Language, Cognition & Computation: Course 096222
Final Project Report
Gender Biases and Debiasing in Language Model Embeddings & The Brain

Idan Horowitz
209723246

Lian Fichman
206238891

Matan Birnboim
313358343

Abstract

Presented is a summary of our findings from our final project in the Language, Cognition & Computation Course. We begin with an overview and results report of the Structured and Semi-structured Tasks, before a description of our research questions, methodologies and findings for the Unstructured Task, in which we explore and quantify gender bias in both language model vector embeddings and fMRI imaging data, and assess the effectiveness of bias removal methods.

1 Introduction

Vector embeddings in the context of language models are numerical representations of words or sentences in a high-dimensional space, where the relationships between the vectors capture semantic meaning. Different embeddings can significantly influence the outcomes of learning tasks, as they encode distinct features relating to meaning and interpretation, impacting the ability to capture fine-grained relationships between words. GloVe vectors are static embeddings created by integrating both the similarity and the co-occurence of a word with its context, producing dense and meaningful vectors (Pennington et al., 2014). BERT embeddings, on the other hand, are obtained from a pre-trained encoder of a Transformer-based model, allowing for a more nuanced context in each vector (Devlin et al., 2018), and is generally considered a very sophisticated model. This project explores the influence of two popular embeddings, GloVe and BERT, on various language tasks in general, and specifically within the realm of gender bias.

In Homework Assignment 3, we replicated Analysis 1 as described in Pereira et al. (2018). We used a decoder model that attempts to predict semantic vectors of 180 predefined concepts using fMRI imaging data obtained from a participant's viewing of the associated word and GloVe vector embeddings. We found that through an 18-fold train-and-test cross-validation, and by using rank-based accuracy[1] as an evaluation metric, found that it generally was successful at decoding brain activity for the relevant concepts. In the Structured and Semi-structured parts here, we extend the scope of this analysis to include replications of Experiments 2 and 3 Pereira et al. (2018), add additional tasks, and integrate BERT vector embeddings as a means of comparison. For the open research task, we draw from Garg et al. (2018) in quantifying gender bias across both vector embeddings as well as the fMRI data. Finally, we attempt to replicate the debiasing task described in Bolukbasi et al. (2016) and assess its effectiveness in light of the tasks performed beforehand.

2 Structured Tasks

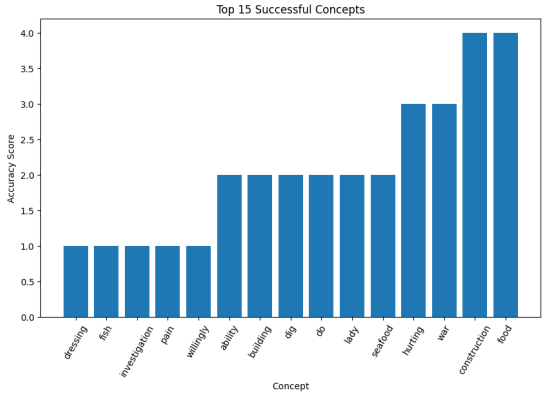
2.1 Repeating the Analysis of Homework Assignment 3 with BERT

The relevant part of Homework Assignment 3 attempted to replicate the results of Analysis 1 from Pereira et al. (2018). Here, we repeat this analysis using BERT vector representations from a pre-trained BERT model on those same 180 concepts. Our initial results showed poor performance, so we attempted to reduce the dimension of the vectors using PCA. We performed cross-validation using various thresholds for explained cumulative variance and found that the best results occurred when taking the principal components that accounted for 85% of the variance in the data, reducing it to 71 components. The PCA was performed on the BERT vectors as well as the fMRI vectors.

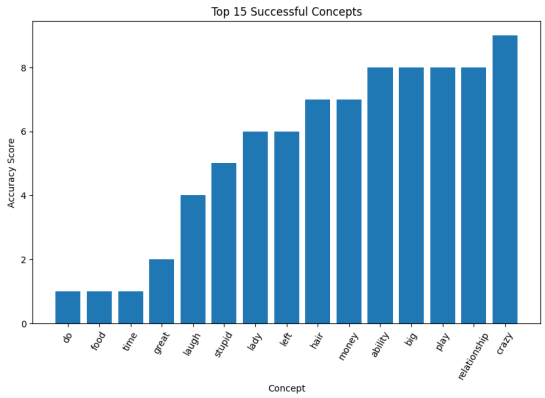
In the same manner as before, we assessed the de-

[1] A decoded vector is assigned a rank based on its similarity to the true concept vector. Thus, a rank of 1 indicates that it is the closest to its true vector, and is the best outcome.

coder performance using the average rank method via 10-fold cross-validation, declaring a concept successful if the rank of the decoded vector was below 90 [2]. The number of concepts for which BERT succeed was 133, while GloVe has 134 successful concepts. However, we believe that this difference alone is not sufficient to determine which model has better results.



(a) BERT



(b) GloVe

Figure 1: Accuracy ranks for Top 15 successful concepts

When comparing the top 15 successful concepts, we observed that although there is some overlap, BERT includes more complex words such as ”investigation”, ”willingly” and ”construction” that are not present in GloVe’s successful concepts. This suggests that BERT performs better with more intricate vocabulary compared to GloVe, which excels in capturing simpler and more standard, commonly used words. This is supported by considering the top 15 failed concepts, as the opposite is true - BERT includes simple and everyday words, for example, ”plan”, ”weak”, ”toy” and ”bag”. On the other hand, GloVe’s failures consist of more complex concepts such as ”argumentatively”, ”mathematical” and ”emotionally”. This again points to GloVe’s success with simpler words while struggling with more complex language, while BERT indicates the opposite.

Considering the accuracy ranks, we see that BERT’s performance was slightly better than GloVe’s. For all top 15 concepts, the rank from BERT did not exceed 4, while GloVe only managed an accuracy below 4 in its top 5 concepts [3]. Also, for the failed concepts, all of the ranks produced from GloVe were above 150, while only 7 from BERT scored above 150. Although the binary metric of successful/failed indicated a similar performance, when assessing the quality of those successes and failures, a picture emerges in which BERT is more impressive. This is by no means to say that GloVe was unsuccessful, but rather that BERT was simply slightly better.

Although the results from BERT were slightly better than those from GloVe, they did not match our expectations. We suggest that this might be owing to the BERT’s architecture - it is a transformer-based model, with the goal of maintaining context. Since this task focused on single-word concepts, perhaps it did not reach our expectations because of the lack of context when considering each word in isolation. We will explore the validity of this claim in the later sections, in which we perform the same task on vectors taken from sentences, not words.

2.2 Overview of Experiments in ”Toward a Universal Decoder of Linguistic Meaning from Brain Activation”, (Pereira et al., 2018)

Pereira et al. (2018) performed 3 different experiments in their analysis. They used individual concepts (words) as stimuli and tested whether the decoder could generalize new concepts. Each concept was presented in the context of a sentence, an image, and in a word cloud with 5 similar words, all intended to reduce ambiguity, and then combined all 3 contexts (paradigms) to produce 1 brain image per participant per word.

Experiments 2 & 3, on the other hand, used text passages as stimuli and tested whether a decoder trained on individual concept imaging data could decode semantic vectors from sentence imaging data. Experiment 2 used a 4-sentence passage de-

[2] i.e., better than random, as there are 180 concepts

[3] Recall that an accuracy rank of 1 is optimal as it indicates the decoded vector was closest to the original vector

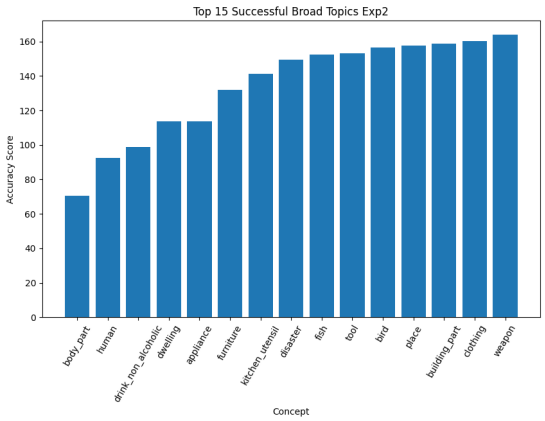
scribing a concept and providing basic information (Wikipedia-style), while Experiment 3 included narrative passages as well. Sentences were presented individually, and the fMRI scans from each sentence were combined together to produce a scan for each participant on each concept. To produce a semantic vector for a sentence, the average was taken across all word embeddings in each sentence (i.e. the true embedding). Evaluation took place in 3 pairwise classification tasks, first taking sentences from different topics (e.g. an animal and a musical instrument), next taking different sentences from different passages within the same topic (e.g. two musical instruments), and finally taking different sentences from the same passage. The model was the one trained in experiment 1.

Ultimately, the difference between Experiments 2 and 3 are minor, being reflected only in the actual stimuli used themselves, and the addition of a narrative passage as a stimulus. However, experiment 1 differs from 2 & 3 both in the type of stimulus (single words as opposed to passages) and in the goal of the analysis (also in the number of participants and the success of the results, but since we are trying to recreate the computational method, this is less relevant to this work). While Experiment 1 trained the decoder and tested its ability to generalise concepts from given brain data, the latter two used a decoder pre-trained on individual concepts and tested whether it could decode brain data for sentences.

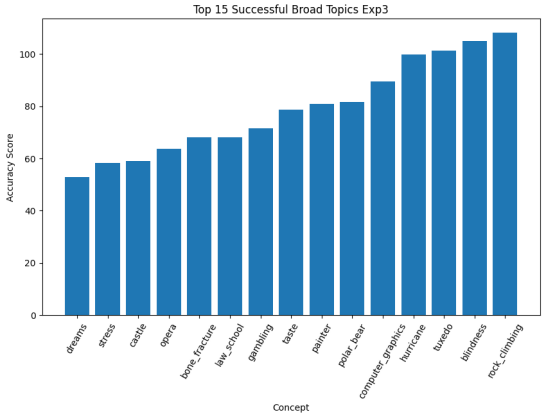
2.3 Testing a pre-Trained Decoder on Sentences

Here we took the pre-trained decoder from above [4], and using the sentences from Experiments 2 and 3 Pereira et al. (2018), we tested the decoder's ability to generalise concepts using GloVe vector embeddings. This means of evaluation was comparable to the above, using a rank-based accuracy to identify the more/less successful topics. The difference here is that each sentence is linked to a broad topic, thus we computed the average rank for each broad topic across all related sentences and classified it as successful if it was better than random.

Experiments 2 and 3 share many similarities in terms of using text passages as stimuli, the primary difference lies in the nature of the stimuli used[5] - Experiment 2 used four-sentence passages that described a concept and provided basic information while experiment 3 included narrative passages as well. Despite these similarities, there are some variations in the outcomes of the decoder using the sentences from the two experiments. For sentences from experiment 2, out of the 24 broad topics, 21 were successfully decoded by the GloVe-based decoder, while 3 failed. The successfully decoded topics seemed to relate to more concrete, physical concepts such as 'body parts', 'humans', 'drinks', and 'furniture'. Sentences from Experiment 3 produced slightly different outcomes. 19 topics were successfully decoded, while 5 failed. Here, the successful topics seemed to be more abstract including 'dreams', 'stress', and 'taste'. It is also interesting to note that on the whole, the accuracy scores of the top 15 successful concepts in experiment 3 were better. Both sets of sentences struggled with less everyday topics, for example, 'beekeeping' and 'pharmacist'. This is in line with our understanding above, where GloVe is successful with general, commonly used topics.



(a) Sentences Exp. 2



(b) Sentences Exp. 3

Figure 2: Accuracy ranks for Top 15 successful broad topics

[4] Trained on all 180 concepts

[5] See Section 2.2

The difference in outcomes between experiment 2 and experiment 3 could be attributed to the addition of narrative passages in the latter. The inclusion of narrative elements might have introduced more complex and diverse linguistic patterns, requiring the decoder to generalize and decode semantic information from a broader range of contexts, and specifically improving its ability with abstract ideas. The general descriptions used for both sets of experiments would intuitively lead to success in decoding physical objects, which are easier to describe. The narratives add a layer of depth that allows for the capturing of less concrete topics, at the cost of a decline in ability to decode those physical objects.

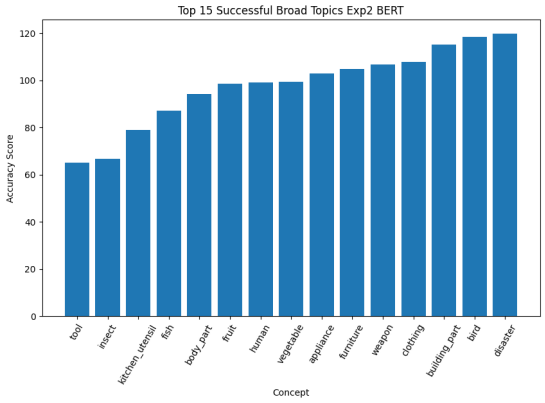
It is important to keep in mind that in both cases, the decoder was trained on individual words, and tested on both sets of sentences. This step-up in the complexity of the task indicates that on the whole, the results are impressive. Overall, both experiments demonstrated the decoder's ability to decode semantic vectors from brain imaging data based on textual stimuli. However, the specific stimuli used in each experiment, particularly the inclusion of narrative passages in experiment 3, led to slightly different outcomes, with experiment 3 yielding better accuracy scores, but succeeding in fewer topics in total.

3 Semi-structured Tasks

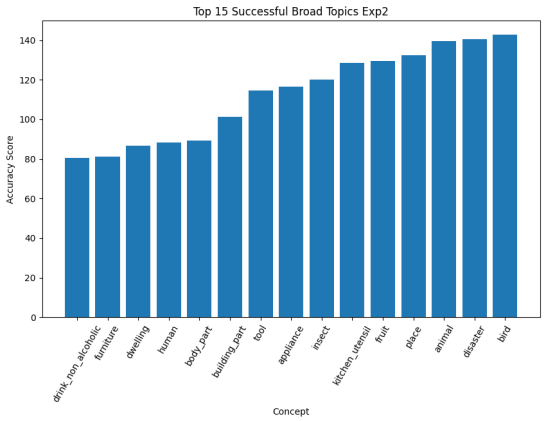
3.1 Training a Decoder on Sentence Representations

Here we repeated the above analysis, but this time training the decoder on the sentences it was to predict. As such, we returned to the k-fold cross-evaluation method, and only ran the analysis on the sentences from Experiment 2. In order to obtain even folds, we split the 384 sentences into 32 folds of 12. Additionally, we compared the performance of GloVe and BERT-based encodings.

Similar to our results with the word decoding, the results of BERT and GloVe lacked any significant differences, with both succeeding in 22 broad topics from 24. However, just as with the word analysis, BERT's actual accuracy scores indicated slightly better results, both in terms of the topics in which it succeeded and failed, and in the average accuracy score for each fold. We had expected, however, for BERT to far outperform GloVe when it came to sentences as it is better designed for maintaining context. We thus decided to perform a PCA decomposition to both sets of vectors to see if it had any impact on the outcomes. Owing to the computational resources required, we did not perform a cross-validation for the best explained-variance threshold as above, but rather used 0.85 measure of cumulative explained variance from the outset.



(a) BERT

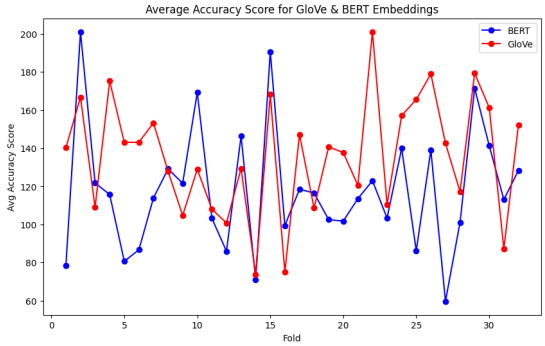


(b) GloVe

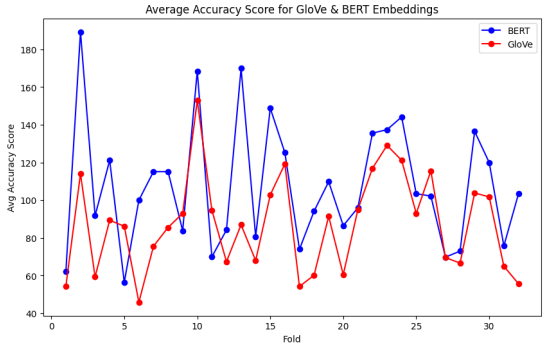
Figure 3: Accuracy ranks for Top 15 successful broad topics

Although the decomposition significantly impacted the results (both succeeding in all 24 topics), they were the opposite of what we expected - GloVe outperformed BERT in terms of accuracy. This indicates that GloVe is clearly a very good model, much better than we had previously given it credit for. Additionally, we can perhaps attribute the results to the manner of evaluation - theoretically, BERT may still be better at maintaining context across sentences, but by averaging across all words in the sentence and taking a single-vector representation of a sentence, it effectively treats the sentence as if it were a single word. This perhaps could be the reason that GloVe and BERT performed similarly across all tasks, whether sentences or words.

We would expect to see a more significant difference in next-word prediction tasks.[6]



(a) No PCA



(b) With PCA

Figure 4: Accuracy ranks for each fold

As a final word, it is also fascinating that the trend we saw earlier - BERT's success with more complicated topics as opposed to GloVe with more 'everyday' ones, was maintained only prior to PCA. Once we took the original components, there was a large overlap among their top-scoring concepts. This may indicate that much of the complexity and nuance that BERT manages to capture extends beyond the principal components.

3.2 Building a Brain-encoder model

We now attempt to reverse the relationship - instead of using brain imaging data to decode concept vectors, we attempt to encode fMRI voxels using word embedding vectors. We produced a simplified version of the brain-encoder model described by Huth et al. (2016). For each voxel in the fMRI data, we fitted a linear regression model. We defined a significant model as one that produced an R^2 score in excess of 0.75. For the significant voxels, we assessed the effectiveness of the predicted voxel from the model via its cosine similarity with the original voxel. The GloVe embeddings found 96.77% of voxels significant, with a mean cosine similarity of the significant voxels of 0.91. Although BERT's predicted voxel vectors had a mean cosine similarity of 0.99, meaning that the vectors predicted were very similar to the originals, most of the R^2 scores were above 1, indicating a very bad model fit. We thus did a PCA reduction on the BERT vectors to match those of GloVe in terms of dimension [7], and found the outcome to be very similar to those of GloVe - 94.62% of voxels were found to be significant, also with a mean cosine similarity of 0.91 for the significant voxels.

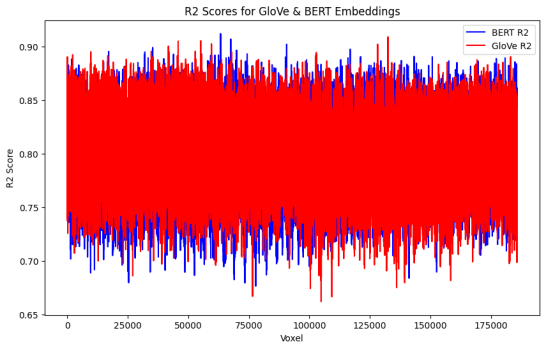


Figure 5: Voxel R^2 scores for each embedding

We believe that owing to the success of both models - the high prediction similarity and the high number of relevant voxels - brain-encoding models are successful and produce significant results. The small difference in the proportion of significant voxels between embeddings may reflect a number of different factors, both neural, linguistic and the manner in which the models encode, and thus further research would be required to account for this difference, which is beyond the scope of this work. Additionally, since the fMRI imaging data was taken from only one test subject, it would not be representative enough to draw significant conclusions about either embedding, beyond the fact that they can be used to accurately capture voxel activations, and are representative of the neural processes occurring.

[6] It is also interesting to note that the average accuracy ranks for this task were much higher than that of the words. This is likely owing to both the complexity in sentence decoding as opposed to single word vectors, as well as the vector averaging mentioned above.

[7] 300

4 Open-ended Task: Gender Bias and Debiasing in Language Models and the Brain

4.1 Introduction

A good language model should, theoretically, contain no biases and be prone to none of the cognitive fallacies that humans are susceptible to. It should capture the semantics and intricacies of a word and its context, without revealing any prejudices or preconceived notions that humans have adopted. However, models are ultimately trained on data generated by humans and these biases can too be reflected in language models.

In this part of the project, we first quantify gender bias in language embeddings, by extending the analysis of Garg et al. (2018). We ask whether biases exist in language models, and if so, how to quantify such bias and if some embeddings are more bias-prone than others, specifically within the context of comparing GloVe to BERT. Additionally, we question if there are also bias-revealing patterns in fMRI imaging data, and how much of a language bias can be attributed to factors inherent to language as opposed to human perception.

Next, to replicate the debiasing methods demonstrated by Bolukbasi et al. (2016) and assess their effectiveness. We ask if it is possible to remove bias from word embeddings, and if so, how effective is it such that it no longer exhibits bias while still preserving the inherent meaning of the word represented?

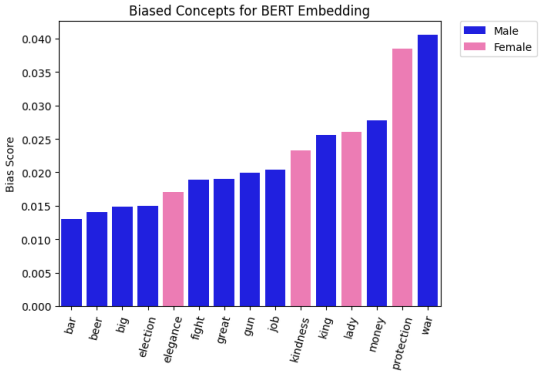
Throughout this section we define the following terms as follows: gender-neutral, referring to words that have no gender, gender-specific, referring to words that have a fixed gender and gender-leaning words, referring to the subset of gender-neutral words which have no inherent gender, but can be associated with a gender. For example, the words 'husband' and 'actress' specifically relate to gender and are thus gender-specific. While 'chair', 'doctor', 'attractive', and 'pageant' are all gender-neutral as they have no inherent gender, the latter three words are all gender-leaning as they can be associated with a specific gender, whereas 'chair' is not. [8]

[8]We place ourselves in no position to make moral judgements about gender biases in words. We are interested in researching the phenomenon in general and handling it in the best manner possible and from a language modelling perspective, believe it to be inaccurate if gender-neutral words exhibit strong tendencies to a specific gender.

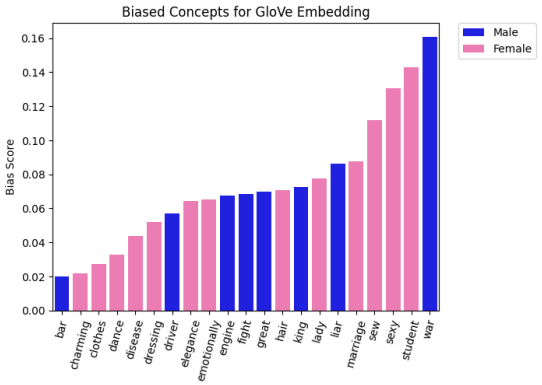
Regarding the vector representation of genders we initially performed our analysis using each model's embedding for the words 'male' and 'female'. This, however, produced erratic results, as these individual vectors do not necessarily capture all associations for each gender. Following directly from Garg et al. (2018), we decided to create a vector of the average of all word vectors associated with or defining each gender.

4.2 Quantifying gender bias

4.2.1 Bias through vector similarity



(a) BERT



(b) GloVe

Figure 6: Biased concepts for each embedding

Our first attempt at quantifying gender bias was by analysing vector similarity differences for each gender. We selected a group of 114 gender-neutral words from the original 180 concepts that were not gender-leaning [9] and defined them as a neutral set of words. We define a word's bias score as follows: $bias(w_i) = sim(w_i, male) - sim(w_i, female)$ where $sim(w, v)$ is the cosine similarity between two vectors any two vectors w and v. The other 66 words, of which 64 we classified as gender-leaning

[9]The full words list, as well as all other word groupings used, appear in the Appendix.

and 2 were gender-specific, were considered biased if their bias score was more extreme than the top 10% of the neutral set in either direction.

We ran this analysis on GloVe, BERT and fMRI vectors.

BERT had the smallest number of biased concepts, 15, with 11 of them towards males, while GloVe produced 21 biased, with only 8 towards males.

For both language embeddings, the results were as expected - the 2 gender-specific words exhibited very high bias to their respective genders, and words relating to violence and work demonstrated male bias, while concepts relating to appearance and timidness dominated the female-biased words, reflecting commonly-known stereotypes.

It is fascinating to note, however, that while both Garg et al. (2018) and Bolukbasi et al. (2016) placed heavy emphasis on bias in occupations, the occupation-related words did not exhibit much bias in our analysis.

The FMRI embeddings demonstrated much higher bias scores for all of the words, and occupations were observed amongst words classified as biased. In total, it had 16 biased concepts, with 9 relating to males - the most even split.

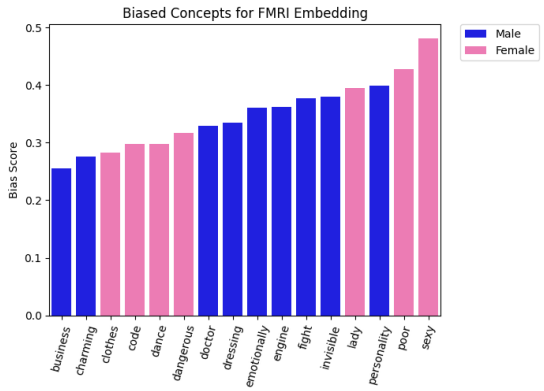


Figure 7: Biased concepts for fMRI vectors

We believe it is possible to account for these differences as follows: since BERT is a context-based model, it is better at identifying the nuances in language and differentiating between the meanings of the concepts. GloVe, on the other hand, is static and only considers co-occurrences, and thus would associate the words with genders alongside which it appears with a significantly higher frequency. We thus observe less bias in BERT, as its vectors can be applied to much broader contexts. The fMRI data does appear to be the most erratic - it is important to consider that the data was only from one participant, and thus we are viewing their personal associations. Additionally, owing to the very high dimension of fMRI data, there is more room for differences between vectors, and we would expect to see larger differences in cosine similarities. It is clear that biases are apparent in the brain imaging vectors, but surprisingly not as much as GloVe.

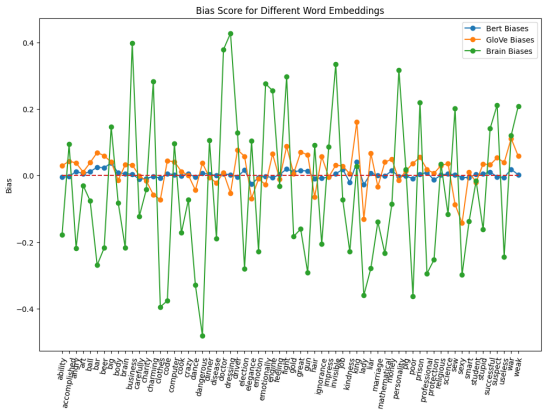


Figure 8: All bias scores

4.2.2 Bias through clustering

Our next analysis focused on clustering - using the K-Means algorithm as a method of grouping similar vectors together, we hypothesised that if biases are inherent to specific vectors, we would obtain clusters of words with each gender that would indicate associations of words to specific genders. In a bias-free environment, gender-neutral words would not be clustered with gender-specific words at all, even if they are gender-leaning. We also defined a test set of 12 selected words, 3 occupations and 3 adjectives previously found Garg et al. (2018) to be closely associated with gender. Since K-Means has a random initialization and runs until convergence, results differ from run to run. Therefore, we ran 10000 simulations and found (See Figure 9) that in an large majority of cases, male and female vectors were clustered together. The differences were stark between GloVe and BERT with the former separating between the genders much more frequently - this again may be as a result of the contextual embeddings from BERT, creating a stronger association between the vectors of gender.

Since any analysis in which male and female vectors are placed in the same cluster is not meaningful for analysing bias, and this occurred very frequently, we decided to change approaches. We set K-Means to run with 3 fixed centroids, one for

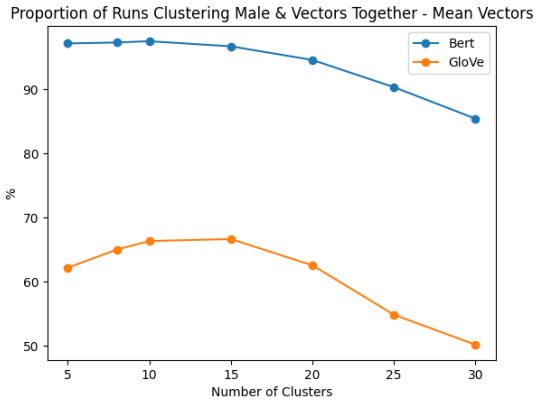


Figure 9: Proportion of clusterings grouping genders together

each gender and a third as the average vector of all words in the neutral set, thereby forcing the cluster separation between the two genders. The results here were the opposite of what we discovered by the similarity bias measure - while GloVe exhibited no bias for gender-neutral words to males and only 5 gender-neutral female words, BERT had 22 female-biased words and 26 exhibited male bias. The test set fed into this clustering were all classified as neutral for GloVe, whereas the BERT vectors were mostly classified with the gender with which they are associated. This is the expected result considering how the clusters formed. It is also interesting to note that except for the GloVe embeddings, the clustering analysis produced more bias than the vector similarity.

In keeping with the theme of this work, we attempted to perform a PCA decomposition to both the BERT and fMRI vectors. While this had little to no effect on the fMRI biases, it actually caused an increase in the number of biased concepts and the bias scores for BERT. This perhaps highlights an important understanding regarding BERT: while the principal components define the majority of the vector's general 'meaning' - as we saw with the decoding tasks in which PCA improved the decoder's ability to predict the vectors - the additional components contribute greatly to the sentiment, context and associations, adding nuance and complexity. That is why removing them may have increased the bias, as now these additional features have been stripped out of the word, leaving only the core 'meaning' components, which are more susceptible to bias.

Embedding	Method	Male	Female
BERT	Similarity	16.67%	6.06%
	Clustering	37.88%	33.33%
BERT PCA	Similarity	19.7%	12.13%
	Clustering	46.97%	7.58%
GloVe	Similarity	12.13%	19.7%
	Clustering	1.52%	9.09%
fMRI	Similarity	13.64%	10.61%
	Clustering	34.85%	1.52%

Table 1: Proportion of words with biases

4.3 Debiasing

4.3.1 Method

Debiasing embeddings is the act of reducing or eliminating gender bias in word representations to reduce stereotypes, promote fairness, and address gender inequalities in natural language processing applications. We attempted to recreate the algorithms described in the article "Man is to Computer Programmer as Woman is to Homemaker", Bolukbasi et al. (2016).

The first step involves identifying a vector subspace, namely a 'gender subspace'. We attempt to find such a k-dimensional vector subspace to account for the components of the word embeddings that represent gender, allowing us to depict the positioning of gender-related words in the embedding space. We did so by taking 'defining sets' - subsets of various words that are used to define or are closely associated with gender. By performing a Singular-values decomposition to a matrix defined by the squared distance of each vector from its subset mean, the top-k singular values return a k-sized gender subspace. That is, defining C as follows: $C := \sum_{i=1}^{n} \sum_{w \in D_i} (w_i - \mu_i)^T (w_i - \mu_i)/|D_i|$ we take the top k rows of SVD(C), where D_i is the i-th defining set, μ_i its mean, and w a word vector.

We can then use this subspace to perform two debiasing tasks - "Neutralize" and "Equalize". Neutralize ensures that words considered gender-neutral have zero influence in the gender subspace, by subtracting from the original vector its components that exist in the gender subspace (obtained by projecting the vector onto the gender subspace). It thus removes any bias associated with those words.

Equalize, on the other hand, forces equidistance between gender-specific pairs for components out-

side the gender subspace. This aims to ensure that the only difference between those vectors should be in the gender definition of the word. For example, the words grandfather and grandmother should only differ within the gender subspace. This would result in any gender-neutral word having the same similarity or distance to both words in the pair.

4.3.2 Results

After successfully recreating the algorithms described, we created a neutralized set and an equalized set, similar to the sets used in the article, by selecting words from a large corpus of GLoVe vectors. In order to assess the effectiveness of the debiasing, we would need to verify that the newly 'debiased' words are not found to be biased by our previous methods, while the semantic meaning of the vectors is still maintained. To do so, we took a selection of the words previously found to be biased and performed the above debiasing on them.

When evaluating for bias via the similarity method used above, the results were remarkable. Only three words exhibited bias: "king" displayed bias towards males, "lady" showed bias towards females, and "hair" exhibited a very low bias towards females. Since the two former vectors are the only two in that set that are gender-specific, it is important that they retained their 'bias', which can be argued is not actually bias at all, but inherent to its meaning. The fact that only one other word exhibited bias, and in a small amount, reflects a success for the debiasing method.

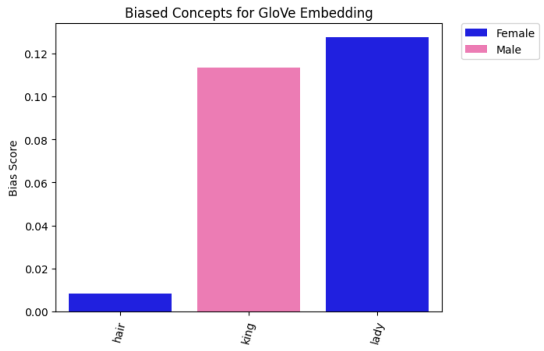


Figure 10: GLoVe biased concepts, post-debiasing

We then performed again the decoder analysis from the structured task on the same set of concepts and fMRI data as in previous experiments. The average scores obtained from each fold were very close to the original results. The top three successful concepts remained identical, while some of the remaining top 15 concepts had similar words but with slightly different rankings. This indicates that the contextual relationships and semantic meanings between words were preserved after debiasing.

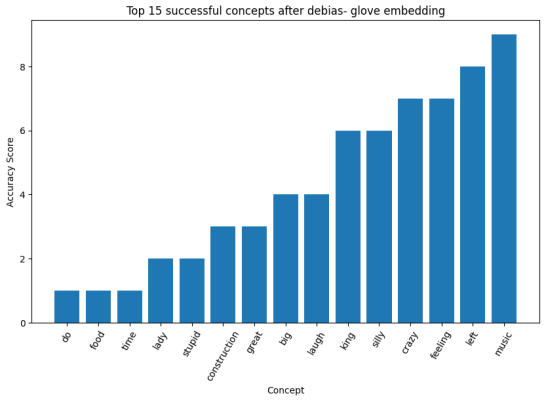


Figure 11: GLoVe successful concepts, post-debiasing

Overall, the results demonstrated the effectiveness of the debiasing algorithms in reducing biases in word embeddings while preserving contextual relationships between words. The results indicated significant improvements in mitigating gender bias and promoting equality and accuracy in language processing tasks.

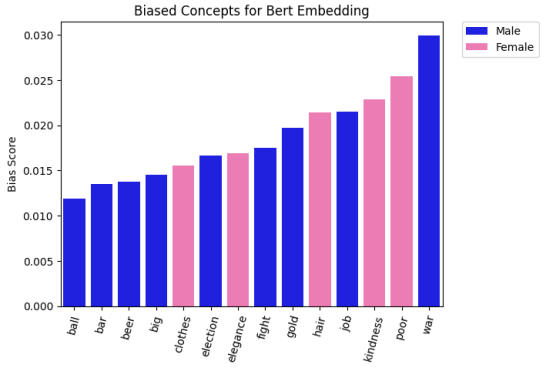


Figure 12: BERT biased concepts, post-debiasing

Finally, we ran the debiasing and analysis again on BERT embeddings. The results here were hardly satisfactory, and the debiasing was not as effective as we would have hoped. Most of the original bias was retained, and new biased concepts were added. Moreover, the two gender-specific words, for which we would hope their bias would be retained, became entirely unbiased, indicating that not only is the debiasing ineffective at removing bias from BERT, but also perhaps that the semantic meanings are not preserved. Alternatively, it could be that there is simply no well-defined gender subspace within the BERT vectors, as more nuance and complexity define the vector's meaning, whereas

GloVe does have such a subspace. Additionally, the performance of the decoder model was significantly worse post-debias, confirming that indeed, those components were necessary to the vector's semantic representation.

In any event, GloVe was the more biased of the two embeddings, and it is thus less bothersome that the debaising was unsuccessful on BERT.

5 Conclusions & Further Research

Over the course of this work, we have analysed the differences between BERT and GloVe embeddings, and have found that in general brain-to-word encoding and decoding tasks, both models generate embeddings that provide excellent results. We did not find a significant advantage to BERT, but this may be owing to the nature of the tasks chosen, which seem better suited to GloVe from the outset. Specifically regarding bias, neither model exhibits exceptional bias, but BERT did exhibit significantly less bias than GloVe. Finally, we showed that Debiasing is an effective method for removing bias, but only on GloVe embeddings.

The limitations of this research relate primarily to the fMRI data, which was taken from only one test subject, and on a limited number of concepts. Ideally, further research would perform the fMRI bias analyses on a larger corpus of words, and averaging over multiple test subjects. Additionally, time limitations prevented us from attempting a soft-debias technique, which aims to preserve more context and meaning in the vector than the hard debias methods described above. It could perhaps be more effective with more complex models such as BERT.

Acknowledgments

We would like to thank the course staff for a meaningful semester in which were challenged positively and learned a fortune. We would particularly like to thank Mr. Refael Tikochinsky for his invaluable advice guiding us throughout the course of this project.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458. Funding Information: This work was supported by grants from the National Science Foundation (NSF; IIS1208203), the National Eye Institute (EY019684), and from the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13.

Appendix

Please note that the entire length of the document without figures is 7 pages.

Code

The code is the work of the authors and may be reproduced for research purposes only. A full code notebook can be found here: `https://github.com/idanh8/glove-bert-debiasing/blob/710e75807882c336ca4ecc931a6fa9ec9f9f1acd/Language_%26_Cognition_Project_Final.ipynb`

Word Lists

The original 180 concepts, as well as the Experiment 2 and 3 Sentences, are taken directly from(Pereira et al., 2018) and will not be listed here. The subset of those 180 concepts we defined as gender-leaning (including the two gender-specific words, king and lady) are the following: ['ability', 'accomplished', 'angry', 'art', 'ball', 'bar', 'beer', 'big', 'body', 'brain', 'business', 'carefully', 'charity', 'charming', 'clothes', 'code', 'computer', 'cook', 'crazy', 'dance', 'dangerous', 'dinner', 'disease', 'doctor', 'dressing', 'driver', 'election', 'elegance', 'emotion', 'emotionally', 'engine', 'feeling', 'fight', 'gold', 'great', 'gun', 'hair', 'ignorance', 'impress', 'invisible', 'job', 'kindness',

'king', 'lady', 'liar', 'marriage', 'mathematical', 'money', 'personality', 'pig', 'poor', 'prison', 'professional', 'protection', 'religious', 'science', 'sew', 'sexy', 'smart', 'student', 'stupid', 'successful', 'suspect', 'useless', 'war', 'weak']

The neutral set is given as follows:['apartment', 'applause', 'argument', 'argumentatively', 'attitude', 'bag', 'bear', 'beat', 'bed', 'bird', 'blood', 'broken', 'building', 'burn', 'camera', 'challenge', 'cockroach', 'collection', 'construction', 'counting', 'damage', 'dedication', 'deceive', 'deliberately', 'delivery', 'dessert', 'device', 'dig', 'dissolve', 'disturb', 'do', 'dog', 'economy', 'electron', 'event', 'experiment', 'extremely', 'fish', 'flow', 'food', 'garbage', 'help', 'hurting', 'illness', 'invention', 'investigation', 'jungle', 'land', 'laugh', 'law', 'left', 'level', 'light', 'magic', 'material', 'mechanism', 'medication', 'mountain', 'movement', 'movie', 'music', 'nation', 'news', 'noise', 'obligation', 'pain', 'philosophy', 'picture', 'plan', 'plant', 'play', 'pleasure', 'quality', 'reaction', 'read', 'relationship', 'residence', 'road', 'sad', 'seafood', 'sell', 'shape', 'ship', 'show', 'deceive', 'silly', 'sin', 'skin', 'smiling', 'solution', 'soul', 'sound', 'spoke', 'star', 'sugar', 'table', 'taste', 'team', 'texture', 'time', 'tool', 'toy', 'tree', 'trial', 'tried', 'typical', 'unaware', 'usable', 'vacation', 'wash', 'wear', 'weather', 'willingly', 'word']

The list of words used to create the male vector is: ['he', 'son', 'his', 'him', 'father', 'man', 'boy', 'himself', 'male', 'brother', 'sons', 'fathers', 'men', 'boys', 'males', 'brothers', 'uncle', 'uncles', 'nephew', 'nephews'], and the female list: ['she', 'daughter', 'hers', 'her', 'mother', 'woman', 'girl', 'herself', 'female', 'sister', 'daughters', 'mothers', 'women', 'girls', 'females', 'sisters', 'aunt', 'aunts', 'niece', 'nieces'] The test set is the following: ['engineer', 'lawyer', 'soldier', 'nurse', 'dancer', 'housekeeper', 'loyal', 'honest', 'strong', 'maternal', 'attractive', 'tidy']. The above 3 lists are all adapted from Garg et al. (2018)

The defining pairs, pairs for equalizing, and set of words to neutralize are given respectively: Defining pairs = [”woman”, ”man”, ”girl”, ”boy”, ”she”, ”he”, ”mother”, ”father”, ”daughter”, ”son”, ”gal”, ”guy”, ”female”, ”male”, ”her”, ”his”, ”herself”, ”himself”, ”mary”, ”john”] Equalizing pairs= [”monastery”, ”convent”, ”spokesman”, ”spokeswoman”, ”dad”, ”mom”, ”men”, ”women”, ”councilman”, ”councilwoman”, ”grandpa”, ”grandma”, ”grandsons”, ”granddaughters”, ”testosterone”, ”estrogen”, ”uncle”, ”aunt”, ”wives”, ”husbands”, ”father”, ”mother”, ”grandpa”, ”grandma”, ”he”, ”she”, ”boy”, ”girl”, ”boys”, ”girls”, ”brother”, ”sister”, ”brothers”, ”sisters”, ”businessman”, ”businesswoman”, ”chairman”, ”chairwoman”, ”colt”, ”filly”, ”congressman”, ”congresswoman”, ”dad”, ”mom”, ”dads”, ”moms”, ”dudes”, ”gals”, ”father”, ”mother”, ”fatherhood”, ”motherhood”, ”fathers”, ”mothers”, ”fella”, ”granny”, ”fraternity”, ”sorority”, ”gelding”, ”mare”, ”gentleman”, ”lady”, ”gentlemen”, ”ladies”, ”grandfather”, ”grandmother”, ”grandson”, ”granddaughter”, ”he”, ”she”, ”himself”, ”herself”, ”his”, ”her”, ”king”, ”queen”, ”kings”, ”queens”, ”male”, ”female”, ”males”, ”females”, ”man”, ”woman”, ”men”, ”women”, ”nephew”, ”niece”, ”prince”, ”princess”, ”schoolboy”, ”schoolgirl”, ”son”, ”daughter”, ”sons”, ”daughters”] Neutralizing words = [”tree”, ”table”, ”paper”, ”wall”, ”floor”, ”door”, ”window”, ”lamp”, ”computer”, ”book”, ”car”, ”chair”, ”plant”, ”clock”, ”shirt”, ”hat”, ”plate”, ”spoon”, ”fork”, ”mirror”, ”cup”, ”picture”, ”painting”, ”brush”, ”knife”, ”blanket”, ”pillow”, ”box”, ”bowl”, ”dog”, ”cat”, ”puppy”, ”kitten”, ”pen”, ”canvas”, ”paintbrush”, ”person”, ”individual”, ”human”, ”being”, ”creature”, ”mortal”, ”figure”, ”body”, ”folk”, ”society”, ”mankind”, ”people”, ”humanity”, ”personage”, ”individuality”, ”soul”, ”individualism”, ”specimen”, ”identity”, ”selfhood”, ”existence”, ”life”, ”personality”, ”essence”, ”consciousness”, ”somesuch”, ”personhood”, ”somebody”, ”lifeform”, ”personification”, ”subject”, ”homosapien”, ”someone”, ”humanoid”, ”character”, ”war”, ”money”, ”job”, ”gun”, ”great”, ”fight”, ”election”, ”big”, ”beer”, ”protection”, ”kindness”, ”elegance”, ”liar”, ”engine”, ”driver”, ”bar”, ”sexy”, ”sew”, ”marriage”, ”hair”, ”emotionally”, ”dressing”, ”dance”, ”clothes”, ”charming”, ”invisible”, ”doctor”, ”business”, ”poor”, ”dangerous”, ”code”, ”gold”, ”crazy”, ”ball”, ”charity”, ”carefully”, ”weak”]

Additional Figures

Structured Task

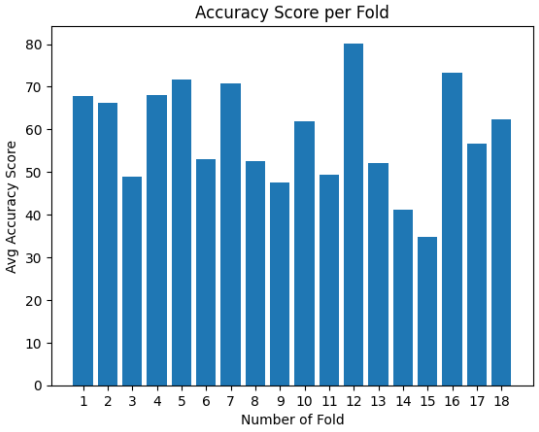


Figure 13: Accuracy Ranks for Each Fold, BERT

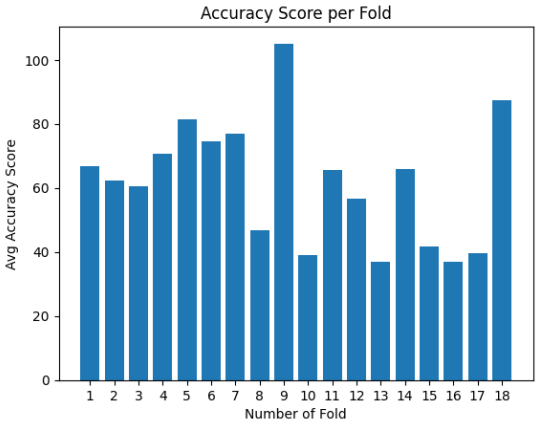


Figure 14: Accuracy Ranks for Each Fold, GloVe

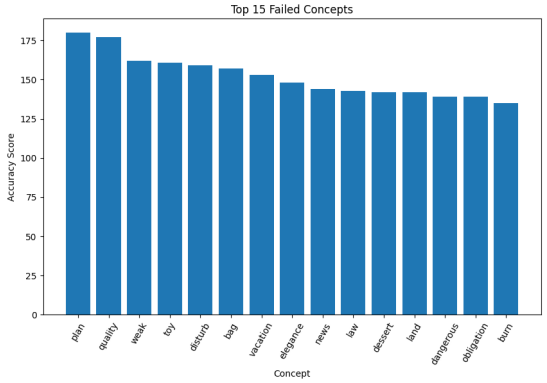


Figure 15: Failed Decoder Concepts, BERT

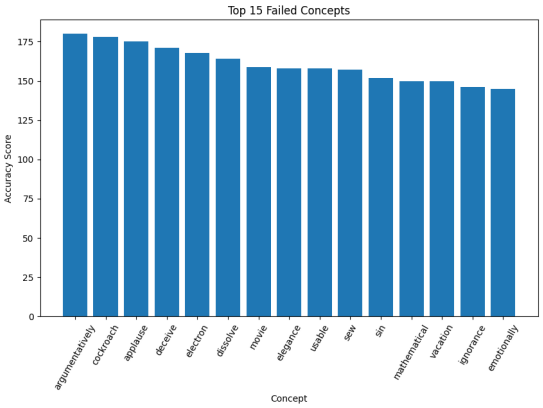


Figure 16: Failed Decoder Concepts, GloVe

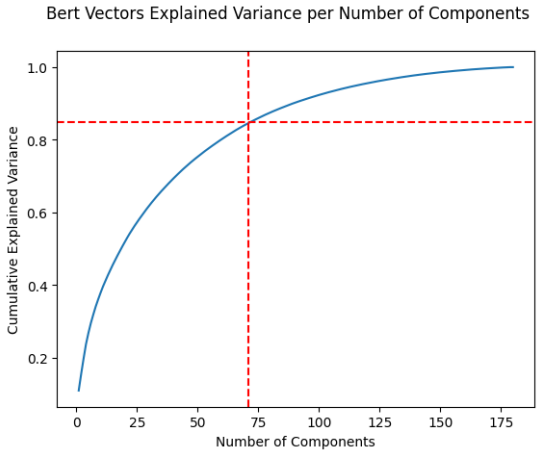


Figure 17: PCA Results, BERT

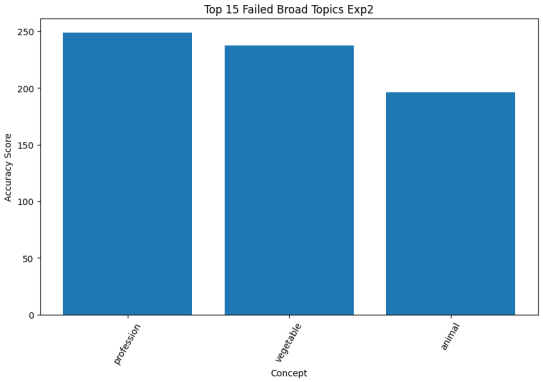


Figure 18: Failed Broad Topics, GloVe, Exp. 2

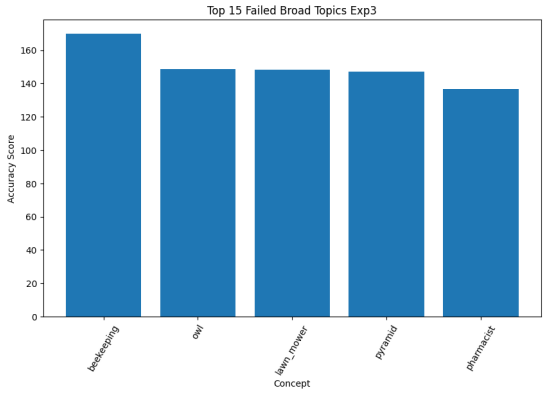


Figure 19: Failed Broad Topics, GloVe, Exp. 3

Semi-structured Task

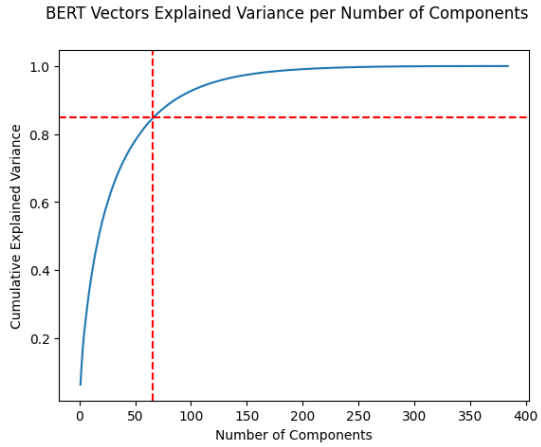


Figure 20: PCA Results, BERT

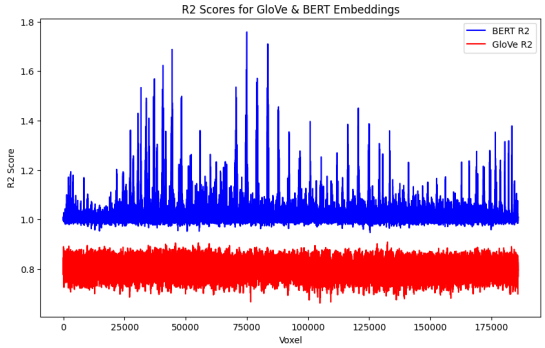
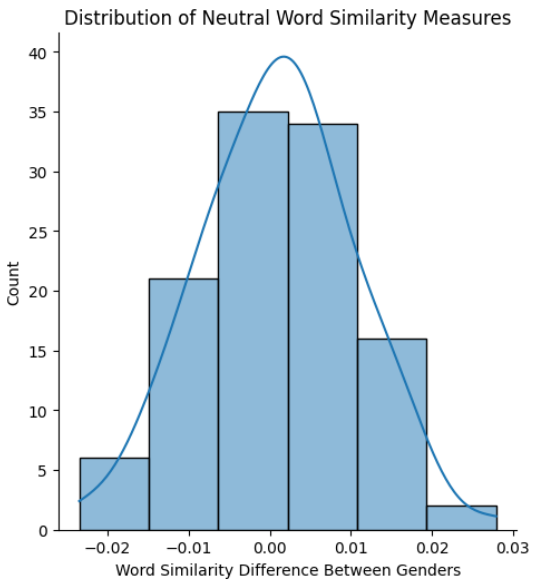
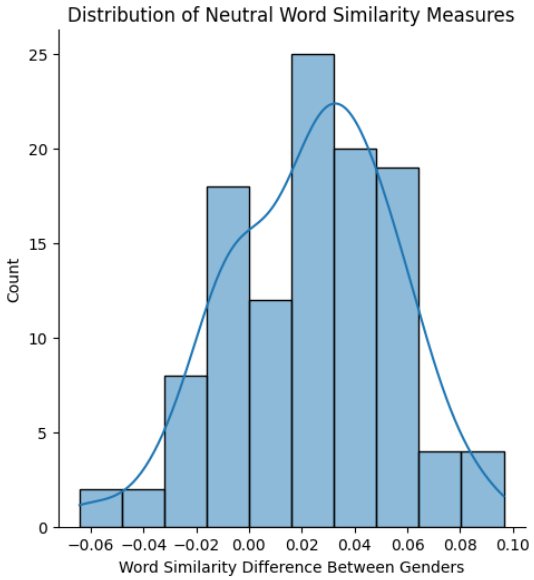


Figure 21: Voxel R^2 score, pre-PCA

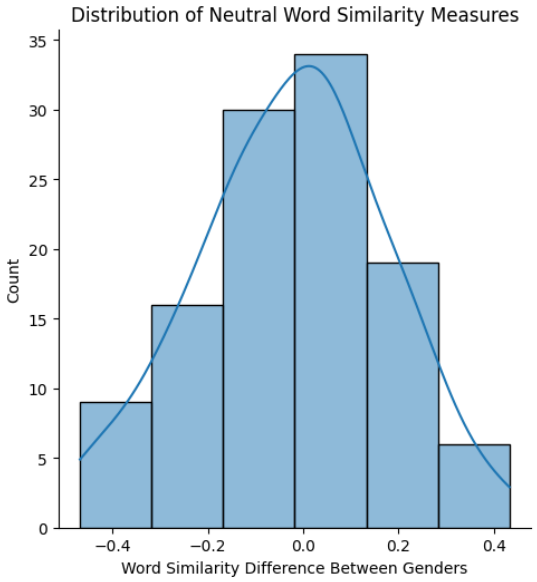
Open Tasks



(a) BERT

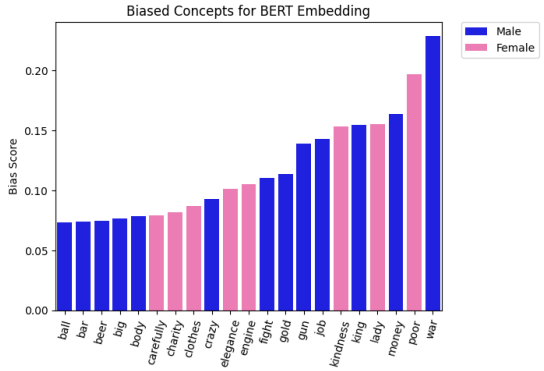


(b) GloVe

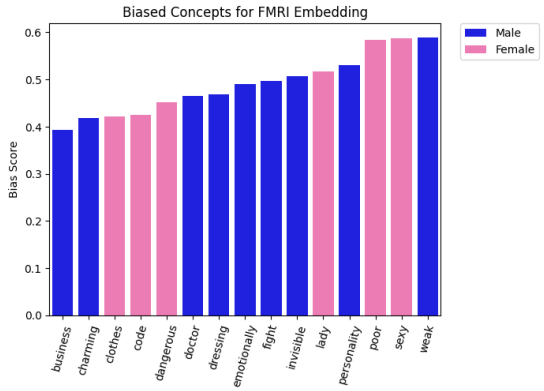


(c) fMRI

Figure 22: Bias Distributions for each embedding



(a) BERT



(b) fMRI

Figure 23: Biased concepts for embeddings, post PCA

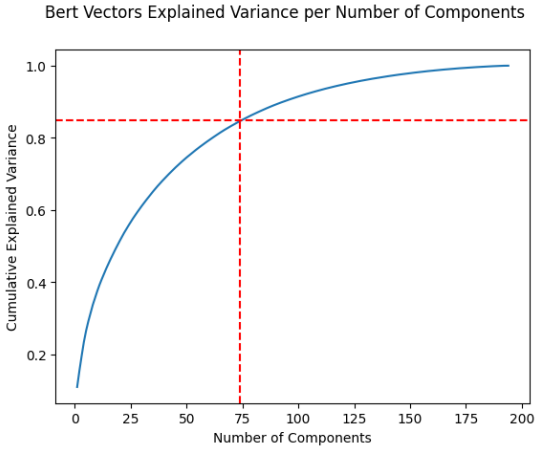


Figure 24: PCA Results, BERT

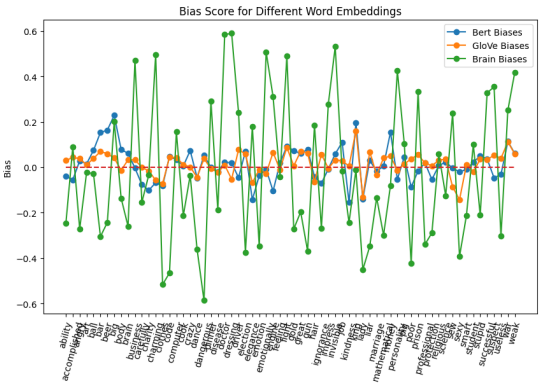


Figure 25: Post PCA Bias Rankings

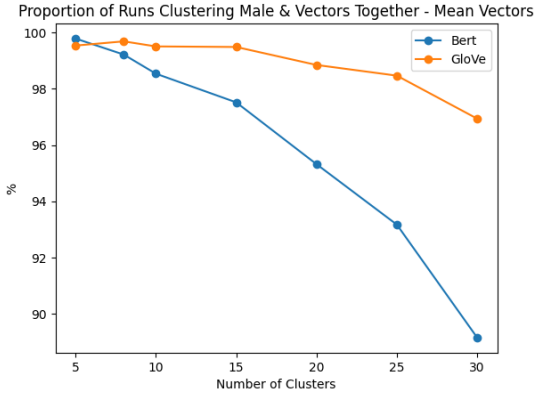
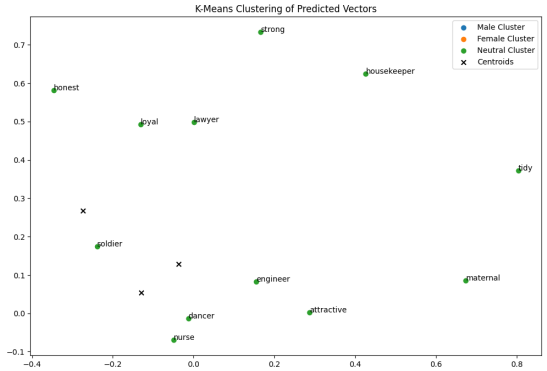
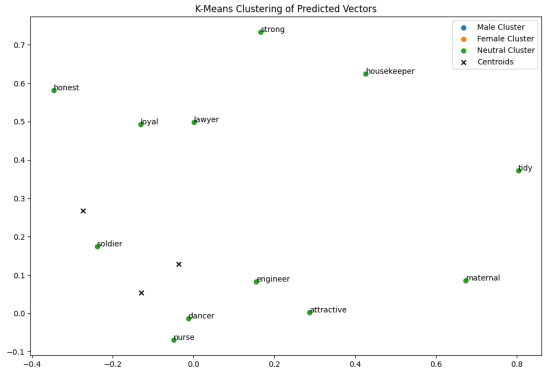


Figure 26: Cluster gender splits on single word gender vectors

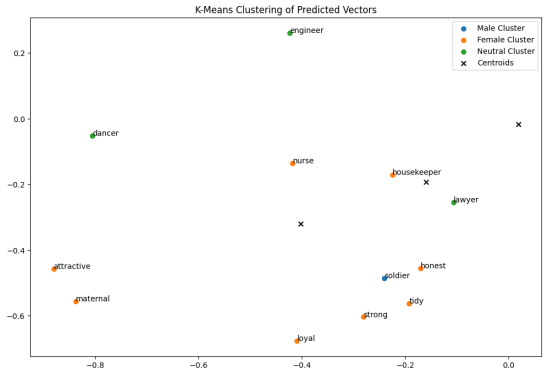


(a) Full Set

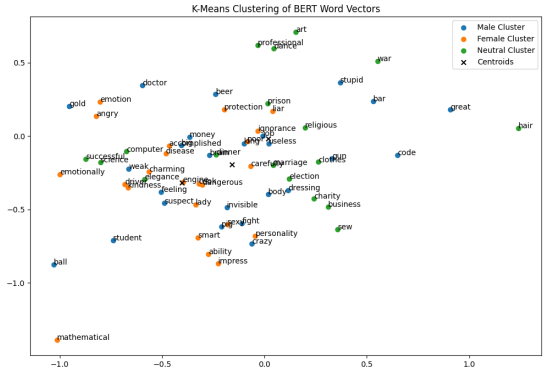


(b) Test Set

Figure 27: Cluster results, GloVe embeddings



(a) Full Set



(b) Test Set

Figure 28: Cluster results, BERT data

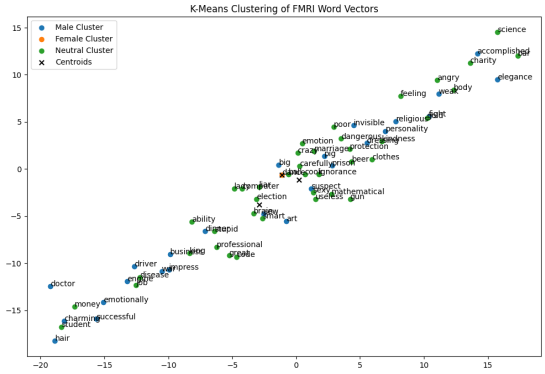


Figure 29: Cluster results, fMRI data

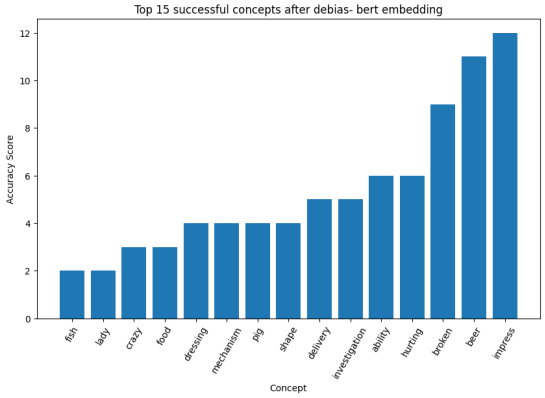


Figure 30: BERT encoder successful concepts post debiasing