Sentiment Sensitive Debiasing: A Learning-Based Approach to Remove Ethnic Stereotypes in Word Embeddings

Audhav Durai *2023adurai@tjhsst.edu*
Thomas Jefferson High School for Science and Technology
Alexandria, 22312, United States

Aditya Vasantharao *2023avasanth@tjhsst.edu*
Thomas Jefferson High School for Science and Technology
Alexandria, 22312, United States

Sauman Das *2023sdas@tjhsst.edu*
Thomas Jefferson High School for Science and Technology
Alexandria, 22312, United States

Abstract

Word vectorization models are used to represent vocabulary in a vector space in a manner that captures semantic relationships between words. However, the state-of-the-art word vectorization models are shown to contain biases in their word embeddings due to ethnic prejudices and underrepresentation in the corpora they are trained on. This paper proposes a novel sentiment-sensitive, learning-based debiasing algorithm for multiclass bias mitigation. In this study, this algorithm is used for ethnic debiasing in CBOW Word2Vec models. Unlike other debiasing algorithms, this methodology accounts for the fact that not all ethnic correlations are biased and proper debiasing should also preserve unbiased ethnic information, such as cultural knowledge. Furthermore, it does not require a pre-defined, finite set of correlations to perform debiasing. Rather, models are penalized for making ethnic correlations towards non-neutral words and are allowed to make ethnic correlations towards neutral words, performing a thorough debiasing without losing ethnic knowledge. This study also proposes a new metric to evaluate bias called S-MAC (Sentiment-Aware Mean Average Cosine Similarity) which accounts for sentiment in bias measurement. We train both the baseline and debiased CBOW models on the WikiCorpus. The Debiased model achieved a**re duction in bias by39.48%** using the S-MAC metric in comparison to the baseline model.

Keywords: Natural Language Processing; Bias Mitigation, Deep Learning, Word2Vec, Sentiment Analysis.

1. INTRODUCTION

Word embedding models (Devlin et al., 2018, Mikolov et al., 2013, Pennington et al., 2014) are extremely effective at understanding semantic and syntactic relationships. Still, research shows that these models learn biases that are present in the corpora they are trained on (Caliskan et al., 2017, Garg et al., 2018). Bias is defined as unfair prejudice against a person, place, or group (Garrido-Muñoz et al., 2021). Bolukbasi et al. (2016) show that Word2Vec models learn to correlate stereotypical professions to genders, declaring that before debiasing, women are more closely associated with “nurse” while men have a higher association to “doctor”. Manziniet al. (2013)argue that most social biases are not a binary issue like gender, showing that word embeddings contain bias towards a number of religions and races (e.g., jew is to greedy and black is to homeless). Furthermore, Manzini et al. show that word embedding models also make positive correlations towards specific social groups, such as Christian and intellectual. This is also an example of a bias because, regardless of positivity, it is a preconceived notion based on a

social group. Caliskan et al. (2017) evaluated racial bias by comparing semantic nearness for European American names and African American names. They found that European American names were correlated more with pleasant terms than African American names. Jentzsch et al. (2019) conducted similar research with common female and male names and found biased gender stereotypes. The bias in word embeddings can also be amplified in common NLP tasks, especially biases against underrepresented groups (Barocas and Selbst, 2016; Zhao et al., 2017).

Existing debiasing methods primarily focus on debiasing word embedding models after learning, doing so by minimizing word projection towards the social subspace that is being debiased (such as gender) (Bolukbasi et al. 2016, Popović et al.,2020, Hube et al.,2020, Kumar et al., 2020).However, learning-based debiasing methods have been shown to prevent the word embedding models from learning the social biases in the first place (Zhao et al., 2018, Lu et al., 2018, Bordia and Bowman, 2019, Maudslay et al., 2019).Yet, all of these methods perform dual-classed, gender debiasing. Our study proposes a learning-based multiclass bias mitigation methodology, where the model is penalized for making biased correlations, and uses this algorithm to perform ethnic debiasing. Additionally, rather than requiring a predefined set of words to debias or debiasing all words, this method uses sentiment analysis to only debias non-neutral words. The algorithm treats both positive and negative correlations as biases and restricts the model from learning such associations. This allows for a thorough debiasing without the loss of ethnic information. Lastly, our study introduces S-MAC (Sentiment-Aware Mean Average Cosine similarity), a novel bias evaluation metric that uses sentiment analysis to measure only biased correlations. A Word2Vec model trained with the debiasing algorithm described in this paper is compared to a standard Word2Vec model. **The Debiased model achieves a S-MAC score of 0.587, 39.48% lower than the standard Word2Vec model of 0.970.**

2. RELATED WORK

2.1 Bias Mitigation

Existing debiasing methodologies focus on dual-classed debiasing (primarily gender debiasing) where the two classes are man-related and woman-related. Bolukbasi et al. (2016) propose a gender debiasing strategy called “hard debiasing”. The hard debiasing algorithm first finds the gender direction using the difference vectors between gender specific words (e.g., $\overrightarrow{he} - \overrightarrow{she}$) and “neutralizes” each word vector by making it perpendicular to this direction. Next, it alter search gendered word pair (e.g., “he” and “she”) to make them equidistant to each word vector. Though this algorithm removes bias, by altering all word vectors, they also end up removing unbiased, important gender information. They determine the gendered words using an SVM trained on a few seed words. This is an unreliable methodology because it extremely reliant on the SVM generalizing to the whole vocabulary.

Manziniet al.(2019) modifies the gender debiasing procedure outlined by Bolukbasi et al.(2016) for multiclass ethnic debiasing, working around the requirement for gendered pairs by adopting a one vs. all methodology for each word in a predefined set of social groups, effectively creating pairs. This predefined set of social groups is compiled by focusing of a few social groups. Furthermore, rather than debiasing all words, they create another predefined set of words to debias. Such a predefined sets prevents the removal of unbiased social information, but it fails to thoroughly remove bias because it is unfeasible to create a list of all words that contain social bias. Furthermore, Gonen and Goldberg (2019) assert that debiasing that is reliant on word vector component removal do not sufficiently remove bias.

Lu et al. (2020) propose a data augmentation approach for learning-based debiasing, using this algorithm specifically for gender debiasing and interchange gender specific terms, such as “he” and “she”, to remove the gender bias from the corpus itself. Maudslay et al. (2019) add onto this approach by balancing first names in the corpus. This method, although effective for gender debiasing, is much more challenging for multiclass debiasing as it is harder to balance ethnic mentions over a large number of ethnic groups, some larger than others. Furthermore, it prevents

word embeddings from creating ethnic correlations that are not prejudiced and remove ethnic information from corpora as a whole. Bordia and Bowman (2019) take a difference approach to learning-based debiasing. They penalize models for projecting word embeddings in the gender direction by modifying the loss function. They choose to define a constant hyper parameter for the importance of bias mitigation and debias all words. Though we also penalize the model during learning to reduce bias, our study introduces asentiment-based penalizing, rather than a constant hyper parameter, to remove ethnic biases while retaining non-biased ethnic information.

2.2 Evaluation Metrics

Word Embedding Association Test (WEAT)

The WEAT test is a bias evaluation test for dual-class debiasing. The test requires two sets of target words (e.g., programmer, engineer, scientist; and nurse, teacher, librarian) and two sets of attribute words (e.g., man, male; and woman, female). In a perfectly Debiased model, the two sets of attribute words will have equal correlations to the two sets of target words. Equations 1 and 2show formulas that calculates the difference in association with the two sets of target words and the two sets of attributes, where X and Y are the two sets of target words and A and B are the two sets of attribute words.

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \qquad (1)$$

$$s(w, A, B) = mean_{a \in A} cos(\hat{w}, \hat{a}) - mean_{b \in B} cos(\hat{w}, \hat{b}) \qquad (2)$$

Mean Average Cosine (MAC) Test

Manzini et al.(2019) propose the MAC test as a modification of the WEAT test to measure bias in a multiclass setting. First, the average cosine similarity between target word and attribute word pairs is determined. The cosine similarity of two vectors is defined as follows:

$$\cos(t, a) = 1 - \frac{t \cdot a}{||t|| \cdot ||a||} \qquad (3)$$

In order to determine the average performance of a Word2Vec model across all such target-attribute word pairs, the MAC formula calculates the cosine similarity of all word pairs. The MAC formula can be defined as:

$$\text{MAC}(T, A) = \frac{1}{|T||A|} \sum_{T_i \in T} \sum_{A_i \in A} \cos(T_i, A_j) \qquad (4)$$

For the optimal Word2Vec model, the MAC score should be close to 0 indicating that the model places no correlation between attribute words (e.g. “black”) and target word (e.g. “criminal”).

3. MATERIALS AND METHODS

3.1 Word Embedding Model

We use the Continuous Bag-of-Word (CBOW) Word2Vec model as our word vectorization model. The CBOW model predicts each word using the representation of the context around that word. We choose the CBOW model over Skip-Gram, a similar, common Word2Vec architecture, because it trains faster and achieves better accuracy on frequently mentioned words. A sliding window, depicted in Figure 1, goes over the corpus. The model aims to predict the target word using all the other words in the sliding window as context. All models used in this study are trained with a Keras implementation using the Adam optimizer. Each model is trained using the NVIDIA Tesla P100 GPU for 100 epochs.

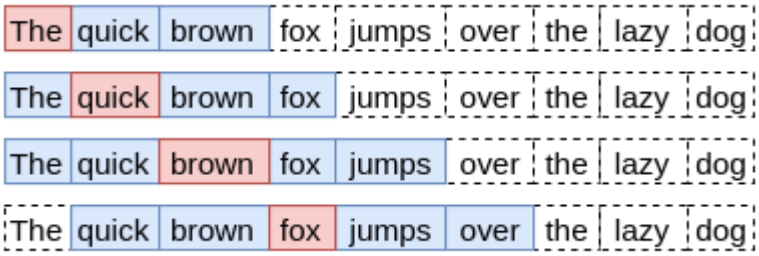

FIGURE 1: Sliding window algorithm with context size five. Context words (blue) are used to guess the target word (red).

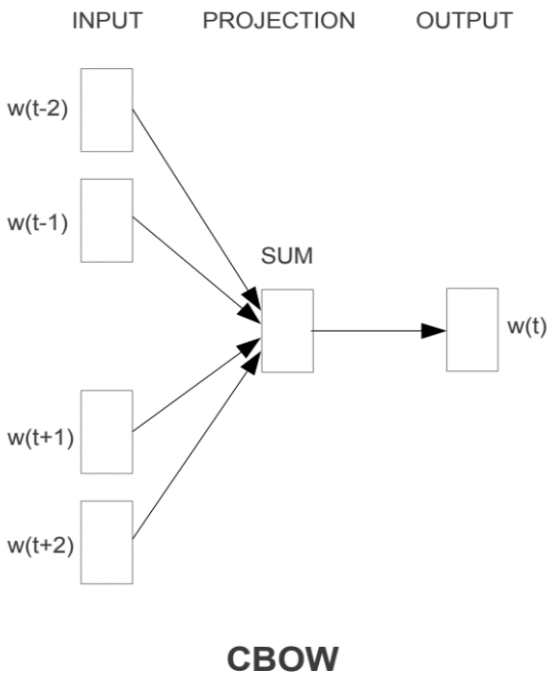

FIGURE 2: CBOW architecture (Mikolov et. al, 2013).

3.2 Dataset

We train our model on the Gensim WikiCorpus (Reese, 2010). Though many works that pertain to ethnic debiasing have elected to train with corpora for social media sites (Manzini et. al, 2019), the WikiCorpus is a dataset that is more representative of the text that state-of-the-art word embedding models are trained on. Furthermore, the WikiCorpus, made up of articles from Wikipedia, inherently contains plainer and more objective knowledge in comparison to social media sites. Nevertheless, our research shows that even when the Word2Vec model learns from such a corpus, it still learns a number of ethnic biases, meaning that there is still much bias to be mitigated in corpora that do not contain significant explicit biases. Such implicit biases are harder to detect and mitigate. Thus, we choose the WikiCorpus to demonstrate that our debiasing algorithm is also able to effectively mitigate implicit bias. Our adopted methodology is inductive in nature, as we look at specific analogies in the Word2Vec model to understand the amount of bias it contains.

Biased Ethnic Analogies	
NegativeSentiment	**Positive Sentiment**
Black→ Criminal	Black → Strong
Russian → Rebellion	Russian → Honor
Indian → Dirty	Indian → Smart
White → Racist	White → Powerful
Chinese → Weak	Chinese → Money

Table 1: Examples of ethnic biases in CBOW model trained on the WikiCorpus.

3.3 Sentiment Based Debiasing

In order to perform ethnic debiasing, we first obtain a list of allethni cities from the United States Census Bureau (Ruggles et. al, 2019).We call this list the "ethnic set" and aim to reduce bias towards these words in the Word2Vec model. To allow the model to learn accurate semantic relationships without the use of ethnic correlations, we use a learning based debiasing rather than manipulating word embeddings after training.

Current debiasing methodologies aim to mitigate bias by either reducing correlation between the ethnic set and all other words or by reducing correlation between the ethnic set and predetermined debiasing words. However, both approaches have limitations that prevent a proper debiasing. By debiasing correlations between the ethnic set and all other words, the word embedding model loses valuable ethnic information that is not biased. For instance, a correlation between Indian and rice is an example of cultural information, not bias. Though creating a predetermined debias set circumvents this issue, a predetermined debias set cannot cover all biased correlations and, thus, does not thoroughly mitigate bias.

In order to perform thorough debiasing while also preserving ethnic information, we take a sentiment-based debiasing approach. We determined that a correlation between an ethnic word to a positive or negative word is a biased correlation, while a correlation between an ethnic word and a neutral word is an unbiased correlation.

When a word in the sliding window is also in the ethnic set, we use VADER (Valence Aware Dictionary and Sentiment Reasoner) (Hutto and Glibert, 2014), a lexicon and rule-based sentiment analysis tool, to find the neutral score (one minus the positive plus negative scores) of all words in the sliding window. We denote the neutral score as α . We multiply the original cross entropy loss by α so that the loss stays the same for neutral contexts and is zero for completely non-neutral contexts. Our algorithm ensures that the model does not learn correlations between words in the ethnic set and non-neutral words while allowing it to learn correlations between ethnic words and neutral words.

$$\text{Ethnic CE Loss}(y, \hat{y}) = -\alpha \sum y_i \cdot \log(\hat{y}_i) \qquad (5)$$

3.4 Evaluation

This study proposes the S-MAC (Sentiment-Aware Mean Average Cosine Similarity) metric. The S-MAC metric is a multiclass bias evaluation metric that takes sentiment account when measuring bias. Not only does it not require a debiasing set, it also does not measure unbiased correlations. The S-MAC metric has the following requirements:

- A sentiment analysis tool (e.g. VADER)
- A set of words that are being debiased (e.g. the "ethnic set").

- A set of distinct words in the corpus that are not in the set of words to be debiased (this study refers to this set as the “debias set”)

The S-MAC metric uses the absolute cosine similarity function (defined in Section 2.2) to measure the correlation between each word in the ethnic set and each word in the debias set. The cosine similarity function returns a value between –1 and 1. A cosine similarity value of 1 means that the two words have a perfect relationship (e.g. “tree” and “tree”) and a value of –1 means that the two words have a perfect opposite relationship. A value of 0 means that the two words have no correlation at all. Since our evaluation metric is measuring all biases, positive and negative, we take the absolute values of all cosine similarity values in this study.

First, each word in the debias set is given a neutral score α calculated using VADER sentiment analysis. We then calculate the absolute cosine similarities between each ethnic word and all words in the debias set. For each ethnic word, the calculated absolute cosine similarities and neutral score for each debias word is used to take a weighted average. Equation 6 shows the formula that calculates the S-MAC score, where D is the debias set and E is the ethnic set.

$$\text{S-MAC}(D, E) = \frac{1}{|D||E|} \sum_{D_i \in D} \sum_{E_j \in E} (1 - \alpha_i) \cos(D_i, E_j) \qquad (6)$$

This weighted average represents the bias for each ethnic word in the ethnic set. To measure the total bias in a Word2Vec model, we take the mean of all measured ethnic biases.

4. RESULTS

We use the S-MAC evaluation to compare the performance of the Baseline model relative to our proposed Debiased model. We use the same ethnic set of 185 words used during the training of the Debiased model to create the target set. In Figure 3, we show the results of the S-MAC test on both models. For all words in the ethnic set, total bias is significantly reduced from the baseline model to the Debiased model. This is further supported by Figure 4, which displays the S-MAC scores for nine highly mentioned ethnicities.

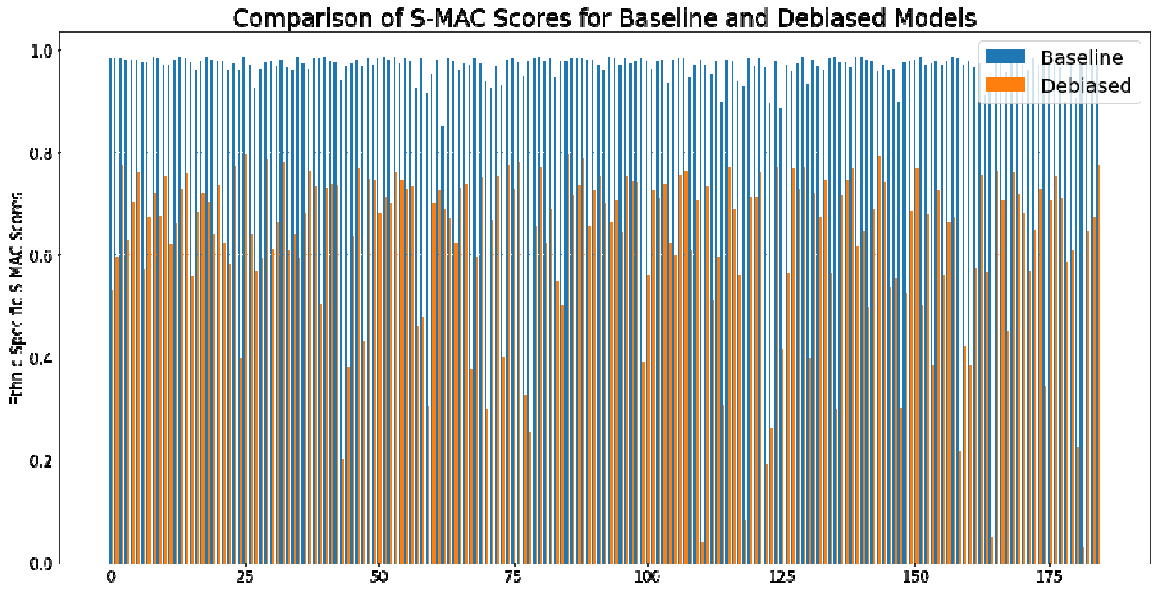

FIGURE 3: Baseline and Debiased model S-MAC scores for each of the 185 ethnic words.

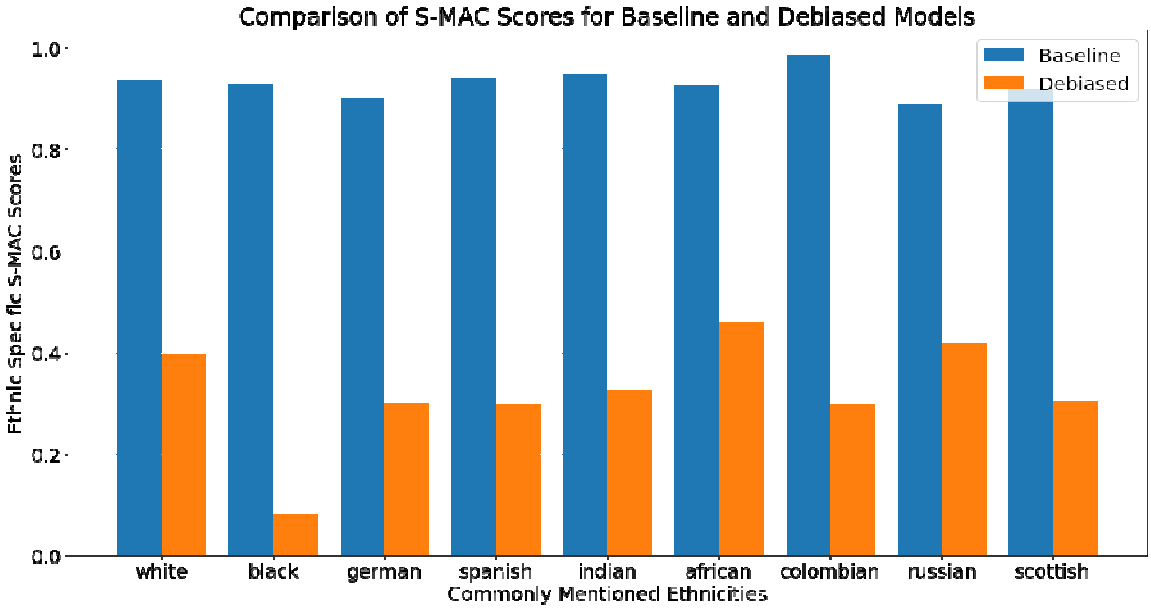

FIGURE 4: Baseline and Debiased model S-MAC scores for commonly mentioned ethnic words

Models	S-MAC
Baseline CBOW Word2Vec	0.970
DebiasedCBOW Word2Vec	**0.587**

TABLE 2: Average S-MAC score across all ethnic words.

The mean S-MAC scores across all 185 ethnic words are shown in Table 1. The Debiased CBOW model reduces the S-MAC score by 39.48%. This indicates that the model successfully reduces the correlation between ethnic words and non-neutral attributes. In comparison, Manzini et al. (2019) evaluates a best of 10.46% change in bias for religious debiasing using the MAC metric described in Section 2.3.

5. DISCUSSION

The results demonstrate that our methodology significantly reduces bias in word embeddings for all ethnicities. This is important for the both ethical and accurate performance of word embeddings in traditional downstream tasks such as part-of-speech (POS) tagging, named entity recognition (NER), POS chunking, and sentiment analysis. The ethical performance of sentiment analysis as a downstream task, in particular, is affected by bias-reduced models; as shown by Siddiqui et al. (2015), sentiment analysis techniques rely on models trained on annotated sentiment analysis corpora, which are prone to being biased and therefore prone to inducing bias in these sentiment analysis classifier models. This is especially important for sentiment analysis because many models, such as those described in Liapakis et al. (2020) and Alhazmi et al. (2013), are trained on unfiltered corpora with tens of thousands of data points to highlight discrepancies in sentiment, which make these models even more prone to bias. With regards to accurate performance on these downstream tasks, though our model modifies certain correlations, it only debiases against non-neutral words, ensuring that all correlations that truly exist are preserved.

Figure 3 shows that while all ethnicities are debiased significantly, there is fluctuation in the total amount of debiasing being done between ethnic groups. We found that the model contains less bias towards ethnic groups that were mentioned frequently in the corpus. It can be seen in Figure

4 that the Debiased model contains less bias towards the commonly mentioned ethnic groups than many of the others shown in Figure 3. This is likely because the model was unable to learn completely unbiased correlations for the lesser mentioned ethnic groups. This limitation can be remedied by training on more diverse corpora that mention all ethnic groups a sufficient amount.

6. CONCLUSION

In this work, we investigate the prevalent question in NLP of generating word embeddings. We attempt to propose a pipeline that generates such embeddings without unwanted correlations. A new training procedure for the Word2Vec model which is one of the most common tools for generating word embeddings is presented. We develop an approach that can remove bias in these word embeddings which is applicable to both binary (e.g. male/female) and multiclass (e.g. ethnicities) debiasing. We automatically extract several ethnic-related words which allowed us to reduce bias correlations for 185 ethnic words. A novel evaluation method called the S-MAC is proposed which specifically measures how well a Word2Vec model disassociates from non-neutral attribute words. Our proposed training approach caused a significant decrease in the final S-MAC scores which implies reduced bias in the generated word vectors.

Although we modified the CBOW Word2Vec training procedure in this work, there are several other methods used in NLP to generate word vectors. Future work should focus on evaluating the debiasing algorithm presented in this work on other existing word embedding models, such as the Skip-Gram Word2Vec model and the BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2018).

We broaden the applicability of the learning-based approach by making it effective for multiclass debiasing as well. Unlike other multiclass debiasing methodologies, our approach does not require a predefined set of debias words, thus allowing our algorithm to conduct a more thorough debiasing. Vector representations of words are the most fundamental components of NLP tasks. By removing unwanted bias at the embedding stage, downstream tasks can be conducted without changing those algorithms. This paper proposes an efficient and effective solution to the important problem of bias in natural language.

7. REFERENCES

Alhazmi, S., Black, W., & McNaught, J. (2013). Arabic SentiWordNet in relation to SentiWordNet 3.0. *International Journal of Computational Linguistics (IJCL)*, *4*(1), 1-11.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California law review, 671-732.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bordia, S., & Bowman, S.R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. *NAACL*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

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

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., & Ureña-López, L. A. (2021). A survey on bias in deep NLP. *Applied Sciences*, *11*(7), 3184.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL*.

Hube, C., Idahl, M., & Fetahu, B. (2020, January). Debiasing word embeddings from sentiment associations in names. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 259-267).

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

Jentzsch, S., Schramowski, P., Rothkopf, C., & Kersting, K. (2019). Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 37-44).

Kumar, V., Bhotia, T. S., & Chakraborty, T. (2020). Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, *8*, 486-503.

Liapakis, A., Tsiligiridis, T., Yialouris, C., & Maliappis, M. (2020). A Corpus Driven, Aspect-based Sentiment Analysis to Evaluate in Almost Real-time, a Large Volume of Online Food & Beverage Reviews. *International Journal of Computational Linguistics (IJCL)*, 11(2), 49-60.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. In *Logic, Language, and Security* (pp. 189-202). Springer, Cham.

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

Maudslay, R. H., Gonen, H., Cotterell, R., & Teufel, S. (2019). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. arXiv preprint arXiv:1909.00871.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Popović, R., Lemmerich, F., & Strohmaier, M. (2020, September). Joint multiclass debiasing of word embeddings. In *International Symposium on Methodologies for Intelligent Systems* (pp. 79-89). Springer, Cham.

Reese, S., Boleda, G., Cuadros, M., Padró, L.,and Rigau, G. (2010).Wikicorpus: A word-sensedisambiguated multilingual Wikipedia corpus.In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources *Association (ELRA).*

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J.,& Sobek, M.(2019). IPUMS USA: Version 9.0 [dataset]. IPUMS.

Siddiqui, M. A., Dahab, M. Y., & Batarfi, O. A. (2015). Building a sentiment analysis corpus with multifaceted hierarchical annotation. *International Journal of Computational Linguistics (IJCL)*, *6*(2), 11-25.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

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