# Stereotype and Unconscious Bias in Large Datasets

Atul Gandhi, Eileen Cho, SunJoo Park, and Elliot Silva

Center For Data Science, New York University

## Problem Formulation

FairFrame is a technology platform which helps users identify words containing unconscious bias in employee evaluations and job descriptions. Their current language checker system employs a direct text-matching algorithm of reviews against a list of problematic gendered words without any linguistic context. The result flags many words not used in a problematic context.

**Objective**: We apply a linguistic context by training contextual embedding representations of sentences containing flagged words, using a small number of annotated reviews in a semi-supervised clustering task.

## Data

| Review | Biased Context? |
|---|---|
| She is a **difficult** person to work with. | Yes |
| She always succeeds at **difficult** tasks. | No |
| We **feel** glad to be working with John. | No |
| We **feel** like he did an amazing job, despite several setbacks. | Yes |
| **Great** work! | Yes |
| He is **great** to work with. | Yes |

**Table 1:** Example reviews with flagged word in bold, showing contexts in which words from FairFrame's list should be flagged.

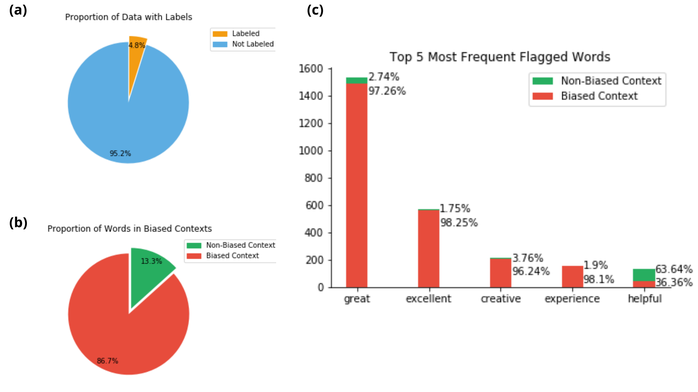


**Figure 1: Data Distribution. (a)** We manually annotated 4.8% of our 110K reviews, enabling us to utilize a semi-supervised learning approach. **(b)** The majority of reviews were biased, with only 13.8% of examples being non-biased. **(c)** Breakdown for top 5 most frequent words in our dataset.

## Model Architecture

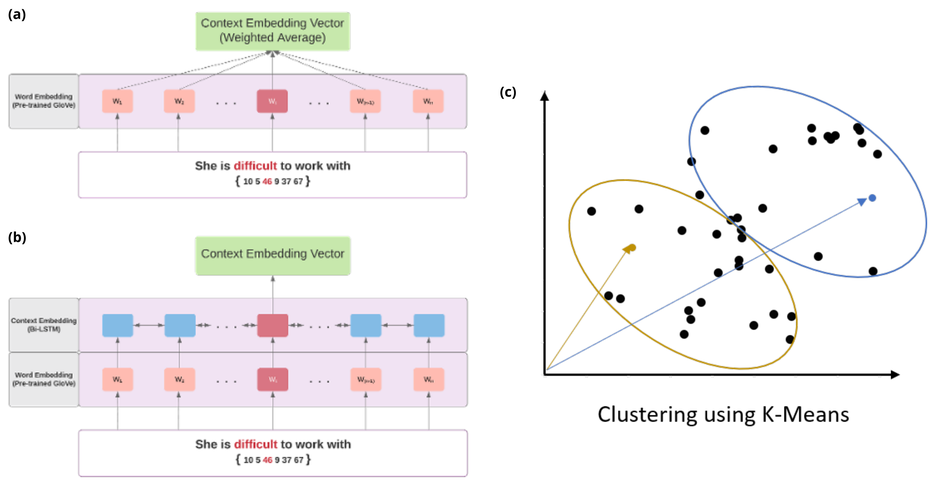


**Figure 2: Models. (a)** Our baseline model produces a representation of the sentence by averaging the embeddings of each word, upweighting that of the flagged word. **(b)** We train a Bi-LSTM and extract the embedding for the flagged word containing context from surrounding words. **(c)** We use semi-supervised methods to cluster embeddings into biased or non-biased clusters.

## Sentence Embeddings

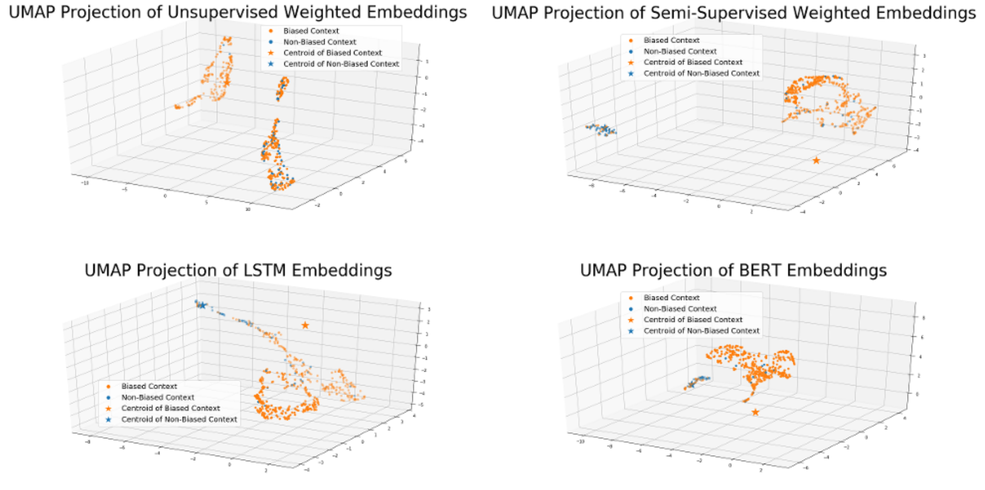


**Figure 3: UMAP projection of trained sentence embeddings. (a)** Unsupervised weighted embeddings. **(b)** Semi-supervised weighted embeddings. **(c)** LSTM model. **(d)** BERT large.

## Results

| Models | F1 Score |
|---|---|
| Weighted Embeddings (Unsupervised) | 70.29 |
| Weighted Embeddings (Semi-Supervised) | 81.28 |
| LSTM | 82.66 |
| BERT large cased | 93.57 |

**Table 2:** F1 scores for each model architecture.

## Discussion

- We were able to train separable embedding representations for both biased and non-biased contexts.
- Due to the small size of the labeled dataset, the model was prone to overfitting. As a result, less complex architectures often performed better in hyperparameter tuning.
- We observed that many of the misclassified sentences would have been difficult for a human to discern, stemming from the ambiguous nature of the task (identifying "unconscious" bias).

## Future Work

- Adapt architecture to produce representations for sentences with problematic phrases, rather than individual words.
- Obtain a more diverse dataset with language more similar to that in typical company employee evaluations.
- Apply our approach to words identified with other types of bias (race, ethnicity, etc.)

## References

Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In:arXiv preprint arXiv:1810.04805 (2018).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In:Empirical Methods in Natural Language Processing (EMNLP). 2014,pp. 1532–1543.URL:http://www.aclweb.org/anthology/D14-1162.

Yazhou Ren et al. “Semi-supervised deep embedded clustering”. In:Neurocomputing 325 (2019),pp. 121–130.ISSN: 0925-2312.DOI:https://doi.org/10.1016/j.neucom.2018.10.016.

## Acknowledgements

We would like to thank our advisors Amy Auton-Smith and Isak Nti Asare from FairFrame, as well as NYU Capstone advisors for their help and support.