Reducing Word Embedding Bias Using Learned Latent Structure

Harshit Mishra
Syracuse University, NY
hamishra@syr.edu

Abstract

Word embeddings learned from collections of data have demonstrated a significant level of biases. When these embeddings are used in machine learning tasks it often amplifies the bias. We propose a debiasing method that uses (Figure 1) a hybrid classification - variational autoencoder network. In this work, we developed a semi-supervised classification algorithm based on variational autoencoders which learns the latent structure within the dataset and then based on learned latent structure adaptively re-weights the importance of certain data points while training. Experimental results have shown that the proposed approach works better than existing SoTA methods for debiasing word embeddings.

1 Introduction

Word embedding is a framework that represents text data as vectors that can be used in many natural language processing tasks such as sentiment analysis [Shi *et al.*, 2018], dialogue generation [Zhang *et al.*, 2018], and machine translation [Zou *et al.*, 2013]. There has been major development over the years in word representation learning [Devlin *et al.*, 2019], [Peters *et al.*, 2018], [Pennington *et al.*, 2014], [Mikolov *et al.*, 2013] which has made word embeddings an essential framework for NLP tasks but they are also not without biases. Word embeddings learned from large amounts of text data have demonstrated a significant level of gender, racial and ethnic biases which in turn impact downstream NLP applications [Bolukbasi *et al.*, 2016], [Caliskan *et al.*, 2017].

AI models should be fair, unbiased, and need to be consistently monitored to make sure any person is not being discriminated against due to a bias present in an AI system [Holstein *et al.*, 2019]. It is no secret that a model is only as good as the data it is trained on and a model trained on biased data will lead to biased algorithmic decisions. Microsoft's AI chatbot Tay [Telegraph, 2016] learned an abusive language from twitter within 24 hours of its release. Previous research has shown that word embedding reflects human-like biases with respect to gender, profession, and ethnicity [Bolukbasi *et al.*, 2016]. For example, "*doctor*", "*programmer*" are considered to be male-related terms have shown to be stereotypically male-biased, whereas "*nurse*", "*homemaker*" are considered to be female-related terms have shown to be stereotypically female-biased. It is important to make sure that word embeddings are debiased before they are used by any machine learning tasks but a debiased word embedding should still retain necessary semantic information (like the vector for *king* should still be close to the vector for *man* and the vector for *queen* should still be close to the vector for *woman*), to be useful for a NLP task while removing information related to discriminative biases.

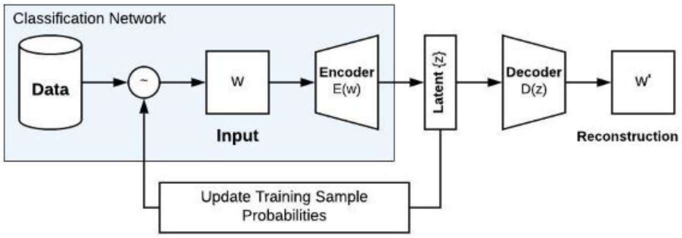

Figure 1: Debiasing Variational AutoEncoder Network. Hybrid architecture combining classifier network along with VAEs

In this paper, we propose a method to debias word embedding using learned latent structure and a training process that adapts in an unsupervised manner to the shortcomings of underrepresented data. This approach learns gender-related information, neutral word information, and stereotypical bias through an underlying latent structure of training data. The algorithm is a combination of classifier and variational autoencoder capable of identifying rare data points in the training dataset.

2 Related Work

[Bolukbasi *et al.*, 2016] proposed an algorithm that achieves fairness by modifying the underlying data. The algorithm focuses on projecting gender-neutral words to a subspace orthogonal to gender direction identified by gender-definitional words. Here, words such as *she,he, daughter, son* are gender-definitional words, rest of the words are gender-neutral and gender direction is identified by combining directions such as $\vec{she}$ - $\vec{he}$ and $\vec{woman}$ - $\vec{man}$. They proposed *hard-*

debiasing and *soft-debiasing* methods with slightly different approaches. In *hard-debiasing* all gender-neutral words are projected to a subspace orthogonal to gender direction. The *soft-debiasing* method preserves inner-products between original word embeddings while also projecting word embeddings into a subspace orthogonal to gender direction. Both methods rely on a SVM classifier to get an expanded list of gender-definitional words and if the classifier incorrectly predicts a stereotypical word as a gender definitional word then it would not get debiased, as gender direction is only evaluated once and remains same for rest of the debiasing process.

[Zhao *et al.*, 2018b] proposed a learning scheme, Gender-Neutral Global Vectors (GN-GloVe) for training word embedding models based on GloVe [Pennington *et al.*, 2014]. This algorithm works by protecting attributes in certain dimensions while neutralizing the others during training. It adds a constraint such that gender-related information is confined to a subvector. During training, a gender-related subvector is maximized while minimizing the neutral words subvector. GN-GloVe learns word embedding from a given corpus and can not be used to debias pre-trained word embeddings.

[Kaneko and Bollegala, 2019] showed a way to use autoencoders to debias word embeddings by changing the way networks process data. They formed 4-dimensional vectors where the dimensions refer to female words, male words, stereotypical words, and gender-neutral words. This vector provides a way for a network to learn classification between male, female words, gender-neutral and stereotypical words. Simultaneously, an autoencoder learns to keep as much semantic knowledge as possible by using a reconstruction loss through the decoder. Due to imbalance in available data of different types they have re-used some words to make up inputs for training. Although this focuses on the class imbalance problem it doesn't use any information available from the structure of latent features [Amini *et al.*, 2019]. It leads to a situation where the type of words more in number are trained much more than the type of words that are less in number.

3 Proposed Method

The work in this paper follows the approach of [Kaneko and Bollegala, 2019] but adds elements from [Amini *et al.*, 2019] which was originally used for image datasets. Auto sampling refers to the process of increasing the relative frequency of rare data points by an increased sampling of underrepresented regions of latent space. Using auto sampling with the approach of [Kaneko and Bollegala, 2019] we get a more robust debiasing structure. The method in [Kaneko and Bollegala, 2019] to propagate data through a network, as a tuple of (female words, male words, stereotypical words, gender-neutral words) and using an architecture similar to [Amini *et al.*, 2019], the encoder portion of the debiasing variational autoencoder network outputs d latent variables given a data point. Two classifiers - male (C_m) and female (C_f) are applied to latent variables of male and female words respectively in order to retain gender-related information. The encoder outputs d activations corresponding to $\mu \in R^{d/2}$,Σ = $Diag[\sigma^2] > 0$ which are used to define the distribution z and the d-dimensional output **w'**. A decoder network is then used to reconstruct the input back from latent space, z. This decoded reconstruction enables unsupervised learning of latent variables during training. Our network is trained end-to-end using backpropagation with a five-component loss.

3.1 Problem Setup

For the remainder of the paper, we will use GloVe embedding [Pennington *et al.*, 2014] as input. This dataset is in a form of tuples having *(female words, male words, stereotypical words, gender-neutral words)* and batches will be drawn from this dataset and sent to our debiased variational autoencoder network. Our output should be able to retain semantic information while removing gender biases present in the dataset. Our goal is to show that we can efficiently mitigate discriminative biases present in word embeddings using adaptive training sample probabilities and a decoder output based on learned latent distribution.

3.2 Formulation

Based on [Kaneko and Bollegala, 2019] approach, we have a pre-trained d-dimensional word embedding having a set of vocabulary V. Our focus is to translate

$$E\colon R^{\mathrm{d}} \to R^{\mathrm{l}}$$

that projects the original word embedding to l-dimensional latent space. No prior information about the pre-trained word embedding or corpora has been used during the construction of this model. Therefore the proposed method can be used solely or as a part of a more complex architecture to debias word embeddings. We propose a debiasing method that models the interactions between values of the protected attribute (male, female, gender) and whether there is a stereotypical bias or not. Given four set of words: *feminine (V_f), masculine (V_m), neutral (V_n), stereotype (V_s)* our proposed method learns a projection that satisfies the following four criteria:

1. For $w_{\mathrm{f}} \in V_f$ we protect its feminine properties
2. For $w_{\mathrm{m}} \in V_m$ we protect its masculine properties
3. For $w_{\mathrm{n}} \in V_n$ we protect its gender properties
4. For $w_{\mathrm{s}} \in V_s$ we remove its gender biases

To explain the proposed method, let's consider a feminine regressor, C_f that predicts the degree of feminineness of the word w, where highly feminine words are assigned values close to 1. Similarly, a masculine regressor, C_m predicts the degree of masculinity of the word w, where highly masculine words are assigned values close to 1. We then learn the debiasing function as Encoder $E\colon R^{\mathrm{d}} \to R^{\mathrm{l}}$ that projects original pre-trained word embedding to a debiased l-dimensional space and Decoder $D\colon R^{\mathrm{l}} \to R^{\mathrm{d}}$ which projects the debaised output back into d-dimensional space.

The way this method differs from [Kaneko and Bollegala, 2019] and other previous word embedding debiasing methods is we are forming a learned latent distribution z from an encoder output. This latent distribution goes through a sampling step, where latent parameters are picked from encoder output based on gaussian distribution using the reparametrization trick (refer equation 3). This learned latent distribution is also used to sample rare data points more frequently during training.

We train the network end to end using backpropagation with a 5-component loss, comprised of male classifier loss, female classifier loss, latent loss, gender-neutral loss, and reconstruction loss. For feminine words and masculine words, we require the encoded space to retain the gender-related information. The squared losses L_f and L_m are defined as:

$$L_f = \sum_{w \in V_f} ||C_f(E(w)) - 1||_2^2 \tag{1}$$

and

$$L_m = \sum_{w \in V_m} ||C_m(E(w)) - 1||_2^2 \tag{2}$$

Then for the latent structure in VAEs, the encoder outputs d output dimensions which are then divided equally into μ,Σ. VAEs utilize reparameterization to differentiate the outputs through a sampling step, where we sample $\epsilon \in (0, 1)$ and compute z, our sampled encoder output as:

$$z = \mu + e^{\left(\frac{1}{2} \cdot \log \Sigma\right)} \circ \epsilon \tag{3}$$

and Kullback-Leibler loss, L_{KL} is given as:

$$L_{KL}(\mu, \sigma) = \frac{1}{2} \cdot \sum_{w \in V} (\sigma_w + \mu_w^2 - 1 - \log \sigma_w) \tag{4}$$

For the stereotypical and gender-neutral words, we need vectors to be embedded into a subspace orthogonal to the gender directional vector. Let Ω be the set of word pairs formed from ($\mathrm{w_f}$,$\mathrm{w_m}$) words. Our gender vector v_g is defined as:

$$v_g = \frac{1}{|\Omega|} \sum_{(w_f, w_m) \in \Omega} (E(w_m) - E(w_f)) \tag{5}$$

Prior work has shown that vector difference between embeddings of male and female word pairs accurately represents the gender direction [Bolukbasi *et al.*, 2016], [Zhao *et al.*, 2018b]. We keep v_g fixed though each epoch and re-estimate between each epoch. Squared inner product between v_g and gender-neutral words or stereotypical words L_g

$$L_g = \sum_{w \in V_n \cup V_s} (v_g^T w)^2 \tag{6}$$

It is important that we preserve the semantic information encoded in the word embedding as much as possible. For this purpose we minimize the reconstruction loss, L_r for autoencoder given by :

$$L_r = \sum_{w \in V} ||D(z) - w||^2 \tag{7}$$

Finally, we define total objective as linearly weighted sum of the above-defined losses as given by:

$$L = \lambda_f L_f + \lambda_m L_m + \lambda_{KL} L_{KL} + \lambda_g L_g + \lambda_r L_r \tag{8}$$

Here the coefficients, λ's are non-negative hyperparameters. We can modulate these parameters to determine importance for each loss.

4 Experiment

4.1 Data Collection

We use the feminine and masculine wordlist created by [Zhao *et al.*, 2018a] as w_f, w_m . To create gender-neutral and stereotypical word list we use the dataset created by [Kaneko and Bollegala, 2019] as w_n, w_s . For gender-neutral words, this dataset has a list of 1031 gender-neutral words and stereotypical word list contains a list of professions associated with one type of gender created by [Bolukbasi *et al.*, 2016]. Pretrained glove embedding obtained is a 300-dimensional word embedding for 322636 unique words.

For example, the feminine and masculine word list contain words such as (*[countrywoman, countryman],[witches, wizards],[actress, actor]*). stereotype word list contains words such as (*[aerobics, tycoon, beauty, colonel, romantic]*), and gender-neutral word list contain words such as (*[abandonment, best, cold, stone]*). The dataset contains 222 gender pair words (pair of feminine and masculine words), 84 stereotypical word pairs, and 1031 gender-neutral words.

4.2 Experiment Setup

Hyperparameter values are set to be λ_f= λ_m= λ_g =λ_{KL}= 0.0001 and λ_r = 1.0 . More penalty is given to reconstruction loss as we want our decoder to accurately reconstruct original word embedding to keep as much semantic knowledge as possible. The training dataset is fed through our encoder network, which provides an estimate of data points based on the frequency distribution of each of the latent variables, then we increase the relative frequency of rare data points by an increased sampling of the underrepresented region of latent space. To do so, we create a histogram of latent distribution and then we sample latent distributions from histograms where density is low. In the dataset of w_f, w_m, w_n, w_s words, there is an imbalance of one type of words being in more in number than other and as we are using a tuple (w_f, w_m, w_n, w_s) for the training, we have reused some w_f, w_m words so that we can make complete tuples for the training

4.3 Implementation

C_f and C_m are both implemented as feed-forward neural networks with one hidden layer and the sigmoid function is used as the nonlinear activation. Both the encoder E and the decoder D of the autoencoder are implemented as feed-forward neural networks with two hidden layers. Hyperbolic tangent is used as the activation function throughout the autoencoder.

At each epoch, all inputs *w* from the original dataset are propagated through a model to evaluate corresponding latent variables *z(w)*. The histograms are updated accordingly. A probability estimate is assigned to each input based on the histogram density of latent variables. During training, we sample a new batch of inputs (batch size is maintained at 32) *w* based on inverted probability estimate, inverted because we want to increase the resampling probability of rare data points. Training on new debiased data batch forces classifier to learn parameters that work better in rare cases with a strong deterioration of performance for common training examples. This sampling is not specified beforehand but purely based on

	SemBias			SemBias-subset		
Embedding	Definition	Stereotype	None	Definition	Stereotype	None
GloVe	80.2	10.9	8.9	57.5	20	22.5
Hard-GloVe	84.1	9.5	**6.4**	25	47.5	27.5
GP (GloVe)	84.3	8.0	7.7	65	15	20
VAE (GloVe)	**87.5**	**5.4**	7.5	**80**	**7.5**	**12.5**

Table 1: Prediction accuracies for gender relational analogies

learned latent variables. The model is trained 25 times and 5 times completely from scratch. Pretraining autoencoders and classifiers help in achieving optimization. We keep the architecture same but train autoencoders and classifiers separately for 300 epochs.

4.4 Evaluating Debiasing Performances

The model is based on GloVe embedding and a list of gender-neutral, stereotypical, male, and female words curated by previous works. We are using 300-dimensional pre-trained word embedding and hidden dimensions are also set to 300 dimensions to obtain 300-dimensional de-biased word embedding.

Baselines and comparisons

1. GloVe- is the pre-trained word embedding used to get a debias embedding
2. Hard-GloVe- is the implementation of hard-biasing by [Bolukbasi *et al.*, 2016] by the authors of [Kaneko and Bollegala, 2019]
3. GP(GloVe) - is the debiased word embedding obtained from [Kaneko and Bollegala, 2019] research
4. **VAE(GloVe)** - is the debiased embedding obtained from our model.

There is a SemBias dataset created in previous work by [Zhao *et al.*, 2018a] to evaluate the level of gender bias in word embeddings. Each instance in SemBias consists of four-word pairs: a gender definition word pair (e.g – waiter, waitress), a gender-stereotype word pair (e.g. – doctor-nurse) and two other word pairs that have similar meanings but no gender relation (e.g. – dog-cat, cup-lid) which evaluates embedding based upon definition, stereotype, and no gender words. SemBias contains 20 gender-stereotype word pairs and 22 gender-definitional word pairs and uses their Cartesian product to generate 440 instances. Among the 22 gender-definitional word pairs, 2 word-pairs are not used as the seeds for training. to test the generalisability of a debiasing method, we use the sub-set (SemBias subset) of 40 instances associated with these 2 pairs. We measure the relational similarity between (he, she) word-pair and a word-pair (a,b) in SemBias using the cosine similarity between the he-she gender directional vector and a-b using the word embeddings under evaluation. For the four word-pairs in each instance in SemBias, we select the word pair with the highest cosine similarity with he-she as the predicted answer. If the word embeddings are correctly debiased, we would expect a high accuracy for definitions and low accuracies for stereotypes and none's. As we see, the proposed **VAE(GloVe)** method achieves high semantic, definition scores while keeping stereotypical and none loss to a minimum.

In previous works, either method has achieved high scores in SemBias set but not able to replicate the same accuracies in SemBias-subset or they have used already debiased word embedding from other methods. The proposed method is able to achieve high accuracies both on SemBias and SemBias-subset using only an unbiased pre-trained word embedding with an easy to use, simple architecture that can be modified based on a task at hand.

5 Future Work

While the embedding achieved from the proposed method works well based on high-level analysis. More research is needed to see how it works with low-level NLP tasks. i.e to see how gender-neutral words change their position while debiased, as we do not want to adversely change their orientation with respect to gender definition words and more complex architecture can be used depending upon downstream applications.

6 Conclusion

We propose a method to remove gender specific-biases from pre-trained word embeddings. Experimental results show that the proposed method can accurately debias pre-trained embeddings, outperforming previous methods while preserving useful semantic information. Similar work can also be done to reduce racial and religious biases. Similarly, it is also interesting to research on biases present in other languages.

References

[Amini *et al.*, 2019] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 289–295, New York, NY, USA, 2019. Association for Computing Machinery. https://doi.org/10.1145/3306618.3314243.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016. http://arxiv.org/abs/1607.06520.

[Brunner *et al.*,] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Michael Weigelt. Disentangling the latent space of (variational) autoencoders for nlp. https://tik-old.ee.ethz.ch/file/9b24a347a3c0b172470c2b800acdf4f6/UKCI2018_DisentVAE_CR.pdf.

[Caliskan *et al.*, 2017] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. https://science.sciencemag.org/content/356/6334/183.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers forlanguage understanding. 2019. https://arxiv.org/pdf/1810.04805.pdf.

[Elazar and Goldberg, 2018] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *CoRR*, abs/1808.06640, 2018. http://arxiv.org/abs/1808.06640.

[Holstein *et al.*, 2019] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 2019. http://dx.doi.org/10.1145/3290605.3300830.

[Kaneko and Bollegala, 2019] Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. https://www.aclweb.org/anthology/P19-1160.pdf.

[Mikolov *et al.*, 2013] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 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*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. "https://www.aclweb.org/anthology/N13-1090".

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. https://www.aclweb.org/anthology/D14-1162.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. 2018. https://arxiv.org/pdf/1802.05365.pdf.

[Schnabel *et al.*, 2009] Konrad Schnabel, Jens Asendorpf, and Anthony Greenwald. Assessment of individual differences in implicit cognition a review of iat measures. *European Journal of Psychological Assessment*, 24:210–217, 01 2009.

[Shi *et al.*, 2018] Bei Shi, Zihao Fu, Lidong Bing, and Wai Lam. Learning domain-sensitive and sentiment-aware word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2494–2504, Melbourne, Australia, July 2018. Association for Computational Linguistics. https://www.aclweb.org/anthology/P18-1232.

[Telegraph, 2016] The Telegraph. Microsoft deletes 'teen girl' ai after it became a hitlter-loving sex robot within 24hours. 2016. https://goo.gl/mE8p3J.

[Zhang *et al.*, 2018] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. https://www.aclweb.org/anthology/P18-1205.

[Zhao *et al.*, 2018a] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. https://www.aclweb.org/anthology/N18-2003.

[Zhao *et al.*, 2018b] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *CoRR*, abs/1809.01496, 2018. http://arxiv.org/abs/1809.01496.

[Zou *et al.*, 2013] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1141.