American Journal of
Education and Practice
(AJEP)

Content Analysis of Textbooks via Natural Language Processing

Mahshad Nasr Esfahani

American Journal of Education and Practice
ISSN 2520-3991 (Online)
Vol.8, Issue 4, pp 36 – 54, 2024 www.ajpojournals.org

Content Analysis of Textbooks via Natural Language Processing

Mahshad Nasr Esfahani
Ph.D. in Language Education and Multilingual University at Buffalo

Article history
Submitted 23.05.2024 Revised Version Received 26.06.2024 Accepted 30.07.2024

Abstract

Purpose: Advanced methods from the field of data science have the potential to shed fresh light on basic concerns in the field of educational research. Natural language processing tools, such as lexicons, word embeddings, and topic models, are used in 15 United States history textbooks that were extensively used in Texas between the years 2015 and 2017.

Material and Methods: This study aims to analyze these textbooks for their portrayal of historically oppressed populations. Latinx individuals are rarely mentioned, but renowned white men are almost always mentioned. People of African descent are characterized as behaving in ways that imply helplessness and lack of control, according to lexical methods. Women are most often addressed in the contexts of the home and the workplace, according to the word embeddings. Issues of a political rather than a social nature are highlighted by subject modeling.

Findings: We also found that textbooks with a smaller representation of women and people of African heritage are more often purchased by conservative nations.

Implications to Theory, Practice and Policy: Our computational toolkit has a rich history of textbook analysis and has recently been distributed as part of our efforts to support new fields of study.

Keywords: *Artificial Intelligence, Case Studies, Content Analysis, Curriculum, Data Science, Gender Studies, History, Natural Language Processing, Race, Textbooks, Textual Analysis*

1.0 INTRODUCTION

Recent methodological advancements in natural language processing (NLP), a branch of artificial intelligence, hold great promise for illuminating several important social and political concerns within the realm of education. Although textbooks are sometimes seen as relics of a bygone cultural system, they have long been an essential tool for comprehending the inner workings of education. We showcase the possible insights that may be gained by using text data science and natural language processing (NLP) to a sample of fifteen of the most popular Texas high school history textbooks. Curriculum study is an area where natural language processing (NLP) methods may greatly benefit. First, we may use these methods to quantify more nuanced ideas with bigger samples, which might provide fresh insight into the breadth and depth of current trends in educational discourse. The ability to better understand texts by analyzing their word-to-word linguistic connections has increased, which has led to an increased emphasis on relational meaning. Now we can go on to the second point. As a third point, we are becoming better at documenting how certain utterances advance various worldviews. Researchers may now access tools that were previously out of reach when it came to analyzing discourse-external variable linkages (Anon, 2022). It is possible because of the capabilities of these measures, namely the ability to employ bigger samples. Computational textual analysis has a lot of promise, but it should never be used in place of more thorough investigations; on the contrary, it should always be used in concert with them. Given the flexibility of NLP tools, we worry that researchers will have a harder time settling on clear conceptual goals for their studies.

The social scientific, policy and practical goals of educational research may be served by demonstrating how these theories might be used to evaluate textbook content. Neither do we conduct an exhaustive investigation into the causes of the selective inclusion of specific content categories in textbooks, nor do we provide a normative assessment of textbooks. Instead, we use Texas history textbooks to illustrate how NLP techniques may answer research concerns about the representation of historically disadvantaged groups.

Textbook editors in the past took a more conventional approach to these problems. Through a combination of descriptive and methodological considerations, we want to pave the way for future research that, with any luck, will pique interest in creating more effective computational tools for this domain. Furthermore, as we will address in the conclusion, these techniques lend credence to content explanations grounded in social science and to evidence-based policy recommendations (Boyd and Schwartz, 2020).

Textbook analysis using Natural Language Processing (NLP) improves educational quality and accessibility. First, natural language processing automates text analysis, making it easier to find key ideas and topics in bigger texts. Summarizing and arranging content helps teachers and students focus on what's essential. Parsing textbooks using natural languages processing methods like topic modeling and sentiment analysis helps students engage and personalize their learning by identifying relevant themes, tracking discourse changes, and measuring emotional tone. Natural language processing may also help personalize instructional content. Natural language processing algorithms may employ adaptive learning methods to identify student issues and adjust course content. This customized approach ensures students get the required support, improving learning and retention. Natural language processing (NLP) may potentially improve classroom accessibility for disabled students. NLP-driven text-to-speech and speech-to-text technologies may convert written content into spoken words and reverse the process, making textbooks accessible to visually impaired and struggling readers. As long as all students have access to excellent education, degree democratization is possible. Due to its text analysis, tailored learning, and accessibility capabilities, natural language processing (NLP) may improve textbooks' pedagogical value. These technologies can improve learning efficiency, personalization, and inclusivity in schools.

Textbook Research

Education research cannot be complete without textbooks, which serve as physical representations of the "intended curriculum." They sit at the crossroads where cultural, political, and social influences interact with individual students. The availability and use of textbooks have a positive effect on student's academic achievement, and textbooks are among the most widely utilized instructional devices worldwide. Having said that, everyone knows that textbooks aren't neutral; rather, the material is up for grabs and reflects society's power dynamics and accepted worldview. These textbooks not only teach pupils real social and cultural values, but they also shape their views on people of many backgrounds. We suggest methods that might be useful for studying how contemporary history textbooks depict gender, race, and ethnicity. Current work in this sector provides the basis for these techniques.

Researchers in the field of education have been unable to fully comprehend the origins of textbook material or the processes that generate and spread various discourses due to the methodologies we have used. A major drawback of conventional approaches is that they can't scale. For most textbook content studies, the gold standard still stands at one researcher reading and coding textbooks by hand. To do this project, you'll need a lot of materials. For instance, to code 80 biology textbooks by hand, Morning (2008) uses keywords to scan the index for pertinent areas. This was achieved in a recent article published in the American Journal of Sociology (Charlesworth et al., 2021). Quite a few programmers investigate potential subjects, check for dependability among raters, and assess the statistical soundness of their assessments. When needed, academics bring in many programmers to read their work. The fact that accurate annotation of nuanced language signals (such as the agency associated with certain verbs) necessitates trained coders who are familiar with linguistic frameworks is another barrier to hand coding's widespread adoption. This implies that general textbook coding attempts can end up focusing on counting or basic sign recognition.

The work of Bromley et al. (2011), for instance, tagged more than 500 international social science textbooks to determine whether any of them addressed "the environment." Because human coding in a large, longitudinal, cross-national sample has its limits, the authors could not provide more nuanced evaluations of environmental education (Chowdhary, 2020). Textbook study using manual coding techniques often requires researchers to read and annotate textbooks by hand, which may be a laborious procedure. Natural Language Processing (NLP) and other efficient approaches are necessary because of the limits of this conventional methodology, notwithstanding its thoroughness. Scalability is a major issue with hand coding. It requires a lot of time and resources to manually analyse big amounts of text, which is a major problem in educational research. Researchers such as Morning (2008) have been forced to use manual keyword indexing, which is not without its flaws and is susceptible to prejudice and human mistake.

However, scalable and efficient solutions are offered by NLP. One example is how natural language processing (NLP) makes use of lexicons and word embeddings to swiftly analyse massive volumes of text data, revealing correlations and patterns that would otherwise go unnoticed by human approaches. Because of how important it is to comprehend how gender, race, and other social aspects are portrayed in textbooks, this is very helpful in that area of study. Compared to conventional coding approaches, natural language processing (NLP) technologies provide a more detailed and all-encompassing picture of the text by methodically capturing how words are employed to support various viewpoints.Also, natural language processing (NLP) helps with word-to-word correlation analysis, which is great for getting to the bottom of what words imply in context. This has the potential to reveal hidden networks and patterns in the text, shedding light on the ways in which textbooks both represent and influence societal and cultural norms. Using natural language processing (NLP), we may see how textbooks frame historical narratives or how they over-represent certain groups.

Computational Approaches

Computational social science has employed natural language processing (NLP) to get insights from educational textual data. These technologies are used for dissertation abstracts, online and in-person class discussions, and interdisciplinary student writing analysis. Learning analytics and education data mining employ NLP technology to characterize text cohesion, complexity, and difficulty. The package includes Reader Bench, Coh-Metrix, and Tool for Automatic Text Cohesion Analysis. These tools may assist instructors assess online student engagement or tailoring courses to various ages. Some natural language processing systems used in educational reading materials include Light SIDE for automated essay grading, TAALES for lexical competence and word choice prediction, and Group Communication Analysis for discussion participant roles.

Educational sociology seldom uses NLP. Machine counting in electronic textbooks was innovative. Lachmann and Mitchell (2014) concluded this after examining media portrayals of wars. No previous NLP research examined how texts reflect gender or other social variables and included schooling. Consider this study: Hoyle et al. (2019) examined one million digital books for gender-related verbs and adjectives, Garg et al. (2018) quantified gender and ethnic stereotypes over a century using word representations from books and newspapers, and Ash et al. (2020) examined how gender bias affects judicial behavior using judge-written text. Media like online literature and journalism have been studied for gender stereotypes (Khurana et al., 2022). This research examines textbook socioeconomic portrayals using natural language processing.

Even while NLP can execute near-human tasks, it is nonetheless prone to errors and bias. Therefore, when drawing judgments on a topic that affects society, like education, prudence is needed. Consider openness and explainability while choosing processes. Lexicon-based methods present tabulated words simply and succinctly, enabling interpretation. Our research has a drawback: many machine learning models or resources are developed or created using non-educational data like news items. Thus, certain models or resources may not work in a new setting. Tailoring these algorithms to future history textbooks is great, but it requires a lot of training data annotation. Our technique section details natural language processing's limitations. Computing methods do not replace or compete with conventional methods for investigating complex social issues. We seek to combine them around shared research objectives and leverage one method's benefits to compensate for another's weaknesses. NLP can only describe the material, not prescribe it. Thus, social scientists, educators, and ethicists must assess the outcomes to ensure they meet curricular and educational objectives (Kleinheksel et al., 2020).

Our Contribution

We provide scalable and quantitative textbook content assessments as our main offering. When employed across several texts in US history instruction, these methods better depict historical events and individuals. Our findings add to the evidence that history textbooks intentionally omitted particular opinions. Texas history textbooks seldom mention Latinx people in descriptions of racial and ethnic groups, whereas most political leaders are White males.

One of the most important contributions that we have made is that we have enabled a relational approach to meaning that applies to textbook research by using natural language processing techniques that identify patterns in the co-occurrence of concepts. In addition to revealing hidden structures and networks of terminology, these tools have the potential to provide a comprehensive picture of how textbooks reflect social meaning. In addition to this, they can assist in answering issues about the substantive character of speech, and more specifically, what meanings are associated with certain words or concepts. Even though there has been a shift towards instructional techniques that center on different viewpoints of the past, we provide evidence that the issues of formal politics continue to continue to dominate history textbooks in the state of Texas. The findings of our studies also indicate that when discussing persons of African descent, words with lower degrees of agency and power are used. This is a result that further underlines the significance of integrating substantive knowledge with computational methodologies. Educational institutions are increasingly embracing NLP to streamline administrative operations, improve accessibility, and enhance learning. An overview of how NLP is affecting education is: Learning Analytics Learning analytics relies on natural language processing (NLP) to illuminate student learning. Text data from student activities like discussion forums, assignments, and review may be analyzed using NLP to identify student participation, comprehension, and problem areas. Teachers may utilize TAACO and Coh-Metrix to assess text complexity, readability, and coherence to better fulfil student needs. Sentiment analysis may also reveal student participation and well-being by assessing message tone.

Text Analysis

NLP is used to analyze educational texts like textbooks and academic journals. Word embeddings and topic modelling may reveal biases in educational materials, track discourse evolution, and identify key topics and ideas. Natural language processing (NLP) may expose biases and improve possibilities in history textbooks by showing social group representation. This skill ensures that educational content is inclusive and properly represents diverse perspectives.

Evaluation by Robots

One of the largest applications of NLP in education is automated grading. Light SIDE uses NLP to score essays for grammar, coherence, and logic. These technologies provide students with rapid feedback, helping them improve their writing without the delays of human grading. Automated grading relieves instructors, allowing them to focus on children who require personalized instruction.

NLP enables customized education by adapting course contents to each student's strengths and shortcomings. Adaptive learning systems analyze student responses and interactions using natural language processing to evaluate performance. Using this information, instructors tailor educational materials and exercises to provide each student with the correct level of challenge and support. Personalized feedback from natural language processing helps students understand their mistakes and improve.

Improvements to Accessibility

By offering resources for special needs students, NLP enhances educational accessibility. TTS and STT can help visually, and reading-impaired students access educational materials. In conclusion, we show that scholars may connect textual patterns to outside social, political, and cultural influences on schooling and its results by doing quantitative analyses of larger samples. This is achieved at a scale that might be more challenging to do with a smaller number of textbooks that undergo intensive hand-coding. We build a relationship between textbook content, district purchasing patterns, and political leanings for example. This is only one of the several possible facets of creating and distributing textbooks. Textbooks with stronger representation of historically oppressed groups are more often acquired by districts in more Democratic counties. The similarities across textbooks outweigh the little differences, but this remains true (Lucy et al., 2020). Several hypotheses, including those concerning the portrayal of social groups in instructional materials, served as inspiration for the research detailed in the article. The researchers based their findings on preexisting frameworks that addressed gender and race in representation. Home, workplace, and politics are a few examples of topics that were chosen because they were pertinent to previous studies on gender representation in history textbooks.

To validate the research and guarantee its reliability and robustness, many measures were taken. An important issue for tiny corpora is the instability of word embeddings; the researchers tackled this by using bootstrapping. Estimates of word similarity were shown to be more robust using this strategy. To keep things consistent across all of the investigations, they also used sentence-level topic modeling. Additionally, the research used cosine similarity to measure semantic similarity between words in the vector space and lexicon-based techniques to evaluate the descriptors' connotations.

These phases in the methodology improved the results' dependability and interpretability by making sure they were based on well-established theoretical frameworks and tested using rigorous computational approaches.

Data

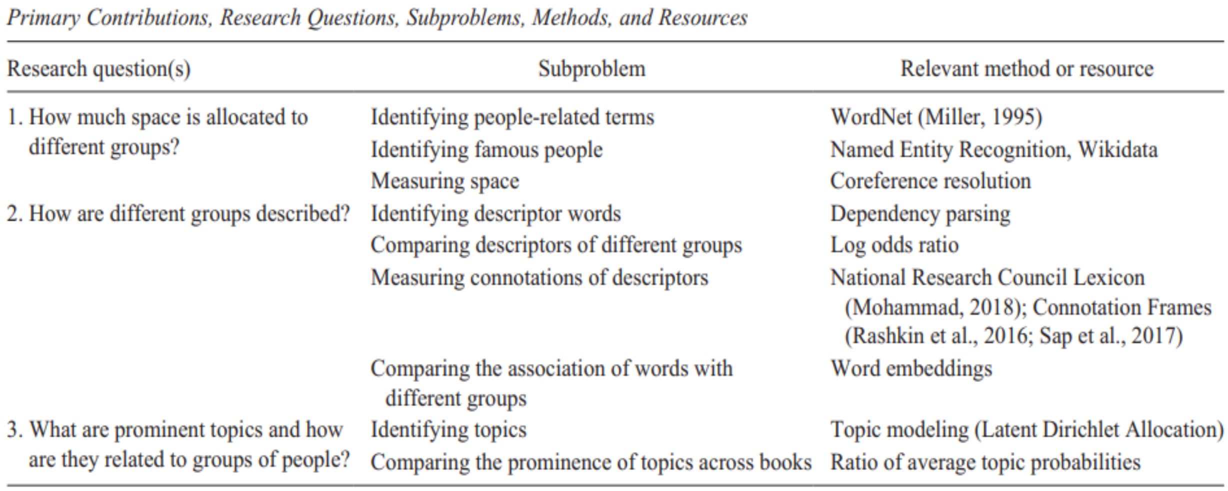
Texas Textbooks

Texas textbooks are our focus for providing district-level textbook purchasing statistics online (Texas Education Agency, n.d.). Due to its second-largest student population—5.4 million in K-12 public education in 2017 Texas influences US textbook content. Thus, textbook publishers find the state lucrative. Over the years, numerous textbook cases have targeted the Texas Board of Education. One example is the 2015 statewide adoption of social studies textbooks, which conservatives pushed but was criticized for curricular bias. Our dataset comprises Texas' 2015–2017 US history textbook sales. Used in ten or more district transactions determines title selection. With six cumulative volumes, Textbook Sources includes fifteen textbooks. All seven volumes were PDFs, so we extracted text. The remaining volumes were digitized using ABBYY FineReader. After parsing the text, we did nothing, as shown in Appendix D. From course materials, we retrieved 7.6 million tokens or continuous characters separated by spaces or punctuation marks.

Demographic Data

We use student demographics and geography to study textbook distribution. National Centre for Education Statistics (n.d.) Common Core of Data collected this data for public school districts in 2016–2017. We also break down the 2016 election results by county and utilize two-party vote shares to determine each county's political lean (Nicolas, Kim and Chi, 2021). Our study uses Democratic vote share estimates to demonstrate the methodologies' external links. School or board demographics may also affect textbook distribution at the district level; future studies may examine these and other factors.

Table 1: (Nicolas, Kim and Chi, 2021)



Primary Contributions, Research Questions, Subproblems, Methods, and Resources

Research question(s)	Subproblem	Relevant method or resource
1. How much space is allocated to different groups?	Identifying people-related terms	WordNet (Miller, 1995)
	Identifying famous people	Named Entity Recognition, Wikidata
	Measuring space	Coreference resolution
2. How are different groups described?	Identifying descriptor words	Dependency parsing
	Comparing descriptors of different groups	Log odds ratio
	Measuring connotations of descriptors	National Research Council Lexicon (Mohammad, 2018); Connotation Frames (Rashkin et al., 2016; Sap et al., 2017)
	Comparing the association of words with different groups	Word embeddings
3. What are prominent topics and how are they related to groups of people?	Identifying topics	Topic modeling (Latent Dirichlet Allocation)
	Comparing the prominence of topics across books	Ratio of average topic probabilities

2.0 MATERIALS AND METHODS

We want to use natural language processing techniques to analyze how textbooks portray historically oppressed populations. Using the question sequence from Table 1, we demonstrate the approaches we thought were most pertinent.

Research Question 1: What is the Distribution of Space among Different Groups?

Here we use frameworks from multicultural curricular studies to measure the amount of textual space that different groups and people occupy, to categorize textbook diversity. The inclusion and discussion of both known, famous persons and more common, generic (nonnamed) people (such as settlers or farmers) have long been the focus of conventional methods for evaluating social studies textbooks. Researchers have attempted to quantify diversity in textbooks by looking at factors such as the relative frequency of mentions of minority groups compared to majority groups, the portrayal of minority roles in texts (i.e., secondary vs. contributing), and the inclusion of well-known figures from minority groups.

Identify People-Related Terms: It is possible to find common nouns that identify unnamed people, like pioneers or Mexicans, by searching WordNet, an English lexical database that records the definitions of words and their interactions (Miller, 1995). We search the database for every conceivable hyponym or subclass for every person, socioeconomic group, and independent entity. One way to evaluate this WordNet-based method is via manual labeling. We use the space program to extract the first words of every noun phrase that appears in every one of our data sets at least 10 times. This process usually results in about 12,000 unique noun heads. After painstakingly searching this database for the common nouns meaning "people," we obtained 2,111 individual phrases. Our automated WordNet-based approach catches more than 95% of all nouns about people that were humanly recognized, except the phrases "group" and "majority," which may symbolize both people and other objects in WordNet. Not only that, it manages to catch 98 out of the 100 most prevalent terms related to humans in our dataset. However, we have included these two words in our list for analysis along with the other 5% of phrases relating to persons that did not get picked up by the WordNet-based approach. The reason is, that we usually take these phrases to mean persons when we read about them in history books (Roblek et al., 2020).

Among the 1,665 unidentified terms about individuals, such as "engineer" or "family," 446 phrases describe a demographic, including both single and plural nouns (Appendix Table A1). We manually sort this compilation by gender and ethnicity so that we may compare the descriptions of various demographic groupings. For words that may be used as adjectives (such as "Navajo community"), we additionally consider their context to see whether they are linked with any specific demographic. An example of a situation containing intersectionality is black women. Black women are not just women, but they are also Black. We limited our gender-based analyses to men and women since our dataset contains very few cases of non-binary gender identities, including just three mentions of transgender individuals.

Identify Famous People: To analyze social groups, every textbook will inevitably use names. While it's true that a textbook may be diversified with as little as a handful of famous people, leaving out even a few of them would be to ignore crucial moments in American history. The named entity recognition (NER) tagger in spaCy enables us to identify individuals based on their names. Things that contain proper nouns describing them, such as persons, locations, or organizations, are automatically labeled. An F1 score of.735, as measured manually on our textbooks, was attained by this NER tagger. While studies on the biases and mistakes made by NER when presented with names from diverse ethnicities and genders are continuing in the field of natural language processing (NLP), our results show that spaCy's pre-trained tagger is accurate enough to inspire further research into making this model suitable for textbook language (Zhao, 2021). We may use the person's official name rather than a pseudonym to avoid counting them twice; for example, the open knowledge source Wiki data includes aliases of notable persons like Franklin D. Roosevelt and Franklin Roosevelt.

One disadvantage is that Wiki data, like its encyclopedic brother project Wikipedia, would not provide enough coverage of underrepresented groups. We also restrict ourselves to the 100 most common NE-detected names, which means we only look at the persons that appear in textbooks repeatedly, because NER tagging isn't foolproof and picks up a long tail of non-human words. In many cases, the only information provided is the last name; for instance, the name "Roswell" might refer to either "Theodore" or "Eleanor," making it impossible to determine the specific individual to whom they pertain. In situations when the coreference spans many paragraphs, coreference resolution errors discussed in the section that follows tend to be the primary cause. We overcome this challenge by comparing partial last names that appear later in the text with those that appear earlier in the text. White people are underrepresented in our database when it comes to racial and ethnic group labels, thus we manually verify these labels in addition to using wiki data to identify racial and gender identities (Anon, 2022).

Research Question 2: About Certain Demographics, What Are the Most Talked-about Subjects?

The next step is to research word-to-topic and topic-to-topic connections in text, building on methodologies that study word-to-word relationships in text. Researchers interested in textbooks typically look at the subjects included and how they relate to one another; this might reveal the author's point of view and the way the textbook is structured. Coders are often allowed to determine the words linked to each subject and how those topics are connected; for instance, Lachmann and Mitchell (2014) employ hand-curated word categories. Annotation bias or coding mistakes caused by missing important elements could cause interrater reliability to be poor. When trying to analyze vast volumes of text, computational approaches that are founded on word cooccurrence patterns may automatically organize words into subjects. This might lead to the discovery of relational meanings more efficiently than human coding. Researchers still need to have strong subject-matter knowledge to provide meaningful interpretations of the automated classifications.

To automatically find themes in a set of texts, topic modeling is an important strategy. Though several varieties of topic models exist, Latent Dirichlet Allocation is by far the most popular. LDA is a representation of the word distribution within subjects and the distribution of themes within texts. Prior work has used such LDA models on a dizzying array of text types and genres. Analysis of student writing and discussion forums in MOOCs (Massive Open Online Courses) has made use of LDA models in educational settings. We use LDA to examine the frequency of various subjects in textbooks and the frequency of terms about various demographics within and across subjects. A set of documents is needed to feed into topic models (Boyd and Schwartz, 2020).

Identifying Topics: By doing topic modeling at the sentence level, we can collect many documents with comparable sizes (17 tokens on average) that may be used to provide consistent estimates of various themes. We start by removing function terms (such as the, it, and have) from the model's vocabulary using a list of stop words contained in MALLET, a commercially available topic modeling tool. Additionally, we use the Snowball Stemmer to carry out stemming5. As an input to MALLET, we create document-to-token counts using our resultant collection of tokens (unigrams and bigrams). Topic models use k=50 topics. We limited textbook topics to fifty, even if they may be large. In trials with several topics (k = 75, k = 100, k = 300), such as numerous topics representing different wars, fine-grained subjects hindered our analysis. The sample size depends on the study objective.

Topics that employ phrases about people without names may reveal social groupings. We examine the top 10 keywords for two topics to create a link. The subject's high-probability words should form a semantic field because we're omitting function phrases. We call this "topical diversity" since a word's debate should be more semantically varied as it is related to more subjects. Books' Topic Predominance Analysis. We calculate the textbook's average chance of a subject appearing in a sentence to assess its pervasiveness. Summing the average probability of connected topics (subject groups) in a phrase determines their prominence. The popularity ratio of the two categories in a book indicates their respective significance. The former method examines topic group pairings across books instead of concentrating on a single subject group, making it more resistant to noise from textbook lexical distributions and topic probabilities. We exclude three books from our thematic judgments because they cover such a limited section of American history (Charlesworth et al., 2021).

3.0 FINDINGS

Research Question 1: What is the Distribution of Space among Different Groups?

Determine Words Associated with People. "People," "women," and "his" are the top three non-normed pronouns in use today. Given how often it appears, it's likely that not all pronouns were resolved with the nouns they relate to while coreference was being done. Although gender and race/ethnicity are not indicated for most phrases, males are referenced at a higher rate than women (Figure 1). Figure 1 shows that among nonwhite racial/ethnic groups, black people are most often mentioned. The fact that these textbooks center on White Americans' history is not surprising, given that a small majority of persons identified by race or ethnicity are not White. Several terminologies (such as "pioneer," "farmer," and "priest") appear to imply or presume Whiteness. People often suffer from this form of "reporting bias," when they refrain from mentioning an entity's most prevalent attributes because they expect the audience would automatically presume the majority demographic.

Figure 1 shows that there are very few Latinx groupings among the ethnicities described in textbooks, which is perhaps the most shocking conclusion. Culturally relevant education, in which students' identities are reflected in the curriculum, has been demonstrated to improve students' learning results, according to previous research. Although Latinx students make up 52.42 percent of Texas's student body, they are only referenced 961 times in all textbooks, making up just 0.248% of the total number of person words and 2.23% of the total number of terms indicated by ethnicity or race (Chowdhary, 2020).

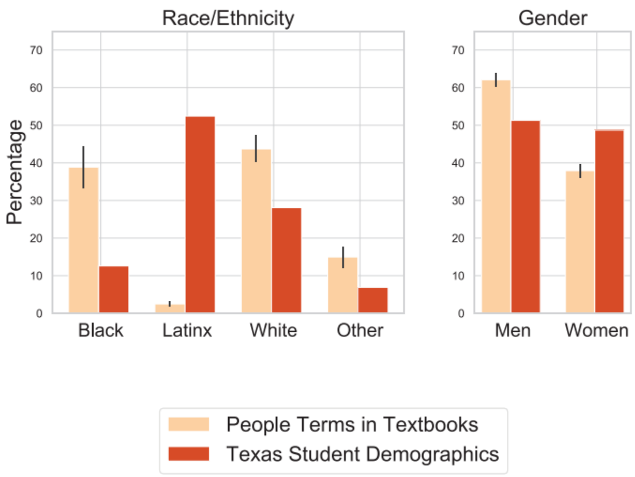
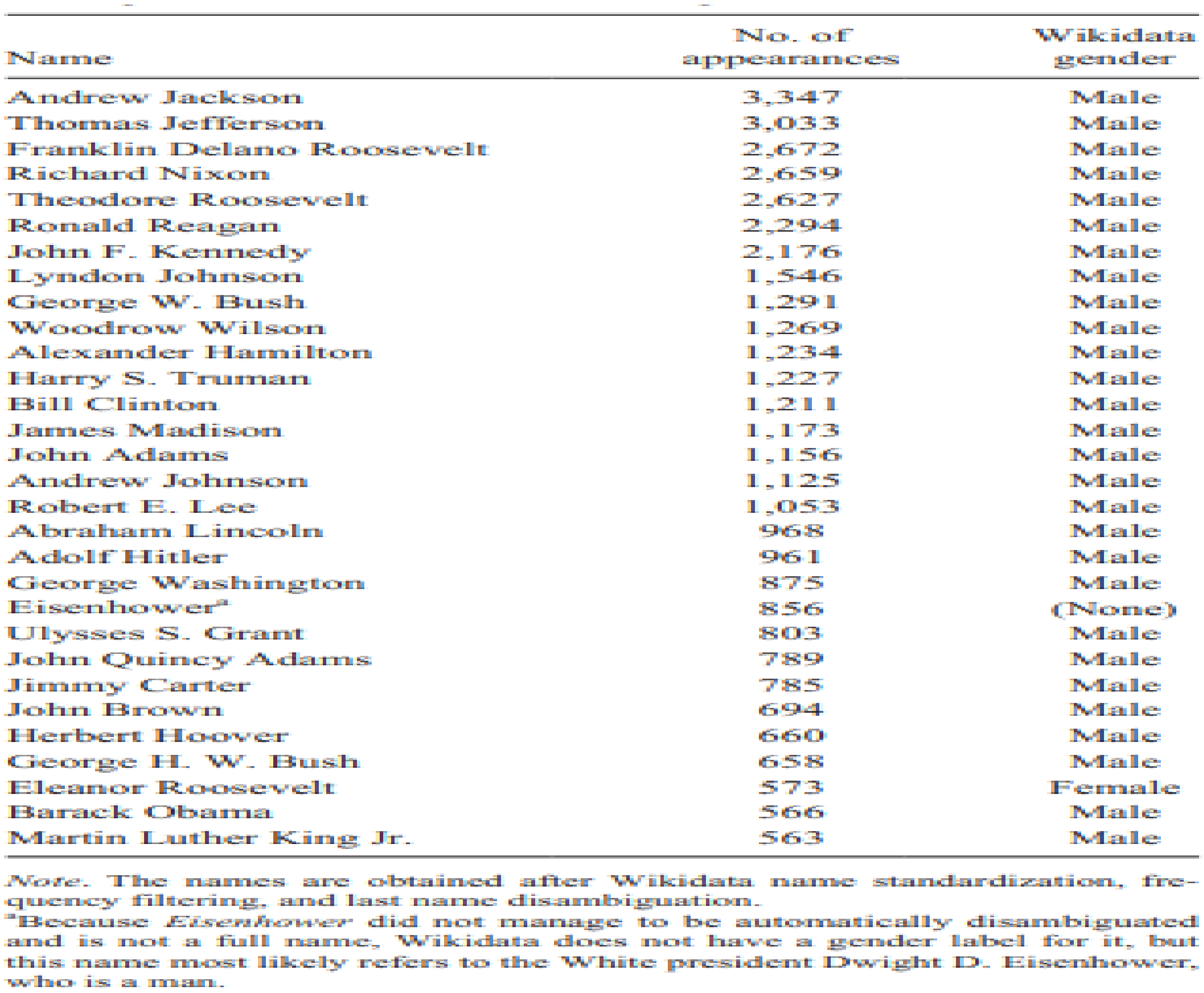


Figure 1: Distribution of Various Demographics in Texas Classrooms and Publications (Chowdhary, 2020).

It is common to see Latinx communities framed within the Mexican-American War and in opposition to the coming White settlers. To avoid Mexicans after abandoning the Oregon Trail, most of the early pioneers remained in the interior, close to the Sacramento River. Native Americans and Asian Americans are underrepresented in literary works. Our findings certainly provide concrete information that might drive future research on the impact of curricula on students, but we should not expect to see representation that is exactly comparable to demographics in the population.

Identify Famous People: Table 2 reveals that most historical figures in books are White male politicians. Since one-fifth of our novels mention the top fifty, we pay close attention to them. Only the 28th-mentioned First Lady Eleanor Roosevelt cracks the top 50. Eleanor Roosevelt has the most pages in US history textbooks, according to prior research. American feminist Jane Addams ranked 54th among school textbooks' most notable women. Few persons of color appear in the top 50, including former slave Dred Scott (42nd), abolitionist Frederick Douglass (44th), and activist Martin Luther King, Jr. (30th). Barack Obama ranks 29th. This implies that, with a few exceptions, one demography dominates textbook space for historical figures. In our subsequent findings, we address the possibility that textbooks emphasized politics over real-world examples or social and cultural trends.

Table 1: The Top 30 Most Common Named People Across All Textbooks (Khurana et al., 2022)



Name	No. of appearances	Wikidata gender
Andrew Jackson	3,347	Male
Thomas Jefferson	3,033	Male
Franklin Delano Roosevelt	2,672	Male
Richard Nixon	2,659	Male
Theodore Roosevelt	2,627	Male
Ronald Reagan	2,294	Male
John F. Kennedy	2,176	Male
Lyndon Johnson	1,546	Male
George W. Bush	1,291	Male
Woodrow Wilson	1,269	Male
Alexander Hamilton	1,234	Male
Harry S. Truman	1,227	Male
Bill Clinton	1,211	Male
James Madison	1,173	Male
John Adams	1,156	Male
Andrew Johnson	1,125	Male
Robert E. Lee	1,053	Male
Abraham Lincoln	968	Male
Adolf Hitler	961	Male
George Washington	875	Male
Eisenhower[a]	856	(None)
Ulysses S. Grant	803	Male
John Quincy Adams	789	Male
Jimmy Carter	785	Male
John Brown	694	Male
Herbert Hoover	660	Male
George H. W. Bush	658	Male
Eleanor Roosevelt	573	Female
Barack Obama	566	Male
Martin Luther King Jr.	563	Male

Note. The names are obtained after Wikidata name standardization, frequency filtering, and last name disambiguation.
[a]Because *Eisenhower* did not manage to be automatically disambiguated and is not a full name, Wikidata does not have a gender label for it, but this name most likely refers to the White president Dwight D. Eisenhower, who is a man.

Research Question 2: About Certain Demographics, What Are the Most Talked-about Subjects?

The phrases with the highest chance are shown in Table 3 for each of the 10 most common topics in the texts. Researchers engaging in topic modeling need to exercise great caution in their interpretive sense of words to provide groups of words with labels or meanings. Upon meticulously reviewing the ten most probable terms for each of the fifty subjects, we discover that seventeen are associated with official politics (including stems like govern, preside, polit, and federal), two with social movements (movement, protest, and civil), and three with ordinary workers (farmer, worker, and soldier). Although some scholars have recommended focusing on themes that center on people's voices, the increased number of subjects indicates that formal politics are given greater weight, since the possibility of any subject happening varies from 1.8% to 2.2%.

How are our themes spread throughout the various non-named persons groups that we defined earlier? This will help us explore further into our topics. The group's thematic variety is examined by counting the topics to which they are related (Table 4). We discover that when we think of gender, two things come to mind: the first is family, and the second is social movements (women's rights).

We discover that white is only used for other ethnicities, which means that Whiteness is not noticeable until minority ethnicities are compared. Native Americans are often mentioned as settlers, Black people in two separate contexts—slavery and civil rights—and Latinx people about territorial claims. These examples show that when people talk about minority ethnicities, they generally bring up issues that include the minority's connection to the majority. The level of multifocality in textbooks might be better understood with the use of more qualitative studies and the subjects (Lucy et al., 2020).

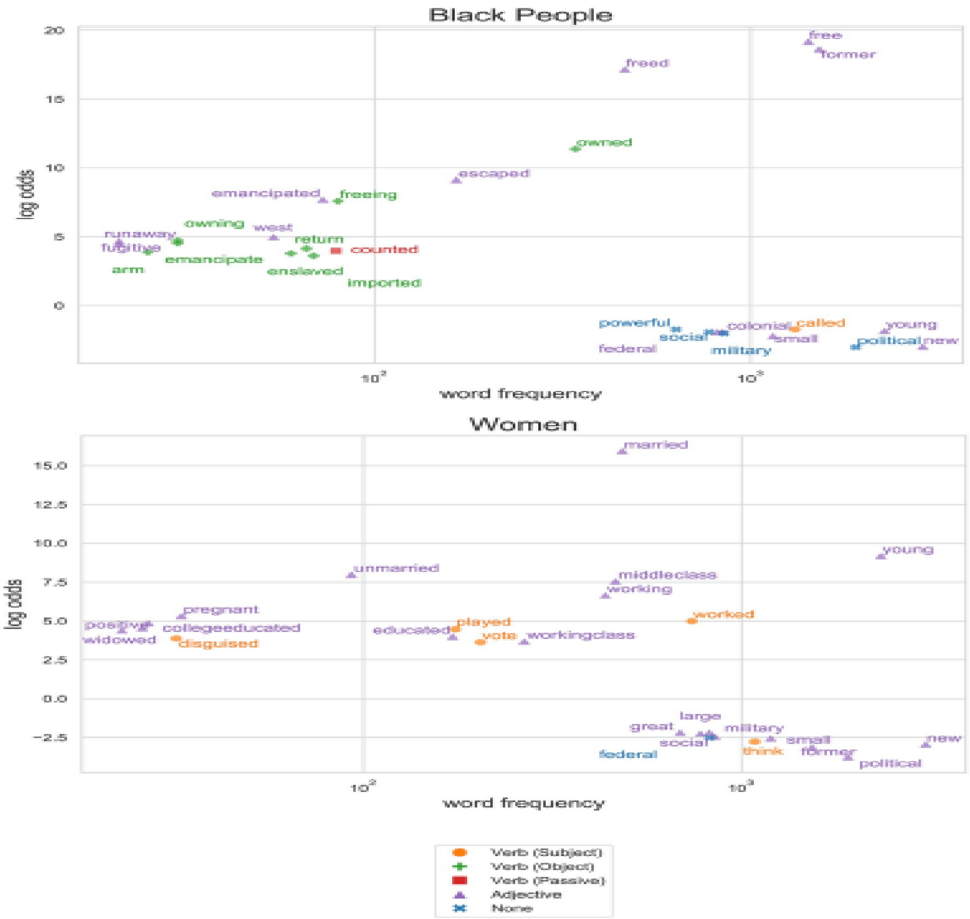


Figure 2: Word Correlations with Black Women and Women in General

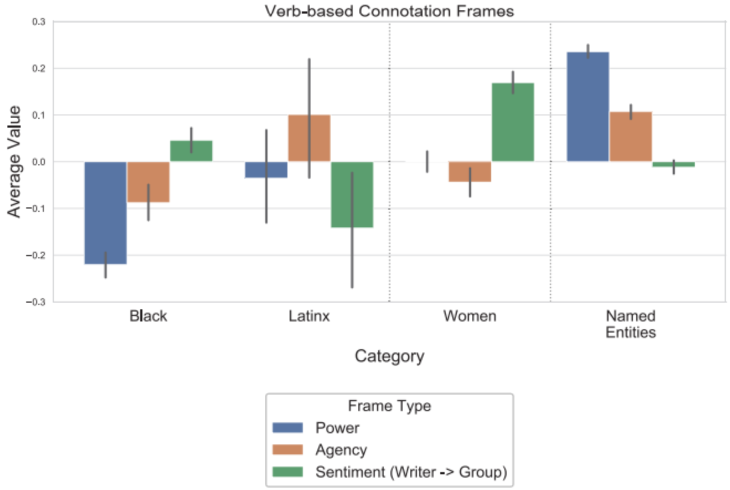


Figure 3: Power, Agency, and Emotion Frameworks for Social Groupings Based on Verb Connotations (Lucy et al., 2020)

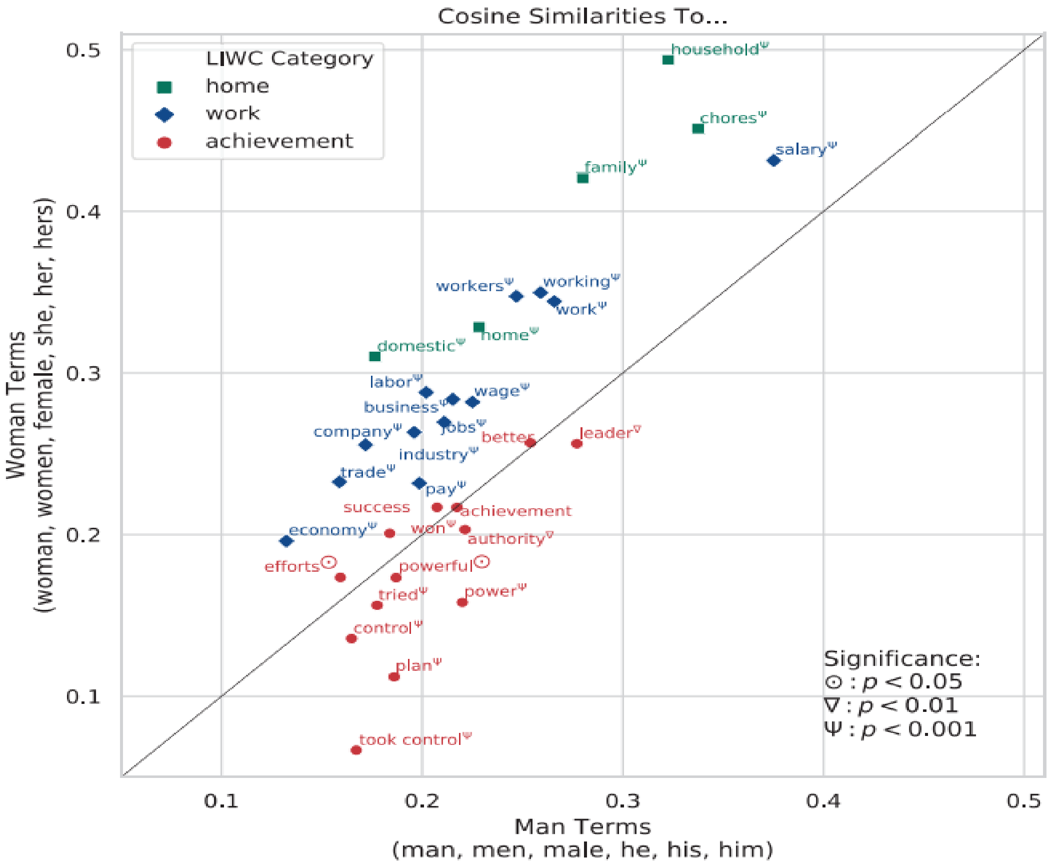
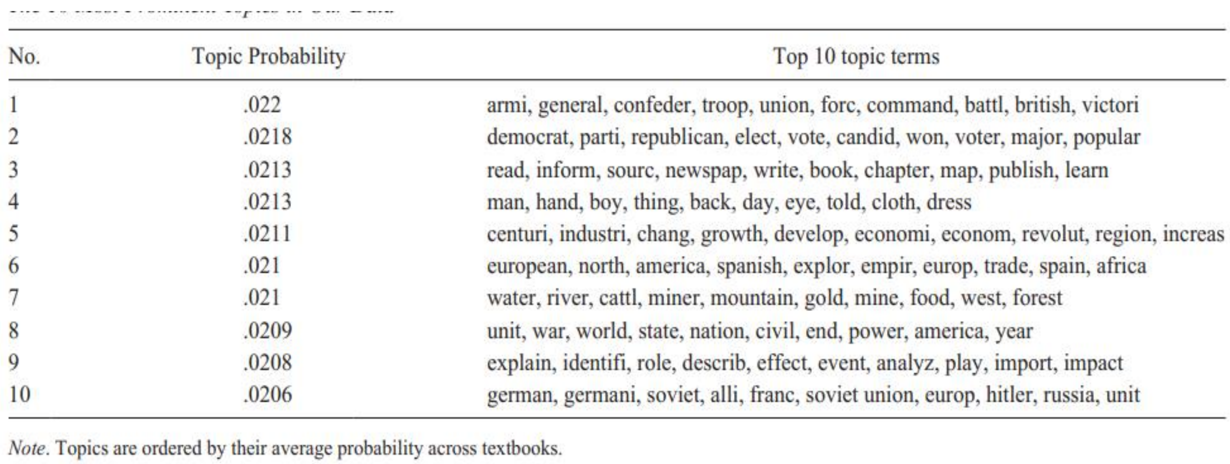


Figure 4: Words about House, Job Success And Gendered Phrases are Cosine Similar

Table 2: Ten of Our Most Common Data Points (Roblek et al., 2020)

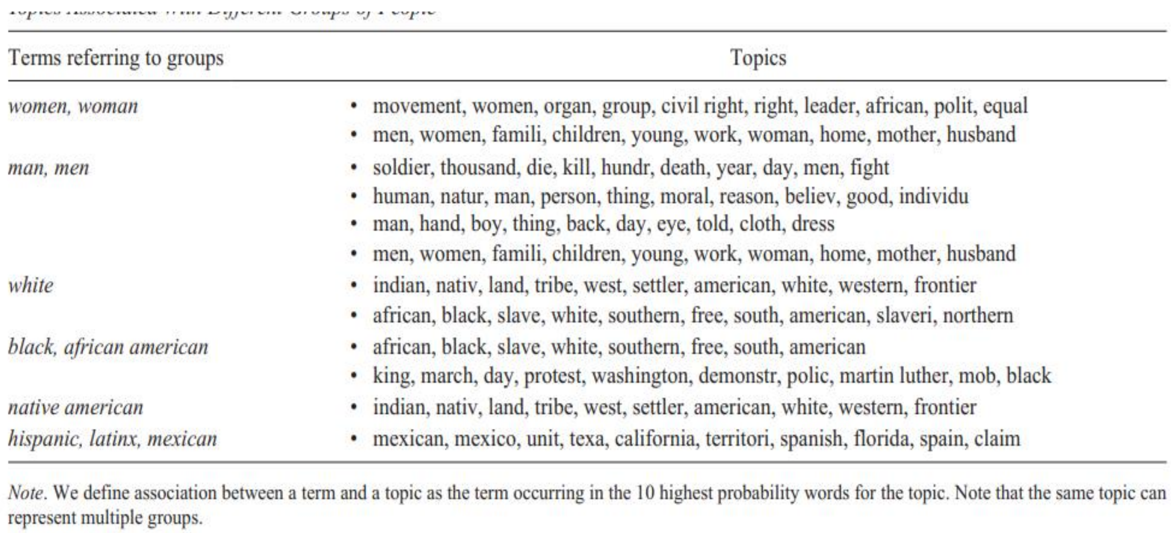


The 10 Most Prominent Topics in Our Data

No.	Topic Probability	Top 10 topic terms
1	.022	armi, general, confeder, troop, union, forc, command, battl, british, victori
2	.0218	democrat, parti, republican, elect, vote, candid, won, voter, major, popular
3	.0213	read, inform, sourc, newspap, write, book, chapter, map, publish, learn
4	.0213	man, hand, boy, thing, back, day, eye, told, cloth, dress
5	.0211	centuri, industri, chang, growth, develop, economi, econom, revolut, region, increas
6	.021	european, north, america, spanish, explor, empir, europ, trade, spain, africa
7	.021	water, river, cattl, miner, mountain, gold, mine, food, west, forest
8	.0209	unit, war, world, state, nation, civil, end, power, america, year
9	.0208	explain, identifi, role, describ, effect, event, analyz, play, import, impact
10	.0206	german, germani, soviet, alli, franc, soviet union, europ, hitler, russia, unit

Note. Topics are ordered by their average probability across textbooks.

Table 3: Discussion Points Relevant to Various Demographics



Topics Associated With Different Groups of People

Terms referring to groups	Topics
women, woman	• movement, women, organ, group, civil right, right, leader, african, polit, equal • men, women, famili, children, young, work, woman, home, mother, husband
man, men	• soldier, thousand, die, kill, hundr, death, year, day, men, fight • human, natur, man, person, thing, moral, reason, believ, good, individu • man, hand, boy, thing, back, day, eye, told, cloth, dress • men, women, famili, children, young, work, woman, home, mother, husband
white	• indian, nativ, land, tribe, west, settler, american, white, western, frontier • african, black, slave, white, southern, free, south, american, slaveri, northern
black, african american	• african, black, slave, white, southern, free, south, american • king, march, day, protest, washington, demonstr, polic, martin luther, mob, black
native american	• indian, nativ, land, tribe, west, settler, american, white, western, frontier
hispanic, latinx, mexican	• mexican, mexico, unit, texa, california, territori, spanish, florida, spain, claim

Note. We define association between a term and a topic as the term occurring in the 10 highest probability words for the topic. Note that the same topic can represent multiple groups.

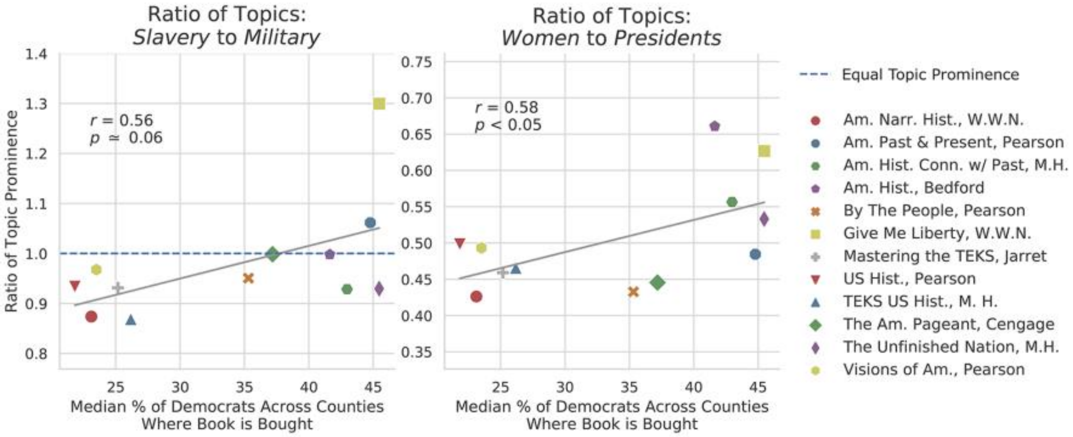


Figure 5: Political Leanings of the Countries That Buy Books and Topic Ratios (Zhao Et Al., 2021)

Additionally, we examine the subjects that have been previously researched about the portrayal of individuals and events in U.S. history textbooks and rank them according to their relative importance. For instance, we look at how often slavery and the military are covered in textbooks and how many times women are discussed compared to presidents. Books bought in more Republican counties, as seen in Figure 5, are more likely to discuss military issues (armi, military) than slavery topics (slave, slavery; r =.56, p ≈.06). There is a significant correlation between the median percentage of Democrats who purchase books and topics about women (women, women) being more prevalent than topics about presidents (r =.58, p <.05). Despite these differences, many of the books focus on male presidents rather than female leaders. There is minimal variation in the depiction of individuals across books, as seen in Figures 3 and 7, but the similarities suggest a shared historical story. This provides further evidence that the paintings are symbolic (Charlesworth et al., 2021).

4.0 CONCLUSIONS AND RECOMMENDATIONS

Conclusion

We observed that Texas's US history textbooks that employed natural language processing for person identification underrepresented Latinx students and labeled mostly white men. The examined word connections suggest that black people are engaged in subordinate tasks, whereas women are addressed in terms of marriage, the home, and the work. Topic modelling shows that minority ethnicity issues include White people and political history is more common than social history. More conservative counties' textbooks had fewer minority representations, although the systematic variance is minimal compared to the textbooks' extensive similarities. Since many of our approaches have been applied to fiction, social media, and news, future methodological work should focus on establishing new models, algorithms, and vocabularies for social science textbooks. As indicated, computational techniques are not enough since this work requires cross-disciplinary cooperation. In-depth qualitative research by education and social science professionals is needed to understand course content and computational outcomes. To increase equality, methods are constantly enhanced. Textbook classification techniques are our main focus. Like Apple and Christian-Smith (2017), we hope our technique will provide legitimacy to future research on book content processes. The makeup of a state's board of education or other characteristics may illuminate the first relationship between political climate and content. Given the ever-changing nature of standards and the absence of a clear response to textbook content, we expect more sophisticated metrics will steer textbook improvement conversations. NLP and other AI and data science approaches might greatly enhance educational research, policy, and practice.

Recommendations

When traditional content analysis methods have long been the standard, like in textbook research, natural language processing (NLP) approaches perform well. Computational methods have the potential to provide discoveries in addition to enabling studies that are broader, more comprehensive, and faster than prior studies. In addition to improving our understanding of how words, people, and problems come together to create conceptions, they may provide new insight into the depth and breadth of discursive tendencies. It is also simpler to examine the connections between the text and outside influences when using larger sample sizes in conjunction with these quantitative data.

REFERENCES

Anon (2022). *APA PsycNet*. [online] psycnet.apa.org. Available at: https://psycnet.apa.org/record/2022-12800-001.

Boyd, R.L. and Schwartz, H.A. (2020). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, 40(1), pp.21–41. doi:https://doi.org/10.1177/0261927x20967028.

Charlesworth, T.E.S., Yang, V., Mann, T.C., Kurdi, B. and Banaji, M.R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, 32(2), pp.218–240. doi:https://doi.org/10.1177/0956797620963619.

Chowdhary, K.R. (2020). Natural Language Processing. *Fundamentals of Artificial Intelligence*, 08(9), pp.603–649. doi:https://doi.org/10.1007/978-81-322-3972-7_19.

Khurana, D., Koli, A., Khatter, K. and Singh, S. (2022). Natural Language processing: State of the art, Current Trends and Challenges. *Multimedia Tools and Applications*, 82(3), pp.3713–3744. doi:https://doi.org/10.1007/s11042-022-13428-4.

Kleinheksel, A.J., Winston, N.R., Tawfik, H. and Wyatt, T.R. (2020). Demystifying Content Analysis. *American Journal of Pharmaceutical Education*, [online] 84(1). doi:https://doi.org/10.5688/ajpe7113.

Lucy, L., Demszky, D., Bromley, P. and Jurafsky, D. (2020). Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks. *AERA Open*, 6(3), p.233285842094031. doi:https://doi.org/10.1177/2332858420940312.

Nicolas, C., Kim, J. and Chi, S. (2021). Natural language processing-based characterization of top-down communication in smart cities for enhancing citizen alignment. *Sustainable Cities and Society*, 66(07), p.102674. doi:https://doi.org/10.1016/j.scs.2020.102674.

Roblek, V., Thorpe, O., Bach, M.P., Jerman, A. and Meško, M. (2020). The Fourth Industrial Revolution and the Sustainability Practices: A Comparative Automated Content Analysis Approach of Theory and Practice. *Sustainability*, 12(20), p.8497. doi:https://doi.org/10.3390/su12208497.

Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K.J., Ajagbe, M.A., Chioasca, E.-V. and Batista-Navarro, R.T. (2021). Natural Language Processing for Requirements Engineering. *ACM Computing Surveys*, 54(3), pp.1–41. doi:https://doi.org/10.1145/3444689.