**Slide 1: Title Slide**

Hi everyone, my name is Grace O’Malley, and today I’ll be presenting my research titled *“Censorship and Narrative Control: A Textual and Comparative Analysis of Florida’s Banned Books.”* This study explores how banned books are linguistically and thematically structured, and what those patterns tell us about censorship across states.

**Slide 2: Agenda**

Here’s a brief overview of what I’ll cover today:

We’ll start with the problem at hand and why it matters.

Then I’ll walk you through my conceptual framework and methodology.

We’ll dive into the key findings from different text analysis techniques.

And finally, I’ll share my conclusions and where I see this research going next.

**Slide 3: Problem Statement**

Book banning is on the rise in the U.S., and Florida has emerged as a focal point. Bans are often justified with vague claims—like books containing themes that are “age-inappropriate”—but these claims rarely engage with the actual content of the texts and we run into the age-old problem of judging a book by its cover.

This study challenges this practice by analyzing both the **metadata** used to justify bans and the **full text** of the books themselves, identifying where the themes align – or misalign – in an effort to coalesce what is on the cover of a book with what is between the cover of a book.

**Slide 4: Conceptual Framework**

This project is grounded in theories of censorship, critical pedagogy, and political sociology. The framework considers the deduced themes of the texts, the published themes of the text, and the published descriptions of the texts, and what kinds of content they contain—particularly focusing on race, gender, history, and identity.

The core idea is that censorship isn’t just about offense—it’s about **narrative control**.

**Slide 5: Data & Methodology**

The dataset includes 49 books—22 banned and 27 not banned.

Full texts were pulled from Project Gutenberg through scripting, and thematic metadata and descriptions were extracted using ChatGPT and validated against Google Books.

The methodology combines traditional text mining—like term frequency, term frequency inverse document frequency—with bigram networks, dictionary-based theme detection, and supervised classification modeling.

**Slide 6: Methodology Overview**

The process involved multiple passes of data cleaning, removing stopwords, lemmatization, and name filtering using Named Entity Recognition.

After preparing the data, I used several analysis techniques that I will dive into in more detail in the following slides. These text analysis techniques were used to detect patterns in content and theme distribution, both by region and by book status.

**Slide 7: Term Frequency Analysis**

In Florida’s banned books, high-frequency terms were emotionally charged—words like “eyes,” “told,” and “heart.”

In contrast, Iowa’s banned books leaned technical and historical, with terms like “platoon,” “engraved,” and “seventeenth.”

This points to **emotional narrative censorship in Florida** versus **epistemic or knowledge-based censorship in Iowa.**

**Slide 8: TF-IDF Results**

TF-IDF analysis shows that:

Florida books surfaced identity-related terms, while Iowa books focused on niche historical references and scientific jargon.

This reinforces the divergence of themes between the two regions and begins to illuminate the fact that **books are banned not just for what they say—but how and to whom they say it.**

**Slide 9: Bigram Network Comparison**

Florida’s bigram graph was sparse and introspective phrases like “shrugged shoulders” or “loving service” suggest emotional, interpersonal, or psychological themes.

Iowa’s network was dense and thematic—phrases like “rare specimens” and “technical evidence” pointed to reference-style content.

So again, we’re seeing that censorship maps onto **narrative style** as much as content – with Florida trending towards emotional themes and Iowa towards the technical

**Slide 10: Thematic Dictionary Application**

After thorough research, I carefully selected terms for a custom dictionary so as to catch key themes, but not to entrap noise in thematic analysis by pulling in words that are too generic – the word race, while applying to potentially racist themes also applies to horse racing.

Themes like **violence, gender, and sexuality** were frequent in full texts but **rarely appeared in Google Books metadata.**

This suggests that publishers may sanitize descriptions—either to protect books from bans or to avoid controversy. Regardless of the motive – the themes are wildly misaligned.

**Slide 11: Predictive Modeling**

I trained two classification models—Ridge regression and XGBoost—on thematic vectors. Ridge was chosen to provide a linear model with variable selection but not elimination, while XGBoost was chosen for its nonlinear properties. This framework was chosen to be able to capture underlying nonlinear themes that may exist while not overfitting our data if the relationship was more linear.

They performed well in training (AUCs > 0.87) but dropped to ~0.52 on test data.

This indicates **overfitting**, likely due to a small sample size and metadata gaps—but the feature importance results were still valuable.

Top predictors included **gender, religion, and violence**—themes that consistently appear in banned books but aren’t always cited in ban justifications.

While these models aren’t yet ready for deployment, they demonstrate the potential of using machine learning to flag censorship patterns.

**Slide 13: Conclusions**

So, what did we learn?

* First, banned books often contain **emotionally resonant, identity-based language**, especially in Florida – aligning with much of the research I read during my preparation for this study.
* Second, bans differ by state—**Iowa’s bans are more about knowledge control**, while Florida’s focus heavily on gender and identity representations.
* Third, public justifications don’t align with actual content—**metadata obscures themes**. Whether this be for nefarious or noble purposes, most of these texts are of underrepresented or marginalized voices and they are being obscured.
* And finally, machine learning can detect thematic patterns—but it needs more robust, diverse training data before being a viable model.

**Slide 14: Future Research**

Future directions include expanding the dataset to include texts, enriching the theme dictionary with new categories like disability or bodily autonomy, and integrating regional metadata like school board politics.

With a larger text corpora, there’s also strong potential for **semi-supervised learning** to capture theme interaction and ambiguity.

**Slide 15: Thank You**

Thank you for your time. This project isn’t just about text—it’s about who gets to tell stories and who decides what stories are too dangerous to hear.