Intersections of Bias in Word Embeddings
CSCI 795 Final Project Proposal

Rebecca Kleinbart Morgan Wajda-Levie

October 19, 2021

1 Description

Natural language processing (NLP) algorithms have made tremendous strides in their ability to learn semantic and syntactic patterns from a corpus of documents. Unfortunately, learning these patterns includes learning underlying biases stereotypes contained in the training documents. Word embedding algorithms, for instance, have been shown to not only capture, but amplify, existing biases in the training corpus[3]. These biases can cause harm, especially to people who are part of marginalized groups.

Understanding underlying biases in word embeddings is an active area of research, with particular focus in binary gender bias. Researchers have proposed a number of ways to evaluate gender bias in word embeddings, and some techniques for reducing the level of bias[3, 2, 4, 1]. Our goal is to expand our understanding of bias beyond the one-dimensional area of binary gender and into a broader, intersectional space of bias and discrimination towards multiple categories such as race, ethnicity, nationality and sexuality. By looking at multiple forms of bias, are we able to find commonalities and correlations that allow us to cast a wide net for understanding, and potentially mitigating bias in our trained model?

Our experiment will start from the work of Brunet, Alkalay-Houlihan, Anderson and Zemel in their paper, "Understanding the Origins of Bias in Word Embeddings." in which they use an effective Word Embedding Association Test (WEAT) for measuring stereotypical gender associations in gloVe word embeddings, and a technique for estimating the impact that an individual document will have on the final bias of a word embedding[2]. Once the impact of a given document is estimated, the corpus can be *perturbed* by witholding that document when training the model, thereby reducing the resultant overall bias in the trained word embeddings. We will use WEAT to measure a large number of other bias features and then, after exploring the data directly, attempt to perform a regression on the bias features to estimate the area of effect on *intersectional bias*, and then evaluate the effectiveness of that estimate in allowing us to reduce a broad range of biases in our word embeddings, without adversely impacting the embeddings' performance.

2 Team Member Contributions

Rebecca Kleinbart

- Research sociological and psychological background for evaluating bias and discrimination
- Further research into word embedding algorithms

- Evaluate results of Brunet algorithm using tests in Gonen and Goldberg[4]
- Identify test sets for national, ethnic and religious biases
- Modify estimation algorithm to perform and record new tests
- Run estimation algorithm on national, ethnic and religious bias sets
- Explore resulting data
- Tune and train regression algorithm on collected bias features
- Use intersectional bias impact score to perturb training corpus and train new word embeddings
- Evaluate individual bias scores of trained word embeddings
- Evaluate NLP performance of trained word embeddings
- Write report

Morgan Wajda-Levie

- Research sociological and psychological background for evaluating bias and discrimination
- Further research into word embedding algorithms
- Dry run WEAT document estimation per Brunet et al[2]
- Identify test sets for economic, gender, sexuality and disability biases
- Modify estimation algorithm to perform and record new tests
- Run estimation algorithm on economic, gender, sexuality and disability bias sets
- Explore resulting data
- Tune and train regression algorithm on collected bias features
- Use intersectional bias impact score to perturb training corpus and train new word embeddings
- Evaluate individual bias scores of trained word embeddings
- Evaluate NLP performance of trained word embeddings
- Write report
- Create video

3 CSCI 795 Related Topics

- In order to better understand the ways in which bias is understood and codified in natural language processing applications, we will be reading a number of papers on *fairness, bias and machine ethics* and using what we learn to select our own bias tests.
- Understanding the specifics of how bias is measured in word embeddings will require learning more about word embedding algorithms. These algorithms include aspects of *regression*, *dimensionality reduction*, *neural network*s and *unsupervised non-parametric learning*.
- Combining a wide range of bias measurements into a single intersectional bias impact is a *regression* task. In particular, because the cost of computing individual features is high, we will be working hard to optimize a λ value and *regularization* function that uses the fewest possible number of bias features while still allowing us to estimate impact for a large number of biases.

4 Dataset

We have chosen to use the same dataset as Brunet et al, the New York Times Annotated Corpus[5]. This dataset collects 1.8 million articles published in the New York Times between 1987 and 2007, and represents a corpus similar to what might be used in a commercial or academic application[2].

5 Timeline

Week Ending	**Task**
10/26/21	Complete literature review Configure WEAT program and perform dry run using experimental data of Brunet et al[2] Evaluate results of estimation algorithm using tests in Gonen and Goldberg[4]
11/02/21	Choose bias features and corresponding tests Modify algorithm to perform and record new tests Begin running bias tests and estimation algorithms
11/09/21	Continue running bias tests and estimation algorithms Explore bias data
11/16/21	Begin tuning and training regression algorithm on collected bias features
11/23/21	Finish tuning and training regression algorithm Evaluate regression algorithm Begin writing report
11/30/21	First draft of report Create video
12/14/21	Final presentation of project

6 Final Deliverables

Data on our experiment, both correlations and relations found in exploratory data analysis and our final evaluations of the trained regression, will be presented as both tables and appropriate graphical plots. We will describe the experiment in detail, as well as our results and conclusions, in a written report, allowing us to present any findings to our classmates.

We will also make a short video to present the information in a slightly less formal and more user-friendly way.

7 Evaluation

We will evaluate the effectiveness of our regression model and our experiment for two concerns. Firstly, we will evaluate the degree to which our model successfully addresses all forms of bias we consider, regardless of whether or not those features were included in the final model. Having trained our regression model, we will remove the documents with the greatest estimated negative impact from our training corpus and train word embeddings on the resulting perturbed corpus. We will measure individual bias features on those word embeddings and compare them to our target. We will also measure the bias-by-neighbors of at least some of our bias features using the method described in Gonen and Goldberg[4].

Secondly, we will evaluate the performance of our perturbed embeddings using the performance tests provided by the gloVe developers. This will allow us to ensure that perturbing our dataset to address biases has not adversely affected overall performance benchmarks.

References

[1] Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. "WEFE: The Word Embeddings Fairness Evaluation Framework". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 430–436. ISBN: 978-0-9992411-6-5. DOI: 10.24963/ijcai.2020/60. URL: https://www.ijcai.org/proceedings/2020/60 (visited on 10/12/2021).

[2] Marc-Etienne Brunet et al. "Understanding the Origins of Bias in Word Embeddings". In: *arXiv:1810.03611 [cs, stat]* (June 7, 2019). arXiv: 1810.03611. URL: http://arxiv.org/abs/1810.03611 (visited on 10/10/2021).

[3] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (Apr. 14, 2017), pp. 183–186. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal4230. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aal4230 (visited on 10/11/2021).

[4] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *arXiv:1903.03862 [cs]* (Sept. 24, 2019). arXiv: 1903.03862. URL: http://arxiv.org/abs/1903.03862 (visited on 10/11/2021).

[5] Evan Sandhaus. *The New York Times Annotated Corpus*. Artwork Size: 3250585 KB Pages: 3250585 KB Type: dataset. Oct. 17, 2008. DOI: 10.35111/77BA-9X74. URL: https://catalog.ldc.upenn.edu/LDC2008T19 (visited on 10/18/2021).