Quantitative Analysis of Gender Bias in English Literatures

Cheng Peng, Jiemin Tang, Zixin Zhang, Ning Zhu, Mingzhi Ye
Department of Computer Science
University of Southern California
Los Angeles, CA
{cpeng606, jieminta, zzhang09, ningzhu, yemingzh}@usc.edu

Abstract

This document demonstrates the team's complete process of analyzing gender bias in English literature. The project utilizes two different approaches to analyze the depiction of male and female roles in classic literature, which are gender classification using BERT and changes in Word2Vec embeddings over different time periods. The details of data collection and processing as well as the construction of the two models will be further discussed in the following sections.

1 Introduction

It is widely recognized that gender bias exists in many genres of literature, especially the ones written in the earlier era when men had dominance in society and females were less appreciated given their social standings. For instance, the female characters in novels are often praised for their beauty, or physical appearance while male characters are praised for being brave and loyal. In order to determine whether the writers indeed tend to portray male characters on their virtues and internal characteristics while describing females in a superficial way, the team has designed a framework that would quantify the extent of gender bias in a wide variety of English literature.

In recent years, Natural Language Processing has gained popularity in conducting research to discover underlying social issues by using text-based information. There are many techniques to automatically process large amounts of texts, understand the context and provide meaningful insights. The goal of the project is to design a framework to extract information regarding gender bias from a corpus of English literature and quantify this issue in the given texts. Two original methods are proposed in order to analyze gender bias within the data from different perspectives. The links to the GitHub repository and presentation recording can be found in Appendix E.

2 Methodology

The team chooses a dataset called *"Classic English Literature Corpus and MetaData"* [1] which contains over 1,000 books and information about their authors. The team adopted a labelling framework proposed by Lucy et al. [2] in their work to analyze gender bias in textbooks as the basis of the project. The team first analyzes gender bias in an aggregated fashion, taking the literature in the corpus as a collection. The team conducts a temporal study by dividing the books into different groups according to the time when they are written, then compares and contrasts the effect of gender bias throughout different time periods.

2.1 Labelling Framework

The team utilized a system in which each sentence or piece of text is categorized using a pair of *(gender_word, categorical_word)* tags, and the tags came from two sets of pre-determined words (Table 1 and Table 2). The *gender_word* refers to the gender of the subject or object in a sentence, and it can be either *Man* (i.e. *he, him, his*) or *Woman* (i.e *she, her*). The *categorical_word* takes on one of the three possible values, *Home, Work* and *Achievement*, and it depends on the occurrence of specific words in it. For instance, if a sentence contains the words "*his" and "success"*, then its label would be *(Man, Achievement)*.

Gender	Related Words
Man	*man, men, male, he, his, him*
Woman	*woman, women, female, she, her, hers*

Table 1: Gender Words

Category	Words in Category
Home	*domestic, household, chores, family*
Work	*work, labor, workers, economy, trade, business, jobs, company, industry, pay, working, salary, wage*
Achievement	*power, authority, achievement, control, won, power, success, better, efforts, plan, tried, leader*

Table 2: Categorical Words

2.2 Data Cleaning and Processing

The corpus [1] contains a total number of 1,087 books. The team breaks the books into individual sentences. Since the purpose of the project is to determine whether there exists evidence of gender bias in literature, sentences without any *gender_word* or *categorical_word* are removed. In order to avoid ambiguity, sentences that contain both genders are also discarded from the corpus.

In addition, different variations of the same word are taken into account by adding them into the list of *gender_word* and *categorical_word*. After cleaning and processing the corpus, a total number of 46,054 sentences are extracted from the corpus, with 36,030 sentences related to males and 10,024 sentences to females. The detailed number of samples in each *(gender_word, categorical_word)* group can be found in Appendix A.

2.3 Aggregated Study: Gender Classification Using BERT

The team proposes to study the gender classification task in order to reveal how gender bias is reflected in a collection of English literature. The idea of using gender classification to illustrate gender bias is that if gender bias did not exist in the literature, the model would predict *Man* or *Woman* with equal probabilities over the three categories (*Achievement, Home* and *Work*), hence a random prediction. It is believed that this is not going to be the case since there is no denying that gender bias exists in the literature. Therefore, by studying the differences between *Man* and *Woman* labels in the predictions, the team can qualify gender bias in the given texts.

Each piece of the gender-interest text is modified by hiding the gender term using a *[MASK]* token and a classifier is trained to classify the *gender_word* based on the given contexts. The BERT model is a suitable choice for this task since it is designed for the Masked Language Modelling task. BERT has gained popularity in the NLP communities since its creation because of its ability to learn contextual information. In addition, BERT is pre-trained on a massive amount of data and can be fine-tuned to achieve high performance in different task-specific domains. Therefore, the team decided to leverage BERT by fine-tuning it on the corpus and performing gender classification to reveal insights regarding gender bias.

2.4 Temporal Study: Word2Vec Embedding of Literature in Different Time Periods

The team investigates if or how gender bias in English literature changes and evolves over time. In order to quantify such changes, it is crucial to represent *gender_word*'s and *category_word*'s meanings and measure their similarities. This goal can be achieved by training a word embedding model to find the association between *gender_word* and each of the *category_word*. The team decides to use Word2Vec as the word embedding model since it is fast and sufficient to handle the task. By comparing the cosine similarity scores computed between the same *category_word* to different gender groups (i.e. compare the similarity scores of *(Man, Home)* and *(Woman, Home) tags*), it will directly reflect the association of *gender_word* with *category_word* in the feature space. If a corpus of text is bias-free, then the similarity for both gender group and *category_word* should be very close since both genders should share common contexts and the word vectors will be close to each other in the feature space. To further study how time affects the terminology used to describe different gender roles, the team uses literature pieces from different time periods to train the word embedding model and get the corresponding similarity scores for comparison.

3 Experiments

After cleaning and processing the data from the corpus, the two tasks are conducted in parallel. The section documents how each task is performed and describes details of the experiments.

3.1 Fine-tune BERT Model for Gender Classification

The team sampled the dataset and created a test set containing 4,000 samples with an equal number of *Man* and *Woman* instances. Then the team performed an 80-20 train-validation split on the remaining data. The text data is then tokenized using BERT's tokenizer and the gender words (i.e *he, she, his, her etc.*) are masked with the *[MASK]* token, and special tokens *[CLS]* and *[SEP]* are added to the beginning and end of the sentences respectively in order to comply with the pre-trained BERT's specifications. The team uses a variation of a pre-trained BERT model called *"bert-base-uncased"*[4] which is a member of the BERT-base family. It is trained on all lower-cased data. The pre-trained model is made of an embedding layer, 12 transformers and an output classification layer, and a total of 1.1 million parameters. The BERT model is used as the encoder to extract contextual information and a layer of fully connected neural networks is used to produce classification outputs.

The model is fine-tuned using the training data with a batch size of 32, a learning rate of *1e-5* and *4* epochs as suggested by the creators of BERT model [3]. The learning curve is shown in Appendix B, and it is shown in Figure A.1 and Figure A.2 that the model performs the best over the validation set in epoch 3 with a loss of approximately 0.035 and starts to overfit after that. Thus the team saves the model weights at epoch 3 as the optimal model.

The test set is then tokenized using the same tokenizer and the gender words are hidden with the *[MASK]* token. The test data is passed to the optimal model to evaluate its performance. The team uses accuracy, precision, recall and F1 scores as metrics, and the evaluation results are shown in Table 3. In addition, the confusion matrix and ROC curve and AUC score can be found in Appendix C.

Accuracy	Precision	Recall	F1
0.97900	0.96057	0.99900	0.97941

Table 3: Evaluation Results of Fine-tuned BERT Model

3.2 Word2Vec Embedding Models

The goal of word embedding is to determine the associations between *gender_word* and each of the *category_word* and the change in such association throughout English literary history. The team first split the dataset according to the death year of the author, which results in six time periods as shown in Appendix A.2. It also shows the distribution of book count in each time period. In the dataset, only sentences that contain relevant gender words are kept for further exploration.

Each of the six sets of text as well as the full dataset are used to train a respective Word2Vec model to learn the word associations. For each model, the word occurrence threshold is set to 5, and the embedding space dimension is set to 100. According to the labelling, the team replaces *gender_word* with the word man or woman while keeping the *category_word*. After models are trained, the team uses cosine similarity to calculate the semantic similarity between *gender_word* and each word in the *category_word* group.

4 Results and Discussions

Both the Gender Classification Study and Embedding Cosine Similarity Study reveal gender bias in data and quantify the extent of bias in *Achievement, Home* and *Work* categories.

4.1 Gender Classification

The dataset used for the classification task suffers from class imbalance since 80% of samples are labelled *Man* and 20% are labelled as *Woman*. However, the team deliberately samples an equal number of *Man* and *Woman* instances to be the test set in order to reflect how the label imbalance in the training data affects the BERT model's ability to predict the two genders respectively. In general, the model achieves high precision, recall and F1 scores, indicating that by fine-tuning the BERT model, it can learn domain-specific knowledge about the task very well and achieve high performance. The confusion matrix in Appendix C shows that the model's *TP* (1,998) is slightly higher than *TN* (1,918) with *Man* being mapped to label 1 and *Woman* to label 0 which means the model learns characteristics of the *Man* labelled samples better than *Woman* labelled ones. In addition, the model's *FP* is *82* which is much higher than the *FN* of *2*, it reveals that the model is biased toward the *Man* label.

The distributions of both genders within *Achievement, Home* and *Work* categories are shown in Figure 1. While the number of *Man* and *Woman* predictions are similar in the *Achievement* category, Woman predictions contribute 20% more in the *Home* category than *Man* and *Man* predictions outrank *Woman* predictions by 20% in the *Work* category. The model predicts both genders with equal probabilities in the *Achievement* category, indicating that gender bias is mitigated in sentences that contain words from this category. However, the gender bias is more significant when it comes to the *Home* and *Work* categories, where the model is biased toward *Women* when it sees words in the *Home* category and biased toward *Man* when it sees words in the *Work* category.

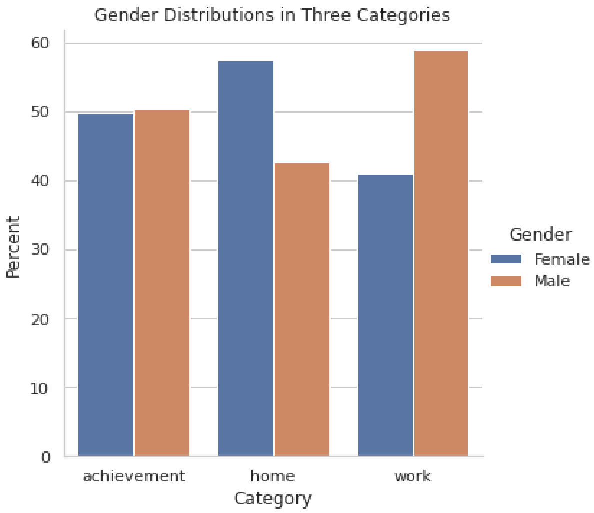

Figure 1: Gender Distributions in Three Categories

4.2 Embedding

Figure 2 shows the difference of *Man*'s and *Woman*'s cosine similarity scores computed by Word2Vec models over the three categories throughout 6 time periods. A positive value indicates that the model is biased toward *Man* in that specific category and vice versa. All three categories of words are more relevant to man before the year 1700. The association between work category and man group became even tighter in this time range, and the similarity score of *(Man, Home)* ranked highest during 1600-1700. In the time period 1700-1800, the cosine similarity score between *Woman* and *Work* category reached its peak. Starting in 1800, the three categories show a split in cosine similarity scores whereas the *Home* category has a high correspondence with *Woman* and the *Work* and *Achievement* category have more association with *Man*. The reason for such a split may be caused by class imbalance; the amount of books available before 1800 is less than books after 1800. In addition, there are approximately three times more samples referring to *Man* than *Woman*, which further enlarges the difference. However, the team decided not to perform sampling methods to solve the class imbalance issue since the variation in the amount of text to both gender groups is evidence of gender bias in literatures.

Figure 3 shows *Man's* average cosine similarity scores versus those of *Woman's*. From the entire dataset's perspective, the *Home* category suffers less from gender bias, which means both gender groups have a similar distance to the *Home* category in feature space. However, the cosine similarity scores between *Achievement* category and *Work* category with *Man* are often higher than with *Woman*. Especially the *Work* category which shows a stronger tendency toward *Man* than the *Achievement* category.

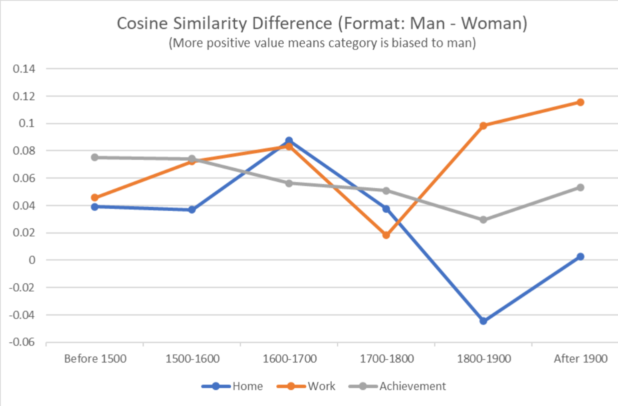

Figure 2: Cosine Similarity Difference Line Chart

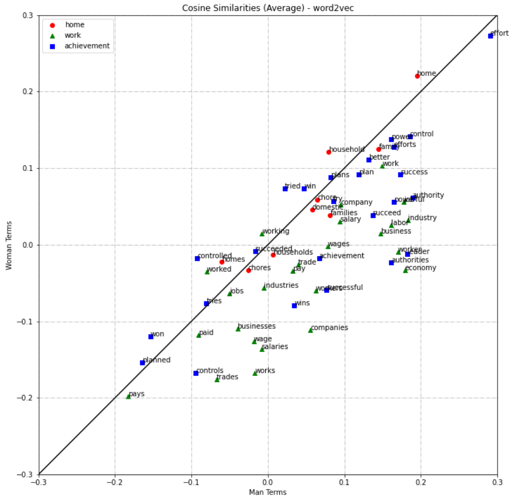

Figure 3: Detailed Cosine Similarity of Full Dataset

Figures in Appendix D show the detailed cosine similarity computed based on Word2Vec model trained using texts from 6 time periods.

5 Conclusion

In conclusion, the project has substantiated gender bias in English literature using BERT classification model and Word2Vec model embeddings. The BERT classification model can accurately predict the gender given different contexts and tends to relate *Man* to the *Word* category and *Woman* to the *Home* category. The Word2vec model embedding similarities reveals that *Man* is closely related to the *Work* and *Achievement* categories, and the scale of this inclination is not significantly different in different time periods. Both models indicate that female characters are more related to domestic settings while male characters to work settings. In addition, the bias does not change significantly in different time periods.

References

[1]RAYNARD JON. 2019. Starter: Classic English Literature.

[2]Lucy, L., Demszky, D., Bromley, P. and Jurafsky, D. Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks. AERA Open; 2020. p. 233285842094031.

[3]Chris McCormick. 2019. BERT Fine-Tuning Tutorial with PyTorch · Chris McCormick.

[4]https://huggingface.co/bert-base-uncased. Hugging Face.

Appendix A: Number of samples in Dataset

Gender	**Category**	**Number of Samples**
Man	**Achievement**	16,743
	Home	5,172
	Work	14,115
Woman	**Achievement**	4,997
	Home	2,117
	Work	2,910

Table A.1: Number of Samples in Each Gender and Category

Year Range	**Author Count**	**Book Count**
Before 1500	20	119
1500 ~ 1600	7	13
1600 ~ 1700	13	70
1700 ~ 1800	21	40
1800 ~ 1900	96	410
After 1900	134	434

Table A.2: Distribution of book count in each time period

Appendix B: Learning Curve of the Fine-tuned BERT Model

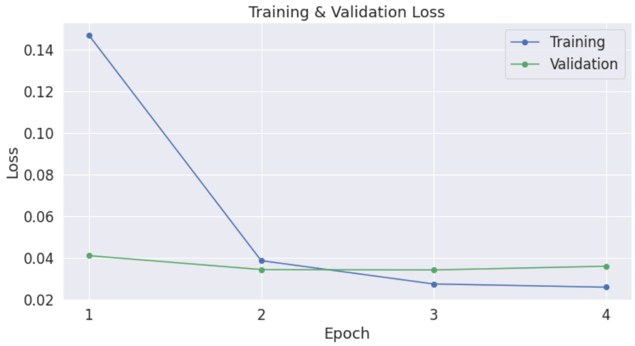

Figure B.1: Training and Validation Loss Over 4 Epochs

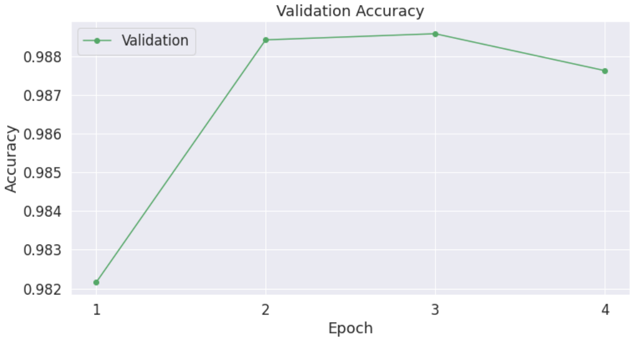

Figure B.2: Validation Accuracy Over 4 Epochs

Appendix C: Confusion Matrix and ROC Curve of Fine-tuned BERT Model

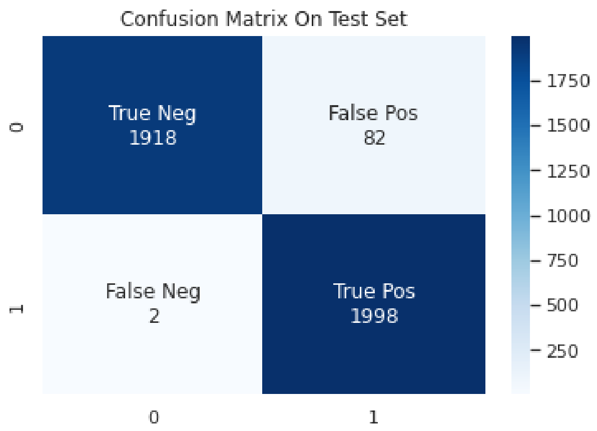

Figure C.1: Confusion Matrix of Test Set Data

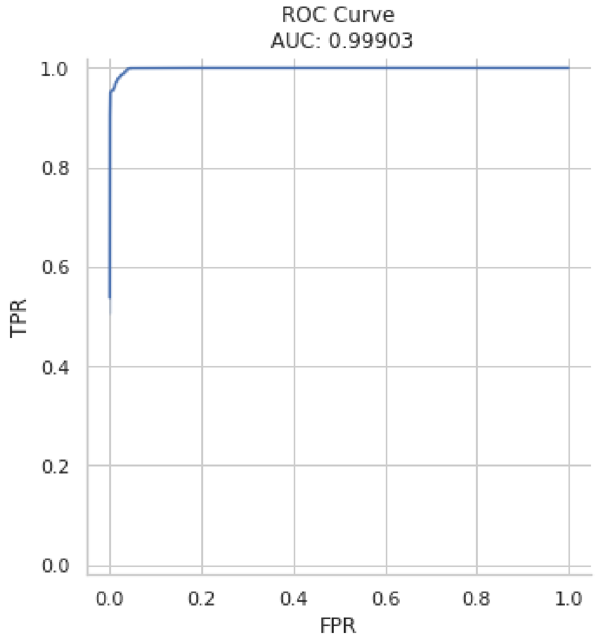

Figure C.2: ROC Curve and AUC Score of Test Set Data

Appendix D: Cosine Similarity Results From Word2Vec Word Embedding Model Over Time Period

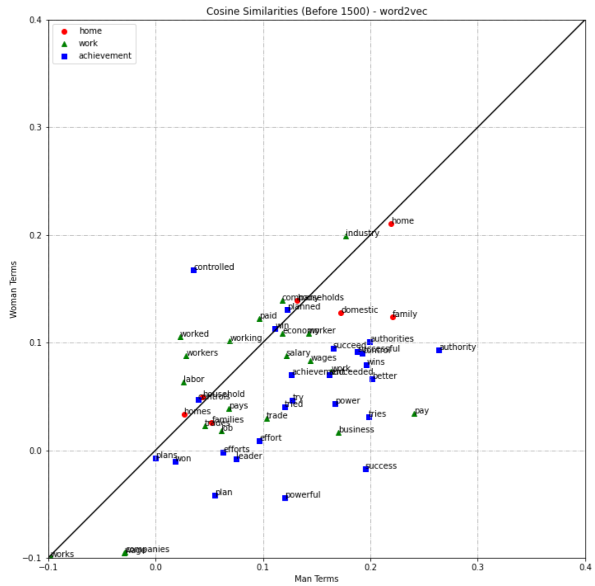

Figure D.1: Cosine Similarity Before 1500

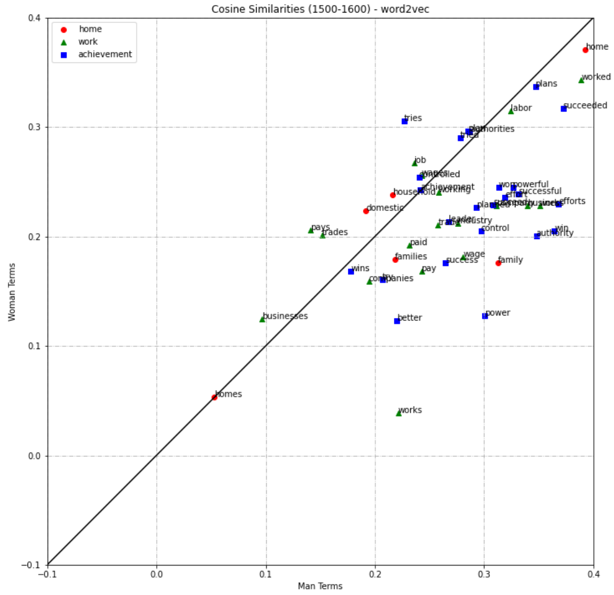

Figure D.2: Cosine Similarity Between 1500-1600

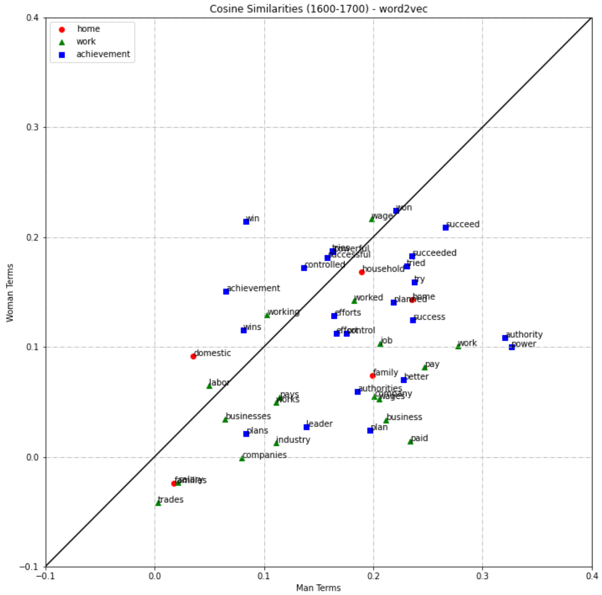

Figure D.3: Cosine Similarity Between 1600-1700

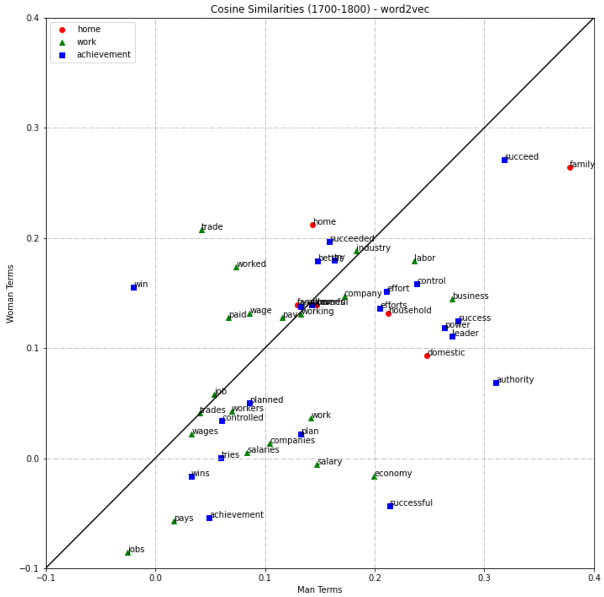

Figure D.4: Cosine Similarity Between 1700-1800

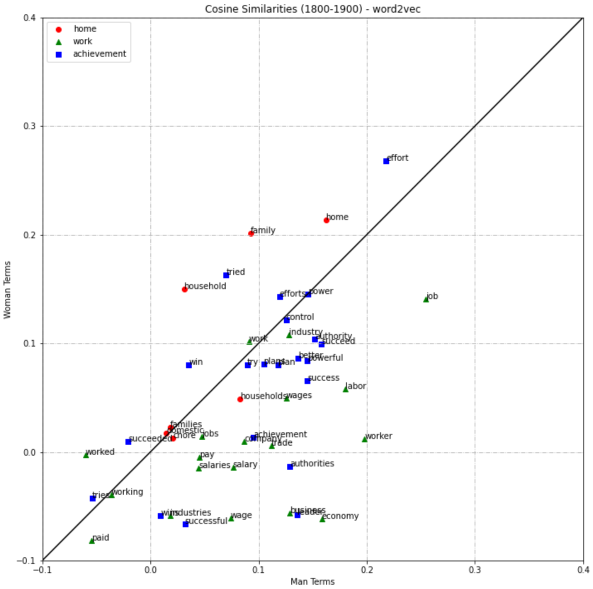

Figure D.5: Cosine Similarity Between 1800-1900

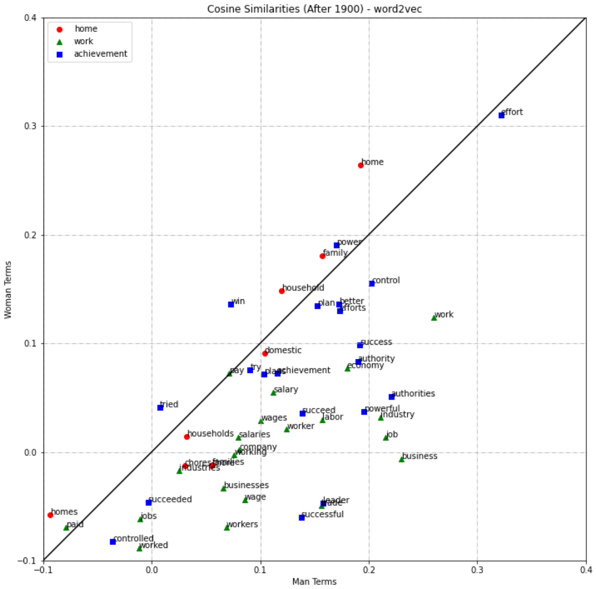

Figure D.6: Cosine Similarity After 1900

Appendix E: GitHub Repository and Presentation Recording

GitHub Repository:
https://github.com/CSCI544-Project-LiteratureGenderBias/ProjectDetails

Presentation Recoding:
https://www.youtube.com/watch?v=MxLaHD5b5Wk