Phonemic Analysis of Bangla Lyrics

RokunuzJahan Rudro
University of Kansas
rudro12356@gmail.com

Ishrak Hayet
North Carolina State University
ihayet@ncsu.edu

Abstract

Bangla music is a treasure trove of cultural heritage that has been prevalent and thriving for more than a century. This paper presents a new Bangla music dataset with unique features that reflect the thematic, phonemic and stylometric evolution of Bangla music from the 20th to the 21st century. The dataset is accompanied by a thorough exploratory analysis to unfold the ever evolving elements of Bangla music from a temporal and lyrical perspective. Additionally, we show that our dataset is a good fit for various classification tasks using deep neural classifiers. We have strategically fine-tuned the BanglaBERT model to achieve an average accuracy of 60% for various phonemic classification and artist identification from the lyrics.

1 Introduction

Despite being a low-resource language, Bangla boasts linguistic diversity that is rich in artistic and cultural significance. This research breaks new ground by conducting a cross-centurial phonemic analysis of Bangla music lyrics, going beyond transcription to uncover hidden gems of rhyme, alliteration, and reduplication. To support our findings, we introduce innovative datasets that capture these phonemic nuances, enhancing our comprehension of Bangla song lyrics.

Existing datasets on Bangla music lyrics like (Khan, 2020) and (Hossain and Al Marouf, 2018) include a single genre per song. However, we found that approximately 21% of the songs in our dataset can belong to multiple genres. As we have provided multi-genre labeling for each song, our dataset has the potential to be used for multi-label genre classification tasks. Additionally, we derive various phonemic features like alliteration, reduplication, and rhyming patterns from the song lyrics which we present as secondary datasets in the paper. We utilize these phonemic features to build classifiers that can recognize these features and learn about this features on a language modeling level. Unlike any previous dataset, we include a "Timeline" feature that represents if the song is from before or after the year 2000.

The addition of these unique and new features enabled us to carry out an in-depth analysis of the phonemic and stylometric features across centuries. We have also demonstrated the applicability of our dataset in various classification tasks using a fine-tuned BanglaBERT model which achieved an average of 60% accuracy. We believe that our dataset along with our insightful analyses will be significantly useful to linguists, lyricists, and musicians alike.

2 Data and Resource Usage

In this research, we utilized music and datasets from various open sources. The use of these resources is in accordance with their respective licenses, which allow for free usage and redistribution. The sources and licenses of the data are as follows:

2.1 Music Data

2.2 Generated Datasets

We also generated datasets for research purposes. These datasets are made freely available for the research community and are released under an open license that allows others to use, modify, and redistribute them. The specific license details for these datasets are as follows:

- **Dataset 1:** (Alliteration, 2023), licensed under the MIT License.
- **Dataset 2:** (Reduplication, 2023), licensed under the MIT License.
- **Dataset 3:** (Rhyme, 2023), licensed under the MIT License.

It is important to note that the open licensing of the data and resources used in this research promotes transparency, collaboration, and the advancement of knowledge in the field.

3 Dataset Preparation

3.1 Data Collection

In this section, we discuss about the preparatory steps and procedures that we followed when creating our primary Bangla music lyrics dataset. We considered an extensive collection of data from various websites. We carefully utilized multiple sources, such as (Lyrics71, 2023), (JioSaavn, 2023), (Wikipedia, 2023), and (Khan, 2020), to ensure a diverse and comprehensive dataset. Our dataset is enriched with various features that enable valuable insights into Bangla song lyrics. These features include essential information such as the title of the song, the performing artist's name, the album to which the song belongs, the complete lyrics, and unique features that represent the timeline of a song and artist (i.e. respectively the "Timeline" and "Mixed" features in the dataset). The music used in our research was obtained from open sources under the MIT License. This license permits the use, modification, and redistribution of the music with certain conditions, such as retaining the original license and copyright information. We have complied with these conditions.

The "Timeline" feature plays a significant role in our analysis as it indicates whether a piece of lyrics was written before or after the year 2000. The "Mixed" feature in the dataset represents whether an artist has published songs both before and after the year 2000. The choice of the year 2000 enables us to compare songs written across two centuries (20th to 21st century). To the best of our knowledge, no existing work consider any timeline when analyzing Bangla music (Khan, 2020)(Hossain and Al Marouf, 2018). This allows us to analyze the evolution of Bangla song lyrics and make cross-centurial comparisons by shedding light on the changing trends and themes of Bangla music over time. Another important aspect of our dataset is the implementation of a one-hot encoded genre labeling for each song, which allows for multiple genre labels to be assigned to a single song. By assigning multiple relevant genres to a song, we are able to capture the categorical essence of a song more accurately. 21% of the songs included in our dataset contain more than one genre. An example song with multiple genre such as Pop, Rock and Patriotic would be Azam Khan's popular song "Bangladesh" (Dataset, 2023).

Regarding the heuristics used in our dataset preparation, it's important to note that the genre assigned to each song was not chosen randomly but rather based on our listening preferences in Bangladesh according to (Music of Bangladesh, 2023). Pop and Rock are currently among the most popular and growing genres in Bengali culture, which is why we included a Pop/Modern genre label to capture modern songs with fast instrumental beats. However, we also recognize the importance of preserving traditional folk music and included a Folk/Urban Folk genre label to capture popular Lalon Sangeets and songs that address societal problems or current affairs such as politics, as seen in some of (Nachiketa Chakraborty, 2023) songs. Overall, our meticulously prepared dataset offers a unique resource for analyzing Bangla song lyrics with potential applications across various domains. In summary, our dataset provides a comprehensive and unique resource for analyzing Bangla song lyrics.

The inclusion of the timeline feature and the multi-label genre labelling task, and the utilization of heuristic techniques during the dataset preparation process greatly contribute to our research significance and potential applications in various domains such as Bengali music lyrical analysis.

3.2 Phonemic Features

From our primary Bangla music dataset, we derive three secondary datasets based on the phonemic features of the lyrics such as alliteration, reduplication, and rhyming. Each dataset has its own specific characteristics and labeling that facilitate the detection of the corresponding phonemic features. We use the algorithms presented in Rakshit et al. (2015) to automatically label the phonemic features.

The (Alliteration, 2023) dataset focuses on capturing the phonemic feature of alliteration in Bengali lyrics. Alliteration occurs when consecutive words or syllables in a line share the same initial sound. For this dataset, we consider all bi-grams of each piece of lyric. Each bi-gram is labeled based on whether it represents an alliteration or not. This binary labeling helps in the detection of alliteration patterns. Our (Alliteration, 2023) dataset has roughly 38,000 observations with columns such as bigrams and alliteration labels.

bigrams	alliteration_lbl	reduplication_lbl
ঘুরে ঘুরে	1	1
মা তাঁর	0	0
তোমার তুলনা	1	0
বাহিরে রে	0	1
যেই জন	1	0

Figure 1: Example of alliteration and reduplication dataset

line_pairs	label
ওরে ও পাগলা ভোলা দে রে দে প্রলয় দোলা	1
রেললাইনের ঐ বস্তিতে জন্মেছিল একটি ছেলে	0
লীন জড়তায় নীল আকাশে ঝড় বাঁধা পড়ে ভাঙা মানুষে	1
রেললাইনের ঐ বস্তিতে বাংলাদেশ বাংলাদেশ বাংলাদেশ	0
ছেলেটি মরে গেছে মা তাঁর পাশে চেয়ে বসে আছে	1

Figure 2: Example of rhyme dataset

The (Reduplication, 2023) dataset with around 38,000 observations aims to identify the phonemic feature of reduplication in Bengali lyrics. Reduplication involves the immediate repetition of a word or a part of a word within a line. Similar to the alliteration dataset, we consider all bi-grams of each piece of lyric. Each bi-gram is labeled based on whether it represents a reduplication or not. This binary labeling facilitates the detection of reduplication patterns.

The (Rhyme, 2023) dataset has roughly 1,40,000 observations which focuses on capturing the phonemic feature of rhyming in Bengali lyrics. Rhyming refers to the similarity in sounds between the last words of two lines in a lyric. For this dataset, we consider all combinations of line pairs from each piece of lyric. Each line pair is labeled based on whether they rhyme or not. This binary labeling helps in the detection of rhyme patterns.

Although it is possible to identify alliteration, reduplication, and rhyming patterns using the algorithms described in the Rakshit et al. (2015), our goal is to incorporate the comprehension of these phonemic features on a language model level. By doing so, we aim to enhance the language model's understanding and generative capabilities in relation to these specific features.

4 Experiments

4.1 Exploratory Data Analysis

We carried out a thorough exploratory data analysis (EDA) to obtain important insights into the linguistic characteristics and thematic variations of our dataset. We delved into the connection between artists and their lyrical output. We used bar charts and heatmaps to visualize the unique lyrical styles, vocabulary choices, and various stylometric features of the lyrics. By performing this analysis, we not only highlighted the diversity among artists but also uncovered potential patterns in their lyrical compositions.

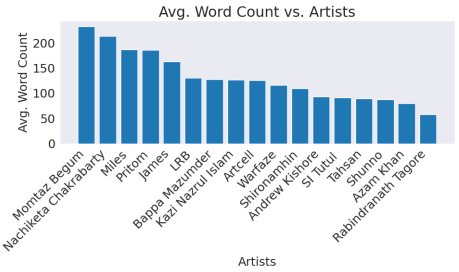

Figure 3: Avg. Word Count by Artists

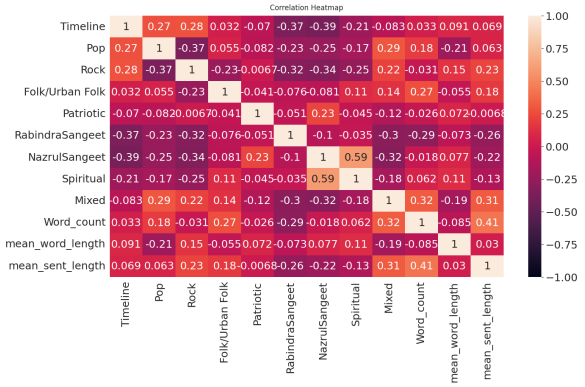

Figure 4: Heatmap of our dataset

Correlation Analysis. In this section, we present the results of our correlation analysis conducted on the Bengali song lyrics dataset. The objective of this analysis was to uncover interesting relationships and dependencies among various linguistic and thematic features of Bengali lyrics. In fig. 4, we visualize the correlation coefficients among different pairs of features. The color intensity of each cell indicates the strength and direction of the correlation. Cells with a darker shade of red indicate positive correlations and negative otherwise.

We observed a strong positive correlation between Nazrul Sangeet and Spiritual genre, suggesting that many songs written by Kazi Nazrul Islam have a religious and spiritual touch in them. This insight could be invaluable for songwriters and artists aiming to create spiritually resonant music. Positive correlation between "mean sentence length" and the "mixed" feature denote that artists who have published songs both before and after the year

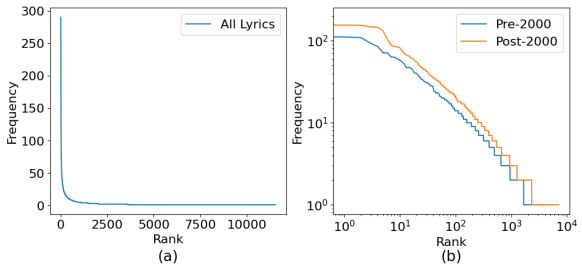

Figure 5: Frequency-Ranked Token Distribution. (a) Shown on linear scale for entire corpus (b) Comparing pre and post 2000 lyrical distribution on a logarithmic scale

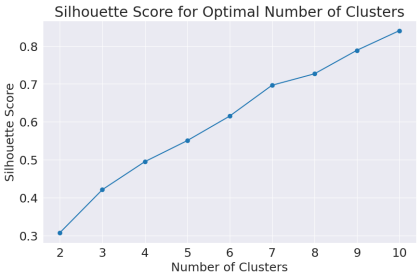

Figure 6: Silhoutte Score

2000, tend to perform longer songs on average. On the other hand, we found a negative correlation between rock and patriotic genre which suggests a clear distinctness between them. Spiritual and rock songs are also negatively correlated showing that there are almost no spiritual songs which have a rock feel in them.

Distribution Analysis. In this section, we analyze the token distribution of the song lyrics based on the frequency rank of the vocabulary. In fig. 5(a), we visualize the token distribution on both linear and logarithmic scales and we have observed that the token distribution of the entire lyrics corpus approximately follows the Zipf's law (Zipf's Law, 2023). In fig. 5(b), we separately observe the frequency-ranked token distribution of pre and post 2000 song lyrics. After carrying out a Kolmogorov–Smirnov test between the pre and post 2000 lyrics distribution, we got a p-value > 0.05, which establishes the null hypothesis with 95% confidence that the frequency-ranked distributions of pre and post 2000 lyrics are identical.

Clustering Analysis. We applied the K-Means clustering algorithm (Pedregosa et al., 2011) to effectively partition data into distinct clusters based on similarity. The goal here was to gain a deeper understanding of how different music genres and artists relate to each other in terms of the stylometric features like average length, word count etc. of the lyrics instead of using audio features as seen in (Jiang and Jin, 2022).

To determine the optimal number of clusters (k) for our dataset, we relied on the Silhouette score (Pedregosa et al., 2011). The Silhouette score is a valuable metric that quantifies the quality of clusters produced by the K-Means algorithm. It helps us to find the balance between compact clusters and well-separated clusters. Our analysis indicated that k=10 yielded the highest Silhouette score, approximately 0.8, suggesting that this configuration best represented the underlying structure within our data. The principal findings of this analysis are as follows:

Cluster 0 is identified as an amalgamation of Rock songs, interwoven with historical undertones before 2000, demonstrating the adaptability of Rock music at the time. Cluster 1 designates a particular collection that is defined by a fusion of patriotic melodies that are primarily centered within the boundaries of post 2000. The fact that Cluster 2 includes a diverse range of pop compositions highlights the diversity of the popular music genres included in this grouping. Cluster 3 appears as a unique repository of pure rock compositions that were created after 2000s, demonstrating an obvious preference for this genre at the time. Nazrul Sangeet is represented by Cluster 5 and exhibits a complex convergence of cultural motifs, with clear inclinations toward themes of patriotism and spirituality. Cluster 7 primarily consists of folk songs which were composed after 2000s, demonstrating the persistent resonance of traditional folk music within this cluster.

4.2 Classification Tasks

We fine-tuned the pretrained BanglaBERT model to classify different phonemic features like alliteration, reduplication, and rhyming patterns and achieved on average 60% accuracy. Additionally, we try to classify 18 artists from our dataset which was an apparently challenging task given the size of our dataset. We achieved 15% classification accuracy when training a BERT model from scratch on the task of classifying artists from the lyrics. However, when we fine-tuned the BanglaBERT model for the same task, we achieved 57% accuracy. Therefore, our dataset is promising to facilitate the difficult task of classification from a small dataset by utilizing transfer learning approaches.

5 Related Work

There is an existing work on Bangla music lyrics classification, as mentioned in the paper (Ahmed et al., 2022). In their study, the authors focused on genre label classification for Bangla music. Our dataset is built upon the motivation of their work, with the inclusion of timeline and multiple genres for the songs. This enables us to use binary encoding for multiple genre labeling. Another notable work in the field of Bangla Music dataset is presented in the paper by (Hossain and Al Marouf, 2018). In this study, a novel dataset is introduced, which is the first stylometric dataset comprising of Bangla music lyrics. None of these datasets used the multilabel genre labeling and did not include the timeline of the songs.

6 Conclusion

We conducted a comprehensive phonemic analysis of Bangla song lyrics, a low-resource language often overlooked. Our research introduces novel datasets, including "Timeline" and "Mixed," to deepen our understanding of linguistic evolution in Bangla music. Additionally, our multi-label genre analysis uncovers the detailed relationship between phonemic patterns and diverse musical genres in Bangla music, enriching the broader field of music analysis. Future direction of exploration include cross-linguistic comparative studies, examining phonemic patterns in Bangla lyrics compared to other languages, and developing interactive songwriting tools that leverage phonemic analysis to aid songwriters in crafting lyrics with specific phonemic attributes.

References

Shafi Ahmed, Md Humaion Kabir Mehedi, Moh Absar Rahman, and Jawad Bin Sayed. 2022. Bangla music lyrics classification. In *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, pages 142–147.

Alliteration. 2023. Alliteration. https://docs.google.com/spreadsheets/d/e/2PACX-1vT00OUbh8Zwplx0aOTJ9HBVJFl-WOznKJGHcm5oFcq_fWG1adfYUU_IWkuOueawRzDsKA8ufwhMix2w/pubhtml. Accessed: Sept 7, 2023.

Music Dataset. 2023. Bangla Music lyrics. https://docs.google.com/spreadsheets/d/e/2PACX-1vRLY_n83tpJaqJO1r6zAIE2mziMkk2l7gkg3Vhld3Xi_lkxOtEBId9Ud3OupeV0b-fECTd4jILiSktO/pubhtml. Accessed: Sept 7, 2023.

Rafayet Hossain and Ahmed Al Marouf. 2018. Banglamusicstylo: a stylometric dataset of bangla music lyrics. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE.

Yanru Jiang and Xin Jin. 2022. Using k-means clustering to classify protest songs based on conceptual and descriptive audio features. In *International Conference on Human-Computer Interaction*, pages 291–304. Springer.

JioSaavn. 2023. Jiosaavn. https://www.jiosaavn.com/.

Shakirul Hasan Khan. 2020. Kaggle. https://www.kaggle.com/datasets/shakirulhasan/bangla-song-lyrics.

Lyrics71. 2023. Lyrics71 bangla song lyrics. https://lyrics71.net/.

Music of Bangladesh. 2023. Music of bangladesh — Wikipedia, the free encyclopedia. [Online; accessed 6-September-2023].

Nachiketa Chakraborty. 2023. Nachiketa chakraborty — Wikipedia, the free encyclopedia. [Online; accessed 6-September-2023].

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, and Gholamreza Haffari. 2015. Automated analysis of bangla poetry for classification and poet identification. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 247–253.

Reduplication. 2023. Reduplication. https://docs.google.com/spreadsheets/d/e/2PACX-1vT00OUbh8Zwplx0aOTJ9HBVJFl-WOznKJGHcm5oFcq_fWG1adfYUU_IWkuOueawRzDsKA8ufwhMix2w/pubhtml. Accessed: Sept 7, 2023.

Rhyme. 2023. Rhyme. https://docs.google.com/spreadsheets/d/e/2PACX-1vRqvHMspm9igtHbZ1o7ZjVFjksJ8g3Xv5t4iCTL-vNyhL7wE6dgCxMrEEooUKO0nhTRG_J6cZMgY09f/pubhtml. Accessed: Sept 7, 2023.

Wikipedia. 2023. Wikipedia. https://www.wikipedia.org/.

Zipf's Law. 2023. Zipf's law — Wikipedia, the free encyclopedia. [Online; accessed 11-September-2023].