Unpacking Social Biases: An Analysis of Sense Embeddings Using the Backpack Model

Stanford CS224N Custom Project

Vedant Garg, Camron Sallade, Molly Cantillon
Stanford University
vedgarg@stanford.edu, camron@stanford.edu, cantillon@stanford.edu

Abstract

Large Language Models are known to contain bias. These biases are based on stereotypes the model inevitably encounters during training, resulting in word associations between areas such as gender, race, and/or socioeconomic status and descriptive words. As a result, the model not only hurts groups with these overgeneralized beliefs, but the accuracy of many NLP tasks are hurt when their foundations are built on biases. Due to the promise the novel Backpack models are showing, it is imperative that we understand and quantify what kind of biases have made its way into the trained model and find ways to mitigate these biases from occurring in embeddings. Various experiments can be run to accomplish this task. In this project, we conduct word association tests for a sample of selected words that contain bias and evaluate it against the model's produced sense vectors. We found that our method accurately detects similar words while still producing biased sense vectors. Our ongoing work aims to refine the method and improve the mitigation of bias in LLMs.

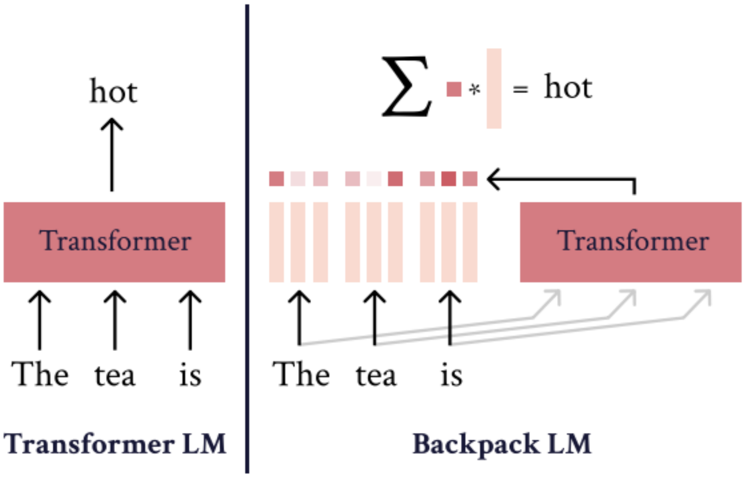
1 Key Information to include

- Mentor: John Hewitt (johnhew@stanford.edu)
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

The ability to accurately discern the appropriate meaning of a word in a sentence is imperative to natural language processing tasks. Words that have multiple senses can carry different meanings across contexts.From sentiment analysis to machine translation and question answering, word ambiguity poses a significant challenge to accurate language understanding. Current approaches, exemplified by Word2Vec, attempt to learn a single vector embedding for each word, but due to their monolithic nature, falls short in adequately capturing the nuanced lexical structure of words.

This paper examines "Backpacks", a novel neural architecture that addresses the challenging problem of disambiguating word meaning based on context. In this model, each word in a vocabulary is a linear combination of a set of non-contextual *sense* vectors that represent distinct learned aspects of the word. For example, different sense vectors for the word "nail" could encode (1) a small metal spike, (2) to expose someone, (3) fingertip specimen, (4) perform perfectly, (5) to tackle, etc. Weights on sense vectors are computed by an expressive network that processes the entire sequence, allowing the Backpack to capture rich lexical structure as well as debias terms predictably in all contexts.



Above, the differences between a traditional Transformer LM and Backpack LM are shown. Instead of applying the transformer to the sentence, the transformer is applied to each word's sense vectors, which then forms an non-contextual output of the sequence.

Taking this novel approach to learning shows significant promise for creating newer, improved models but it is a major priority to analyze whether there are more insidious forms of bias that this newer architecture is more prone to learning. Through this paper, we will attempt to stress test this model using a myriad techniques to detect bias in particular against particular ethnic groups. To get a better sense of what the true bias is, we will also use a set of multiple large datasets to perform these analyses.

In this study, we evaluated the biases of sense vectors on the backpack language model. We focus on the performance of all sense vectors and compare it to the Word2Vec model's embeddings. Using the Google Analogy testset, which contains a diverse range of semantic and syntactic relationships, we assess the performance of the backpack sense vectors in terms of their bias towards certain relationships. We examined gender and racial biases in these models by analyzing their representations of relationships across various contexts and demographics.

For gender bias evaluation, we used an occupational dataset, the US Bureau of Labor Statistics to study the models' associations between gender and professions. We looked at the similarity scores assigned to word pairs such as "nurse" and "woman" and compared them to the scores for pairs like "man" and "nurse." The results indicated much more bias towards gender than ethnic groups for all sense vectors shown in [5.3]. Especially for occupations such as secretary, assistant, and care-taker, the model's embeddings for all sense vectors showed a significant bias toward female associations with those occupations.

3 Related Work

The language modeling bias landscape reveals a gap in literature for models that offer strong modeling performance, interpretability and control. To fill this gap, we analyze the Backpack model that obtains both rich lexical structure and interventions with contextual performance in a single model, in doing so, enabling the reduction of bias.

At a high level, a Backpack representation is a *bag-of-words* sum of non-contextual senses. By learning multiple non-contextual sense vectors, Backpacks are able to represent words as context-dependent, non-negative linear combinations of sense vectors that all bear different semantic meaning.

A Backpack representation o_i of a word x_i for any sequence $x_{1:n}$ is a weighted sum of predictive sense vectors $C(x)_1, ..., C(x)_k$ for a finite vocabulary V with $x \in V$. These sense vectors are simply a multi-vector analog of classic word representations through word2vec and GloVe.

$$o_i = \sum_{j=1}^{n} \sum_{l=1}^{k} \alpha_{lij} C(x_j)_l \tag{1}$$

where $\alpha_l ij$ are contextualization weights of a Backpack are themselves defined by a nonlinear contextualization function of the sequence.

The Backpack model then applies a softmax parameterization to the Backpack representation, resulting in the following log-linear probabilistic model defined over output space Y.

$$p(y|o_{1:n}) = softmax(Eo_{1:n}) \tag{2}$$

In the quest to measure bias, methods to measure semantic associations between words and attributes is fundamental. The Implicit Association Test (IAT) [6] as well as Word Embedding Association Test (WEAT) [2] leverage word embeddings to measure the strength of association between sets of target words and attribute words. The association is measured through cosine similarity and detailed below. These two tasks do a sufficient job of exploring problematic associations.

Overall, our review of related work highlights the limitations and motivations that point to why our Backpacks model is a promising next step in the field of neural language modeling. The model offers improved interpretability and controlability compared to existing methods, and has the potential to be applied to a range of NLP tasks.

4 Approach

4.1 Detecting Bias

The first step in mitigating bias is to declare methods for detecting it. In this paper, we discuss multiple methods for evaluating bias in the language model, including a word association test to test for bias in certain areas such as gender or ethnic groups that contain overgeneralized beliefs. For example, the word teach may have a bias towards one ethnic group or gender that won't be accurate. The word association test was done manually on a small scale to confirm the model produces some sort of bias in sense embeddings. A

4.2 Analyzing Bias

The first step in mitigating bias is to declare methods for detecting it. A word association test can be done to test for bias in certain groups that contain overgeneralized beliefs. For example, the word teach may have a bias towards one race or gender that won't be accurate. The word association test was done manually on a small scale to confirm the model produces some sort of bias in sense embeddings.

4.3 Data

- Google analogies dataset: The Google analogies dataset was used to test the models ability to make connections with seemingly no bias. The Google Analogy Test Set is a benchmark dataset developed by Mikolov to evaluate the performance of language models on tasks dealing with semantic and syntactic relationships. It consists of 19,544 question pairs, including semantic questions and syntactic questions. These questions span across 14 types of relations, including 9 morphological and 5 semantic relations. To assess a model's bias, the Google Analogy Test Set can be utilized to evaluate whether the model is consistent in answering questions and identifying relationships across various contexts, subjects, and demographics. The dataset provides a robust testing ground for identifying potential biases in the model's understanding and representation of different concepts, as well as its ability to generalize relationships across diverse scenarios.
- SentiWordNet 3.0.0 was used as a comprehensive lexical resource on top on WordNet. Not only is SentiWordNet built on top of a reliable resource, but it covers a wide range of English words, including nouns, verbs, adjectives, and adverbs. This extensive coverage allows us to obtain sentiment information for a more diverse set of attribute words, which is crucial

for a thorough analysis of potential bias in the word embeddings. It also provides positive and negative sentiment scores for each synset, allowing us to differentiate between positive and negative attributes with more accuracy. This fine-grained sentiment information enables us to perform a more detailed analysis of how the model associates different ethnic groups with positive or negative words.

- The wordsim353 goldstandard dataset word similarity and relatness dataset was used to test the models ability to rate similarity between words for each of the sense vectors.
- The US Bureau of Labor Statistics dataset for occupations was also used to test the models sense vectors abilities to handle significantly different associations between gender and occupation. Proportions of bias were observes based on occupation

5 Experiments

5.1 Categorical Bias Scores for Ethnic Bias

With categorical bias scores based on mean differences between cosine similarities of ethnic group words and attribute words, we propose a comprehensive way of evaluating the model's bias towards each ethnic group. A categorical bias score can be calculated based on a language model's embeddings for certain stereotyped groups. We basically analyze the model by calculating a similarity score between words representing ethnic groups to a large dataset of positive words to get a sense of what the model associates with particular ethnic groups.

The Average Distance is the mean value of the cosine similarity between the ethnic group and the attribute words. We also calculate the variance of this parameter using the typical tools. A visualization of the distribution of ethnic words and a fixed positive/negative word list is shown below.

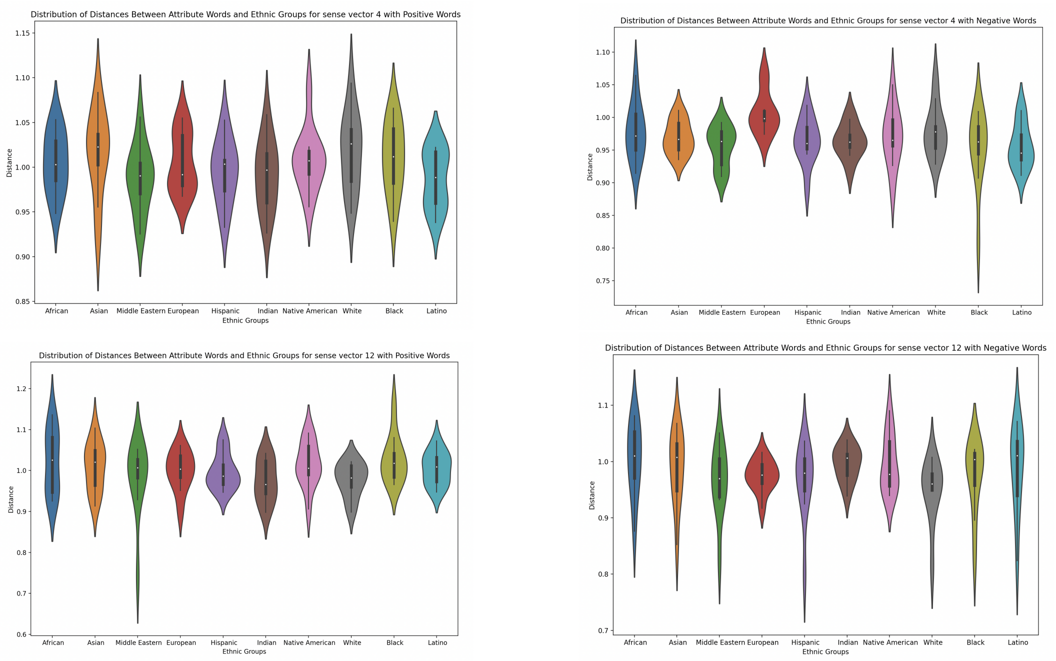


Figure 1: Four different violin graphs representing the distribution of cosine distances (scaled to 0 to 2) between attribute words and ethnic groups

The model's sense vectors 4 and 12 showed the most promise for not having bias in previous tests, along with sense vector 4 performing well against the baseline word2vec model in the google analogy test set. However, both sense vectors show a bias towards certain ethnic groups among positive and negative words. Sense vector 12 shows a bias towards Middle Eastern groups of people in terms of distance from positive words, while showing a bias toward many groups when tested with negative words. This demonstrates an unfair relation between ethnic groups and negative words from the

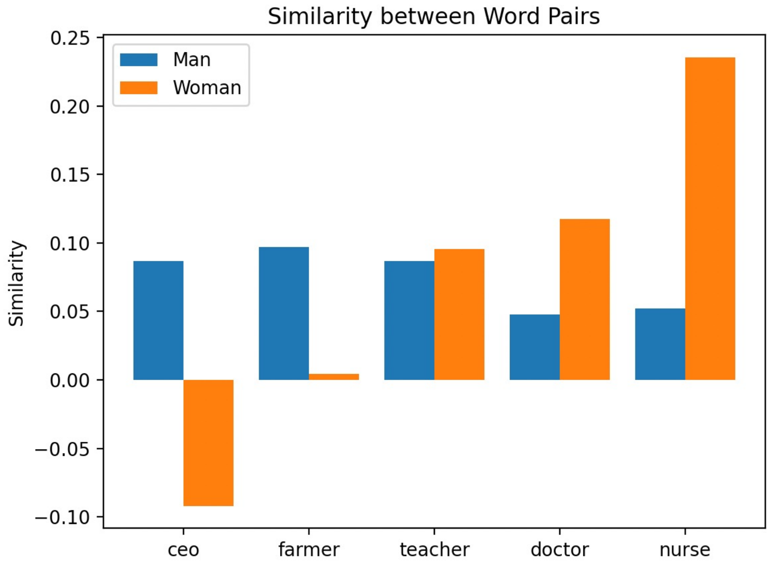
SentiWordNet 3.0 dataset for sense vector 12. For sense vector 4, there is a clear bias toward the black ethnic test group when tested against negative words. Sense vector 4

Additionally, a WEAT score was calculated to measure bias in ethnic groups with respect to attribute words. The method takes a set W to be a list of words representing different ethnic groups, a set A to be a list of positive attribute words, and a set B to be a list of negative attribute words. A higher absolute value of s would indicate a stronger association between the ethnic group words and one of the attribute word sets (positive or negative), which could suggest bias in the model's embeddings. This is not an entirely indicative predictor of bias, but a suggestion that the relation between ethnic words and some attribute words is innaccurate.

$$s = \sum_{w \in W} \frac{(mean_{a \in A} cos(w, a) - mean_{b \in B} cos(w, b))}{\sqrt{\sum_{w \in W} (mean_{a \in A} cos(w, a) - mean_{b \in B} cos(w, b))^2}} \tag{3}$$

However, the WEAT score provides a measure of association between the words, but the interpretation of the results requires caution. It is essential to carefully select the words in the sets W, A, and B to understand how the results could exhibit bias.

5.2 Comparative Bias Scores



A bias test for comparing purpose is to detect biases in the language model by comparing the similarity between the given adjectives and ethnic words. The function returns a list of tuples containing the adjective, two ethnic words, and the difference in similarity scores, sorted in descending order by the absolute value of the difference. Difference in performance of sense vectors was evaluated with:

$$biasindicator = \frac{1}{|W|} \sum_{w \in W} |cos(w, g_{bias})| \tag{4}$$

where The direct bias measure calculates the projection of words onto the gender bias direction (e.g., the difference between the word embeddings for "he" and "she") in the word embeddings.

5.3 Analogy Test

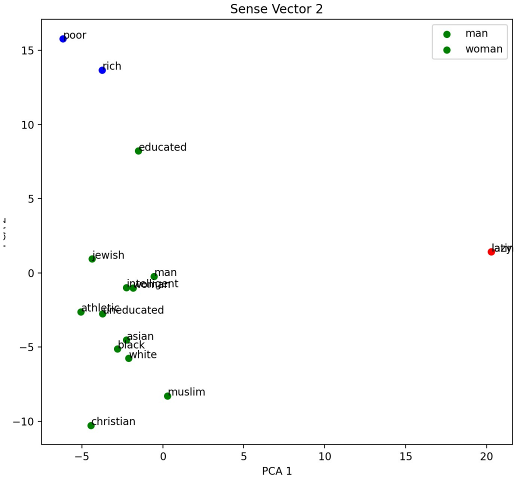
An analogy test on the google analogy dataset was performed, where the performance of backpack model's sense vectors compared with word2vec models word embeddings was compared. For each line in the google analogies dataset, a list of all the sense vectors was obtained for each word and the cosine similarity was computed between The accuracies of each sense vector was kept track of along with the accuracy of the word2vec model as a baseline. The word2vec dataset achieved an accuracy score of 28 percent, while sense vector 2 performed significantly better with a score of 70 percent.

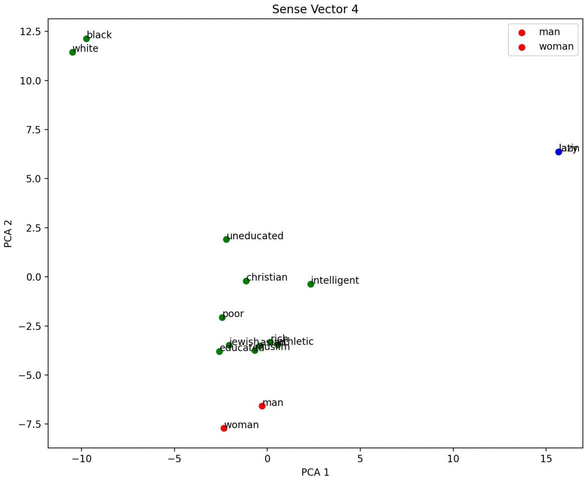
5.4 Relatedness Test

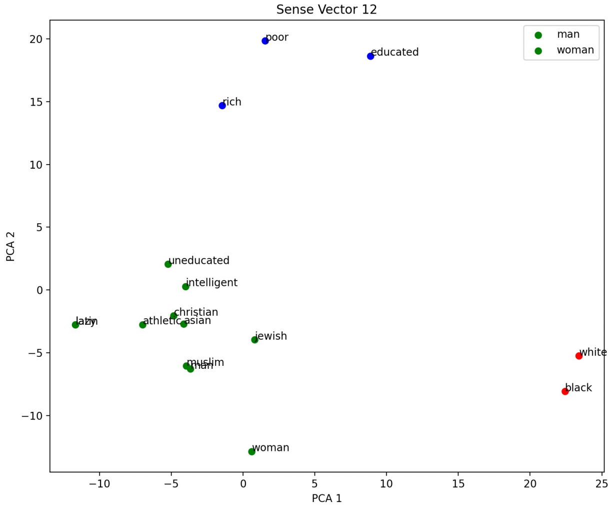
The wordsim353 dataset was used to test the models ability to find similarity between words. A function was used that computes the cosine similarity between sense vectors of pairs of words for a given language model and tokenizer. The resulting distribution of cosine similarities among word pairs in the wordsim353 dataset were used as input for testing similarity and correlation.
Relatedness Dataset - Correlation: 0.5462, p-value: 0.0043
Similarity Dataset - Correlation: 0.7326, p-value: 0.003
The results showed that sense vector 2 showed the strongest correlation between pairs in the dataset according to the cosine similarities.

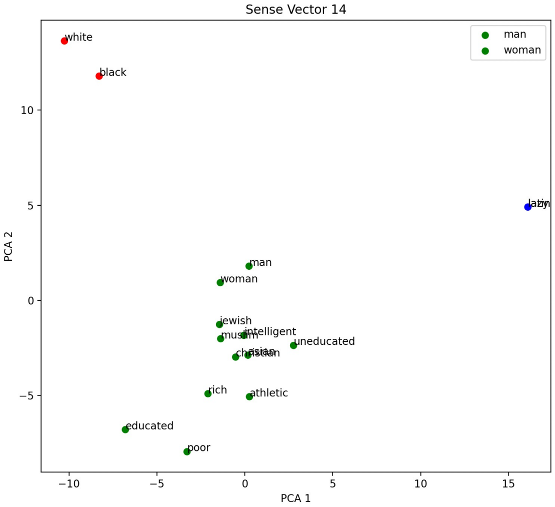
5.5 Results

After a comprehensive analysis of the sense vectors and their biases in the backack model, including cosine distance distribution visualization and average scores, Spearman's rank correlation between ethnic words and a fixed attribute word list, and a variance calculation based on average ethnic group bias across positive and negative words in the SentiWordNet 3.0 dataset. The gist of the idea is that we understand the variation difference to be the mean of the sum of squared differences between the differences.
$VariationDifference = \frac{1}{m-1}\sum_{j=1}^{m}(Difference_j - Difference_i)^2$
Where m is the total amount of ethnic groups tested The average difference in the cosine similarities between ethnic group i and the positive/negative word list was calculated with:
$AverageDistanceEthnicGroup = \frac{1}{n}\sum_{i=1}^{n} Distance_i$









- Categorical Bias Scores showed a bias in the model's sense vectors 4, and 12, with sense vector 2 only showing some bias in the white ethnic group for positive and negative words. Sense vector 4 showed a bias in positive words only slightly in the asian group, but showed a significant bias for the black ethnic group shown in the graph in [5.2]. Sense Vector 12 showed significant bias for positive words only in the middle eastern ethnic group, while showing an equally strong bias for all ethnic groups but European, Indian, and Native American groups. Although each sense vector exhibits different forms of bias, through stronger correlations between words and different ethnic groups, there still exists the pattern of bias for every sense. However, it was found that 2 exhibited a much smaller amount of bias as shown below.

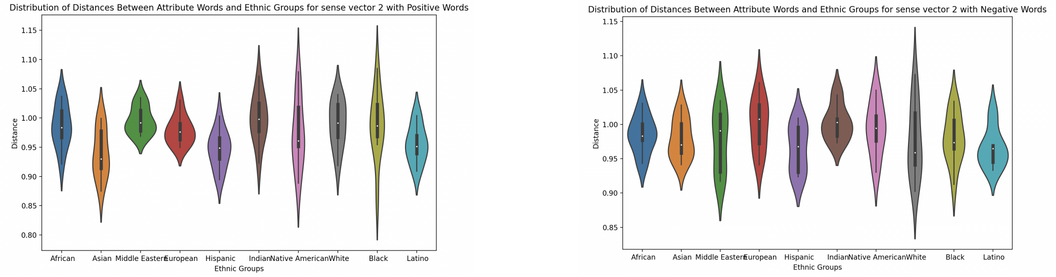


Figure 2: Two violin graphs showing the lesser biases in sense vector two

- The Comparative Bias Scores showed a bias in occupations between genders, especially between those of farmer for man and nurse for woman. Doctor also showed a significant bias for women rather than man, which contradicted proportions found in the US Bureau of Labor Statistics.

- The Analogy Test showed a significant performance over the word2vec model when run against the google analogy dataset, with sense vector 2 showing a 40 percent improvement over the word2vec embeddings.

- The Relatedness Test was run against the wordsim353 dataset where cosine similarity was calculated between pairs of words in the wordsim353 dataset. We found that sense vectors 9 and 12 showed the greatest distribution in cosine similarities between pairs in the wordsim353 dataset.

6 Analysis

Overall, our analysis of sense vectors in the backpack model, showed a significant bias in different categories, with each sense vector performing differently in certain gender and ethnic categories. Sense vector 2 performed best overall among all categories, only exhibiting slight bias in gender and ethnic categories. Sense vectors 4, 14, and 12 oexhibited significant bias in gender and ethnic categories, while also performing poorly in the google analogy testset, showing that they would not be best for general nlp tasks and the models sense vectors are for the most part not generalizable. Sense vector 2 however, proved generalizable for all tasks, showing minimal bias in gender and ethnic groups. This is significant, as not only did it outperform baselines such as word2vec, but it serves as a foundation for bias mitigation among general tasks as its weight can be adjusted slightly to account for its slight biases in categories such as female and male occupation stereotypes and white ethnic category stereotypes. As a result, Backpack models can be used to not only perform well for general nlp tasks, but versatile across multiple sense vectors and ethnic and gender groups.

7 Conclusion

In this paper, we endeavoured to stress test the Backpack model in its ability to unpack natural language biases. Through our use of sense vectors, we uncovered a host of deeply ingrained stereotypes that continue to plague the field of natural language generation. With a simple method, we were able to demonstrate bias across a state-of-the-model. In the future, we hope to examine and mitigate bias among other novel models using more sophisticated methods.

While our findings shed important light on the current state of bias in natural language processing, it is crucial to acknowledge that Backpacks represent just the tip of the iceberg. A nascent model, their potential to revolutionize semantic analysis and unlock new frontiers in language generation is enormous. However, work in not only mitigating problematic bias but intervening predictably remains to be done before this potential can be fully realized.

Through this work, we have shown the presence of significant bias in the model which is not atypical of a model of this sophistication. A useful follow up of our work would be to analyze what exactly is causing these differences in learning bias in backpacks as compared to other models that use sense vectors. A thorough analysis of the learning processes of backpacks would be crucial towards understanding what causes these bias and how to mitigate them.

In order to truly harness the power of Backpacks and other emerging language models, further deep analysis and development is needed. By pushing the boundaries of our current understanding and relentlessly challenging the status quo, we can begin to build a more equitable and inclusive future for natural language processing.

8 References

1. Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." Advances in neural information processing systems. 2016.
2. Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186
3. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2002). Placing search in context: The concept revisited. Proceedings of the 10th ACM international conference on Information and knowledge management, 406-415.
4. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations (ICLR).
5. Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. "Nuanced metrics for measuring unintended bias with real data for text classification." Companion Proceedings of The 2019 World Wide Web Conference, 2019.
6. Jeffery Yen, Kevin Durrheim, Romin W. Tafarodi. (2018) ‘I’m happy to own my implicit biases’: Public encounters with the implicit association test