Lyrics Generation Based on Artist and Genre Using Enhanced Spotify Dataset

Fatma M. AbdelHadi, Aya Kandil, Jessica Ehab Bassily, and Demiana Yacoob

***Abstract*—Automatic song lyrics generation is a challenging Natural Language Processing task that requires linguistic coherence, creativity, and stylistic control. While music platforms such as Spotify provide large-scale datasets with rich metadata, they typically lack access to complete song lyrics, limiting their usefulness for lyrics generation tasks. This paper addresses this limitation by enhancing a Spotify-based playlist dataset through the integration of links to the Genius lyrics platform, enabling access to full lyrical content while preserving artist and genre metadata. The resulting dataset provides a foundation for controlled lyrics generation conditioned on artist identity and musical genre and supports future modeling and evaluation efforts.**

I. Introduction and Motivation

THE most recent advances in Natural Language Processing have enabled significant progress in creative text generation tasks, including poetry, storytelling, and dialogue systems. Among these tasks, automatic song lyrics generation represents a particularly challenging problem due to the need to balance linguistic coherence, stylistic consistency, creativity, and thematic relevance. Lyrics generation systems are expected not only to produce grammatically correct text, but also to reflect artistic style, emotional tone, and genre-specific conventions.

Music streaming platforms such as Spotify provide large-scale datasets containing rich metadata about songs, including artist names, genres, popularity metrics, and audio features. However, many publicly available music datasets focus primarily on metadata and lack access to complete lyrical content. This limitation restricts the ability of NLP models to learn meaningful patterns between artist identity, genre characteristics, and lyrical structure. As a result, existing lyrics generation approaches often rely on limited or fragmented lyric corpora, which constrains their expressive capacity and generalizability.

To address this gap, this project focuses on enhancing a Spotify-based music dataset by integrating external lyric resources. Specifically, links to the Genius lyrics platform are added as an additional dataset column, enabling access to full song lyrics. This enrichment allows for more comprehensive textual analysis and supports the development of generation models that condition output on both artist identity and musical genre. By bridging metadata-rich music datasets with complete lyrical content, this work aims to provide a stronger foundation for controlled and context-aware lyrics generation.

II. Literature Review

Early research on song lyrics within the NLP community primarily focused on lyrical content analysis rather than generation. Studies applying topic modeling and statistical analysis have demonstrated that lyrics encode recurring semantic themes that vary across musical genres and cultural contexts. For example, large-scale analyses of metal music lyrics using Latent Dirichlet Allocation (LDA) revealed strong correlations between lyrical topics and audio features such as perceived darkness and aggression [1]. These findings highlight the importance of textual data in understanding musical expression and motivate the inclusion of lyrics in music-related NLP tasks.

Statistical investigations of music metadata have further emphasized the role of genre as a defining attribute of musical identity. Longitudinal studies examining genre popularity trends show that genre labels capture meaningful shifts in listener preferences over time [2]. Although such studies do not directly address lyrics generation, they establish genre as a critical conditioning variable that generation models should account for.

With the rise of neural text generation, researchers began exploring sequence modeling approaches for lyrics generation. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models were among the earliest architectures applied to this task, demonstrating the feasibility of generating stylistically coherent lyrical lines. However, these models often struggled with long-range dependencies and thematic consistency, leading to repetitive or semantically shallow outputs [3]. In parallel, research on lyrics-based classification has shown that lyrical text alone contains sufficient stylistic signals to support tasks such as genre and artist identification. For instance, studies using statistical and machine learning classifiers on song lyrics demonstrate that linguistic features capture meaningful genre-related patterns [10]. Although these works focus on classification rather than generation, they reinforce the importance of full lyrical content for modeling stylistic variation in music-related NLP tasks.

More recent work has explored advanced neural architectures for creative text generation, including Variational Autoencoders (VAEs) and Transformer-based models. These approaches have been shown to improve fluency and structural coherence in generated lyrics by capturing latent stylistic representations and leveraging self-attention mechanisms [4]. In music-related contexts, some studies have investigated multimodal generation, incorporating musical structure, rhythm, or melody alongside textual input [5].

A notable direction within this line of research involves the use of Variational Autoencoders for artist-conditioned lyrics generation. Vechtomova and Bahuleyan propose a VAE-based framework that integrates artist-specific embeddings derived from audio features to guide lyrics generation [11]. By conditioning the generative process on stylistic representations, their approach aims to produce lyrics that better reflect individual artist characteristics. This work highlights the potential of latent-variable models for controlling stylistic aspects of generated lyrics, while also illustrating the dependency of such models on the availability of high-quality lyric corpora.

Despite these advancements, many lyrics generation systems rely on limited or curated lyric datasets, often restricted to specific genres or languages. Moreover, the majority of studies assume direct access to full lyrical corpora, without addressing how such data is obtained or maintained at scale. This presents a practical limitation for projects that aim to combine large music metadata collections, such as Spotify datasets, with textual generation objectives.

Recent evaluations of neural creative systems also raise concerns regarding controllability and interpretability. While models can generate fluent text, enforcing constraints related to artist style or genre remains an open challenge [6]. This further emphasizes the need for datasets that explicitly link metadata attributes with complete lyrical content.

Publicly available music datasets vary significantly in scope and structure. Spotify-based datasets typically provide extensive metadata, including artist information, genre tags, and audio descriptors, making them valuable for music recommendation and analysis tasks [7]. However, these datasets rarely include full lyrics due to licensing and copyright restrictions. As a result, researchers often resort to external lyric sources or small-scale annotated corpora.

Several studies highlight the importance of dataset quality and coverage in creative NLP tasks. Missing or incomplete textual data can introduce bias and limit a model's ability to generalize across artists and genres [8]. In the context of lyrics generation, the absence of full lyrics prevents models from learning stylistic patterns unique to individual artists, such as vocabulary usage, rhyme schemes, and thematic focus.

Efforts to bridge this gap have involved web-based lyric collection and alignment strategies. Platforms such as Genius provide structured access to song lyrics and metadata, making them a common resource for music-related NLP research [9]. However, systematic integration of such resources with large-scale music metadata datasets remains underexplored in existing literature.

Based on the reviewed literature, a clear gap emerges at the intersection of music metadata and lyrical content availability. While prior work demonstrates the effectiveness of neural models for lyrics generation and analysis, limited attention has been given to dataset enrichment strategies that enable scalable and controllable generation.

This project addresses this gap by augmenting a Spotify playlist dataset with Genius lyrics links, enabling access to complete song lyrics while preserving rich metadata such as artist identity and genre. This enhanced dataset supports deeper textual analysis and provides a foundation for developing lyrics generation models conditioned on meaningful musical attributes. By focusing on dataset enhancement as a core contribution, this work aims to improve the quality and flexibility of lyrics generation systems and facilitate future research in music-oriented NLP.

III. Dataset Description and Analysis

The improved music dataset utilized in this study was first created using Spotify metadata and then expanded with complete song lyrics gathered from the Genius platform. The complete textual lyrics of the song, the artist's name, the song title, and a link to the corresponding lyrics on Genius are all included in each entry in the dataset. The main objective of this enhancement is to combine rich musical metadata with complete lyrical material, enabling deeper text analysis and aiding future lyrics creation applications.

Every song is linked to full textual material, according to a preliminary analysis of the dataset, which revealed no missing lyrics. A number of preprocessing techniques were used to guarantee consistency and enhance the quality of the data. English contractions were enlarged (for instance, "they're" was changed to "they are") to maintain semantic meaning during normalization, and all text was turned to lowercase to decrease vocabulary variety. Minor spelling errors in song titles were fixed, and white space was standardized while maintaining line breaks to preserve the original lyrical structure. Crucially, the original lyrics were kept intact, providing room for further experimentation and study.

After preprocessing, the dataset consists of 57,651 songs performed by X unique artists. This indicates that the dataset represents a wide variety of artists while maintaining a clean and consistent lyrical format suitable for large-scale textual analysis.

To analyze the textual characteristics of the dataset, descriptive statistics were computed based on the length of song lyrics measured in word count. On average, each song contains approximately 216 words, with the shortest songs having around 60 words and the longest exceeding 747 words. This wide range reflects the diversity of lyrical structures within the dataset, spanning from short, repetitive compositions to longer, narrative-driven songs.

A severely skewed distribution of song contributions is shown by an artist-level examination. The majority of artists are represented by a small number of tracks, while a comparatively small number of musicians provide a significant share of the songs in the dataset. The artist frequency distribution, which shows that a small number of extremely frequent artists predominate in the sample, amply illustrates this disparity. Although these patterns are prevalent in actual music collections, they have significant effects on model evaluation and training, especially with regard to bias and generalization.

The variation in song structure is further highlighted by the variety of lyrics length. The existence of outliers with remarkably short or long lyrics suggests that any sequence-based modeling strategy must be able to handle significant variance in input length, even when the majority of songs fall within a moderate range of word counts.

A. Insights

The dataset analysis yields several important findings. First, adding whole lyrical material to Spotify metadata greatly raises the analytical usefulness of the collection. Access to full lyrics facilitates the construction of generative models that can immediately learn stylistic and structural trends from lyrical data as well as deeper textual study.

Second, if remedial action is not taken, the apparent imbalance in artist representation raises the possibility that artist-conditioned lyrics generation models will become biased in favor of the styles of highly represented musicians. This finding emphasizes how crucial it is to use weighting or balanced sample algorithms during training in order to preserve stylistic variance across artists.

Third, models for producing lyrics must be able to handle variable-length sequences due to the broad range of lyrics' lengths. This directly impacts later architectural choices, including the application of padding, truncation techniques, or attention-driven algorithms to handle extended sequences while maintaining contextual consistency.

Ultimately, repeated phrases and refrains are instances of lyrical elements that reveal the fundamental organization of song lyrics. Such repetitions may result in outputs that are overly repetitive, despite enhancing artistic authenticity. This emphasizes the necessity of generating methods that strike a balance between controlled repetition, coherence, and inventiveness.

B. Dataset Limitations

The dataset contains a number of drawbacks that should be noted despite its advantages. Most significantly, there are no explicit genre designations in the improved dataset. At this point, direct genre-based analysis is limited, despite the fact that genre is a crucial conditioning factor in the overall project objective. As a result, artist identification now reflects stylistic variation, and upcoming projects will include genre metadata.

Additionally, the dataset exhibits a significant imbalance in artist representation, which may restrict stylistic diversity and bias generative models toward artists who occur often. Moreover, the lyrical content includes structural noise like repetitive phrases, informal language, and inconsistent formatting that, if not adequately handled during preprocessing, could negatively impact feature extraction and model performance. Finally, copyright and licensing regulations regarding song lyrics limit dataset redistribution and scalability, creating challenges for replication and wider adoption of research.

References

[1] I. Czedik-Eysenberg, O. Wieczorek, and C. Reuter, "'Warriors of the Word' – Deciphering Lyrical Topics in Music and Their Connection to Audio Feature Dimensions Based on a Corpus of Over 100,000 Metal Songs," arXiv preprint arXiv:1911.04952, Nov. 2019, doi: 10.48550/arXiv.1911.04952.

[2] A. M. Petitbon and D. B. Hitchcock, "What Kind of Music Do You Like? A Statistical Analysis of Music Genre Popularity Over Time," *Journal of Data Science*, vol. 20, no. 2, pp. 168–187, Apr. 2022, doi: 10.6339/22-JDS1040.

[3] P. Potash, A. Romanov, and A. Rumshisky, "GhostWriter: Using an LSTM for Automatic Rap Lyric Generation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, L. Màrquez, C. Callison-Burch, and J. Su, Eds., Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1919–1924, doi: 10.18653/v1/D15-1221.

[4] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, Jun. 2017, doi: 10.48550/arXiv.1609.04747.

[5] M. Mayerl, S. Brandl, G. Specht, M. Schedl, and E. Zangerle, "Verse Versus Chorus: Structure-Aware Feature Extraction for Lyrics-Based Genre Recognition," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[6] "LyricLure: Mining Catchy Hooks in Song Lyrics to Enhance Music Discovery and Recommendation," *ACM*, accessed Dec. 26, 2025. [Online]. Available: https://dl.acm.org/doi/epdf/10.1145/3640457.3688049

[7] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[9] Genius Media Group, "Genius Lyrics Dataset," 2023.

[10] A. Girase, A. Advirkar, C. Patil, D. Khadpe, and A. Pokhare, "Lyrics Based Song Genre Classification," in *Proceedings of the International Conference on Advances in Computing, Communication and Informatics (ICACCI)*, 2018.

[11] O. Vechtomova, H. Bahuleyan, A. Ghabussi, and V. John, "Generating Lyrics with Variational Autoencoder and Multi-Modal Artist Embeddings," arXiv preprint arXiv:1812.08318, Dec. 2018, doi: 10.48550/arXiv.1812.08318.