NLP Analysis of Environmental Themes in Phish Lyrics Across Concert Locations and Years

Anna Farzindar and Jason Jarvis
Loyola Marymount University
1 LMU Drive Los Angeles, California, 90045
anna.farzindar@lmu.edu , jason.jarvis@lmu.edu

Abstract

This work studies the application of advanced AI and natural language processing (NLP) techniques, to analyze the lyrics of Phish, a renowned American jam band known for their groundbreaking improvisational live shows and eclectic lyrics. Focusing on environmental themes within their extensive repertoire, this paper aims to uncover latent topics pertaining to environmental discourse, by using the topic modeling and environmental classifier to filter out the list of topics present within their songs. Through meticulous preprocessing, modeling, and interpretation, our findings shed light on the multifaceted portrayal of environmental issues in Phish's lyrics. In this study, our primary contribution lies in lyrical analysis, as well as visualization and interpretation of the topics their lyrics cover, over the forty plus years the band has existed. Our lyrical visualizations aim to facilitate an understanding of how Phish selects the timing and location for their live performances in relation to the themes present in their music.

1 Introduction

As the planet plunges headlong into 1.5 degree Celsius warming, public support for protecting the biosphere and altering our relationship with the non-human world is critical. To this end, art and music can play a significant role in raising public awareness. We contend that Phish is an important band leading this charge in the United States. Phish, a jam band renowned for their psychedelic rock performances, has connected with their fans over the years through their distinctive approach to improvisation.

Phish is an iconic act in the "jam band" genre of music that has its roots in bebop, bluegrass and rock. Jam bands have several critical characteristics: (1) they allow live recording and sharing of their concerts by fans, (2) they engage in extensive improvisation during live concerts and (3) concert setlists that vary from show to show. Consequently, a jam band never plays the same show twice.

Phish is one of the most significant bands in America despite having almost no radio airplay or music industry awards over their 41-year history. Nonetheless, in 2023 they were the 10th highest grossing rock music touring act in America: selling 597,000 tickets across 41 shows with a total gross of $76.8 million dollars (Frankenberg, 2023). For context, ninth place went to the Red Hot Chili Peppers while Coldplay and Elton John were the first and second respectively.

One of the things Phish is known for is their regular use of environmental themes. This was evident at Phish's 2024 residency in the Las Vegas Sphere. Over the course of the four night stand, Phish organized their shows around a progression of environmental themes "from solid to liquid to gas to plasma" (Renner Brown, 2024). Only the second band to play the venue after U2, the concerts were a groundbreaking success, famously causing attendee (and host of The Price is Right) Drew Carey to say that Phish's shows made U2 "look like a bar band" and that he "wanted to call U2 and get my money back" (Simpson, 2024).

Also on April 23, 2022 (Earth Day), the band performed what is now a legendary show at Madison Square Garden built around the theme of water that included whale and dolphin drones circling the arena while bathed in blue light that simulated the ocean with kelp descending from the ceiling as the band played. Phish does more than just create images of nature in their songs. Notably, Phish's non-profit organization, The Waterwheel Foundation specifically lists environmental issues as a key area that donations support[1].

In 2023, the band held two benefit concerts in New York raising $3.5 million for people suffering due to the extensive flooding experienced by

[1] https://www.waterwheelfoundation.org

residents of New York and Vermont (wat, 2023). Donations are still being taken for this project as of the writing of this essay. In this research, we analyze the album titles, song titles and lyrics of Phish to consider their support for environmental protection and preservation. We believe that Phish rhetorically advocates for environmental protection in a range of ways that are both obvious and subtle. To test this theory, we use NLP techniques. We explored the application of topic modeling and machine learning technique, to dissect and interpret the environmental discourse embedded within Phish's lyrical compositions. Our research study tries to focus on the environmental topics of Phish, as well as to understand the popularity of these themes over time and location. To our knowledge, this study marks the first application of AI and NLP techniques for processing this lyrical information. Our objective is to offer comprehensive statistical insights into the presence of environmental themes in Phish's live performances, considering the dates and locations of their concerts. Our research is based on the utilization of NLP algorithms made feasible through data compiled from the Phish.net[2] lyrics collection created by the laborious work of Dr. Ellis Goddard, Associate Professor of Sociology at California State University, Northridge.

2 State of the Art

Several studies have explored Phish's relationship with their community and culture, covering topics such as public participation, copyright law, engagement, gender and racial diversity among fans, cultural practices of fans and music therapies (Carlson, 2020), (Kushner, 2020), (Marshall, 2003),(McClain, 2016), and (Rothstein, 2023). Additionally, ongoing research was gathered and discussed at the Phish Studies Conference, in 2019 and 2024 (Farzindar et al., 2024), organized by Oregon State University.

Our study focuses on Phish lyrics and song titles and includes both original Phish songs as well as cover songs written and performed by other bands. Phish regularly performs cover songs and the decision to include them is based on the fact that covers are a choice by the band. Consequently, they represent the musical and lyrical interests of the band as they perform unique songs by other artists.

Since lyrical texts consist of a sequence of words, multiple NLP methods could be applied to it. However, NLP methods are not always as effective for lyrics due to the differing nature of lyrics compared to traditional texts. These challenges have several reasons, such as creativity, ambiguity, variability, and emotional depth in texts. The text of lyrics often contains non-standard language, creative expressions, slang, and poetic devices. This makes it harder for NLP models, which are usually trained on more formal, standardized text, to interpret the meaning accurately. The presence of ambiguity, figurative language and metaphors are bold in lyrics, making it difficult for NLP models to accurately process the intended meaning. Lack of context and using very short sentences in lyrics are another leading factors in correct language processing and interpretation of emotional nuances.

Several studies explore themes in songs, often using LDA, a probabilistic topic modeling method (Liew et al., 2020). More recent work utilizes the Bidirectional Encoder Representations from Transformers (BERT), a pre-trained deep learning model by Google that generates word or sentence embeddings, capturing contextual and semantic meaning. Specific BERT models, such as MusicBERT, are tailored for NLP tasks involving both text and music (Rossetto and Dalton, 2020).

3 Methodology

In this research, we employ the topic modeling technique for lyrical information analysis. The goal is to identify clusters of words that frequently co-occur, aiming to represent topics related to the environment within the text and visualize them. For this purpose, we utilize topic modeling to classify Phish album titles, song titles, and lyrics as either "Environmental" or "Non-Environmental." Data classified as "Environmental" was subjected to further analysis, including mapping the time and location of live performances. The steps undertaken in this study consists of Web Scraping and Preprocessing, Topic Modeling, Environmental Classification, Evaluation and Visualizing. We utilized BERTopic, which captures nuanced semantic relationships between words and documents, leading to more accurate and interoperable topic clusters.

3.1 Web Scraping and Preprocessing

3.1.1 Web Scraping

Song lyrics and live performance data were scraped from Phish.net, providing a comprehensive dataset

[2] phish.net https://phish.net/

for analysis. Phish.net is an online community for fans of the Phish band, offering set lists, show reviews, and analysis of their music and performances.

Phish took a two-year hiatus starting in October 2000, resuming performances in December 2002, only to disband again in August 2004. This resulted in a gap in their concert data until their official reunion in March 2009, following an announcement in October 2008. Consequently, the dataset for Phish concerts is empty for the years 2005 through early 2009.

From 1980 to March 2024, a total of 1052 songs, with lyrics and performance data, was scraped from the website.

3.1.2 Exclusion of Instrumental Tracks

As noted earlier, our corpus includes both original Phish songs and songs Phish covers at their concerts. However, only songs with lyrics are included in the analysis, omitting instrumental tracks to focus solely on lyrical content. Instrumental tracks are songs that do not have any verbal lyrics and are completely consisting of instrumental music. After filtering out instrumental songs in our dataset, we were left with a total of 645 songs that contained full lyrics.

3.1.3 Tokenization and Data Processing

Data preprocessing included sentence tokenization and stop-word removal to ensure the quality and consistency of textual data. In the sentence tokenization process, each line of lyrics is segmented into individual sentences, breaking down the text into smaller units for analysis. This allows the BERTopic model, used in the next stage, to understand the context and meaning of each sentence independently, facilitating the clustering of similar topics or themes within the lyrics.

3.2 Topic Modeling

3.2.1 Embedding Generation

Embedding refers to the vector representations of words or sentences generated by pre-trained transformer models like BERT. These embeddings capture semantic meaning and context, allowing models such as BERTopic to analyze similarities between words or sentences based on their vector representations. By leveraging embeddings, BERTopic can cluster text data effectively, identifying topics or themes within the corpus.

In our study, the input for these models is the set of song lyrics, and output is the embeddings needed for the BERTopic models. The most popular models of Sentence Transformers for the English language were utilized to generate embeddings for the lyrics[3], namely model all- MiniLM-L6-v2 and model all-mpnet-base-v2.

3.2.2 Topic Modeling

BERTopic is a topic modeling tool that utilizes BERT embeddings to cluster documents based on their semantic similarity. Unlike traditional topic modeling methods like Latent Dirichlet Allocation (LDA), BERTopic captures nuanced semantic relationships between words and documents, resulting in more accurate and interpretable topic clusters.

The BERTopic model generated an output as a list of topics, with each list containing the key words highlighted within that topic. Each topic has its associated list of documents that represents that topic. In this study, we obtained a list of 34 topics generated from BERTopic model over the dataset of Phish lyrics.

3.3 Environmental Classification

An environmental classifier was employed to filter out topics related to environmental subjects from the total list of topics generated from BERTopic. The classifier used was: ESGBERT/EnvRoBERTa-environmental[4]. After classification, three topics were identified as containing environmental subjects from the list of 34 topics. We labeled the automatically selected topics from the topic modeling module, as described in the previous section, as **Water**, **Planets** and **Living things**. These topics include the following concepts:

- **Water**: rolls, away, water, sea, bouncing, flowing, wind, sky, room, light

- **Planets**: planet, slippin, flip, way, time, space, oh, world, soul, easy

- **Living things**: bug, hear, living, wind, quiet, sound, ringing, peeping, pane, frustration

Out of a total of 645 lyrics, 200 songs, making up approximately 31% of the corpus, were classified

[3] SentenceTransformers in Python framework for sentence, text and image embeddings https://huggingface.co/sentence-transformers?sort_models=downloads#models

[4] ESGBERT/EnvRoBERTa-environmental https://huggingface.co/ESGBERT/EnvRoBERTa-environmental

as belonging to environmental topics. This significant percentage highlights Phish's substantial engagement with environmental discourse throughout their music catalog.

3.4 Evaluation of models

3.4.1 Evaluation of topic modeling

For evaluating our topic modeling modules and select the best models, we used the C_V coherence score. The score is between $0 < x < 1$ and a higher score indicates that the top words in the topic frequently appear in similar contexts, suggesting that the topic is coherent and meaningful. Topic modeling for lyrics is challenging due to the poetic and metaphorical language used, which often conveys abstract themes rather than concrete topics, and words in lyrics can have multiple meanings or shift dramatically in context. However, in this study, a coherence score of C_V equal to 0.46 was obtained, indicating medium coherence. To further assess the medium coherence, we manually examined the three selected topics concerning environmental topics and concluded that this performance was sufficient to meet our objectives.

3.4.2 Evaluation of classifier

The automatic binary classifier analyzed 645 songs, classifying 194 as environmental. For evaluating the performance of classifiers, we needed a labeled dataset to check the precision of the machine's output, but we did not have any annotated data. For this purpose, we manually labeled a random sample of 143 lyrics as environmental and non-environmental. In addition to the classification task, we consider the confidence level of the classifier, which indicates its certainty about the predictions it makes.

In this study, considering an 80% confidence level in the classifier's labeling as environmental, the precision is 0.634 and recall is 0.866.

3.5 Visualization and Interpretation

To demonstrate the result of topic modeling techniques and automatic classification of topics, we develop an interactive visualization showing the distribution of topics related to the environment in Phish lyrics.

In our analysis of the Phish datasets, we employed several visualization techniques to gain insights. One approach involved mapping the time and location of the environmental songs' performed, with a specific focus on North America.

Additionally, we utilized statistics to track the occurrences of environmental songs over time and location. The Fig 1 revealed the percentage of environmental songs were shared with audiences across various cities hosted concert venues.

Furthermore, we calculated the trends of environmental themes in Phish live performances. Trends were calculated using a specific formula designed to measure changes relative to a baseline. Specifically, the trend for each value is calculated as follows:

$$\left(\frac{\text{Current value} - \text{Reference value}}{\text{Reference value}}\right) \times 100\%$$

This formula expresses the change as a percentage of the reference value, providing a standardized way to understand shifts over time or between datasets. In this analysis, the year 2019 was chosen as the reference value for trends over the years. The choice of 2019 as the baseline is strategic; it serves as a solid reference point because it is the most recent complete year before the disruptions caused by the COVID-19 pandemic. By using 2019 as the reference, it ensures that the data compared is from a period of relative normalcy, thereby providing a clear picture of how metrics have evolved from a pre-pandemic standpoint to the present. Fig 2 summarizes a comprehensive picture and confirms that the band has been increasingly active in promoting and addressing environmental topics through their concerts over years.

4 Conclusion

The analysis of Phish's lyrics using advanced natural language processing techniques reveals a strong and consistent presence of environmental themes in the band's music. By leveraging topic modeling, we identified clusters of lyrics focused on elements such as water, planets, and living things, demonstrating Phish's engagement with environmental discourse over time and across various locations. Our findings indicate that a significant percentage of Phish's lyrical content relates to environmental topics, highlighting the band's commitment to raising awareness through their music. This study provides a novel contribution to both musicology and environmental studies by using AI-driven techniques to quantify and visualize the influence of environmental themes in live performances. Moreover, the alignment between the band's concert locations and thematic content suggests that Phish

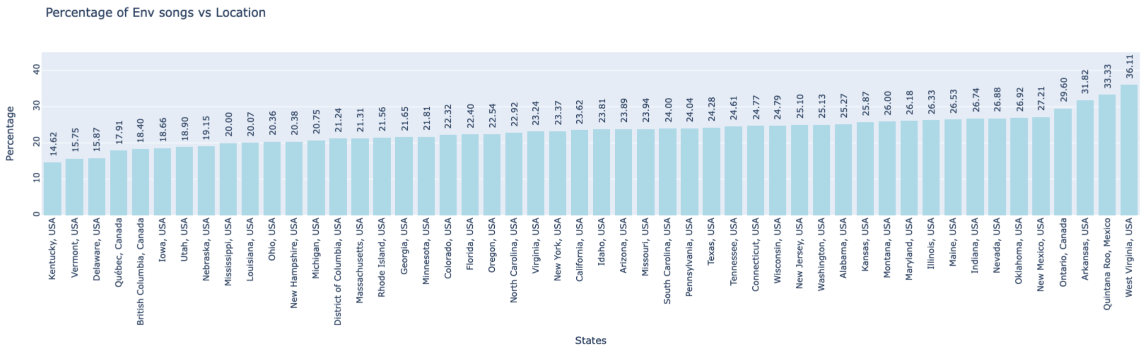

Figure 1: Percentage of Environmental songs vs location

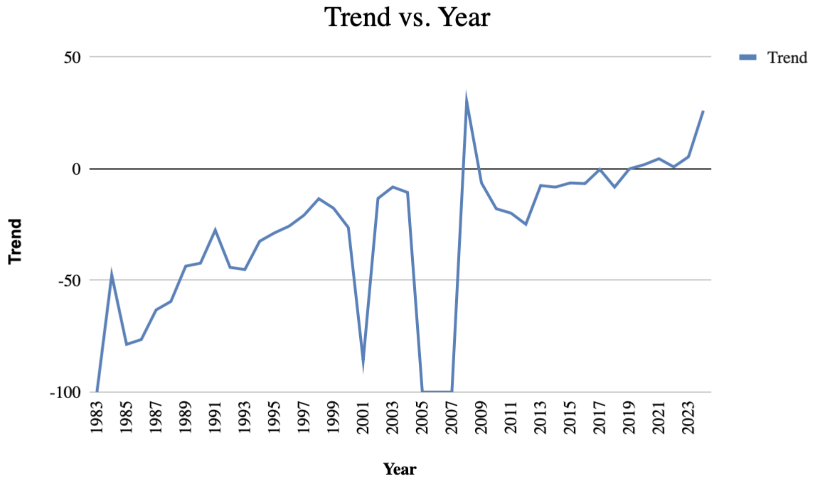

Figure 2: Trends of environmental themes in Phish live performances over Years

strategically incorporates environmental advocacy into their performances, further solidifying their role as a cultural force for environmental awareness. Future research could expand this framework to explore other thematic elements in their lyrics or apply similar methods to other artists and genres. The adaptability and scalability of these NLP techniques make them valuable tools for any study aimed at understanding the intersection of art, culture, and societal issues. As demonstrated in the Phish case study, the use of topic modeling, classification, and visualization can uncover latent themes in artistic works, providing insights into how artists communicate with their audiences and contribute to broader cultural movements.

References

2023. The WaterWheel Foundation - Phish concerts raise more than $3.5 million for flood relief in Vermont and New York.

Dennis Carlson. 2020. *A History of Progressive Music and Youth Culture: Phishing in America*, 1st edition edition. Peter Lang Inc., International Academic Publishers.

Anna Farzindar, Jason Jarvis, and Deepak Jayan. 2024. Divided sky: The environmental rhetoric of phish. In *Phish Studies Conference 2024*, Oregon State University, Corvallis, Oregon.

Eric Frankenberg. 2023. Top 10 Highest-Grossing Rock Tours of 2023.

Scott Kushner. 2020. Collecting and media change, or: Listening to Phish via app. *Convergence*, 26(4):969–989. Publisher: SAGE Publications Ltd.

Kongmeng Liew, Yukiko Uchida, Nao Maeura, and Eiji Aramaki. 2020. Classification of nostalgic music through LDA topic modeling and sentiment analysis of YouTube comments in Japanese songs. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 78–82, Online. Association for Computational Linguistics.

Lee Marshall. 2003. For and Against the Record Industry: an Introduction to Bootleg Collectors and Tape Traders. pages 57–72.

Jordan M. McClain. 2016. Framing in Music Journalism: Making Sense of Phish's "Left-Field Success Story". *The Journal of Popular Culture*, 49(6):1206–1223. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpcu.12489.

Eric Renner Brown. 2024. Phish Sphere Review: Jam Band Masters Four-Show Residency.

Federico Rossetto and Jeff Dalton. 2020. MusicBERT - learning multi-modal representations for music and text. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 64–66, Online. Association for Computational Linguistics.

Caroline Rothstein. 2023. *I've Been Wading in the Whitest Sea: REFLECTIONS ON RACE, JUDAISM, AND PHISH*. Penn State Press. Google-Books-ID: bTrUEAAAQBAJ.

Michael Lee Simpson. 2024. Drew Carey Says Seeing Phish at Sphere Made U2 'Look Like a Bar Band'.