711
A baseball player throws the ball	−0.10576
Obama takes offensive against Romney in debate	−0.10341
Tampa Bay manager Lou Piniella, bench coach John McLaren and right fielder Aubrey Huff were ejected for arguing after Huff was called out on strikes to end the ninth	−0.10028
A shirtless man catches a football	−0.10253
Gunners fire but fail to advance in Champions League	−0.09914
Russians rally against Putin's rule	−0.09817

Table 15 Top ten most biased sentences in both female and male directions in MultiNLI according to bias score

Sentence	Bias Score
I wore a swimsuit	0.08836
She looked hideous in her frumpy gown	0.08015
yeah exactly be everything be supermom exactly	0.07856
Cooking is one of my passions – particularly baking delectable goodies!	0.07431
And knitting and	0.07424
Lewinsky supposedly would wear a sexy dress to lure the president's eyes	0.07365
But the blinking eyes in his mechanical ballet are heavy with mascara, while the sexy mouth shines with lipstick	0.07203
Lipstick lesbians wear heavier makeup than straight women	0.06573
Martha is finally being treated as the CEO of a company called Omnimedia, not as a bitchy hausfrau	0.06495
and quilting and um	0.06431
Parcells' teams commit fewer penalties than almost any team in the league	−0.11650
The NFL teams consisted only of competent players	−0.11425
The Sox have never played the Indians	−0.10777
Senor Juanito said that	−0.10611
Some famous players spoke out against the majority's judgment that Woods is the best player in the world	−0.10507
It was dedicated the the legendary Hulk Hogan	−0.10423
And if things go well, you were behind the commander in chief all the way	−0.10004
The book's introduction was drafted by Appiah	−0.09771
okay pro football i like two teams one the New York Giants and the second is the Raiders	−0.09731
Asoka defeated his enemies but failed as a conquerer	−0.09537

Acknowledgements We are grateful to Letizia Tanca for her advice during the definition of this study and the continuous support during the experimental and writing phase of this work.

Authors' contributions All authors contributed to the conception and design of the work and its methodology. Data analysis, coding and execution of the experiments were performed by Tommaso Dolci and FA, with substantial support from MT. All authors contributed to the definition of the experiments and the interpretation of the results. The first draft of the manuscript was written by TD and all authors contributed on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Not applicable.

Availability of data and materials All data used in this work are referenced and publicly available on either the official website of the corresponding project, or the original authors' repository. In particular, the dataset mainly analysed in this paper, i.e., the Stanford Natural Language Inference corpus (SNLI), is available on the Stanford NLP Group website, https://nlp.stanford.edu/projects/snli/SentEval, alongside the training corpora used in their system, can be downloaded by following the authors' instructions available in the following repository: https://github.com/facebookresearch/SentEval

Declarations

Conflict of interest Not applicable.

References

1. Agirre E, Banea C, Cardie C, et al (2014) Semeval-2014 task 10: Multilingual semantic textual similarity. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 81–91
2. Basta C, Costa-jussà MR, Casas N (2019) Evaluating the underlying gender bias in contextualized word embeddings. In: Proceedings of the first workshop on gender bias in natural language processing (GeBNLP), pp 33–39
3. Bender EM, Friedman B (2018) Data statements for natural language processing: toward mitigating system bias and enabling better science. Transact Associat Comput Linguist (TACL) 6:587–604
4. Bender EM, Gebru T, McMillan-Major A, et al (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT), pp 610–623
5. Bhatia V, Rawat P, Kumar A, et al (2019) End-to-end resume parsing and finding candidates for a job description using bert. arXiv preprint arXiv:1910.03089
6. Blodgett SL, Barocas S, Daumé III H, et al (2020) Language (technology) is power: A critical survey of "bias" in nlp. In: Proceedings of the 58th annual meeting of the association for computational linguistics (ACL), pp 5454–5476
7. Bolukbasi T, Chang KW, Zou JY, et al (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Adv Neural Inform Process Syst. 29
8. Bordia S, Bowman SR (2019) Identifying and reducing gender bias in word-level language models. In: Proceedings of the 2019 Conference of the north american chapter of the association for computational linguistics: Human Language Technologies - Student Research Workshop (NAACL-HLT SRW), pp 7–15
9. Bowman SR, Angeli G, Potts C, et al (2015) A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)
10. Brunet ME, Alkalay-Houlihan C, Anderson A, et al (2019) Understanding the origins of bias in word embeddings. In: International conference on machine learning (ICML), pp 803–811
11. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186
12. Cer D, Diab M, Agirre E, et al (2017) Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055
13. Cer D, Yang Y, Kong Sy, et al (2018) Universal sentence encoder. arXiv preprint arXiv:1803.11175
14. Chang KW, Prabhakaran V, Ordonez V (2019) Bias and fairness in natural language processing. In: Proceedings of the 2019 conference on empirical methods in natural language processing and 9th international joint conference on natural language processing (EMNLP-IJCNLP): Tutorial Abstracts
15. Cheng P, Hao W, Yuan S, et al (2021) Fairfil: Contrastive neural debiasing method for pretrained text encoders. In: International conference on learning representations (ICLR)
16. Conneau A, Kiela D (2018) Senteval: An evaluation toolkit for universal sentence representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC)
17. Conneau A, Kiela D, Schwenk H, et al (2017) Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP), pp 670–680
18. Devlin J, Chang MW, Lee K, et al (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human Language Technologies (NAACL-HLT), pp 4171–4186
19. Dolan W, Quirk C, Brockett C, et al (2004) Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th international conference on computational linguistics (COLING)
20. Dolci T (2022) Fine-tuning language models to mitigate gender bias in sentence encoders. In: 2022 IEEE eighth international conference on big data computing service and applications (BigDataService), IEEE, pp 175–176
21. Du Y, Fang Q, Nguyen D (2021) Assessing the reliability of word embedding gender bias measures. In: Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP), pp 10,012–10,034
22. Garrido-Muñoz I, Montejo-Ráez A, Martínez-Santiago F et al (2021) A survey on bias in deep nlp. Appl Sci 11(7):3184
23. Gira M, Zhang R, Lee K (2022) Debiasing pre-trained language models via efficient fine-tuning. In: Proceedings of the second workshop on language technology for equality, diversity and inclusion (LT-EDI), pp 59–69

24. Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies (NAACL-HLT), pp 609–614
25. Gupta U, Dhamala J, Kumar V et al (2022) Mitigating gender bias in distilled language models via counterfactual role reversal. Find Associat Comput Linguist: ACL 2022:658–678
26. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 168–177
27. Kaneko M, Bollegala D (2019) Gender-preserving debiasing for pre-trained word embeddings. In: proceedings of the 57th annual meeting of the association for computational linguistics (ACL), pp 1641–1650
28. Kaneko M, Bollegala D (2021) Debiasing pre-trained contextualised embeddings. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics (EACL), pp 1256–1266
29. Lan Z, Chen M, Goodman S, et al (2020) Albert: A lite bert for self-supervised learning of language representations. In: 8th international conference on learning representations (ICLR)
30. Liang PP, Li IM, Zheng E, et al (2020) Towards debiasing sentence representations. In: Proceedings of the 58th annual meeting of the association for computational linguistics (ACL), pp 5502–5515
31. Liu Y, Ott M, Goyal N, et al (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
32. Lu K, Mardziel P, Wu F, et al (2020) Gender bias in neural natural language processing. In: logic, language, and security. Springer, p 189–202
33. MacAvaney S, Yates A, Cohan A, et al (2019) Cedr: Contextualized embeddings for document ranking. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 1101–1104
34. Manzini T, Chong LY, Black AW, et al (2019) Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, pp 615–621
35. Marelli M, Menini S, Baroni M, et al (2014) A sick cure for the evaluation of compositional distributional semantic models. In: Proceedings of the ninth international conference on language resources and evaluation (LREC), pp 216–223
36. May C, Wang A, Bordia S, et al (2019) On measuring social biases in sentence encoders. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies (NAACL-HLT), pp 622–628
37. Meade N, Poole-Dayan E, Reddy S (2022) An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In: Proceedings of the 60th annual meeting of the association for computational linguistics (ACL), pp 1878–1898
38. Mikolov T, Chen K, Corrado G, et al (2013a) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
39. Mikolov T, Sutskever I, Chen K, et al (2013b) Distributed representations of words and phrases and their compositionality. Adv Neural Inform Process Syst 26
40. Mikolov T, Yih WT, Zweig G (2013c) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies (NAACL-HLT)
41. Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL), pp 271–278
42. Pang B, Lee L (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL), pp 115–124
43. Park JH, Shin J, Fung P (2018) Reducing gender bias in abusive language detection. In: proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)
44. Park S, Shim HS, Chatterjee M, et al (2014) Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In: Proceedings of the 16th international conference on multimodal interaction, pp 50–57
45. Parraga O, More MD, Oliveira CM, et al (2022) Debiasing methods for fairer neural models in vision and language research: A survey. arXiv preprint arXiv:2211.05617
46. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
47. Poria S, Hazarika D, Majumder N, et al (2019) Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 527–536
48. Qian Y, Muaz U, Zhang B, et al (2019) Reducing gender bias in word-level language models with a gender-equalizing loss function. In: Proceedings of the 57th Annual meeting of the association for computational linguistics: student research workshop (ACL SRW), pp 223–228
49. Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 3982–3992
50. Reimers N, Gurevych I (2020) Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP). association for computational linguistics, arXiv: 2004.09813
51. Rudinger R, May C, Van Durme B (2017) Social bias in elicited natural language inferences. In: Proceedings of the first workshop on ethics in natural language processing (EthNLP), pp 74–79
52. Sanh V, Debut L, Chaumond J, et al (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108
53. Socher R, Perelygin A, Wu J, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP), pp 1631–1642
54. Stoyanovich J, Howe B, Jagadish H (2020) Responsible data management. Proceedings of the VLDB Endowment 13(12)
55. Voorhees EM, Tice DM (2000) Building a question answering test collection. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp 200–207
56. Wang T, Lin XV, Rajani NF, et al (2020) Double-hard debias: Tailoring word embeddings for gender bias mitigation. In: Proceedings of the 58th annual meeting of the association for computational linguistics (ACL), pp 5443–5453
57. Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Language Resour Evaluat 39(2):165–210
58. Williams A, Nangia N, Bowman S (2018) A broad-coverage challenge corpus for sentence understanding through inference. In:

Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, Vol 1 (Long Papers), pp 1112–1122
59. Xu G, Hu Q (2022) Can model compression improve nlp fairness. arXiv preprint arXiv:2201.08542
60. Yang Z, Feng J (2020) A causal inference method for reducing gender bias in word embedding relations. In: Proceedings of the AAAI conference on artificial intelligence, pp 9434–9441
61. Ye W, Xu F, Huang Y, et al (2021) Adversarial examples generation for reducing implicit gender bias in pre-trained models. arXiv preprint arXiv:2110.01094
62. Zhao J, Wang T, Yatskar M, et al (2018a) 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 language technologies (NAACL-HLT), pp 15–20
63. Zhao J, Zhou Y, Li Z, et al (2018b) Learning gender-neutral word embeddings. In: proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP), pp 4847–4853