DSA4212: Word Embeddings

Bao Jiaqi, A0211257N, bao.jiaqi@u.nus.edu

April 12, 2024

1 Introduction

In the context of a collection of documents and a vocabulary consisting of a set of words, an intriguing problem arises: finding an embedding vector of dimension d, denoted as $\mathbf{v}_w \in \mathbb{R}^d$ to represent each word w in the vocabulary. We denote the embedding matrix, with its rows representing words' embedding vectors, as $\mathbf{V}$. The objective is to ensure that words with similar semantic meanings are positioned closely to each other in the embedding space. Embedding serves as the foundation for many applications in various downstream natural language processing tasks.

This report is organized as follows. In section 2, we will introduce 3 methods that aim to infer the word embeddings by solving 3 distinct optimization problems. Subsequently, in subsection 3.2, we will train the embeddings on a particular vocabulary based on 2 different movie review corpora. We will then evaluate the obtained embeddings using 2 tasks: word similarity calculation and document sentiment classification. Furthermore, in subsection 3.3, we will compare the biases present in the embeddings towards sentiment words and the gender bias between the two corpora. Finally, in subsection 3.4, we will interpret the learned embeddings and utilize them to cluster words.

2 Methods

2.1 Skip-Gram with Negative Sampling (SGNS)

Consider a word and context pair (w, c) within a collection of W words and C contexts. Here a context is also a word and we assume that the set of words and contexts are identical in later applications. Skip-Gram [MCCD13] models the probability of (w, c) co-occurring in a context, such as within a window size of length 5 or in a sentence, with the word embedding $\mathbf{v}_w$ for w and the context embedding $\mathbf{e}_c$ denoted as follows:

$$p(y = 1|w, c) = \sigma(\mathbf{v}_w \cdot \mathbf{e}_c)$$

Here, $\sigma(x) = \frac{1}{1+\exp(-x)}$ and $y = 1$ indicates the co-occurrence of (w, c) in a context, while $y = 0$ corresponds to the absence of co-occurrence. Thus, we also have $p(y = 0|w, c) = \sigma(-\mathbf{v}_w \cdot \mathbf{e}_c)$. While we can only obtain word pairs with $y = 1$ from a corpus, negative sampling generates word pairs with $y = 0$.

For each co-occurrence (w, c), [MSC+13] proposes generating k negative samples of the form $(w, c_N, y = 0)$, where c_N is independently drawn from the categorical distribution Categorical(θ). θ is a distribution over all words, with the recommended value given by $\theta_w \propto f_w^{3/4}$, where $f_w = \frac{\#w}{D}$ represents the empirical frequency of word w among all D word-context pairs. Finally, the loss function is the negative log-likelihood of all observed and sampled pairs $\{(w_i, c_i, y_i)\}_{i=1}^n$, and can be expressed as:

$$\ell(\{\mathbf{v}_w\}_{w=1}^W, \{\mathbf{e}_c\}_{c=1}^C) = -\sum_{i=1}^{n} y_i \log(\sigma(\mathbf{v}_{w_i} \cdot \mathbf{e}_{c_i})) + (1 - y_i) \log(1 - \sigma(\mathbf{v}_{w_i} \cdot \mathbf{e}_{c_i})) \tag{1}$$

To solve for the word embedding $\{\mathbf{v}_w\}_{w=1}^W$, we can employ stochastic gradient descent schemes. For a single observation (w_i, c_i, y_i), the gradient w.r.t. $\mathbf{v}_{w_i}$ can be derived as $\mathbf{e}_{c_i}(\sigma(\mathbf{v}_{w_i} \cdot \mathbf{e}_{c_i}) - y_i)$, and the gradient w.r.t. $\mathbf{e}_{c_i}$ can be derived as $\mathbf{v}_{w_i}(\sigma(\mathbf{v}_{w_i} \cdot \mathbf{e}_{c_i}) - y_i)$.

2.2 Spectral method and the shifted PPMI matrix (SPPMI)

In their work, [LG14] offers theoretical insights into Skip-Gram with Negative Sampling (SGNS). They demonstrate that SGNS implicitly performs factorization on a word-context matrix. It is derived in the paper that the embedding vectors $\mathbf{v}_w$ and $\mathbf{e}_c$ that minimize Equation 1 satisfy the following equation:

$$\mathbf{v}_w \cdot \mathbf{e}_c = \log(\frac{\#(w,c)D}{\#w\#c}) - \log k = \text{ PMI}(w,c) - \log k$$

where $\text{PMI}(w,c) = \log(\frac{\#(w,c)D}{\#w\#c}) = \log(\frac{\#(w,c)/D}{(\#w/D)(\#c/D)})$ represents the empirical pointwise mutual information of word w and context c. Consequently, SGNS can be viewed as performing factorization on the shifted PMI matrix $M \in \mathbb{R}^{W \times C}$, where $M_{wc} = \text{PMI}(w,c) - \log k$.

However, factorizing the shifted PMI matrix presents a computational challenge. This is mainly due to the fact that many entries of matrix M are expected to be $-\log 0$, as it is highly likely that $\#(w,c) = 0$. Consequently, M becomes dense and even ill-defined in numerous entries. To address this issue, a common approach is to use the positive PMI matrix, where all negative PMI values are replaced by zero, emphasizing positive correlations. To proceed, we perform singular value decomposition (SVD) on the shifted PPMI matrix M, where M where

$$M_{wc} = \max(0,\ \text{PMI}(w,c)) - \log k$$

It is important to note that the w-th row of M can also be treated as an embedding of dimension C for word w. Thus, we refer to this approach as SPPMI. Let d denote the embedding dimension, and consider $M_d = U_d \Sigma_d V_d^T$, which represents the rank d SVD approximation of M. M_d can be considered as the matrix of rank d that best approximates M in the Frobenius norm. In other words, $M_d = \arg\min_{\text{rank}(M')=d} ||M' - M||_F^2$

[LGD15] has identified two scenarios where SVD performs poorly. In the first scenario, the rows of $U_d(\Sigma_d)^\gamma$ and $(\Sigma_d)^\gamma V_d$ yield word and context embeddings, respectively, with γ as a tuning parameter. It is advisable to set a small value of $\gamma = 0$. Intuitively, when $\gamma = 1$, context embeddings are orthogonal, while word embeddings are not. However, context and word embeddings are expected to be symmetrical. Empirical observations indicate that setting $\gamma = 1$ leads to unfavorable results in word similarity tasks when comparing cosine similarity with human-assigned similarity scores. Furthermore, it has been observed that increasing the value of k does not enhance the performance of SVD on the SPPMI matrix. Therefore, it is recommended to apply SVD on the PPMI matrix. However, it is observed empirically that increasing k improves the performance of Shifted SPPMI and SGNS.

2.3 Global vectors for word representation (Glove)

SGNS trains on individual local context windows, which limits its ability to effectively utilize global corpus statistics and the abundance of repetitions. In contrast, Glove [PSM14] introduces a weighted least squares problem that leverages global word-word co-occurrence counts to make more efficient use of corpus statistics.

Consider a collection of W words, and let X denote the word-word co-occurrence matrix, where X_{wc} represents the number of times word c occurs in the context of word w. If the context is defined as co-occurrence within a window size of length 5 or within a sentence, then X is symmetric. Glove aims to find embeddings $\{\mathbf{v}_w\}_{w=1}^W$ and $\{\mathbf{e}_c\}_{c=1}^W$ by minimizing the following objective function: $\{\mathbf{v}_w\}_{w=1}^W, \{\mathbf{e}_c\}_{c=1}^W$ by minimizing

$$\sum_{w=1}^{W}\sum_{c=1}^{W} f(X_{wc})(\mathbf{v}_w \cdot \mathbf{e}_c + b_w + b_c - \log(X_{wc}))^2 \tag{2}$$

where the biases b_w and b_c are optimized, and the recommended weighting function $f(X_{wc})$ is defined as $f(x) = (x/x_{\max})^\alpha$ if $x < x_{\max}$; otherwise, $f(x) = 1$. Stochastic gradient descent can be used to optimize the objective function by updating the biases and embeddings. The gradient w.r.t. embedding such as $\mathbf{v}_w$ can be easily computed as $2f(X_{wc})(\mathbf{v}_w \cdot \mathbf{e}_c + b_w + b_c - \log(X_{wc}))\mathbf{e}_c$. The The gradient w.r.t. bias such as b_w can be easily derived as $2f(X_{wc})(\mathbf{v}_w \cdot \mathbf{e}_c + b_w + b_c - \log(X_{wc}))$. Finally, the embedding for word w is obtained as $\frac{\mathbf{v}_w + \mathbf{e}_w}{2}$.

3 Applications to movie reviews corpora

3.1 Data and parameter settings

We utilize the Chinese movie review dataset[1] and collect two distinct corpora. Each corpus consists of reviews associated with a movie and includes ratings ranging from 1 to 5. We classify reviews with a rating score of ≥ 4 as positive sentiment and those with a rating score ≤ 3 as negative sentiment. Corpus 1 comprises a random sample of 10,000 positive reviews and 10,000 negative reviews on the movies "Tiny Times 1.0" (2013) or "Tiny Times 3.0" (2014). Similarly, corpus 2 consists of a random sample of 10,000 positive reviews and 10,000 negative reviews on the movie "La La Land" (2016).

Initially, we preprocess the data by splitting the reviews with punctuations (including ,.!?) and tokenizing the text using Jieba[2]. We consider a context window size of 5 and define the context set to be the same vocabulary as the word set. From the corpus, we select $W = 829$ words that have at least 2 Chinese characters and appear at least 200 times in both corpora.

We will train 2 embeddings for all 829 words based on 2 corpora. Here are some details about the parameter settings for each algorithm. For SGNS, we set $k = 5$ and learning rate equals 0.05. For SVD and SPPMI, we experiment with $\gamma = 0$ and $k = 1, 5$. For Glove, we set $x_{\max} = 100, \alpha = 0.75$ and learning rate equals 0.1. For all methods except SPPMI, we choose embedding dimension size $d = 50$.

3.2 Evaluation

3.2.1 Word similarity

We will evaluate our learned embedding vectors by comparing them to a pre-trained Chinese word embedding text2vec[3]. This pre-trained embedding has a dimension of 768 and has been trained on a large-scale corpus. To perform the evaluation, we will calculate the cosine similarity between any two words within each of our embeddings $\mathbf{V}^i$, denoted as $\{\text{cosine similarity}(\mathbf{v}_w^i, \mathbf{v}_c^i)\}_{w<c}$. Then for each $\mathbf{V}^i$, we calculate

pearson correlation$(\{\text{cosine similarity}(\mathbf{v}_w^i, \mathbf{v}_c^i)\}_{w<c}, \{\text{cosine similarity}(\mathbf{v}_w^{\text{text2vec}}, \mathbf{v}_c^{\text{text2vec}})\}_{w<c})$ in Table 1.

	SGNS	Glove	SVD ($k = 1$)	SVD ($k = 5$)	SPPMI ($k = 1$)	SPPMI ($k = 5$)
Corpus 1	0.0691	-0.0197	0.0479	0.0451	**0.1546**	0.0305
corpus 2	0.0828	-0.0093	0.0489	0.0454	**0.1426**	0.0263

Table 1: Word similarity evaluation. Each entry represents the pearson correlation of pairwise word similarity with the external text2vec embedding.

We can observe that the SPPMI embeddings align most closely with the external text2vec embedding, and the SGNS embeddings also exhibit relatively high correlation with the text2vec. However, Glove embeddings align least closely with the text2vec embeddings. Additionally, the performance of SVD and SPPMI embeddings in this word similarity task both show a decline when using a larger negative sample size k.

3.2.2 Sentiment classification

One application of word embeddings is semantic classification. In this project, each corpus consists of positive and negative reviews. After removing reviews that contain zero words present in our vocabulary, we obtained 9,128 positive reviews and 8,640 negative reviews in Corpus 1, and 8,913 positive reviews and 9,255 negative reviews in Corpus 2. For each corpus, we split all reviews into training and test sets with an 8:2 ratio. Following a conventional approach, we represent each review by taking the average of the word embeddings of the words within the review, excluding any words not in our vocabulary. Additionally, we consider One-Hot encoding as a word embedding with a dimension of $W = 829$.

[1] https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/dmsc_v2/intro.ipynb
[2] https://github.com/fxsjy/jieba
[3] https://github.com/shibing624/text2vec

In addition, we explore the use of Latent Dirichlet Allocation (LDA) [BNJ03] for document or review representation, using the Gibbs sampling package[4]. For each corpus, we select 50 topics and estimate the topic distribution to represent each review. Unlike other methods, LDA does not rely on local word-word co-occurrence information in small windows, but instead exploits document or review-level information. We experiment with two classification models: Logistic Regression[5](LogReg) without ℓ_2 penalty and K-Nearest Neighbors[6] (KNN) with 10 neighbors. We train the LogReg and KNN classifiers on the training set obtained from each method of review representation. Subsequently, we evaluate the trained models on the corresponding test set.

	SGNS	Glove	SVD ($k=1$)	SVD ($k=5$)	SPPMI ($k=1$)	SPPMI ($k=5$)	One-Hot	text2vec	LDA
Corpus 1 (LogReg)	0.6775	0.6848	0.6913	0.6637	0.7326	0.7290	0.7304	**0.7341**	0.6536
Corpus 2 (LogReg)	0.6970	0.7003	0.7025	0.6824	0.7179	0.7176	0.7132	**0.7220**	0.6719
Corpus 1 (KNN)	0.6631	0.6789	**0.6857**	0.6699	0.6845	0.6713	0.6637	0.6769	0.6041
Corpus 2 (KNN)	0.6730	**0.6788**	0.6469	0.6330	**0.6788**	0.6505	0.6576	0.6706	0.6219

Table 2: Sentiment classfication evaluation. Each entry represents accuracy on the test set of the classifier trained from the training set.

Table 2 presents a comparison of the binary sentiment classification test accuracy for each corpus. The results indicate that embeddings derived from Glove, SPPMI ($k=1$), SVD ($k=1$), and text2vec are effective document representations for predicting sentiment. However, increasing the negative sample size k leads to less effective representations in this sentiment classification task.

Additionally, we observe that the One-Hot representation is effective in predicting sentiment with LogReg, but its performance is the lowest with KNN. This discrepancy may be due to One-Hot embedding relying solely on word indices, which challenges its ability to capture geometric distance and similarity between reviews. In contrast, other word embeddings are not affected by the transition from LogReg to KNN, as they capture semantic distances between reviews based on word embeddings.

Furthermore, it is clear that the review topic distribution representation derived from LDA may not be a reliable indicator of the review's sentiment. There are several potential reasons for this. Firstly, the distance metric in KNN is typically the standard Euclidean distance, which may not accurately measure the difference in topic distribution. Secondly, LDA does not leverage local word context, which could impact its ability to capture sentiment. Lastly, LDA may face challenges when dealing with short documents [YGLC13], which is a common characteristic of our review data. The average length of reviews (number of words in our vocabulary) in Corpus 1 is 12.4 words and in Corpus 2 is 14.8 words. Therefore, topic distribution may not be an effective representation for short documents, while word embeddings that leverage local context can address the sparsity issue and representing short documents effectively.

3.3 Embedding bias

The literature contains numerous studies that investigate the differences in association strength, also known as embedding bias, between a set of target words (e.g., "nurse" and "engineer") and two sets of attribute words (e.g., "female", "male"). According to gender stereotypes, the embedding of "nurse" is expected to be more associated with women-related words than men-related words, while the embedding of "engineer" is expected to be more associated with men-related words than women-related words. If the gender stereotypes exist in the training corpus, then it is natural that trained embeddings also exhibit the bias.

It is important to note that word embeddings trained from different corpora may reflect varying levels of bias. For instance, [GSJZ18] demonstrates that changes in embedding bias across corpora from different decades can reflect societal shifts, such as the US women's movement in the 1960s. Additionally, [HLJ16] constructs historical word embeddings from various time periods and discovers that words with low frequency and polysemy have higher rates of semantic changes. Furthermore, [YR21] shows that embeddings trained from censored and uncensored corpora exhibit different associations with concepts such as democracy, freedom, and equality. The study also reveals that different embeddings affect the results of embedding-based sentiment classification related to these concepts.

[4] https://github.com/lda-project/lda/

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

In this subsection, we will examine the bias existing in the trained embeddings, specifically focusing on gender and sentiment biases.

3.3.1 Aspect based sentiment comparison

In this project, we explore two embedding spaces created from two different corpora. Our main focus is investigating the association or bias between *target* words related to different aspects of movies, such as directing and acting, and *attribute* words that reflect positive sentiments (e.g., "5 stars" and "recommend") and negative sentiments (e.g., "awful" and "low score") within two sets of movie reviews.

Beyond the context of movie reviews, this technique of quantifying embedding bias can also be applied to aspect-based sentiment analysis. For example, if we have two competing items (e.g., movies and restaurants) and want to compare user sentiment on different aspects (targets) such as directing and acting in movie reviews, or location and taste in restaurant reviews, we can train embeddings from the reviews on these two items separately. We can then select specific target and attribute words to examine their association, allowing us to compare different aspects between the two items in terms of user preference.

type	class name	example words
target	acting	acting, leading actor, performance
target	writing	script, plot, screenwriter
target	directing	director, shooting, cinematography
target	appearance	costume, look, clothes
target	life	youth, love, growth
attribute	positive sentiment	five stars, good, recommend
attribute	negative sentiment	one star, bad, terrible

Table 3: Summary of target and attribute words selected.

To evaluate the association or bias between target words and attribute words in different embeddings, we employ evaluation statistics in [CBN17, YR21]. We manually select 5 classes of target words and two classes of attribute words (positive and negative). The summary of the selected words is provided in Table 3, and the full list can be found in the accompanying python notebook. The evaluation statistics is defined as follows. Let $\mathbf{V}^1$ and $\mathbf{V}^2$ represent the word embeddings trained from two different corpora, respectively. Denote each class of target words as $\mathcal{T}_i$, the set of positive attribute words as $\mathcal{P}$, and the set of negative attribute words as $\mathcal{N}$. For each target word $t \in \mathcal{T}_i$, we compute:

$$s(t, \mathcal{N}, \mathcal{P}, \mathbf{V}^1) = \sum_{p \in \mathcal{P}} \text{cosine similarity}(\mathbf{v}_p^1, \mathbf{v}_t^1)/|\mathcal{P}| - \sum_{n \in \mathcal{N}} \text{cosine similarity}(\mathbf{v}_n^1, \mathbf{v}_t^1)/|\mathcal{N}|$$

Similarly, we can compute $s(t, \mathcal{N}, \mathcal{P}, \mathbf{V}^2)$ for each target word $t \in \mathcal{T}_i$. The *effective size* of the difference in word associations across the word embeddings for the target class $\mathcal{T}_i$ is defined as

$$\frac{\text{mean of } \{s(t, \mathcal{N}, \mathcal{P}, \mathbf{V}^1) | t \in \mathcal{T}_i\} - \text{mean of } \{s(t, \mathcal{N}, \mathcal{P}, \mathbf{V}^2) | t \in \mathcal{T}_i\}}{\text{standard deviation of } \{s(t, \mathcal{N}, \mathcal{P}, \mathbf{V}^1) | t \in \mathcal{T}_i\} \cup \{s(t, \mathcal{N}, \mathcal{P}, \mathbf{V}^2) | t \in \mathcal{T}_i\}} \tag{3}$$

The numerator of Equation 3 is referred to as *test statistics*. A larger effective size indicates that the target words in $\mathcal{T}_i$ are more strongly associated or biased toward positive sentiment words in the embedding obtained from corpus 2 than in the embedding obtained from Corpus 1. The associated one-side p-value of the permutation test can be calculated as

$$\frac{\#\text{permutations s.t. the resulting test statistics} > \textit{test statistics}}{\#\text{permutations}}$$

We follow the implementation[7] from the author of [YR21], where a permutation refers to a random shuffle of the embeddings ($\mathbf{v}_w^1$ and $\mathbf{v}_w^2$) for each target word w.

[7] https://github.com/EddieYang211/TrainingDatasetCensorship/blob/main/3.Word_embdding_test_baidubaikevswiki.R

target class name	SGNS	Glove	SVD ($k = 1$)	SPPMI ($k = 1$)
acting	-0.4306 (0.029*)	-1.412 (0.553)	-0.4493 (0.311)	-1.096 (0.683)
writing	0.0197 (0.013*)	-1.0360 (0.878)	-0.0407 (0.011*)	-0.6820 (0.866)
directing	-0.5076 (0.131)	-0.8704 (0.639)	-1.0981 (0.618)	-1.3364 (1.000)
appearance	-0.4573 (0.123)	0.6333 (0.631)	0.6579 (0.250)	0.6116 (0.472)
life	0.1016 (0.041*)	-0.4153 (0.497)	-0.9797 (0.850)	-1.2210 (1.000)

Table 4: Effective sizes calculated between the embeddings obtained from Corpus 1 and corpus 2. The number in parentheses associated with each effective size represents the associated p-value.

Table 4 presents a summary of the results, focusing on the performance using $k = 1$, which demonstrates good performance in both the word similarity and sentiment classification tasks (see Table 1 and Table 2). Notably, the p-values associated with SGNS are the lowest among the 4 methods, indicating a significant difference in the association between target and attribute in the 2 corpora when using SGNS embeddings. Additionally, we observe that words related to acting and writing exhibit a stronger bias towards positive sentiment words in the reviews on "La La Land" compared to "Tiny Times", while words related to life exhibit a stronger bias towards positive sentiment words in the reviews on "Tiny Times" compared to "La La Land".

3.3.2 Gender bias

We adopt the methodology described in [BCZ+16] to ascertain gender in the embedding space. To begin with, we select 5 antonym pairs of gendered words $\{(m_i, f_i)\}_{i=1}^{5}$ (e.g., "male" and" "female", "boy and "girl", see details in the python notebook) and obtain five directions $\{\mathbf{v}_{m_i}^j - \mathbf{v}_{f_i}^j\}_{i=1}^{5}$ from each embedding $\mathbf{V}^j$. Subsequently, we compute the first component (first eigenvector) associated with the largest eigenvalue of the matrix whose five columns are the five gender directions. Consequently, we can calculate the fraction of variance explained by this first component for each method in each corpus. The average fraction of variance explained from both corpora is reported for each method in Table 5. We observe that the first component associated with the SGNS embedding explains the largest fraction of variance, indicating that the 5 directions are roughly aligned, and the first component can be considered as the *gender direction*. Therefore, we utilize the SGNS embedding to quantify and compare the gender bias across the 2 corpora.

SGNS	Glove	SVD ($k = 1$)	SVD ($k = 5$)	SPPMI ($k = 1$)	SPPMI ($k = 5$)	One-Hot	text2vec
0.6398	0.5390	0.5178	0.5738	0.4924	0.5534	0.5000	0.7227

Table 5: The average fraction of variance explained by the first principal component from both corpora is reported. For text2vec and One-Hot embeddings, we apply a single unique embedding.

Furthermore, we project each target word (see Table 3) onto the gender direction. Specifically, for each corpus, we calculate the cosine similarity between the corresponding SGNS embedding of each target word and the gender direction. Additionally, we project the text2vec embedding of each target word onto the gender direction associated with text2vec. A visual comparison can be found in Figure 1.

The first notable observation is that all target words align more closely with the male direction in text2vec, suggesting the existence of gender bias within the text2vec embedding space. The biased embedding is likely a reflection of the gender stereotypes present in the training corpus. On the other hand, the gender bias differs between the two corpora. Firstly, the appearance target words exhibit a bias towards females in both corpora. However, the life target words remain neutral in Corpus 2 while being biased towards females in Corpus 1. Additionally, the acting and writing words in Corpus 2 tend to be gender-neutral, while in Corpus 1 they exhibit a slight bias towards males. Interestingly, the directing target words are biased towards females in Corpus 2.

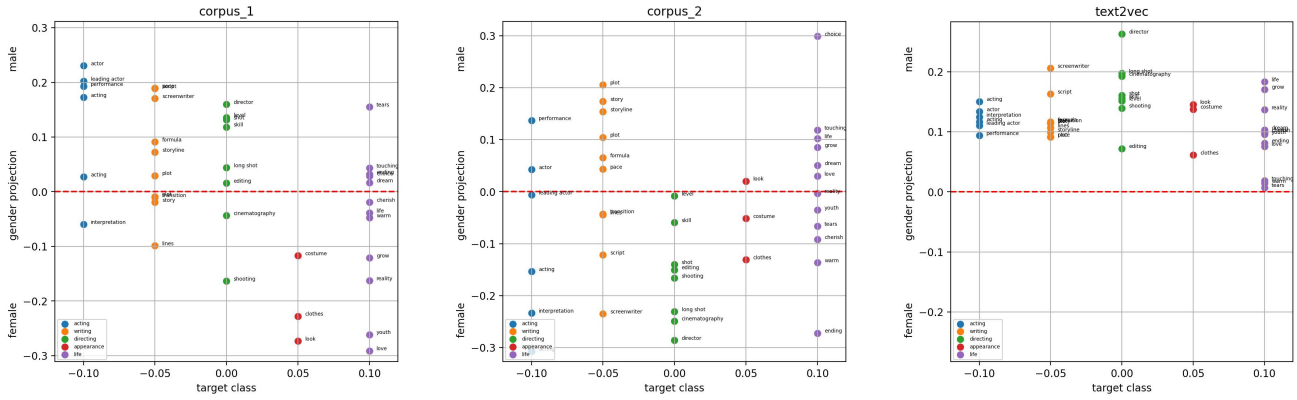

Figure 1: Gender projections on various embedding spaces. The cosine similarity between the SGNS embedding of each target word and the gender direction is plotted.

3.4 Words clustering

Embeddings capture meaningful semantic information of words; however, interpreting them directly can be challenging. One commonly used technique for interpretation is the projection of word embeddings in lower dimensions using Principal Component Analysis (PCA). The leading components of the projected embeddings are easier to interpret, and PCA aid in unsupervised word clustering.

We focus on all target words selected in Table 3. We extract the associated embedding matrix $\mathbf{V}$ and perform 2-dimensional PCA on $\mathbf{V}$. Subsequently, we apply k-means clustering[8] with 5 clusters (corresponding to the number of target classes) to the word projections on the first 2 PCA components. For each embedding method and corpus, we report the Normalized Mutual Information (NMI) score between the estimated labels of the words and their true labels, which are their corresponding target classes, in Table 6. A higher NMI score indicates that the estimated word labels align more closely with their target classes.

	SGNS	Glove	SVD ($k=1$)	SVD ($k=5$)	SPPMI ($k=1$)	SPPMI ($k=5$)	One-Hot	text2vec
Corpus 1	0.4455	0.2894	0.3817	0.2461	**0.5583**	0.3012	0.1117	0.4496
Corpus 2	0.4483	0.3896	0.4295	0.4283	**0.5325**	0.2744	0.1418	0.4496

Table 6: NMI scores for each corpus and each embedding methods on clustering the target words.

The results show that the word clusters derived from SPPMI embeddings align most closely with the true target classes of the words. In Figure 2, we visualize the SPPMI ($k=1$) embeddings of all target words from the two corpora on the first 2 principal components. Based on the first component, we observe clear separation between the words in the life target class (associated with the theme and feeling of movies) and the other 4 classes, which are related to the technical aspects of movies, such as acting and directing. Moreover, the second component enables a tentative classification between the words in the writing target class and those in the acting and directing classes, particularly in Corpus 2.

4 Conclusion

In this report, we have observed that word embeddings can quantify the similarity between different words and different documents, making them effective representations for both words and documents. Additionally, we have interpreted the semantic features of embeddings and applied them to clustering words in an unsupervised manner. Furthermore, we have investigated the presence of gender-wise and sentiment-wise biases within the trained embeddings.

Consequently, debiasing the embeddings emerges as a critical task and a potential future direction for this project. Existing approaches, such as the one proposed in [BAHAZ19], aim to identify documents whose removal

[8] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

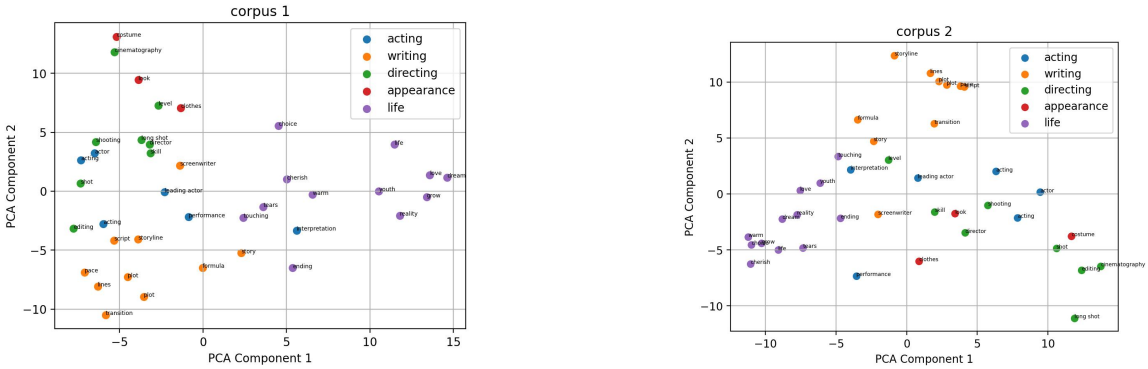

Figure 2: Two dimensional PCA projection of the SPPMI ($k = 1$) embedding for each target word.

would reduce the bias metric in defined Equation 3 significantly in an computationally feasible way. Another possible approach involves generalizing the optimization problem in Glove by reserving specific embedding dimensions for gender [ZZL+18].

A Code availability

The datasets and the Python notebook can be found at https://github.com/maggie980000/embedding-bias.

References

[BAHAZ19] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[CBN17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[GSJZ18] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[HLJ16] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501. Association for Computational Linguistics, 2016.

[LG14] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

[LGD15] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225, 2015.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543, 2014.

[YGLC13] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456, 2013.

[YR21] Eddie Yang and Margaret E Roberts. Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 537–548, 2021.

[ZZL+18] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.