Digital Cowboy: A Computational Analysis of Country Music

Ari Gurovich[⋆], Eric Steinberg[⋆], and Joshua Kim[⋆]

[⋆]Emory University, Atlanta, GA

1 Introduction

Country is a distinct and popular musical genre, particularly common in rural and southern areas of the United States. Known for its mix of bittersweet and uplifting melodies, country music often explores themes of heartache, resilience, and faith, woven into imagery that reflects classic Americana and relatable, working-class experiences. While most people can easily identify country music when they hear it, pinpointing the specific features that make a song "country" is a more complicated task. This project aims to uncover the defining linguistic and thematic characteristics of the genre: what makes country lyrics distinctly "country", distinguishing them from those of other genres? We do so by employing several natural language processing methods that we have learned over the course of our QTM 340 class to a large dataset containing lyrics from country and non country songs. These methods include topic modeling, log odds analysis, vocabulary richness metrics, and sentiment analysis. From these findings we reveal insights into the emotional tones, thematic patterns, and lexical sophistication of country music. These findings contribute to the growing field of computational musicology, provide a template for the analysis of other genres, and a foundation for the generation of country song lyrics down the line.

2 Related Research

We drew inspiration from three key research projects that shaped our approach to analyzing country music lyrics using natural language processing (NLP) techniques. The project *What's in a Song? Using LDA to Find Topics in Over 120,000 Songs* (Tim, 2021) applies Latent Dirichlet Allocation (LDA) to uncover thematic structures across various genres, including pop and rock. While this study also employs topic modeling, it doesn't focus on country music as we do. Our research goes further by applying additional methods, such as log odds analysis and sentiment analysis, to isolate the unique linguistic and thematic features of country lyrics.

The article *Heavy Metal and Natural Language Processing* (Iain, 2016) analyzes heavy metal lyrics, delving into themes like aggression and rebellion. Although this study also uses NLP methods to analyze genre-specific themes, our work highlights the narrative and emotional aspects of country music, emphasizing its connection to blue-collar life—a contrast to the more qualitative focus of their analysis. Lastly, the project *Sentiment Analysis and Lyrics Theme Recognition of Music Lyrics* (Pudding, 2020) investigates emotional nuances across genres, revealing sentiment patterns and cultural attitudes. While we also use sentiment analysis, our goal is more targeted: to pinpoint the defining characteristics that make country lyrics uniquely "country."

By extending these studies, we aim to deepen the understanding of country music's lyrical identity. Our work contributes to computational musicology by uncovering emotional tones and thematic patterns specific to the genre and setting the stage for future analysis of other musical styles.

3 Dataset and Pre-Processing

The dataset for this study was sourced from the publicly available "Song Lyrics" dataset on Hugging Face, published by user Amishshah in 2023, which contains song titles and corresponding lyrics from 2.94 million songs across 6 genres: rock, pop, rap, country R'n'B, and a 'misc' genre. Due to the large size of the dataset and computational constraints, a random sample of 25 percent of the dataset was taken for analysis. Its distribution was 348,547 pop songs, 241,143 rap songs, 158,148 rock songs, 38,608 R'n'B songs, 35,398 miscellaneous songs, and 21,706 country songs.

To ensure a robust analysis, we split the data into two subsets: one comprising 20,000 randomly sampled country songs and the other containing 4,000 songs from each of the following genres—pop, rap, rock, R'n'B, and a miscellaneous category. This allowed for a diverse yet balanced representation of country and non-country lyrics for comparative analysis.

We thoroughly cleaned the raw lyrics so as to standardize them for downstream methods. First, we removed non-lyrical annotations such as "[Chorus]" or "(Repeat)", along with filler words and other non-informative terms. Contractions were preserved (e.g., "don't" as "dont") to retain the original emotional tone. We then lemmatized words to their base forms using the NLTK WordNet Lemmatizer, and stripped away non-alphabetic characters. Additionally, we normalized repetitive letters in words (e.g., "loooove") to their standard spelling.

4 Methods

4.1 Topic Modeling

We utilized topic modeling to identify recurring themes and patterns in country lyrics, by clustering related words into coherent topics. This process is particularly useful in analyzing large textual datasets where patterns may not be immediately evident. We applied Latent Dirichlet Allocation (LDA), a widely used algorithm for topic modeling, to treat each document (in this case, a song's lyrics) as a mixture of topics. LDA represents each topic as a distribution of words, finding statistical patterns between word usage.

We first tokenized our lyrics into words and removed stopwords and infrequent terms. We created a dictionary to map words to unique IDs, and generated a bag-of-words representation. Using LDA, we identified 12 latent topics, each described by a set of key terms and their associated probabilities. We then tuned the hyperparameters, such as the number of passes, iterations, and distribution priors, to enhance the model's accuracy. To evaluate the quality of the generated topics, we calculated coherence scores, measuring the interpretability of each topic based on word co-occurrence patterns. A higher coherence score indicates that a topic is more meaningful and likely to be understood by a human. Additionally, we generated PyLDAVis visualizations to illustrate the distribution of topics across the dataset and provide an intuitive way to explore the relationships between topics.

4.2 Log Odds

We conducted log-odds analysis to identify words disproportionately used in country lyrics compared to other genres. Log-odds are a useful metric for the frequency of a particular word's use in one dataset relative to another, adjusted for overall word frequencies. Using a CountVectorizer, we calculated word counts for each dataset, and derived log-odds ratios for each word in country versus non-country lyrics. Words with high positive log-odds ratios, such as "truck" and "heartland," are indicated as having a strong association with country music, while negative log-odds ratios are indicated as less characteristic for the country genre.

4.3 Vocabulary Richness

We analyzed vocabulary richness using three metrics to assess the lexical diversity and complexity of lyrics: Type-Token Ratio (TTR), Hapax Legomena Ratio, and Average Word Length. TTR measures the proportion of unique words to total words in a song, serving as an indicator for vocabulary diversity within individual lyrics. A higher TTR suggests a broader range of vocabulary used within a song relative to its length, which can be indicative of linguistic creativity or thematic complexity. Hapax Legomena Ratio quantifies the proportion of words that appeared only once within a song, a useful metric for lexical uniqueness and the detection of rare or context-specific terms that contribute to the distinctiveness of a song's narrative. Average Word Length serves as a proxy for linguistic complexity by measuring the mean length of words used in a song, with longer words typically suggesting a more sophisticated or formal tone. We computed these metrics for each song and genre-averages, allowing for comparisons across different musical styles.

4.4 Sentiment Analysis

We used sentiment analysis to explore the emotional tone present in country music lyrics and compare it to other genres. Specifically, we utilized the NRCLex library to classify lyrics into ten distinct emotional categories—joy, sadness, anger, fear, trust, surprise, disgust, anticipation, positivity, and negativity. We tokenized each song's lyrics and mapped individual words to their corresponding emotional categories using a predefined lexicon, and used NRCLex to assign weighted scores to words based on their frequency and relevance to each emotion. Words strongly associated with trust, such as "faith" or "family," receive higher trust scores in a song. We summed these individual word scores to create an emotional profile for each song, and aggregated them for all songs in the dataset to produce a comprehensive profile for each genre for comparison.

4.5 'Country-ness' Composite Method

We hoped to be able to integrate sentiment analysis, vocabulary richness, log-odds ratios, and topic modeling coherence scores into one metric reflective of the country genre: a 'country-ness' score. With this score, we aimed to rank the country songs in our dataset, to find which ones were the most 'country'. To achieve this, we attempted to normalize each metric and weigh them according to a qualitative assessment of their importance to defining country music. However, we encountered several technical difficulties combining the different metrics together, and ultimately chose to focus our effort elsewhere.

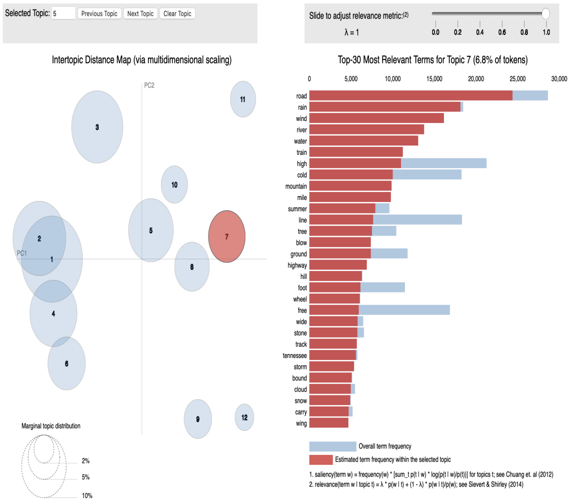
5 Results

5.1 Topic Modeling

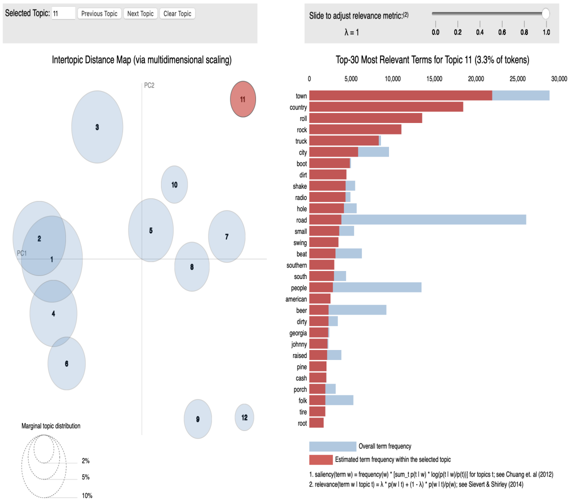
Our topic modeling analysis identified 12 unique themes in country music according to the LDA model. We validated these topics by generating a coherence score of 0.427, indicating a mix of meaningfulness and less coherent themes. Themes centered on love and relationships (e.g., “love,” “heart,” “baby”) and rural imagery (e.g., “backroad,” “rain,” “whiskey”) provided clear insights into the genre’s subject matter. However, some topics, such as those dominated by generic descriptors (e.g., “like,” “aint”), offered less thematic clarity, underscoring the challenges of capturing a genre’s nuances through automated methods. The corresponding distribution values for each topic indicate their prominence within our lyrics data, making it clear that some topics are far more consistently prominent in country music than others.

Topic	Distribution	Key Terms
1	0.164	like, dont, baby, little, good, aint, girl, want, right, know
2	0.041	light, night, shine, coming, star, christmas, burn, bright, waiting, bell
3	0.050	long, gone, song, sing, wrong, wish, play, miss, youre, lonely
4	0.066	lord, soul, heaven, river, water, hand, train, moon, child, jesus
5	0.030	rain, wind, cold, wild, deep, ground, young, blow, blood, stone
6	0.063	home, town, road, come, mama, roll, daddy, rock, truck, line
7	0.053	aint, drink, high, dance, night, round, nothin, whiskey, goin, beer
8	0.043	came, took, bring, seen, gave, lady, called, today, rose, fell
9	0.049	blue, life, ride, black, summer, white, gold, thank, wife, green
10	0.020	country, sweet, mountain, cowboy, city, hill, fuck, west, horse, climb
11	0.332	love, know, time, heart, away, dont, come, life, tell, thing
12	0.089	said, going, better, friend, hell, head, people, money, getting, told

The most meaningful themes we found featured religious undertones and depictions of nature and rural life. The visualizations we generated, such as intertopic distance maps, also illustrated certain overlaps in themes. Family-oriented and religious topics showed very strong interconnections, while topics like "road," "river," and "storm" pointed to country music’s deep connection to nature and adventure.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



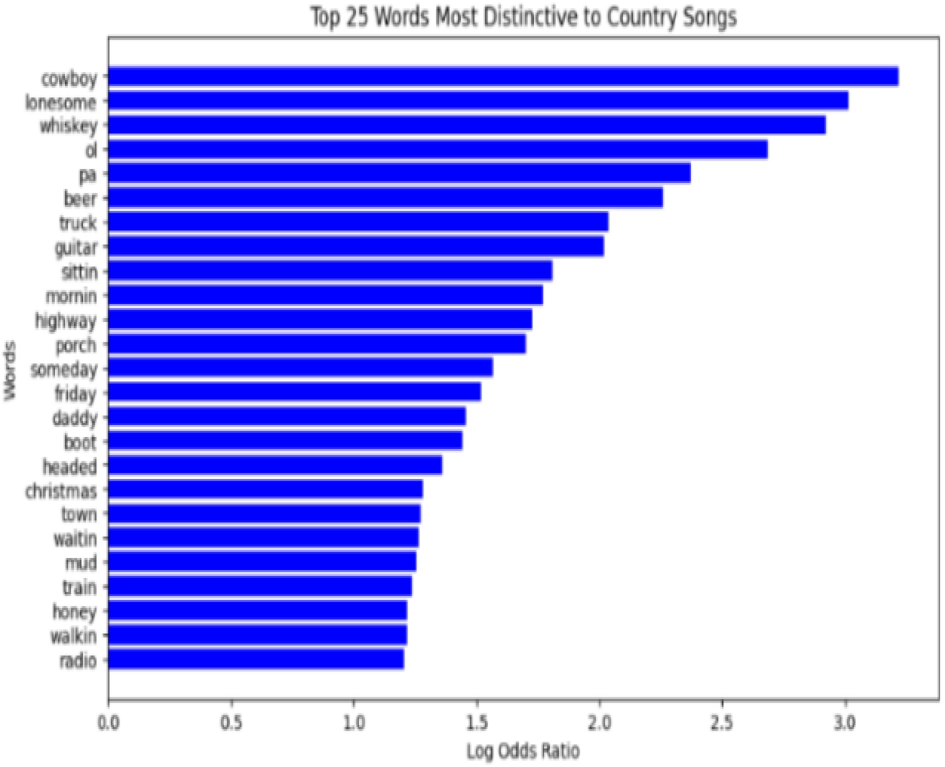
Selected Topic: 11

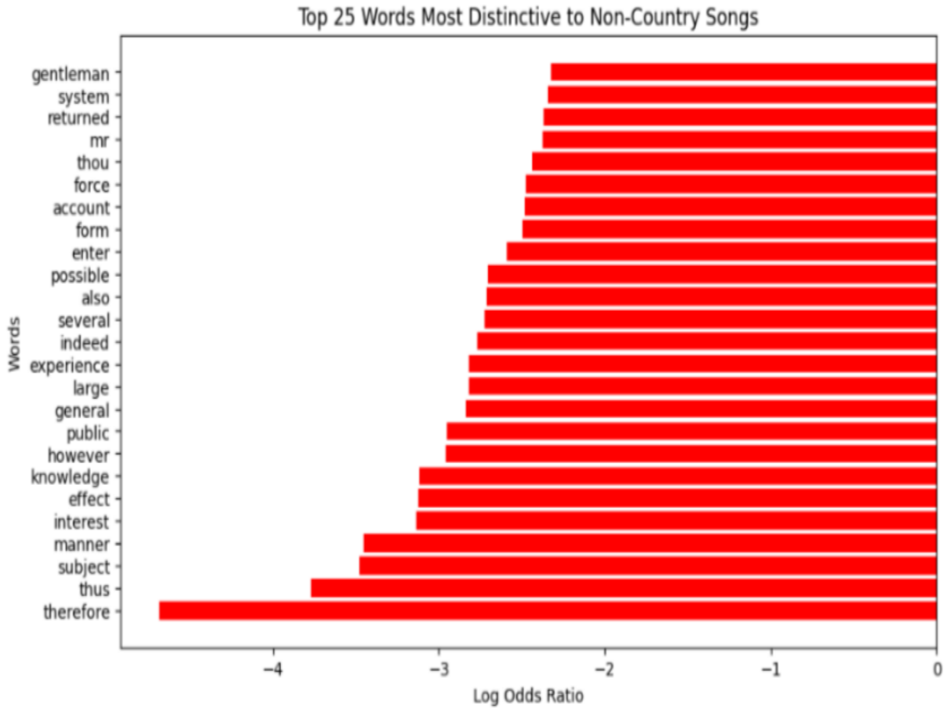
Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

5.2 Log-Odds

Our log-odds analysis highlighted terms that are especially characteristic of country music relative to other musical genres. These words are pretty much exactly as one would expect from country lyrics. We found that terms like "cowboy," "lonesome,", and “whiskey," were strongly associated with country songs, encapsulating the genre’s emphasis on Americana tropes and darker, melancholic emotional underpinnings. Interestingly, many of the terms most distinct to non country songs appear to be random and notably word, although it is difficult to understand why this is the case.



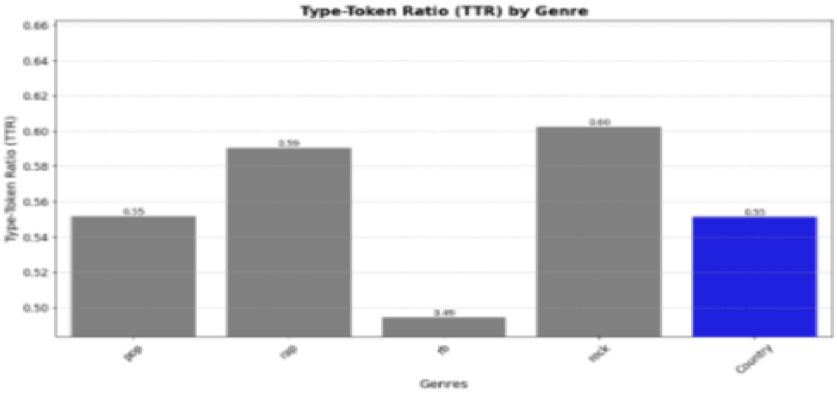


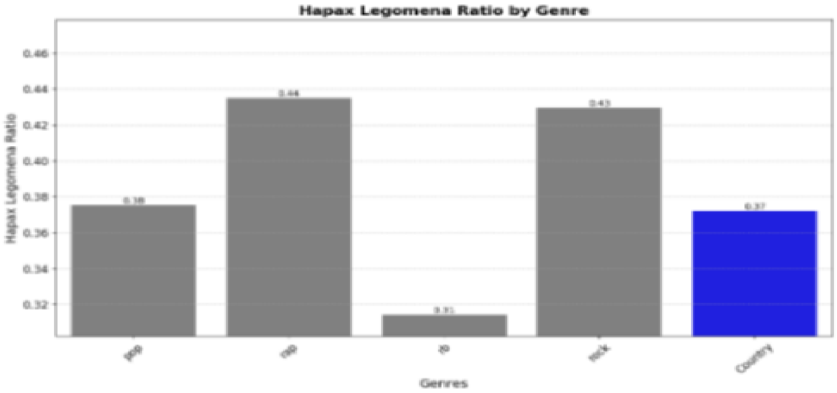
5.3 Vocabulary Richness

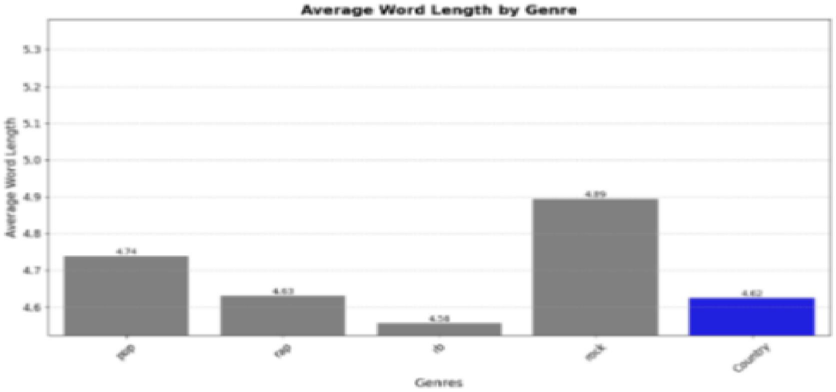
While rock and rap are marked by increased vocabulary diversity, country music still clings to relatively restrained Type-Token Ratios, lower Hapax Legomena Ratios, and average typical word lengths. These measures indicate a tendency in country music to restate key motifs to drive home its points, invoking a lexicon that is both basic and rich. The lower Hapax Legomena Ratios also suggests a lack of variation in the unique usage of words, which may point to a greater concern with story continuity than linguistic range.

Genre	TTR	Hapax Legomena	Avg Word Length
Pop	0.551766	0.375204	4.737637
Rap	0.590445	0.435096	4.630106
R&B	0.494333	0.314448	4.556044
Rock	0.602635	0.429460	4.893277
Country	0.551639	0.372071	4.624234

While the metrics provide a useful comparison, they do not account for stylistic elements or cultural significance of word choices. For instance, country music's lower TTR and Hapax Legomena ratios may result from its intentional use of repeated motifs and imagery to create emotional resonance and thematic coherence. So while these methods are informative, they should probably be contextualized with other kinds of insights.







5.4 Sentiment Analysis

Our sentiment analysis found ten quite distinct emotional categories that characterize country music.

Average Emotion Scores for Country Songs:

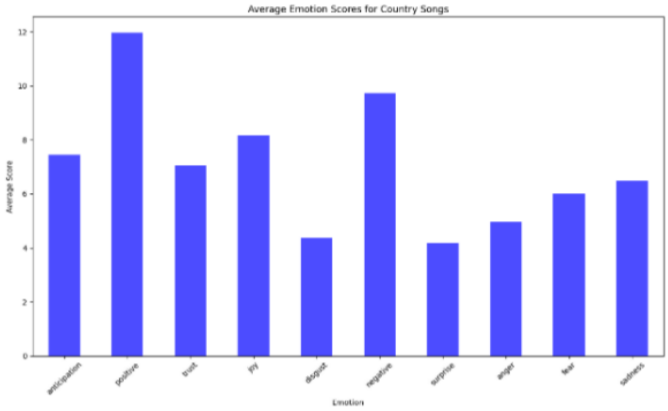
- Anticipation: 8.0
- Positive: 12.0
- Trust: 8.5
- Joy: 7.0
- Disgust: 5.0

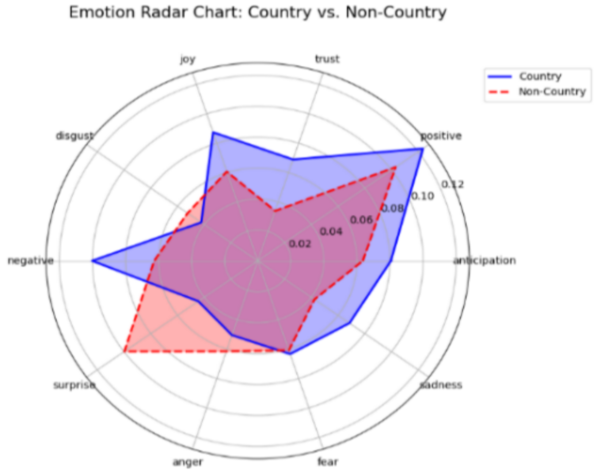
- Negative: 9.5
- Surprise: 6.0
- Anger: 4.5
- Fear: 7.0
- Sadness: 8.0

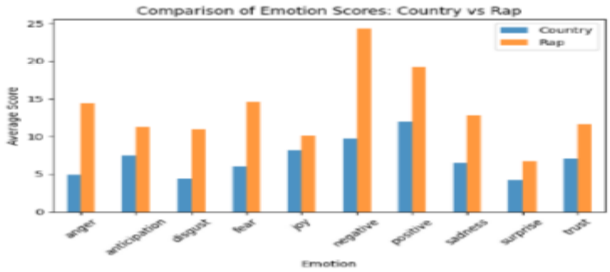
Relative Emotional Trends

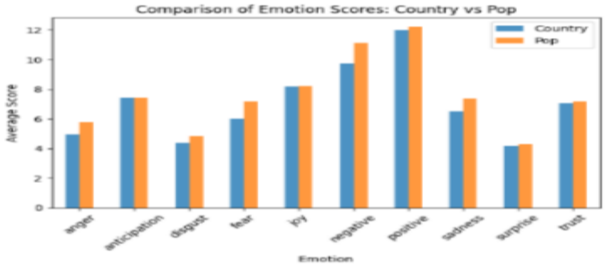
- Country vs. Pop: Higher "trust" and "joy" relative to pop, which leaned more heavily on "positive" sentiments overall.
- Country vs. Rap: Rap displayed higher "anger" and "fear," contrasting with country's more balanced emotional tone.
- Country vs. Rock: Rock exhibited stronger "negative" emotions, while country emphasized "trust."

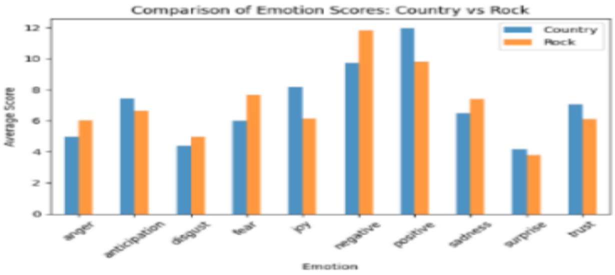
We found sentiment analysis to have some limitations, such as risks of oversimplifying subtle expressions to reductive dichotomies and failing to account for genre-specific context like dialect or irony. Moreover, subtle emotional language is de-emphasized in comparison with stronger language in other genres. Furthermore, a notable issue emerged: charts indicated that emotions prominent in country music also appeared more frequently in other genres. This raises questions about potential data or methodological biases, and as such it appears that the most important takeaways from these genre comparisons are the relative differences in scores rather than their exact scores. Nevertheless, our data uncovers useful insights in relative differences in emotions between genres.


Average Emotion Scores for Country Songs
Average Score
0
2
4
6
8
10
12
anticipation
positive
trust
joy
disgust
negative
surprise
anger
fear
sadness
Emotion


Emotion Radar Chart: Country vs. Non-Country
Country
Non-Country
joy
trust
positive
anticipation
sadness
fear
anger
surprise
negative
disgust
0.02
0.04
0.06
0.08
0.10
0.12


Comparison of Emotion Scores: Country vs Rap
Country
Rap
Average Score
0
5
10
15
20
25
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
Emotion


Comparison of Emotion Scores: Country vs Pop
Country
Pop
Average Score
0
2
4
6
8
10
12
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
Emotion


Comparison of Emotion Scores: Country vs Rock
Country
Rock
Average Score
0
2
4
6
8
10
12
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
Emotion

6 Discussion

Our results point to the themes of rural life, family, and emotional pragmatism as the most defining for country music. Topic modeling illuminated how these themes interweave into narratives, with rural imagery and emotional storytelling forming the backbone of the genre. Log-odds analysis further emphasized the unique linguistic markers of country music which evoke strong cultural and emotional associations. Vocabulary richness analysis highlighted how country lyrics achieve a balance between accessibility and depth. And sentiment analysis underscored the diverse and prominent emotions underpinning the genre, compared to others.

This project is situated within a broader genre of work in quantitative validation of qualitative observations around music lyrics. It extends the genre in its use of computational methods to identify and measure thematic and emotional patterns in country music, in particular. In particular, we verify patterns in country music's structural language and themes, bridging traditional musicology with computational analysis. These results underscore the utility of computational techniques in understanding genre identity and development.

However, this project does face particular limitations within the context of its design and dataset. Aggressive cleaning removed structural elements like verse-chorus patterns, which could have provided deeper insights into how song structure amplifies thematic and emotional content. Additionally, the dataset's lack of related subgenres, such as folk and blues, limited opportunities to explore country music's uniqueness within its broader musical family. These factors contextualize the findings and underscore areas where future projects could build upon this work by incorporating richer datasets or alternative methodologies.

7 Conclusion

This research presents a novel approach to quantifying "country-ness" in song lyrics, combining sentiment analysis, vocabulary richness metrics, log odds analysis, and topic modeling into a unified framework. The findings highlight country music's unique linguistic and emotional traits, offering valuable insights into the genre's cultural and artistic significance.

Future research should incorporate metadata to explore trends over time and examine regional variations within the genre. Additionally, integrating musical features such as tempo, melody, and instrumentation with lyrical analysis could provide a more holistic understanding of country music. Finally, the methods developed in this study could be adapted to analyze other genres, paving the way for broader applications in computational musicology.

8 References

Iain. (2016). *Heavy metal and natural language processing.* https://www.degeneratestate.org/posts/2016/Apr/20/heavy-metal-and-natural-language-processing-part-1/

Pudding. (2020). *Hip hop words.* https://pudding.cool/2017/09/hip-hop-words/

Tim, D. (2021). *What's in a song? lda topic modeling of over 120,000 lyrics.* https://tim-denzler.medium.com/whats-in-a-song-using-lda-to-find-topics-in-over-120-000-songs-53785767b692

9 Code and Dataset

Code: https://colab.research.google.com/drive/1EKD7TR-qGx07UzNW5HGYcbbETOh4XpgA
Dataset: https://huggingface.co/datasets/amishshah/song_lyrics