Analysis of Bias in U.S. History Textbooks Using BERT

Stanford CS224N Custom Project

Grace Lam
Department of Computer Science
Stanford University
gslam@stanford.edu

Marilyn Zhang
Department of Computer Science
Stanford University
zmarilyn@stanford.edu

Abstract

U.S. History textbooks have a profound influence on childrens' social understanding of the United States. This is the reason activists and social scientists analyze textbooks for issues on bias and representation. Computational NLP methods can provide more holistic analyses compared to traditional qualitative studies. Our research supplements prior word2vec analyses of gender word relations in 15 U.S. History textbooks used in Texas by taking advantage of BERT's versatility with two studies. First, we compare BERT's embeddings between gender and interest words (related to home, work, and achievement). Second, we perform a gender word prediction study, where we mask out the gender word in each context containing interest words and evaluate BERT's ability to predict the correct gender in different contexts. We repeat both studies on fine-tuned and pretrained BERT. Our analysis is done with all textbooks taken as a collective, as well as stratified by historical time period discussed. Overall, we find that the textbooks contain idiosyncrasies that tend to associate women with "home" and "work" contexts more strongly than "achievement" contexts and that these trends stay relatively constant over historical time periods discussed.

1 Key Information to include

- Mentors: Dorottya Demszky, John Hewitt
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Because of their profound influence on childrens' social understanding of the United States, K-12 U.S. History textbooks have long been scrutinized by activists and social scientists alike [1, 2, 3]. Perhaps most recently, The New York Times analyzed textbooks by the same publisher used in California versus in Texas and found alarming differences with regards to the acknowledgement of white supremacy agendas, discourse on climate change and environmental policies, perspectives on immigration, and more [1]. The traditional qualitative approach to textbook analysis is essential for uncovering specific differences on how issues and people are represented. However, it is difficult to systematically capture broader trends and themes that are found in textbooks without quantitative tools.

Recently, NLP has been used to conduct social scientific analyses of textbooks with the goal of supplementing qualitative research [4, 5]. Our work expands on this line of research by leveraging the versatility of BERT [6] in particular to understand gender bias in Texas U.S. History Textbooks. To the best of our knowledge, the process itself of using BERT to analyze textbooks is a novel contribution to social scientific research on textbooks.

In particular, our research uses the same set of Texas U.S. History Textbooks as [4], fine-tunes BERT on this dataset, and builds on [4]'s analysis of gender bias in two key studies. Our first study uses contextualized BERT embeddings. Whenever a gender word (she, her, him, etc) and interest word (related to home, work, or achievement) co-occurs in the same context, we measure the BERT cosine similarity between the interest and gender word. We find that woman words are more similar to work and home category words than achievement words when aggregating all textbook contexts together. However, when we bucket contexts by the 50-year historical time period being discussed, "work" has no gendered difference in cosine similarity whereas "workers" is more similar to woman words. BERT's attention maps display a similar trend of "workers" attending to gender words with heavier weight than "work" does in the contexts they occur in. Our second study uses similar contexts as the first, but masks out the gender word and asks BERT to predict the gender word. We find that BERT is statistically significantly more accurate when predicting the gender word when it is a woman word. Furthermore, when the interest word is related to home or work (as opposed to achievement), BERT is more accurate when the gender word is a woman word. Finally, we run all our experiments on both pretrained BERT and fine-tuned BERT in an attempt to separate out the idiosyncracies of BERT's pretraining vs our specific textbook data; our results, however, can not be definitively attributed to the textbook alone in all cases.

3 Related Work

Lucy and Demszky have done a prior study applying a wide variety of NLP techniques (lexicons, word embeddings, topic models) to U.S. History textbook data, in order to investigate the depiction of historically marginalized groups in textbooks [4]. In particular, the study analyzes cosine similarities of static word2vec embeddings between gender words and other words related to "home," "work," and "achievement" to reveal that women are more closely associated with work and home, while men are more closely associated with achievement. We build on this work by exploiting the versatility of BERT to experiment with a variety of methods to analyze gender bias.

Significant work has been done on investigating bias found in word embeddings, as embeddings are a prevalent framework used to represent language in NLP tasks [7]. Biases present in the embeddings are commonly attributed to the datasets used to train them [8], and embeddings trained on books, newspapers, and other texts have been used to quantify historical trends in gender and ethnic stereotypes [9]. While earlier work has focused on probing static word embeddings, the rise in popularity of BERT has led to new ways to quantify bias. We draw inspiration from a recent study that explores the use of BERT's contextualized embeddings to inform their understanding of lexical semantic change over time [10]. Rather than shifts in word senses over time, our work aims to perform a similar contextualized analysis on word relations over different historical time periods discussed, across U.S. History textbooks. Another study proposes querying the underlying masked language model in BERT to measure gender bias present in a particular token [11], and we use a modified version of this method in our study to detect any association between gender words and our words of interest. Finally, attention maps are a vastly popular interpretability tool used to understand BERT's behavior and predictions [12]. Our study seeks to use attention maps to inform us more deeply about how these gender biases are learned from the U.S. History textbook data.

4 Approach

We use some of [4]'s open-sourced textbook preprocessing code [13] and fine-tune from the open-sourced BERT-base-uncased model [14], but everything else in our approach is original (including further preprocessing the textbook for temporal stratification) and can be found at [15].

4.1 Terminology

Throughout, we use *[category-name] words* to refer to words in woman (ie, she, her), man (ie, male, him), home (ie, domestic, chores), work (ie, economy, jobs), and achievement (ie, power, leader, plan) categories (ex: "woman words"). The full list of categories and words is presented in Table 2 in the Appendix. Home, work, and achievement category words were selected to be consistent with [4], which used the Linguistic Inquiry and Word Count (LIWC) lexicon as a starting point and additionally filtered based on word usage in the textbook data. We use *gender words* to denote

words in the woman and man categories, and *interest words* to denote words in the home, work, and achievement categories. Finally, we refer to contexts of textbook data where at least one gender word and one interest word occurs as *gender-interest contexts*.

4.2 Preprocessing Data: Temporal Aggregation and Temporal Stratification

We perform analysis of the textbook in two formats: temporally aggregated and temporally stratified. *Temporally aggregated* means that we combine and analyze results for all textbook data as a whole, across all historical time periods discussed. *Temporally stratified* means that we first group together contexts by the approximate 50-year historical time period discussed. We then take all contexts in a 50-year bucket and analyze the results for each time bucket separately. Because we cannot assume that textbooks are written in chronological order, we design a heuristic-based script to segment each textbook by chapter, determine each chapter's covered time range, then concatenate the chapters across all textbooks within 50-year buckets from 1300-2050.

4.3 Two Approaches for Quantifying Bias in Word Relations

We repeat both of these studies with BERT fine-tuned on our textbook data as well as with BERT-base-uncased.

Embeddings Cosine Similarity Study. We feed each gender-interest context through BERT and extract embeddings corresponding to the interest and gender word. We then calculate the cosine similarity between the interest and gender word, as well as extract the attention weights (for all 12 layers and 12 heads) from the interest word to the gender word. This process is depicted in Figure 1.

For each interest word, we calculate the mean cosine similarity between the interest word and all woman words vs all man words (see Table 2). This mean cosine similarity is reported for each 50-year temporally stratified bucket of contexts, as well as one mean from contexts temporally aggregated over the entire textbook. We additionally calculate the average of all attention weights from an interest word to a man word, and separately, to a woman word (aggregated over the entire textbook).

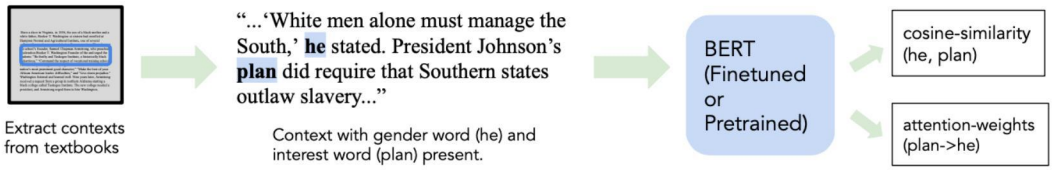


Figure 1: Embeddings cosine similarity study experimental process.

Gender Word Prediction Study. For each gender-interest context, we preprocess the text by replacing the gender word with a [MASK] token. We feed the modified context into BERT and examine the softmax probabilities in order to investigate how well the model predicts the correct gender, obtaining a woman probability and man probability in return (see Section 5.2 for details). This general workflow is shown in Figure 2. We analyze the average woman probabilities vs man probabilities for each context category when temporally aggregated across the textbook data. For each interest word, we also compare average woman probabilities vs man probabilities for each 50-year temporally stratified bucket of contexts.

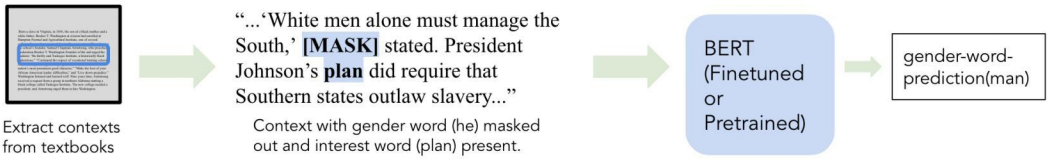


Figure 2: Gender word prediction study experimental process.

4.4 Baselines

Because we could not train BERT from scratch with only our textbook data, we run all our experiments on BERT-base-uncased as a baseline for our results from BERT fine-tuned. This partially helps us account for whether our findings are due to the textbook data or due to other data that BERT was pretrained on.

Additionally, for our temporally aggregated embeddings cosine similarity study, we use [4]'s static word2vec embeddings cosine similarity study as a baseline. Because our temporal aggregation averages cosine similarities across the entire textbook, this directly allows us to get a sense of how BERT embeddings compare to static word2vec embeddings.

5 Experiments

5.1 Data

We use a dataset of 15 U.S. History Textbooks widely used in Texas between 2015 and 2017. Each of the 15 textbooks selected had been purchased in at least 10 districts in Texas. This dataset is the same one used in Lucy and Demszky's paper [4]. In total, the textbook dataset contains 7.6 million tokens.

5.2 Evaluation method

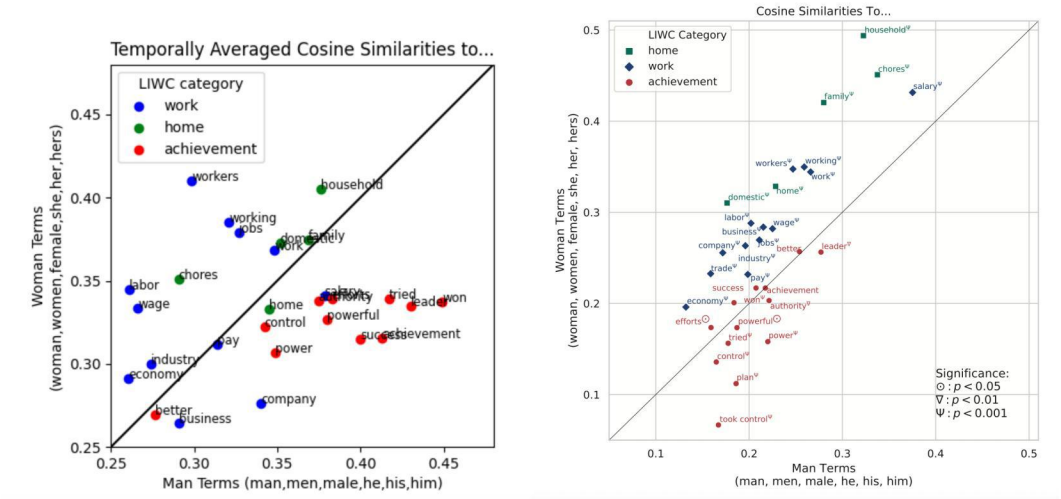
Embeddings Cosine Similarity Study. (1) Temporally aggregated: for each interest word, we perform two-tailed t-tests to test for differences in the average cosine similarity between the interest word and all man words vs the interest word and all woman words. We report corresponding p-values. (2) Temporally stratified: for each interest word, for each 50-year bucket, we plot the average cosine similarity between the interest word and all man words, and the interest word and all woman words, along with one standard deviation error bars. (3) Follow-up attention (temporally aggregated): we take all attention weights of an interest word attending to woman, man or all gender words, from all 12 layers x 12 heads. We perform two-tailed t-tests to compare average attention weights of different types and report t-values and p-values.

Gender Word Prediction Study. After feeding a gender-interest context into BERT with the gender word masked, we measure how much BERT prefers to predict the correct gender by taking the probability BERT assigns to the correct gender normalized by all gender word prediction probabilities. We sum together the softmax probability outputs for all woman words, and separately aggregate the probabilities for all man words. We define *woman probability* as the gender-normalized conditional probability of predicting a woman word given the gold token for [MASK] being a woman word. Similarly, we define *man probability* as the gender-normalized conditional probability of predicting a man word given the gold token for [MASK] being a man word. We follow the same general evaluation approach as the embeddings cosine similarity study with temporally aggregated and temporally stratified analyses. For our temporally aggregated experiment, we perform two-tailed t-tests on each of the home, work, and achievement categories instead of each interest word, so we could see higher-level trends.

5.3 Experimental details

BERT Fine-tuning. We perform an 80-10-10 train-dev-test split on the chapters found in each textbook. Then, we use the BERT-base-uncased tokenizer and data collator to prepare the data for fine-tuning under the masked language modeling objective. We use a block size of 512, learning rate of 5e-05, and batch size of 32. We train for 3 epochs and evaluate the model every 500 steps, with a total training time of 2.5 hours. The cross-entropy loss and perplexity scores on our held-out test set are lower post fine-tuning on our textbook data, as seen in Table 3 in the Appendix. This serves as a helpful sanity check that our fine-tuned model has properly learned the textbook data and produces contextual embeddings that reflect the language found in the history textbooks.

Extracting Contexts for Analysis. We use sentence-length gender-interest contexts for our embeddings cosine similarity study, and 512-token length gender-interest contexts for our gender word prediction study.



(a) With BERT embeddings

(b) With Word2vec embeddings

Figure 3: Averaged cosine similarities from BERT fine-tuned vs static word2vec embeddings (copied from [4] for comparison). Words to the left of the 45 degree line are closer to woman.

6 Results and Analysis

6.1 High-Level, Temporally Aggregated Word Relations

Embeddings Cosine Similarity Study.

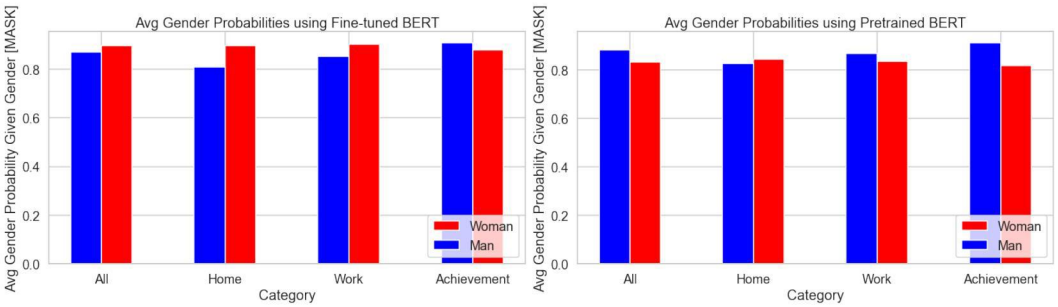
Result 1: Woman word BERT embeddings are more similar to home and work than achievement category word embeddings, consistent with word2vec embeddings baseline results (see Figure 3).

The association between woman and home words did not surprise us, as women are most often discussed in domestic settings in U.S. History textbooks. Similarly, the association between man and achievement words was not too surprising: we hypothesize that achievement category words rarely occur in direct relation to woman words as compared to man words. For example, the average number of words between woman and achievement words in each context was 1226, compared to 316 for men. Figure 16 in the Appendix also shows consistently fewer words between man and achievement words in every historical time period discussed compared to woman words. Hence even if woman words appeared close to an achievement word, it would very likely be in a separate phrase rather than in the same one, which would likely yield a lower BERT embedding similarity. Significantly, the higher association between woman and work words is undercut by the lower association between woman and achievement words: the textbooks are not necessarily discussing woman and work as leaders or with power (which are achievement words). The association of women with work not necessarily implying feminist views is an idea that is also discussed in [4].

BERT's embeddings are determined by a combination of specific contexts in the textbook and learned biases in BERT's generation of embeddings. We found that "household," "work," and "economy" are statistically significantly closer to woman words in textbook-fine-tuned BERT but not statistically significant in pretrained BERT, which implies our textbook contexts did play a role in these results. For other interest words, because BERT's embeddings are still partially determined by the context fed into it, and not just the model itself, it is still possible that some of the other results are due to particular textbook contexts rather than pretrained BERT embeddings. See Table 7 in the Appendix for full pretrained BERT results.

Gender Word Prediction Study.

Result 2: Fine-tuned BERT more accurately predicts woman words in general. The fine-tuned model predicts woman words more accurately than man words when the interest word is in the home or work category, and less accurately when the interest word is in the achievement category (see Figure 4a, and Table 4 in the Appendix for numerical results).



(a) With Fine-tuned BERT

(b) With Pretrained BERT

Figure 4: Differences in woman probability vs man probability using fine-tuned vs pretrained BERT, all $p < 0.00001$ except pretrained BERT on "home" which is not statistically significant.

Comparison of [MASK] to interest word attention	t-value	p-value
low vs high correct gender probability	-7.121	0.000000
low vs medium correct gender probability	-3.109	0.001881
medium vs high correct gender probability	-2.778	0.005472

Table 1: Differences in BERT's attention weights from [MASK] to interest word in contexts corresponding to predicting the correct gender with low, medium, and high probabilities.

Result 3: Pretrained BERT more accurately predicts man words in general. The pretrained model predicts man words more accurately than woman words when the interest word is in the work or achievement category. The difference in man vs woman probabilities in the home category is not statistically significant (see Figure 4b).

We find it intriguing that fine-tuned BERT predicts woman words more accurately than man words, which is the opposite of what we initially expected and what we see in pretrained BERT. One confounding factor we considered when running these experiments was the high frequency of man words relative to woman words in the textbooks. However, the lower frequency of woman words did not hinder fine-tuned BERT's ability to predict woman words in the gender-interest contexts. We note that fine-tuned BERT appears to associate woman words with contexts involving home and work, which agree with the results we found from our temporally aggregated embeddings cosine similarity study. Moreover, it is fascinating that pretrained BERT exhibits drastically different behavior, as it has a significantly stronger ability to predict man words correctly than woman words. We hypothesize this may be due to idiosyncracies in the textbook that are different from pretrained BERT data, such as work words being more associated with woman, likely due to increased discussions of woman labor movements as opposed to general male-dominated workplaces discussed in pretrained BERT data.

Result 4: Contexts that give higher gender probabilities tend to correlate with higher attention weights from [MASK] to interest word (see Table 1). We further investigate how much BERT's ability to predict a gender word correctly can be attributed to the interest word in the gender-interest context that is fed in. To do this, we split the contexts into three groups: low, medium, and high gender probability. Contexts are assigned to a group based on the gender probability BERT assigns: <0.25 for low, between 0.45 and 0.55 for medium, and >0.9999 for high. We notice that BERT's gender prediction ability does correlate with how much [MASK] attends to the interest token, which tells us that BERT potentially looks at the interest token to inform its gender prediction.

6.2 Specific Words and Temporally Stratified Word Relations

Embeddings Cosine Similarity Study.

Result 5: Throughout all historical time periods discussed, "workers" is more similar to woman words whereas "work" is equally similar to woman and man words (see Figure 6 and Figure 7). Interestingly, although "workers" appears gendered whereas "work" is not when stratified temporally, both words are statistically significantly more similar to woman words when we aggregate the cosine

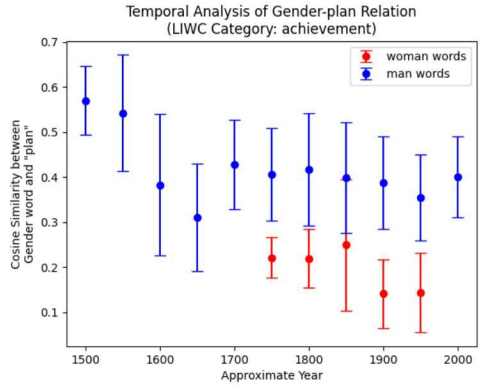


Figure 5: Gender words' similarity to "plan"

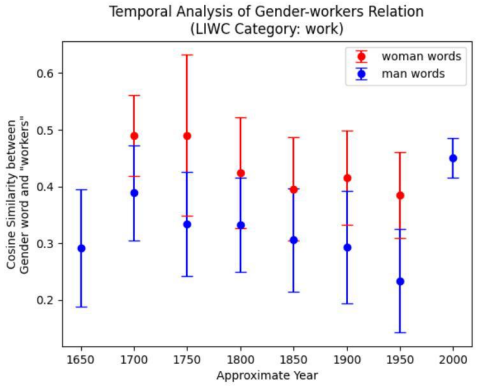


Figure 6: Gender words' similarity to "workers"

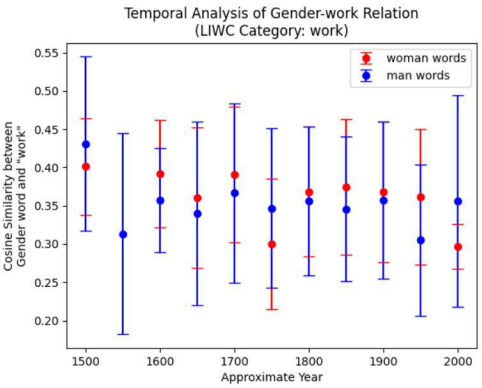


Figure 7: Gender words' similarity to "work"

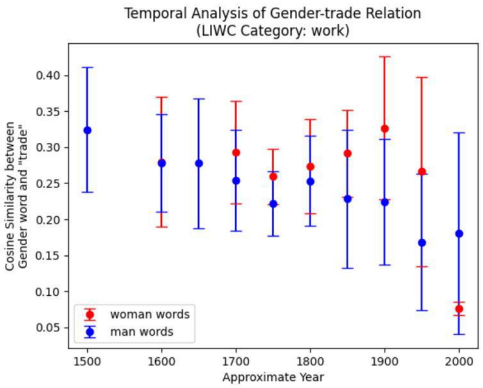


Figure 8: Gender words' similarity to "trade"

similarities and disregard the historical time periods discussed (see Table 7 in the Appendix and Figure 3), with "work" definitively so (compared to pretrained BERT) as discussed in 6.1.

Part of the reason for this discrepancy could be that BERT has learned to associate "workers" with gender words more so than with "work," likely due to a combination of the context and BERT's pretraining and fine-tuning process. As shown in Table 6 in the Appendix, although BERT attends from "work" and "workers" to woman words with heavier weight than to man words, BERT also attends from "workers" to any gender word with heavier weight than it attends from "work" to any gender word. We note, however, that attention weights in BERT are far from well-understood and this correlation does not definitively capture how BERT produces embeddings.

Result 6: Word relation cosine similarity trends are generally consistent regardless of historical time periods discussed. As shown in Figure 5 and 6, "plan" is more similar to man words whereas "workers" is more similar to woman words independent of the time period discussed, respectively. Furthermore, cosine similarities overall do not seem to have much variation from time period to time period; Figure 8 with "trade" had the most variation we saw.

For some words, the lack of temporal variation in cosine similarity is expected: "plan," for example, might be closer to man words because of male generals' plans and presidents' plans being discussed throughout historical time periods. On the other hand, we expected "work" and "workers" to increase in similarity to woman words over time as discussion of women's work transitioned from domestic to industrial settings. It's possible that the interest word embeddings co-evolved with gender word embeddings from context to context, which is why trends are consistent despite different historical time periods being discussed.

Finally, pretrained BERT results gave similar temporal plots as the ones from fine-tuned-BERT, although in some cases with larger error bars. Hence we cannot definitively conclude that these are textbook-specific results; rather they are most likely a combination of the way the textbook is written

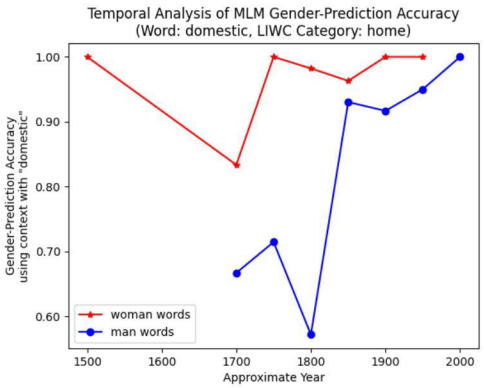


Figure 9: MLM Accuracy in "domestic" contexts

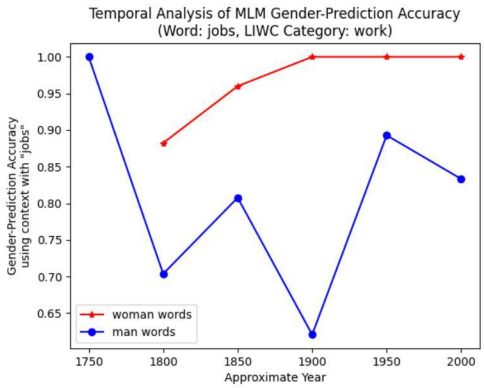


Figure 10: MLM Accuracy in "jobs" contexts

as well as language from BERT's pretrained data. (See Figures 11, 12, 13, 14 in the Appendix for fine-tuned vs pretrained comparisons for "plan," "workers," "work," and "trade," respectively.)

Gender Word Prediction Study.

Result 7: As shown in Figure 9 and 10, fine-tuned BERT predicts woman words more accurately in "domestic" and "jobs" contexts. There are no clear trends in how gender prediction accuracy varies over time.

The results found for "domestic" and "jobs" contexts fall in line with the temporally aggregate results for the "home" and "work" categories, correspondingly. We run the same experiment on pretrained BERT and interestingly notice that for "domestic" contexts, there is no clear difference between the prediction accuracy of woman words vs man words (see Table 15 in the Appendix). This aligns with our temporally aggregate study results in Table 5, where we see that pretrained BERT does not show statistically significant differences between woman and man probabilities for the "home" category.

7 Conclusion

While we attempt to distinguish between the influence of BERT's pretraining process vs our textbook data on our results by running each experiment with both pretrained and fine-tuned versions of BERT, future work can be done where pretrained and fine-tuned BERT did not differ significantly in results. We would like to further investigate whether these similarities in results were caused by the textbook contexts being fed in rather than the pretrained effects of BERT. Additionally, in our gender word prediction study, an interesting follow-up would be to isolate the association between gender and interest words from the effects of other words in the context.

In this paper, we explored different BERT methods to analyze bias in Texas U.S. History textbooks, a novel contribution to social scientific analyses. We ran experiments on both pretrained BERT and BERT fine-tuned on our textbooks. We found that fine-tuned BERT produced embeddings where woman words are closer to "home" and "work" words than "achievement" words, with little variation over time when we bucketed by 50-year historical time periods discussed. These findings were only slightly different from pretrained BERT's. Furthermore, we masked out the gender word in sentences with "home," "work," and "achievement" words, and found that fine-tuned BERT generally predicted the correct gender for women more accurately than for men, while pretrained BERT predicted the correct gender for men more accurately. Overall, we found that the textbooks contain idiosyncrasies that seem to associate women with work more than expected, but given a lack of strong association with achievement words, this is not necessarily a feminist trend.

References

[1] Dana Goldstein. Two states. eight textbooks. two american stories. *The New York Times*, Jan 2020.

[2] Frances Fitzgerald. *America Revised: History Schoolbooks in the Twentieth Century*. Random House, Inc., 1980.

[3] Patrick Riccards. It's time to start decolonizing america's history textbooks. *Rethink Together*, Nov 2020.

[4] Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. Context analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas u.s. history textbooks. In *American Education Research Association (AERA) Open Journal*, 2020.

[5] Richard Lachmann and Lacy Mitchell. The changing face of war in textbooks. *Sociology of Education*, 87(3):188–203, 2014.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[8] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

[9] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. In *Proceedings of the National Academy of Sciences*, 2018.

[10] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020.

[11] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[12] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. In *Association for Computational Linguistics*, 2019.

[13] https://github.com/ddemszky/textbook-analysis. *GitHub*.

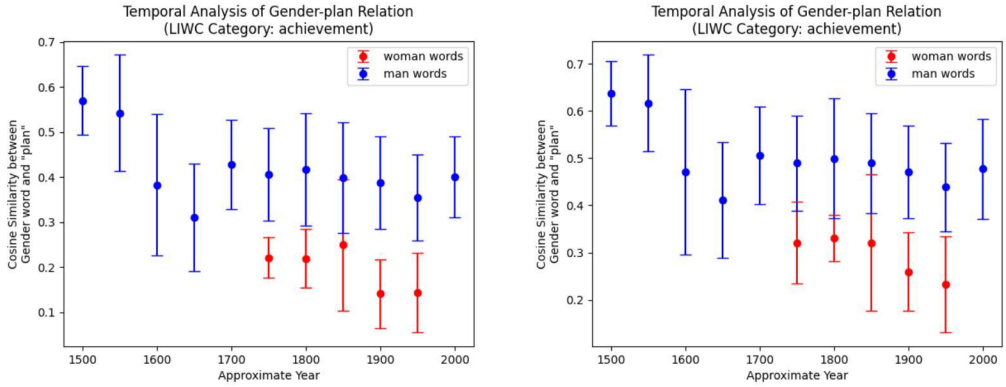
[14] https://huggingface.co/bert-base-uncased. *Hugging Face*.

[15] https://github.com/grace-lam/nlp-textbook-bias. *GitHub*.

A Appendix

Category	Words in Category
woman	woman, women, female, she, her, hers
man	man, men, male, he, his, him
home	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, powerful, success, better, efforts, plan, tried, leader

Table 2: Categories of words referenced in our studies.



(a) With Fine-tuned BERT

(b) With Pretrained BERT

Figure 11: "Plan" similarities to woman and man words: textbooks-fine-tuned vs pretrained BERT results.

Comparison	t-value	p-value
"work" vs "workers" average attention to all gender words	-11.15	0.000001
"work" average attention to woman vs man words	15.68	0.000001
"workers" average attention to woman vs man words	23.81	0.000001

Table 6: Differences in "work" and "workers" attending weights to gender words.

Model	Cross-Entropy Loss	Perplexity
BERT-base-uncased	3.39	29.70
BERT-base-uncased + fine-tuning	2.09	8.06

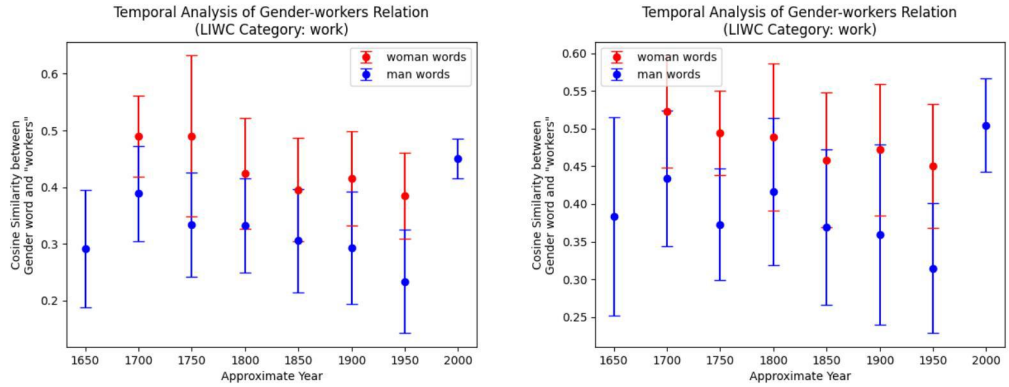
Table 3: BERT evaluation results on test set

Category	Avg Woman Probability Given Woman [MASK]	Avg Man Probability Given Man [MASK]	t-value	p-value
All	0.895	0.871	5.811	0.000000
Home	0.897	0.809	8.036	0.000000
Work	0.901	0.853	8.119	0.000000
Achievement	0.880	0.909	-4.335	0.000015

Table 4: Differences in woman probability vs man probability using fine-tuned BERT.

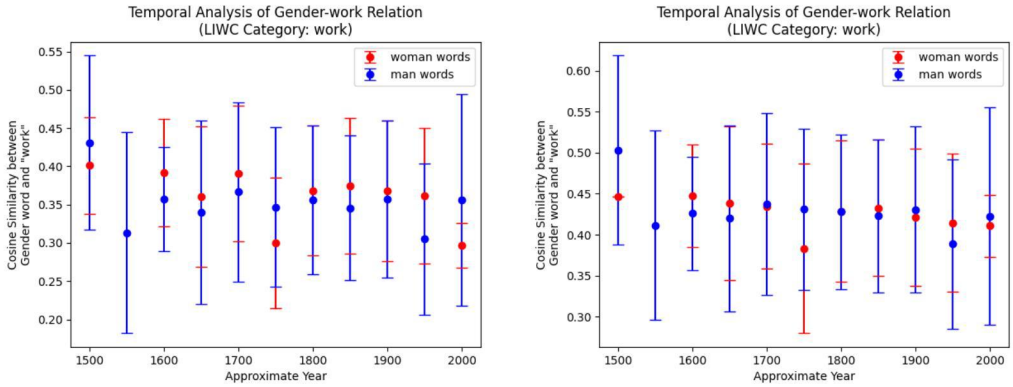
Category	Avg Woman Probability Given Woman [MASK]	Avg Man Probability Given Man [MASK]	t-value	p-value
All	0.833	0.882	-12.441	0.000000
Home	0.844	0.827	1.553	0.120589
Work	0.835	0.868	-5.867	0.000000
Achievement	0.818	0.913	-14.20	0.000000

Table 5: Differences in woman probability vs man probability using pretrained BERT.



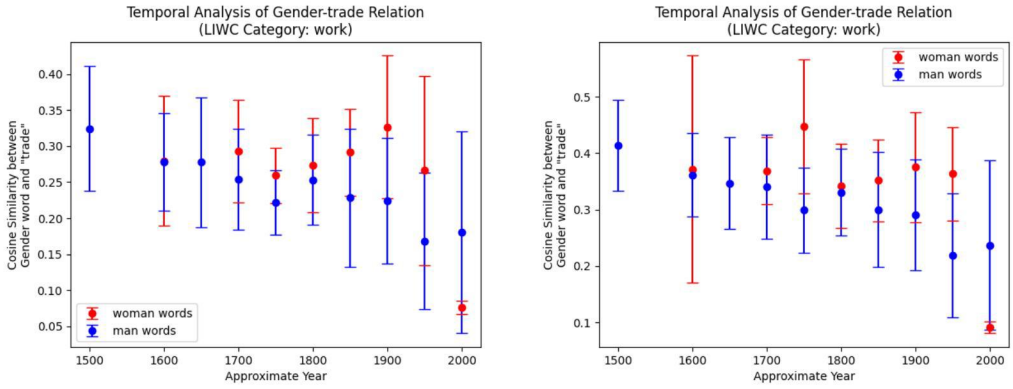
(a) With Fine-tuned BERT (b) With Pretrained BERT

Figure 12: "Workers" similarities to woman and man words: textbooks-fine-tuned vs pretrained BERT results.



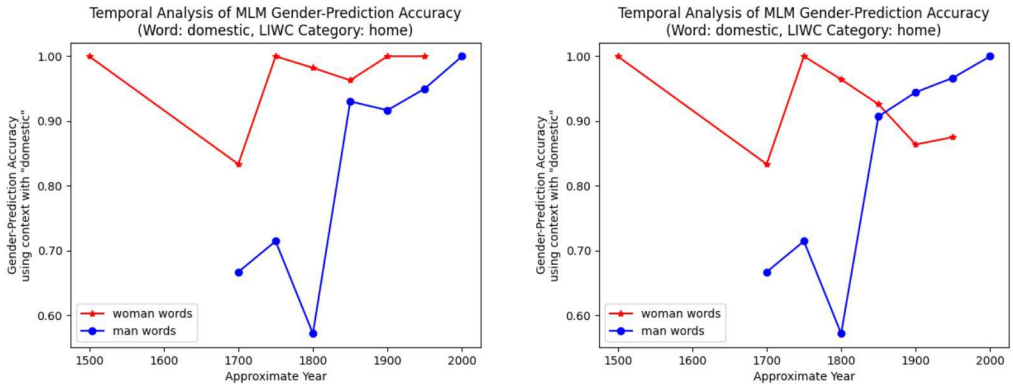
(a) With Fine-tuned BERT (b) With Pretrained BERT

Figure 13: "Work" similarities to woman and man words: textbooks-fine-tuned vs pretrained BERT results.



(a) With Fine-tuned BERT (b) With Pretrained BERT

Figure 14: "Trade" similarities to woman and man words: textbooks-fine-tuned vs pretrained BERT results.



(a) With Fine-tuned BERT

(b) With Pretrained BERT

Figure 15: MLM Accuracy in "domestic" contexts: textbooks-fine-tuned vs pretrained BERT results.

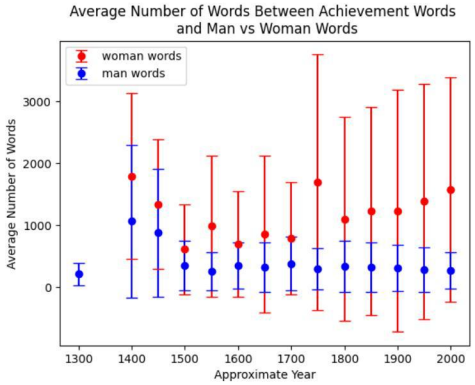


Figure 16: Number of words between achievement words from male vs woman words in each time period discussed.

Table 7: Two-tailed t-test p-values (upper bounds) between average man vs woman word-similarities to each query word. Blanks denote statistically insignificant p-values.

Query	**LIWC Category**	**Textbook Fine-tuned BERT**	**Pretrained BERT**	**Textbook Trained word2vec ([4])**
home	home	0.05	0.01	0.001
domestic	home	0.01	0.01	0.001
household	home		**0.05**	**0.001**
chores	home			0.001
family	home			0.001
work	work	**0.0001**		**0.001**
labor	work	0.0001	0.0001	0.001
workers	work	0.0001	0.0001	0.001
economy	work	**0.0001**		**0.001**
trade	work	0.0001	0.0001	0.001
business	work	0.0001	0.001	0.001
jobs	work	0.0001	0.001	0.001
company	work	0.0001	0.0001	0.001
industry	work	0.01	0.01	0.001
pay	work			**0.001**
working	work	0.0001	0.0001	0.001
salary	work			**0.001**
wage	work	0.0001	0.001	0.001
power	achievement	0.0001	0.0001	0.001
authority	achievement	0.01	0.001	0.01
achievement	achievement	**0.0001**	**0.0001**	
control	achievement	0.05	0.0001	0.001
won	achievement	0.0001	0.0001	0.001
powerful	achievement	0.0001	0.0001	0.05
success	achievement	**0.0001**	**0.0001**	
better	achievement			
efforts	achievement	0.001	0.0001	0.05
plan	achievement	0.0001	0.0001	0.001
tried	achievement	0.0001	0.0001	0.001
leader	achievement	0.0001	0.0001	0.01