Censorship and Narrative Control: A Textual Analysis of Florida’s Banned Books

|  |  |  |
| --- | --- | --- |
|  | Conie O’Malley  American University  co1984a@american.edu |  |

**Abstract**

*This study examines the increasing censorship of books in Florida, identifying key themes among banned books using text mining methodologies. By employing term frequency (TF), TF-IDF, n-grams, dictionary-based analysis, and machine learning classification, this research aims to uncover the socio-political motivations behind book banning practices. The analysis compares Florida’s banned books with those from Iowa to identify regional censorship trends. This study also investigates the role of demographic patterns in book censorship, particularly the impact on authors from marginalized communities. The findings provide insights into the broader implications of book bans on intellectual freedom, educational equity, and the politicization of literature. By leveraging computational text analysis, this research contributes to the ongoing discourse on censorship, power, and the control of narratives in contemporary society.*

**Keywords:** books, censorship, ban

# 1. Introduction

Book banning has become an increasingly pressing issue across the United States, with Florida standing out as a focal point for literary censorship in educational institutions. This phenomenon reflects deeper ideological struggles over access to information, identity representation, and political discourse. Throughout history, book bans have served as mechanisms for controlling narratives, limiting exposure to controversial ideas, and reinforcing dominant social ideologies. The 2021–2022 school year saw an unprecedented rise in book challenges, prompting widespread debate about intellectual freedom, educational equity, and the role of politics in shaping curriculum decisions (Goncalves et al., 2024). This study aims to examine the themes of banned books in Florida, identifying recurring patterns of censorship and assessing their broader implications. By analyzing the justifications for book bans and exploring their sociopolitical contexts, this research seeks to contribute to the ongoing discourse on censorship, education, and democratic access to diverse perspectives.

# 2. Problem Statement

The problem is that there is a significant research gap in how book censorship is analyzed. While thematic analysis offers insight to the most dangerous trend – judging a book by its cover and discovering the reasons why banned books are often banned or objected to before they even leave the shelves, there is a limited amount of research connecting the cover themes and summaries to the textually analyzed themes between the covers. Decisions to ban are frequently based on the cover text, including the title, summary, and keywords that appear on the front or back cover raising critical concerns about the motivations behind censorship and how superficial judgments dictate access to literature. This project aims for a full textual analysis that can provide insight into the themes of banned books and compare them to those in the summary and online, shedding light on linkage between the themes that may cause many of these books to be banned without ever being read.

Book bans have become a widespread phenomenon in the United States, with Florida emerging as a critical battleground for literary censorship. The issue extends beyond the removal of individual books; it reflects deeper ideological conflicts regarding race, gender identity, and historical narratives in education. Knox (2020) highlights that book bans often suppress perspectives that challenge dominant cultural norms, reinforcing systemic inequalities. Lycke and Lucey (2018) argue that such censorship limits students’ ability to engage critically with democratic principles, reducing their exposure to diverse viewpoints. Florida’s censorship policies have disproportionately targeted works by authors of color and LGBTQ+ themes, raising concerns about the erasure of marginalized voices in educational settings (Goncalves et al., 2024).

This study seeks to address this research gap by identifying the keywords and themes that trigger book bans before the books are even opened. Using text mining methodologies, this project will analyze patterns in book descriptions to determine the linguistic elements most associated with censorship decisions. By shedding light on the role of superficial language in book bans, this research will contribute to the broader discourse on intellectual freedom, educational equity, and the sociopolitical mechanisms behind literary censorship.

**3. Purpose Statement**

The purpose of this research is to conduct a textual analysis of Florida’s banned books to identify thematic trends in censorship and explore the socio-political motivations behind these restrictions. By leveraging text mining methodologies, this study will systematically categorize recurring themes in descriptions and summaries of banned books and assess how these themes align with broader discussions on censorship, anti-intellectualism, and political mobilization. This research aims to provide empirical insights into the narratives that are being systematically excluded and to contribute to the broader discourse on intellectual freedom and educational equity.

# 4. Conceptual Framework

This study draws upon theories of censorship, political sociology, and critical pedagogy to examine the phenomenon of book banning. The analysis considers the themes of banned books, focusing on recurring topics such as race, gender identity, historical accuracy, and political ideology. These themes often reflect broader societal tensions, revealing how censorship operates as a means of controlling discourse and restricting access to certain narratives. Furthermore, the study investigates the stated justifications for banning books, which include reasons provided by school boards, policymakers, and parents. These justifications often cite concerns about age-appropriateness, offensive language, or controversial subject matter, yet they may mask underlying ideological motives. Additionally, the study examines demographic patterns, analyzing the identities of authors and the targeted readership of banned books. A significant proportion of censored works are authored by individuals from marginalized communities, suggesting that book bans disproportionately impact diverse voices. Lastly, this research explores the political and social contexts that influence book banning trends at both local and state levels, considering the role of legislative policies, community activism, and political affiliations in shaping censorship practices. By understanding these dynamics, this study seeks to provide a comprehensive perspective on the intersection of literature, power, and societal control.

Fig. 1 – Conceptual Framework

# 5. Methodology

To systematically analyze the themes of Florida’s banned books, this study employs text mining and machine learning techniques. The following methodologies will be used:

* **Key Word Frequency (TF/TF\*IDF)** – Individual words from book descriptions will be analyzed using Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF\*IDF) to identify the most prominent terms while correcting for any commonly used words to prevent inadvertent data skewness.
* **Key Phrase Frequency** – Bigrams (N-grams) will be used to detect frequent multi-word phrases within book descriptions.
* **Subgroup Comparisons** – A comparative analysis will be conducted between banned books in Florida and those in Iowa to identify regional differences in censorship patterns.
* **Dictionary-Based Analysis** – A dictionary will be developed in R to categorize and interpret themes based on their frequency of appearance.
* **Classification Model** – A machine learning classification model will be trained to predict whether a book is likely to be banned based on its description, using features extracted from text analysis.

# 6. Research Questions

Understanding the underlying themes and patterns of book censorship is crucial for framing this analysis. By examining linguistic trends in banned book descriptions, this study will uncover recurring motifs that may indicate broader sociopolitical influences on censorship decisions.

* How do TF/TF\*IDF measures highlight key themes in censored books compared to non-censored books?
* How do the themes of banned books in Florida compare to those in Iowa, and what regional differences emerge from this comparison?
* How do the stated justifications for banning books align with the actual themes detected through text analysis?
* What textual characteristics are most predictive of a book being banned?

# 7. Data Collection

7.1 **Data Types**

This study uses two primary types of textual data: full-text book content and thematic metadata. The full-text data comprises books that have been banned in Florida and Iowa, sourced from the public domain repository, Project Gutenberg. These texts provide the raw material for computational analysis, allowing for comparisons of linguistic structures, themes, and keyword frequencies. In addition to full-text analysis, thematic metadata is collected from Google Books. This metadata includes assigned themes, genres, and summary descriptions, providing additional context for understanding the reasons behind book censorship. The combination of these data sources enables a comprehensive investigation into censorship trends, allowing for statistical and machine learning-based text analysis.

**7.2 Data Description**

The dataset consists of 20 books, with 10 books selected from each state’s banned books list. The books from Florida include Wuthering Heights, The Taming of the Shrew, Helen of Troy, Leonardo Da Vinci, The Road, The Pirate, The Dark Tower, Redeemed, Monster, and The Heir. Similarly, the Iowa dataset contains The Picture of Dorian Gray, The Talisman, The Great Return, Smoke, The Best Man, Glass, Redeemed, The Bridge, Monster, and Dead End. The selected titles represent a mix of classic literature, contemporary fiction, and young adult novels, reflecting diverse genres and themes. This sample allows for a comparative analysis of censorship patterns across different literary styles and subject matters.

The textual data has been assigned to two data frames: ia\_book\_texts\_df for the Iowa books and fl\_book\_texts\_df for the Florida books. These structured datasets facilitate preprocessing tasks such as tokenization, keyword extraction, and machine learning classification. The thematic metadata from Google Books will be assigned to corresponding variables, creating a structured representation of the books' themes for further analysis.

**7.3 Data Retrieval Process**

The full-text books were obtained from the Project Gutenberg library using the gutenbergr package in R, which allows for seamless access to public domain texts. Initially, an attempt was made to retrieve a completely random sample, but due to limitations in full-text availability, the selection process was adjusted to include only books from each state’s banned books list that were available in the library[[1]](#footnote-1). Six out of the twenty books were retrieved directly using gutenbergr, while the remaining fourteen required custom functions to scrape the texts directly from the Project Gutenberg website. This additional step introduces complexities in data cleaning but ultimately enhances the analysis by providing a more comprehensive dataset.

Each book was retrieved using its unique Gutenberg ID and stored in data frames, one for Florida (fl\_book\_texts\_df) and one for Iowa (ia\_book\_texts\_df). The next step in the process is data cleaning, which will involve several key tasks to ensure the dataset is structured and ready for computational analysis. This includes removing HTML spacing left over from direct web downloads, eliminating unnecessary whitespace, stripping chapter numbers and labels, and filtering out stopwords, punctuation, and other extraneous symbols that could distort the analysis.

Additionally, once the initial cleaning is complete, further preprocessing steps will be applied. This includes normalizing text formatting to maintain consistency across all documents, converting all words to lowercase to standardize frequency counts, and segmenting sentences to ensure optimal phrase extraction. After these transformations, trigram vectors will be generated to capture meaningful three-word sequences, allowing for deeper thematic exploration. Furthermore, term frequency and inverse document frequency calculations will be performed to determine the significance of specific words within each dataset.

By following this structured approach to data preparation, the study will ensure that the dataset is well-processed for subsequent machine learning applications, leading to more accurate insights into censorship patterns and textual analysis.

Given the importance of thematic metadata in my analysis, I explored multiple retrieval methods to ensure comprehensive data collection. My initial approach involved developing a custom web scraping function to extract book descriptions and assigned thematic categories from online sources. This method aimed to automate data extraction directly from publicly available book repositories. However, security measures, dynamic page loading, and inconsistencies in site structure rendered this technique unreliable, preventing complete retrieval of the necessary metadata.

To address these limitations, I pursued an API-based solution, utilizing an API key with structured query parameters to access thematic metadata programmatically. Despite successful authentication and data requests, the API responses either lacked the required level of detail or restricted access to critical fields due to licensing constraints. After repeated query adjustments and attempts to refine data retrieval, it became evident that this method would not yield the comprehensive dataset needed for my analysis.

Recognizing these challenges, I turned to ChatGPT as an alternative solution. Through carefully structured prompts, I successfully extracted detailed thematic metadata, including book descriptions, assigned categories, and keyword associations. The retrieved data were formatted into a structured CSV file, ensuring compatibility with my preprocessing pipeline for text mining and machine learning analysis. This iterative process underscores the adaptability required in data collection, demonstrating how AI-assisted retrieval can serve as a viable complement to traditional scraping and API-based techniques when access constraints arise. By leveraging multiple methodologies, I was able to secure a complete and structured dataset, allowing for a more robust examination of censorship themes in banned books.

# 8. Data Preprocessing

Effective text analysis requires a structured preprocessing pipeline, and in this study, multiple iterations were necessary to refine the data. The initial steps involved normalizing the text—converting it to lowercase, removing punctuation, special characters, and unnecessary whitespace. Stopwords were then filtered out to reduce noise and improve the relevance of the extracted terms.

Once the text was tokenized and lemmatized, a preliminary analysis of the cleaned dataset exposed several persistent issues. The Gutenberg boilerplate text, which contains licensing information and non-relevant metadata, remained embedded in several documents. Additionally, chapter headings and excessive whitespace disrupted the dataset’s structure, interfering with pattern recognition.

To address these issues, the text-cleaning function was modified and re-run. After each iteration, manual inspection of the output was conducted to ensure the adjustments effectively removed unwanted elements. It took three full iterations before the dataset reached a suitable state for analysis. Each revision involved refining regular expressions and improving filters to strip residual noise while preserving meaningful text.

This iterative process applied to both the full text of banned books and their descriptions. By carefully validating each cleaning cycle, the dataset was transformed into a structured format, enabling reliable feature extraction, machine learning classification, and visualization. This approach underscores the importance of iterative refinement in text preprocessing, ensuring that only relevant content remains for analysis.

# 9. Data Analysis and Visualization

To analyze censorship themes in banned books, I constructed term frequency matrices (TFM) and document-term matrices (DTM) for both the Florida and Iowa datasets. These matrices provided a structured representation of word occurrences across the dataset, allowing for the identification of frequently used terms within censored texts. By creating DTMs, I was able to quantify the presence of words in each document and prepare the dataset for feature extraction and text classification.

Once the DTMs were generated, I moved to data visualization to better understand the distribution of word frequencies. However, I quickly discovered a significant issue: a large number of NA values were present in the matrices. These missing values disrupted frequency calculations and required me to return to data preprocessing to investigate the root cause. Through additional iterations of text cleaning, I refined the handling of empty cells and missing text entries, ensuring that all documents contained valid text data before reprocessing the matrices.

With the cleaned dataset, I recalculated and visualized TF-IDF scores to determine the importance of specific terms relative to their frequency across multiple documents. However, even at this stage, I noticed NA values persisted in the TF-IDF output, affecting the weighting calculations. This prompted another round of data preprocessing adjustments, where I carefully re-examined stopword filtering, tokenization, and text normalization procedures. After implementing additional cleaning functions, I successfully removed the remaining NA values, allowing for a more accurate and structured representation of term importance within the dataset.

Ultimately, these iterative refinements ensured that the final TF-IDF analysis provided meaningful insights into censorship patterns, highlighting key words and themes disproportionately targeted in banned books. The structured approach to data correction significantly improved the accuracy of my text mining models, ensuring that the results were reliable and interpretable.

A graph of a number of words

AI-generated content may be incorrect.

Fig. 2 – Florida Term Frequency Histogram – top 20

A screenshot of a computer screen

AI-generated content may be incorrect.

Fig. 3 – Florida TF-IDF Bar Chart – top 20

# 10. References

Abraham, F. (2021). Gun Violence Corpus. *GitHub*. https://github.com/fabraham/gun-violence-corpus

Astrolabe Diagnostics. (2022). Biblical Lexicon for NLP. *GitHub*. https://github.com/astrolabediagnostics/biblical-lexicon

Baker, C. F., et al. (2020). FrameNet. *International Computer Science Institute*. https://framenet2.icsi.berkeley.edu

Banjoemon. (2019). US History Keyword Lists. *GitHub*. https://github.com/banjoemon/us-history-keywords

Barrett, M. (2020). UNESCO Education Glossary. *GitHub*. https://github.com/malcolmbarrett/unesco-glossary

Basile, V. (2019). HurtLex: Multilingual lexicon of offensive words. *GitHub*. https://github.com/valeriobasile/hurtlex

Bird, J. (2020). Religion in FrameNet. *GitHub*. https://github.com/julia-bird/religion-in-frameNet

COIN Project. (2021). Anti-Science and Censorship Discourse Corpus. *GitHub*. https://github.com/COIN-project/anti-science-discourse

Da Silva, D. (2022). Domestic Life in FrameNet. *GitHub*. https://github.com/daniel-dasilva/domestic-life-framenet

EdTech Hub. (2021). Education Policy Lexicon. *GitHub*. https://github.com/edtechhub/education-policy-lexicon

Fahlgren, N. (2021). World History Corpus. *GitHub*. https://github.com/nfahlgren/world-history-corpus

Goncalves, M. S. O., Langrock, I., LaViolette, J., & Spoon, K. (2024). Book bans in political context: Evidence from US schools. *PNAS Nexus, 3*(6), Article pgae197. https://doi.org/10.1093/pnasnexus/pgae197

Goodspeed, B. (2022). Political Framing Lexicons. *GitHub*. https://github.com/besongoodspeed/political-framing-lexicons

Halterman, A. (2019). Media Frames Corpus and Policy Frames Codebook. *GitHub*. https://github.com/ahalterman/mediaframes

Hendricks, D. M. (2021). Gender Gap Lexicon. *GitHub*. https://github.com/dmhendricks/gender-gap-lexicon

Human Rights Campaign Foundation. (n.d.). Glossary of Terms. https://www.thehrcfoundation.org/professional-resources/glossary-of-terms

Human Rights Campaign Foundation. (n.d.). Glossary of Terms: Chosen Family & Adoption. https://www.thehrcfoundation.org/professional-resources/glossary-of-terms

Knox, E. (2020). Books, censorship, and anti-intellectualism in schools. *Library Trends, 69*(2), 287–306. https://doi.org/10.1353/lib.2020.0001

Kno10. (2020). Historical Keywords (Wikipedia-based). *GitHub*. https://github.com/kno10/historical-keywords

Lee, J. T. (2022). Self & Identity Lexicon (Psychology NLP). *GitHub*. https://github.com/jtlee99/self-identity-lexicon

Lycke, K., & Lucey, T. (2018). The messages we miss: Banned books, censored texts, and citizenship. *Journal of Social Studies Education Research, 9*(3), 1–26. https://www.jsser.org/article/view/5000217740

NLP4PI. (2021). Political Discourse NLP Lexicon. *GitHub*. https://github.com/NLP4PI/political-discourse

NxG, N. (2021). Intersectionality Lexicon. *GitHub*. https://github.com/napsternxg/intersectionality-lexicon

Open Source Gender Project. (n.d.). Gender-Inclusive Dictionary. *GitHub*. https://github.com/gender-inclusive/gender-inclusive-dictionary

OpenAI. (2025). ChatGPT (Feb 18 version). *OpenAI*. https://openai.com/chatgpt

OpenVA. (2020). Violence Detection Corpus. *GitHub*. https://github.com/openva/violence-detection

PEN America. (2023). Banned in the USA: Censorship Report. https://pen.org/report/banned-in-the-usa-growing-movement-to-censor-books-in-schools/

Pew Research Center. (n.d.). Religious Glossary. *Pew Research Center*. https://www.pewresearch.org/religion/religious-glossary/

Pi, S. (2021). Chicago Tribune Hate Speech Lexicon. *GitHub*. https://github.com/stevenpi/chicagotribune-hate-speech

Project Gutenberg. (n.d.). Project Gutenberg. Retrieved February 18, 2025, from https://www.gutenberg.org

Rowan, Z. (2020). Race Bias Lexicon. *GitHub*. https://github.com/rowanz/race-bias

Satola, K. (2022). Labeling Theory Lexicon. *GitHub*. https://github.com/ksatola/labeling-lexicon

Schweinberger, M. (2020). Indoctrination Lexicon. *GitHub*. https://github.com/schweinberger/indoctrination-lexicon

Then, S. (2022). LGBTQ+ Hate Lexicon. *GitHub*. https://github.com/seanthen/LGBTQ-Hate-Lexicon

Tolà, T. (2022). Implicit Race Bias Corpus. *GitHub*. https://github.com/tolatola/implicit-race-bias

Zach, M. (2022). Family Lexicon: Structure and Socialization. *GitHub*. https://github.com/melissazach/family-lexicon

Zambrano, D. (2022). LGBTQ Lexicon. *GitHub*. https://github.com/DanielaZambranoG/LGBTQ-Lexicon

1. This will be noted as a limitation of the study as the project progresses [↑](#footnote-ref-1)