tAIylor's version: Using LLMs to Classify and Generate Taylor Swift Lyrics

Tilina Alzaben, Marion Haney, Divya Rao

December 11, 2023

1 Introduction

Taylor Swift stands out as one of the most popular songwriters in history, with an extensive discography and an even bigger fanbase known as Swifties. She currently has 10 unique studio albums ranging from 2006 to 2022, each representing a different era, or stage in her life. Her early albums started in the country genre, and later transitioning to pop. Common themes in her work feature lyrics about relationships, love, and relatable "edgier" topics which include loss of innocence, friendship, isolation, and family. Recently, Taylor Swift has embarked on a massive international tour, The Eras Tour, where she performs songs from every album and pays homage to her growth and transformation as an artist.

We are interested in the lexical and semantic differences between each Taylor Swift era. Our analysis focuses on 9 of her studio albums: "Taylor Swift", "Fearless," "Speak Now," "Red," "1989," "reputation," "Lover," "folklore," and "evermore". Given that there were measurable differences between each era, we then used a large language model (LLM) to classify lyrics by era and generate lyrics given a prompt.

In this study, we will investigate the following research questions:

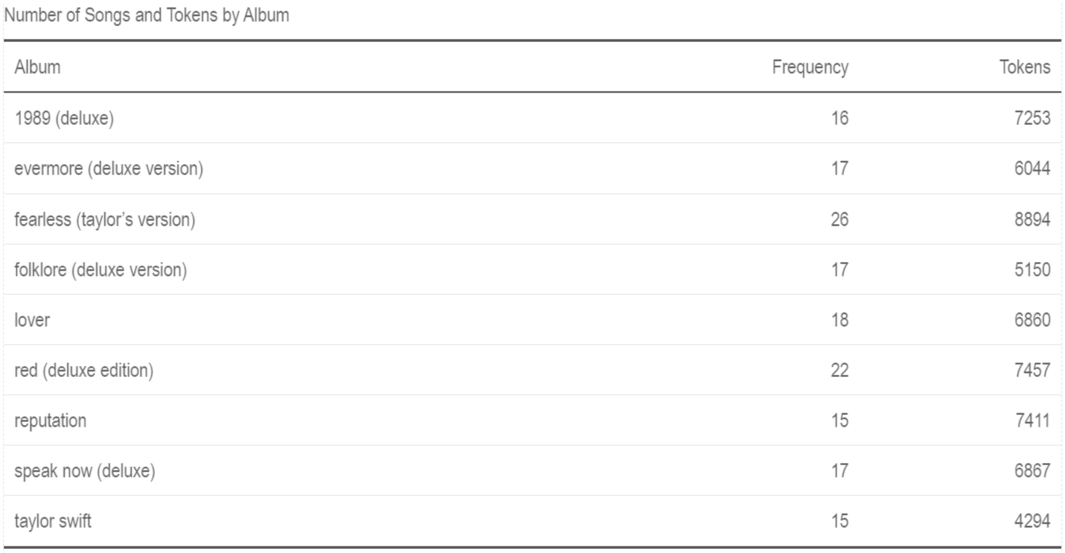
1. Using Biber's multidimensional analysis and sentiment analysis, how lexically similar are eras?
2. Given that there differences between eras, how well can a LLM classify lyrics from an era?
3. Can a LLM accurately capture Taylor Swift's writing when given lyric generation tasks?

2 Data

We compiled our dataset from Taylor Swift lyrics, sourced from Kaggle and extracted from the widely-used lyrics database, Genius.com. We have details about each song from her 9 albums, which includes the track number, title, and the lyrics, and are divided and numbered line by line.

A potential limitation of our data is the size of the documents within it. While Taylor Swift has one of the largest discographies of any modern artist, songs are inherently shorter than other types of documents. Using song lyrics as a context for lexical, semantic, and topical analysis may pose issues because of the length of each song. Also, song lyrics contain words such as "ooh" and "mm", which are not found often in writing and have different usage in speech. Another limitation of our data for the purposes of classification and generation is that we have a finite amount of training data. A larger amount of training data may be required for properly fine-tuning LLMs.

Overall, our corpus includes 9 albums, 163 songs, and 60230 tokens, where each album has between 4294 and 8894 tokens, as shown in the figure below.



Number of Songs and Tokens by Album

Album	Frequency	Tokens
1989 (deluxe)	16	7253
evermore (deluxe version)	17	6044
fearless (taylor's version)	26	8894
folklore (deluxe version)	17	5150
lover	18	6860
red (deluxe edition)	22	7457
reputation	15	7411
speak now (deluxe)	17	6867
taylor swift	15	4294

Figure 1: Albums, Number of Tracks, and Tokens

3 Biber's Multidimensional Analysis

Biber's Multidimensional Analysis (MDA) was used in our study to identify lexical similarities across Taylor Swift's different eras [Bib12]. This language analysis technique examines text in different ways, checking variables such as word usage, the formality of language, and the complexity level of the writing. Through Biber's tagging, we were able to create a stickplot, a graph that is used to visualize similarity between documents.

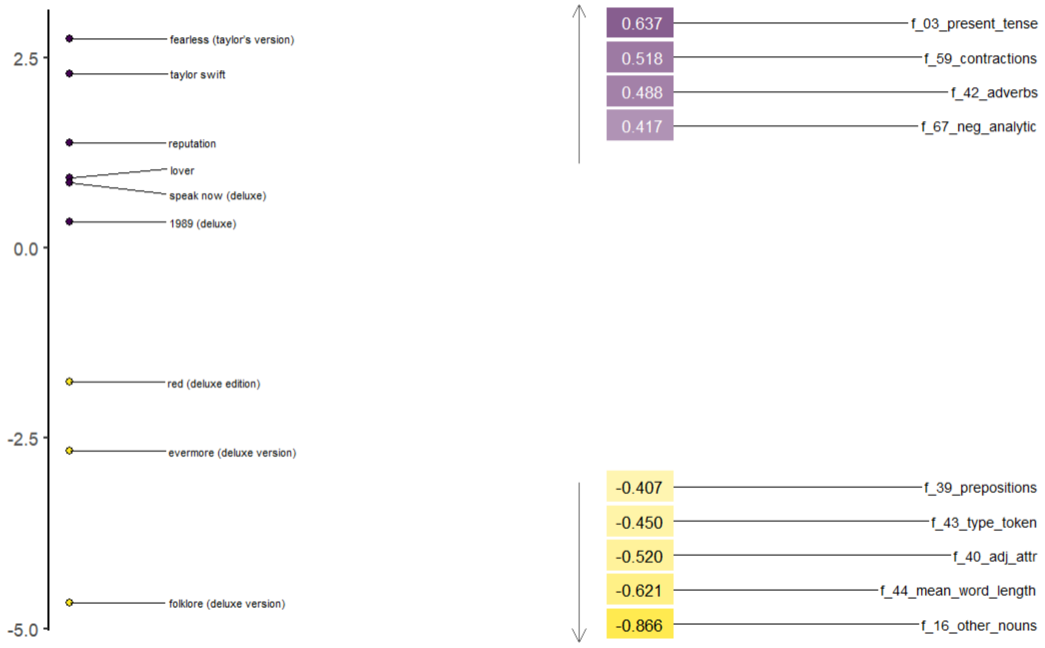


Figure 2: Stickplot of Lexical Features Across Taylor Swift's Eras

According to Figure 2, the stickplot helps us see the language differences in Taylor Swift's eras. In "Fearless" and "Taylor Swift," her oldest albums, she often uses present tense words, portraying events as if they are occurring in the current moment, suggesting she is singing about events happening in her life at that time. In "reputation," the stickplot shows more negative feelings in the lyrics, suggesting a change in the emotional tone, expressing more anger and disappointment. In the album ”evermore,” there are significantly fewer prepositions. On the other hand, in ”folklore,” there is a substantial drop in the frequency of words used, especially in the type-token ratio which suggests there is a wider range of words. ”folklore,” which is one of Taylor Swift’s newer albums, adds to the analysis that she uses a more diverse vocabulary in her later albums.

The stickplot confirms the general Swiftie (a name given to Taylor Swift fans) knowledge about similarities between albums. ”Taylor Swift” and ”Fearless” are positively similar to each other, which is expected because they are her first two albums within the country genre. Along the y-axis, ”reputation,” ”Lover,” and ”1989” are grouped together, which is also anticipated since these albums begin Taylor Swift’s transition into the pop genre. Surprisingly, ”Speak Now” is also in this group, even though it is considered one of her more country-inspired albums. Next is ”Red,” followed by ”evermore” and ”folklore”. We hypothesized that ”evermore” and ”folklore” would be lexically similar to each other since they are the two most recent albums, and fans have noted a more indie feel to them instead of pop. The unexpected finding was that ”Red” is more lexically similar to ”evermore” than ”evermore” is to ”folklore”. A reason why ”Red” and ”evermore” might be more similar could be that they both address similar themes or use similar words, even though we thought ”evermore” would be more like ”folklore” because they were released closer together. This shows that the language used in songs does not only depend on time period or genre of the album and that word choice might be influenced by shared themes or other artistic choices.

4 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), following Biber’s tagging, offers a distinct advantage over MDA in our study. While MDA allows us to explore various linguistic dimensions, LDA focuses on optimizing the separation between eras by reducing the dimensionality of our data. By doing so, LDA identifies linear combinations of lexical features that best capture the variance between different periods.

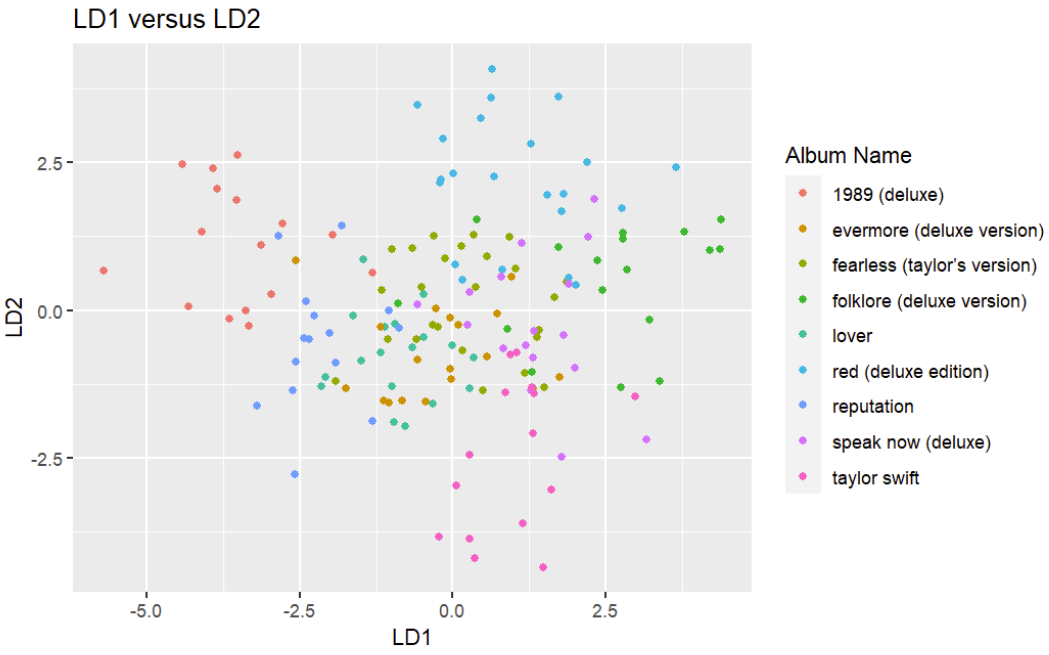


Figure 3: Scatterplot of LD1 versus LD2

The figure above gives an overview of the distribution of lexical features across Taylor Swift’s albums, revealing both distinct clusters and overlapping points which suggest shared linguistic characteristics. To dive deeper into these specific differences, we transition to a biplot which precisely identifies the lexical features responsible for the distinctions.

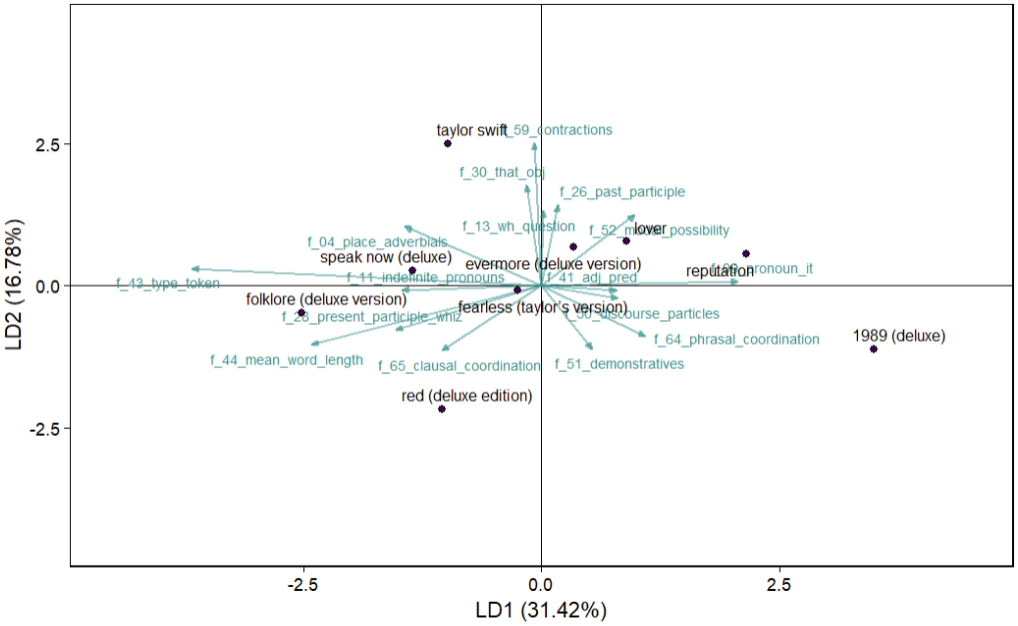


Figure 4: Biplot of Lexical Features Across Taylor Swift's Eras

Figure 4 displays similar information to the stickplot but provides a more detailed exploration of the relationships between each era and individual song. For example, "1989" contains high phrasal coordination and demonstration, suggesting an emphasis on clarity. In contrast, "Speak Now" relies on the use of adverbs, emphasizing detailed descriptions. "Taylor Swift" features a conversational tone with more contractions and questions, while Red uses complex sentences to make its structure more detailed. Notably, "Lover" and "reputation" stand out for their use of pronouns, reflecting a shift in the storytelling approach.

5 Sentiment Analysis

To understand the general sentiment within each of the Taylor Swift eras, we performed sentiment analysis. After partitioning the data by album, we grouped lines of the songs into sentences and calculated the sentiment for each chunk. In order to get the overall sentiment tone of the album, we took the average of the sentiment scores across each of the calculated segments.

Album	Sentiment Score
Taylor Swift	1.49
Fearless	2.22
Speak Now	-1.91
Red	1.06
1989	-3.64
reputation	-2.99
Lover	-3.30
folklore	-4.81
evermore	-1.54

Figure 5: Average Sentiment Scores of Taylor Swift Albums

From Figure 5, we see that some of her earlier albums, "Taylor Swift", "Fearless", and "Red" have positive or highly positive scores, indicating positive emotions or themes. In contrast, we see that "Speak Now", "1989", "reputation", "Lover", "folklore", and "evermore" all have strongly negative scores, suggesting more negative themes or emotional elements. It is interesting to note that the albums from "1989" (which was released in 2014) to her latest album, "evermore" (released in 2020), all have negative sentiment scores. This further suggests a shift in the emotional tone and themes within her music during this period.

In addition to understanding the general sentiment of an era, we also wanted to analyze how the sentiment has changed over time. From Figure 6, we see that the sentiment varies significantly across the different eras. We can see that over the years, Taylor Swift's music becomes more diverse, exhibiting a negative slope in sentiment with occasional increases.

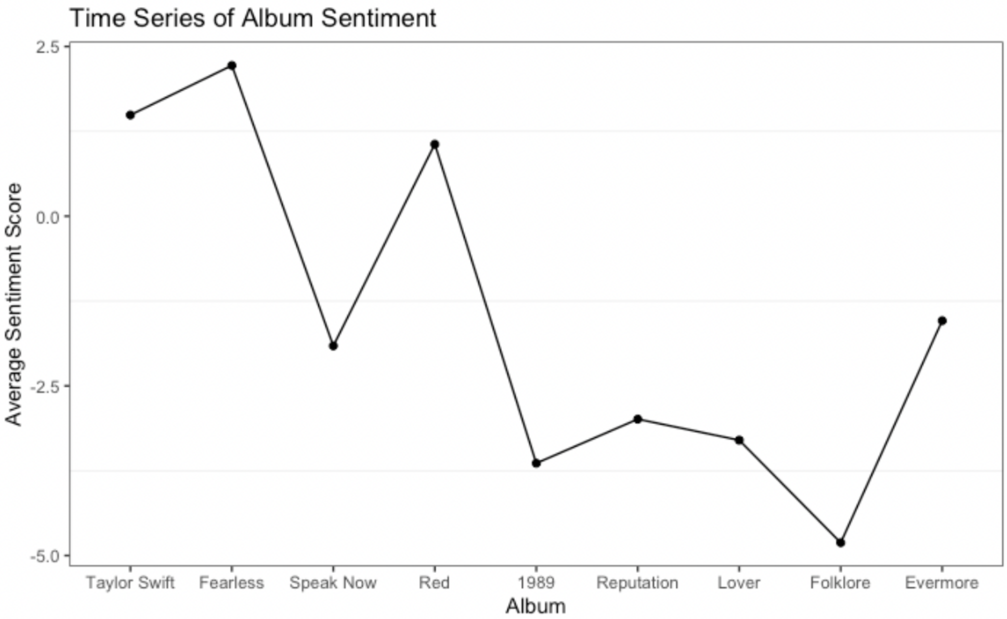


Figure 6: Time Series Across Taylor Swift's Eras Albums

6 BERT for Classification

Having proof of measurable lexical differences between Taylor Swift eras, our next step was to fine-tune a LLM for classification. BERT, or Bidirectional Encoder Representations from Transformers, is a language representation model proposed in 2018 [JD19]. Since its origin, it has proved to be a useful open-sourced tool for language generation and classification. We utilized the bert-base-cased model from HuggingFace, a BERT model that is able to take in text with capitalization [BER].

We prepared our fine-tuning data by formatting it into rows which contained the text of four or less lines of Taylor Swift lyrics and a corresponding integer representing the album from which the lyrics came from. We chose four as the number of lines in each training example because four lines in song lyrics generally equates to half of a stanza. Also, combining lyrics into sets of four or less lines aligns with our goals by allowing us to predict classes from a relatively small amount of textual information. We randomly split our data into training and test sets. 80% (1719 rows) of our data were used to train BERT, leaving 20% (430 rows) to test model performance. Before training, the data underwent a tokenization process using the corresponding bert-base-cased tokenizer, which was also provided by HuggingFace.

The fine-tuned model takes in a text prompt, ex. a few lines of lyrics, and outputs the integer representing the Taylor Swift album most likely to contain those lyrics. Our label mappings are as follows:

Given the task of classifying between 9 distinct labels, the model performed quite well, with an accuracy 0.72 and loss of 0.493. Key hyperparameters include the length of lyric sections, batch size, and the number of training epochs (the number of iterations it takes to pass the training data through the model). This accuracy and loss was achieved through training over 8 epochs, a batch size of four, and approximately four lines per training sample.

Furthermore, evaluation metrics like an evaluation loss of 1.351, an accuracy rate of 72.3%, a precision of 0.741, a recall of 0.723, and an F1 score of 0.724, highlight the model's performance and

Label Mappings	
Taylor Swift	0
Fearless	1
Speak Now	2
Red	3
1989	4
reputation	5
Lover	6
folklore	7
evermore	8

Figure 7: Label Mappings of Taylor Swift Albums

effectiveness. This demonstrates that the model not only successfully distinguished between the 9 distinct labels but also shows a high level of precision and recall in its predictions.

7 GPT2 for Generation: tAIylor's version

Next, we enlisted GPT-2, a Generative Pre-training Transformer created by OpenAI. OpenAI's GPT family, namely ChatGPT based on GPT-3 and GPT-4, rose to fame because of its ability to generate large sequences of comprehensive text [AR18]. In our tAIylor's version model, we fine-tuned a condensed version of GPT-2 containing 124 million parameters. For reference, the largest GPT-2 model on HuggingFace is GPT-XL, which contains 1.5 billion parameters [GPT].

To fine-tune GPT-2, we used data similar to the classification task. Key differences were that the generator-training data did not have a training and test split and each training example was 8 or less lines long instead of four or less lines. So, all of our available data– the 9 full albums – were used for fine-tuning. We used sections of 8 lines, which is generally the size of one stanza, to increase the model's ability to generate longer text. We chose not to have each training example be an entire song because this would decrease the amount of training data available and the flexibility of the model.

Using GPT-2's model and tokenizer from HuggingFace, we created tAIylor's version: a fine-tuned GPT-2 model on Taylor Swift Lyrics. tAIylor's version was trained using 8 epochs and a batch size of 4, achieving a training loss of 1.73. We were careful to avoid overfitting of the model, as an overfit generation model in this context would simply output existing Taylor Swift songs.

We tasked tAIylor's version with creating 15 songs to create an AI-generated Taylor Swift inspired album. The prompts used to generate each song were provided by ChatGPT to provide for consistent prompt structure. In an attempt to encourage creativity, we used top-k ($k = 50$) sampling in our generation pipeline as opposed to adjusting a temperature parameter. Top-k sampling involves selecting from the top-k most likely next words based on their probabilities, potentially leading to more diverse outputs. This approach contrasts with adjusting the temperature parameter, which modulates the randomness of the sampling process. For examples of prompts and AI-written songs, see Appendix I.

8 Analysis of tAIylor's version Album

In our experience with generating song lyrics, tAIylor's version was able to capture musical asides represented within parentheses and non-word vocalizations such as "la la la", "ah", and "oh oh oh". We noticed that the lyrics generated could be categorized within a few groups. The first group consists of lyrics which were grammatically correct but nonsensical in terms of content (they had nothing to do with the prompt passed in). The second includes lyrics that were heavily inspired by existing Taylor Swift lyrics and contained phrases from the chorus of select songs such as "Picture to Burn", "Look What You Made Me Do", and "cowboy like me". The last group consists of lyrics that impressed us by being both grammatically correct, sensible, and pertaining to a consistent theme or context. In particular, Track 12 of our AI-generated album was successful in relaying a story about falling in love with a Hollywood starlet. Also, tAIylor's version did not know how to properly end a song, likely because instead of passing the entire song as a training example we split it into sections of 8 lines. Our resulting generated tracks are each 366 words long, which is the mean number of words in the songs in our data.

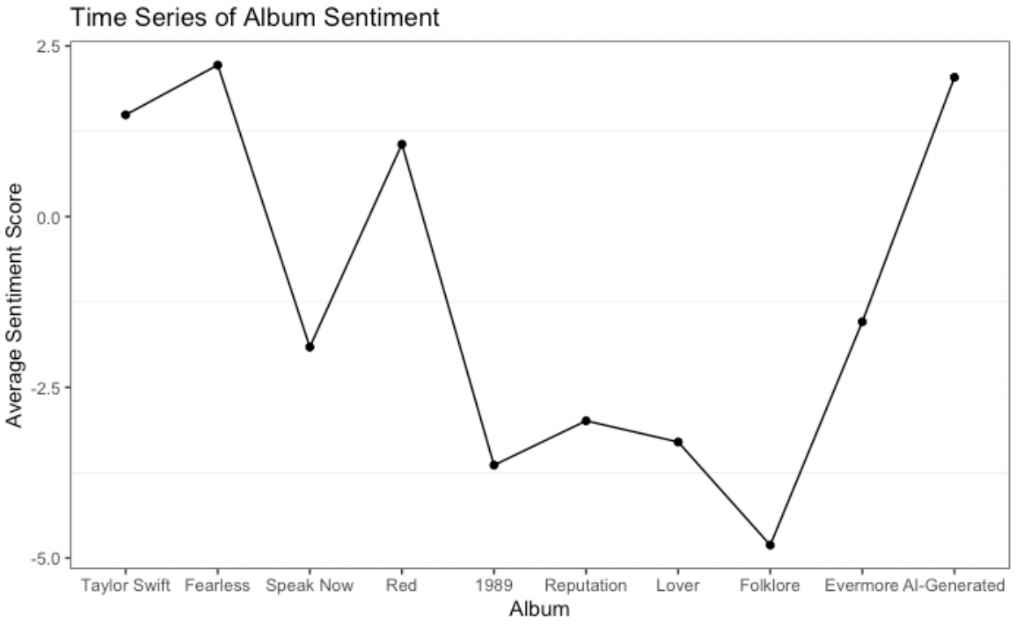


Figure 8: Time Series Across Taylor Swift's Eras Albums (with AI-Generated Album)

Following the sentiment analysis procedure from Section 4, we found that the AI-generated album has a sentiment score of 2.04, which indicates that the tone of the album is more positive. It was interesting to see that this score differed from the other albums in Figure 7. In terms of lexical features, the AI-generated album was the most similar to "reputation" and "Lover".

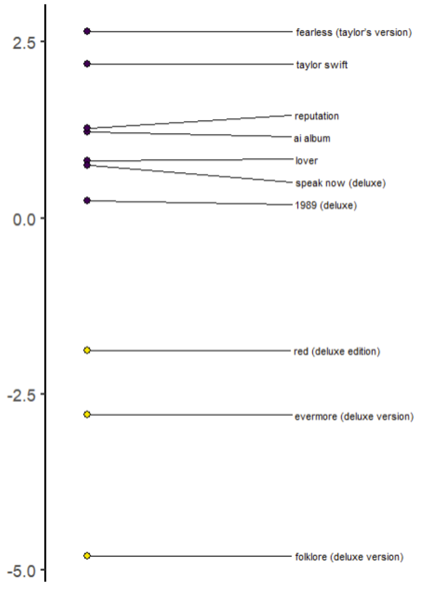


Figure 9: Stickplot of Lexical Features Across Taylor Swift's Eras (with AI-Generated Album)

9 Discussion

Throughout the generative tasks, we observed behavior that we believe is unique to generation for song lyrics. Taylor Swift generally uses similar themes or phrases so tAIylor's version often struggled to generate lyrics outside of those themes. In our training iterations, we noticed that an over-trained model would output verbatim Taylor Swift lyrics, line-by-line, rather than generate its own.

In addition to this, a higher accuracy for classification could be achieved by training with more epochs and a smaller batch size. Higher accuracy may also be achieved if larger sections of lyrics or a larger GPT-2 model were used. However, using larger sections of lyrics (as opposed to sections of 4 or 8 lines at a time) would require longer text prompts to receive an accurate classification. Utilizing a larger GPT-2 model would require a significantly greater amount of computational power and storage.

Future work could include expanding our data with the ”Midnights” album, Taylor's Version releases of ”1989”, ”Speak Now”, ”Red”, and future re-recordings of ”reputation” and ”Taylor Swift”. We could also apply similar methods to different artists with comparably sized discographies such as The Beatles, The Rolling Stones, Billy Joel, and ABBA or to other forms of creative expression like poetry.

In summary, our conclusion is that LLMs are able to capture the general voice of an artist given past works and can be helpful with creation or creative tasks. Also, LLMs are successful in classifying lyrics into categories such as which album they came from with a moderate to high level of accuracy.

10 Bibliography

References

[AR18] Rewon Child David Luan Dario Amodei Ilya Sutskever Alec Radford, Jeffrey Wu. Language models are unsupervised multitask learnerse understanding. 2018.

[BER] Hugging Face: BERT base model (cased).

[Bib12] Conrad Susan Reppen Randi Byrd Pat Helt Marie Biber, Douglas. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 31(1):9 – 48, 2012.

[GPT] Hugging Face: GPT-2.

[JD19] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2019.

11 Appendix

11.1 I. Generated Songs and Prompts

AI Album: Track 12

Prompt: Craft a country song about a lone traveler navigating city streets. Capture the anonymity, solitude, and hidden stories whispered by the urban landscape with a country twist.

Wind in my hair, I moved from New York to LA
This was a long and hard road
We were on the phone when we said, ”Hello”
The conversation turned to our hometown problems
And I pointed out where we stood in the fight for love
And she called back and said, ”Just give me a day and think it over”
But you would've been so kind
Had I known about the fate of you when we first met?
You had a pretty girl crush, but that girl's reputation was rosy
And your reputation was revving up your wild side
and you were the superstar on the road to glory
With a wide wide-eyed girl crush on a big name brand name

She was making movies and your body was getting better
And you were the superstar on the road to glory
And your reputation was glowing like the stars that flashed in the dim light
And you were the superstar on the road to glory
Oh, darling, don't you let it go
You made one of those films so that women might see what you do
And she was a star, now her name will forever hang around
In pictures like a Hollywood star who never leaves the scene
You were the superstar on the road to glory (Oh)
You were the superstar on the road to glory
She was a star, now her name will forever hang around (Oh)
You just keep on winning like you did before
You're the superstar on the road to glory
You're the superstar back in the door
Dancing with the crowds, dancing like you did before
You're the superstar on the road to glory
I'm a cowboy like you were back when we were young (Oh)
I'm the cowboy like you were back when we were young (Mmm)
I'm the cowboy like you were back when we were young (Back when we were young)
Back when

AI Album: Track 10

Prompt: Immerse yourself in the whimsical world of childhood memories through pop lyrics. Capture the innocence, laughter, and imagination of youth, creating a nostalgic picture of carefree days.

And bring it all back to me
And save the day by taking your time
I think you should, think you should be celebrated
Think you should be celebrated
like the Champagne fountains would be celebrated
Would you like my Champagne fix?
I'd like it in your face, oh darling
Let me know if you'll be taking it anywhere
But don't ask me to do things for you, just let me know
I'm not a princess and you're not my queen
Would you like my Champagne fix? Would you like my Champagne fix?
I'd like it in your face, oh darling
Girl, just let me know I'm not a princess and you're not my queen
I'm not a princess and you're not my queen, oh darling
I'm not the kind of princess I'd want to lose, oh darling
Would you like my Champagne fix?
Oh darling, let me know that you'll be a part of my story
I'm not a princess and you're not my queen
It was a bad summer, but I'd be Miss Misery again
And that was the ending I was looking for
So don't ask me to do things for you, just let me know
I'm not a princess and you're not my queen
I'm not the kind of princess I'd want to lose, oh darling
Can I just keep you company?
The ladies' room at home
I know our secrets about each other, I know that your laughter will impress me
The cascade of lights that blind us is a cruel mistress
Are you ready for it? I'm a hostage now
'Cause she wasn't the one I wanted
But I'm fine with it, honey, I think you should
Give me a reason
Why I'm not a princess and you're not my queen

If you could stay

AI Album: Track 13

Prompt: Write pop verses that delve into the complexity of human connections. Explore the threads that bind friendships together and the sadness and feelings of loss when friendships fade away.

And bring this love to yourself
If you're not familiar with the sound of it pouring out of your windows
These walls of stone are raze now
And life has given up on you
It's like a million little stars spelling out your name
They tell us that things will change, but we never take them very far
This slope is treacherous
This path is reckless
This path is reckless (Are you okay?)
This slope is treacherous (Are you okay?)
This slope is treacherous (Are you okay?)
Touch my face, will you please slow it?
Will you please slow it? (Would you please slow it?)
Worry not, my love
Will you please slow it?
It rains in your bedroom (Does it have to be that bad?)
Does it have to be that bad?
Rain in your bedroom (Does it have to be that bad?)
Do you remember, all along, that I had this love for you?
This love was illicit, illicit, illicit, illicit, illicit
(Does it have to be that bad?)
So it's thrown the spotlight back on me, spotlight on me, spotlight on me (Yeah)
So it's thrown the spotlight back on me, spotlight on me, spotlight on me (Yeah)
And the spotlight goes back on you, spotlight on you (Yeah)
So it will, it will, it will, it will (Ah)
And I will, it will, it will (Yeah)
It will be forever, oh, oh (Yeah)
It will be forever, oh, oh (Oh)
Ah, it will be, oh, oh (Yeah, oh)
Oh, oh, oh (Ooh, oh, oh, oh)
The way you whisper in the hallway (Silence)
'Cause you were a part of my past (Silence)
You were a part of