Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25483

Conditional Rap Lyrics Generation with Controlled Thematic Coherence using Fine-tuned Transformer Models

Peihang Lyu

Xiamen University Malaysia (XMUM), Kuala Lumpur, Malaysia
cys2209189@xmu.edu.my

Abstract. Rap lyric generation is a specialized task that demands both thematic consistency and stylistic creativity. Traditional language models often struggle to generate coherent rap lyrics that align with specific themes. This paper addresses this challenge by proposing a conditional text generation approach based on a fine-tuned BART transformer model, guided by explicit thematic control tokens. Through a method of data collection, thematic annotation, and model fine-tuning, the paper explores how controlled generation can enhance the quality of rap lyrics. The study constructs a multi-label rap dataset by scraping and annotating lyrics with twelve common themes such as love, street, and storytelling. The model leverages both content word extraction and theme-specific prompts to generate context-aware lyrics. Experimental evaluation, including both qualitative review and quantitative classification using F1 scores, shows strong performance: validation F1 (micro: 0.992, macro: 0.980) and test F1 (micro: 0.96, macro: 0.86). The paper finds that thematic prompting significantly improves coherence and controllability in generated lyrics, offering new insights into the application of transformers in stylized conditional text generation tasks.

Keywords: Rap Lyrics Generation, Conditional Generation, Thematic Control, Transformer, Multi-label Classification

1. Introduction

Rap lyric generation is a natural language generation subfield that results in output which embodies a mixture of artistic expression, linguistic rhythm, and cultural nuance. The general text generation task does not involve such special considerations and challenges as maintaining intricate rhyme schemes and preserving stylistic consistency [1]. Another difficulty that general text creation does not encounter is long-range dependency created by the verbosity of the rap lyrics [2]. Most models lack the required semantic depth and expressive flexibility to deal with the dense slang and stylistic diversity inherent in rap lyrics.

Conventional approaches usually work on unconditional generation which includes generation-based methods and retrieval-based methods [3-4]. In these cases, the models generate lyrics autonomously and without control which may result in incoherence or thematic inconsistency with the output. Earlier works have asserted that such models cannot relate content to a particular theme, which is why their application in creative writing is restricted [5]. To address these shortcomings,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ25483

the researchers have promoted the development of conditional generation strategies with the support of external control signals.

Existing conditional generation models, such as Rapformer, have improved rhythmic flow in rap verses using Transformer architectures and rhyme improvement mechanisms [6]. Such models continue to lack clear thematic control, limiting how well they can align generated content according to specific semantic intent.To overcome this restriction, a theme-regulated approach is proposed based on a BART model fine-tuned , with added explicit control tokens for theme representations of rap. Through multi-label thematic annotation, content word extraction, and conditional decoding, the model makes controlled lyric production possible with improved coherence and theme fit.

This study adds to the growing body of controllable text generation. It aims to demonstrate that well-tuned Transformer models can grasp accurately the balance between content fidelity and stylistic diversity in such an important creative domain as music. The method proposed has both theoretical and practical value. Theoretically, it extends the application of sequence-to-sequence learning in stylized text domains,and practically, it enables personalized, theme-aware rap lyric generation for creators and entertainment platforms.

2. Literature review

The issue of automatic rap lyrics creation has emerged as a target of increasing research attention since it poses unique requirements on rhythm, rhyme, and stylistic expression. Early attempts such as Ghostwriter employed LSTM-based models for generating rap lyrics with a focus on syntactic coherence [3]. The models failed to possess long-term semantic planning, however, resulting in locally coherent but topically fragmented outputs.

Follow-up work has employed Transformer-based architectures to enhance the quality of generative output. Rapformer introduced a denoising autoencoder architecture specifically tailored to rap lyrics, reconstructing lines from raw content words and integrating a BERT-based rhyme enrichment module [6]. The model showed improved performance in rhythmic fluency and rhyme density—achieving a reported 10% improvement in rhyme density—demonstrating structural augmentation can increase stylistic similarity. But Rapformer lacked direct mechanisms of thematic control, and therefore there was little flexibility while it was generating content on user-specified themes or narrative coherence.

Parallel to this, other models like Plug-and-Play Language Models (PPLM) [5] have made controlled text generation possible by injecting latent attributes at inference via gradient-based updates. While helpful for controlling overall text styles and sentiment, PPLM's indirect conditioning produces unstable or sporadic results in highly structured domains such as rap lyrics. Moreover, utilization of backpropagation gradient computation introduces inefficiencies and hampers real-time production of lyrics.

Topic-aware generation work has been limited in music. Most existing methods either depend on implicit topic modeling or use template-based generation, which lacks expressiveness. There is a wide gap in the integration of explicit multi-label thematic conditioning with a Transformer-based generation pipeline that captures both linguistic and stylistic information.

In order to bridge this gap, this paper suggests a theme-controlled conditional generation model based on a fine-tuned BART model. The proposed model inserts thematic tokens into the input format in an explicit way such that the encoder can inject theme semantics into the latent space directly. Additionally, a specially annotated multi-label rap dataset of twelve high-level themes (e.g., "love", "street", "pain") is constructed in order to enable fine-grained thematic supervision.

DOI: 10.54254/2755-2721/2025.BJ25483

Briefly, while previous models like Rapformer can achieve stylistic attributes like rhythm and rhyme, they lack the ability to provide explicit and consistent thematic control. The methodology outlined here makes a new contribution by using multi-label thematic control over a powerful sequence-to-sequence Transformer model to facilitate the generation of rap lyrics not just stylistically fluent but semantically consistent with user-specified themes.

3. Methodology

3.1. Dataset construction

First, songs were taken from the Genius site to collect 50 tracks from each of five rap artists resulting in 12,493 samples that could be used. These lyrics were stored as separate JSON files with each file having the title and text for a song.

In the preprocessing stage, text cleaning procedures were applied to remove non-lyrical content and metadata (e.g., [Chorus],contributor lines) as well as duplicate lines.Texts were filtered through a language detection module to keep only English-language lyrics and those with sufficient content. Individual songs were divided into verses and lines; short or noisy lines were filtered out.

Content words were extracted from each lyric line by removing stopwords, punctuation, and non-alphabetic tokens. To tag the themes, a rule-based keyword matching technique was applied. For each of the twelve pre-determined thematic labels (e.g., love, street, motivation, pain), a manually selected set of representative keywords was established based on domain knowledge and frequency analysis of words. A lyric line was assigned one or more themes if it displayed lexical matches with the corresponding keyword lists.

To facilitate overcoming the limitation of rule-based annotation alone that can cause mislabeling due to polysemy or context uncertainty, a post-processing manual verification step was introduced to ensure label quality. That is, a random 15% subset of the annotated content was sampled and manually verified by two independent reviewers. Reviewers were requested to verify whether the assigned themes were semantically justified from the full verse context. In cases of disagreement or uncertainty, group agreement was reached through discussion. Manual checking helped to clarify lists of keywords and exclude systematic label errors, thereby strengthening the overall reliability of the dataset.The resulting data set was saved as CSV file. There were several fields in each row. These were artist, song title, verse, content words, and topics. The dataset was divided into three parts. 81% of the data was used for training. The validation set is at 9% and the test set is the remaining 10%. Stratified sampling was employed on the most dominant themes to achieve even balance over the sets.For fine-tuning, conditional inputs were constructed in the form of <topic1> <topic2> ... <CTRL> content_words, and were paired with the corresponding target lines, resulting in a dataset suitable for sequence-to-sequence learning.

3.2. Model architecture

A BART model was used to implement a sequence-to-sequence paradigm for the conditional generation task. BART employs a bidirectional encoder with an autoregressive decoder, which allows for highly effective contextual comprehension alongside fluent text generation. Fine-tuning experiments were conducted with the base version of BART which has 6 layers in both the encoder and decoder. The base model was used in this case because it was more resource-efficient for computation and training.

DOI: 10.54254/2755-2721/2025.BJ25483

To enable explicit control over themes, a sequence of topic tags was prepended to each input which altered the original format of the input. When a verse carries the tags, “love” and “pain” for instance, the input is encoded as <love> <pain> <CTRL> content_words. This enables the model to pay attention to the primary themes lexically during generation.

The corresponding verse was set as the target output for each training instance. As the default decoding strategy, a teacher-forcing approach was used. Model loading and fine-tuning were done on the BART model using the HuggingFace Transformers library [9] and all experiments were run on a single GPU.

Lyric generation is made controllable through instruction-based signalling at the decoder level using this architecture. Apart from its strong generative capabilities, the BART backbone can also respond flexibly to input structural enforcement which steers the semantics of the output to be aligned with user-defined themes.

4. Evaluation metrics and results

4.1. Automatic evaluation

To determine how well the generated rap lyrics align with their assigned themes, an automatic evaluation was conducted. In this case, a multi-label classifier was trained specifically as an evaluator. The classifier was created using a TF-IDF vectorizer and a One-vs-Rest logistic regression model. Each generated line was converted into feature vectors representing unigrams and bigrams. The labels were encoded as multi-hot vectors of thematic tags.

Apart from the automated evaluation, a manual evaluation was also performed that assessed the congruency of the produced lyrics with the given prompts thematically. This work was concluded by applying the classifier to predict themes for the generated lyrics and assessing them against conditioning topics used during generation. The micro-F1, macro-F1, and sample-F1 scores were used to measure thematic similarity between generated song lyrics and provided themes.

Micro-F1 calculates precision and recall for metrics that have been symmetrically stratified at theme-appropriate points. It demonstrates the model performance, especially on shared themes. Macro-F1 computes F1 metrics for each individual theme (love, pain, street et al.), then averages them out without weighting based on occurrence frequency thus focusing on rare themes.

Sample-F1 computes the F1 score of each generated lyric line as a multi-label instance and then averages over all samples, which reflects the extent to which each line aligns with its target theme set.

Table 1 shows that the classifier achieved almost perfect precision and recall on dominant themes such as general, love, and storytelling and slightly less so on low-frequency themes such as party and pain. This reflects the model’s strong ability to reproduce salient thematic structures, but performance on low-resource categories remains a weak point.

The overall results in Table 2 still attest to the model performance in generating text conditioned on themes with micro-F1 score of 0.992 and macro-F1 of 0.980 on the validation set. The high macro-F1 score confirms the model’s performance on infrequent themes is strong, which can be attributed to the multi-label conditioning technique used.

Table 3 showcases the results from the test set, where the model achieved micro-F1 0.96, macro-F1 0.86 and sample-F1 0.96. These results indicate the model’s ability to generalize and maintain thematic coherence throughout generation is performed at sample level, not just at label level.

DOI: 10.54254/2755-2721/2025.BJ25483

Table 1. Per-label multi-label classification performance on validation set

theme	precision	recall	f1-score	support
drugs	1.00	0.96	0.98	24
faith	1.00	1.00	1.00	9
family	1.00	1.00	1.00	14
general	0.99	1.00	1.00	818
love	1.00	0.98	0.99	56
motivation	1.00	1.00	1.00	2
pain	0.77	0.91	0.83	11
party	1.00	0.93	0.97	15
politics	1.00	1.00	1.00	2
social	1.00	0.96	0.98	50
storytelling	1.00	1.00	1.00	18
street	1.00	1.00	1.00	11
violence	1.00	1.00	1.00	17

Table 2. Overall multi-label classification metrics

Metric	value
Micro-F1	0.992
Macro-F1	0.980
Sample-F1	0.990

Table 3. Generated test set evaluation results

Metric	value
Micro-F1	0.9597
Macro-F1	0.8572

4.2. Human evaluation

Along with the automated assessment, the researchers conducted a human assessment concentrating on theme alignment. Three different people were invited to assess whether the topics assigned to the lyrics matched those intended. Each assessor was given a sample of 50 verses which were generated based on 1-2 theme tags provided as input. All of the verses were generated irrespective of the underlying themes.

The participants were instructed to evaluate each verse based on a binary logic where 1 denotes adherence to the topics while 0 signifies no adherence. The final figure of accuracy was calculated as the percentage of verses which a majority of evaluators voted as thematically relevant to the topics. It was found that on average, 87 percent of the generated verses were deemed appropriate. These results corroborate the results of automated evaluation and conclude that the model operates within human defined constraints generating topic relevant lyrics.

DOI: 10.54254/2755-2721/2025.BJ25483

4.3. Examples of generated lyrics

To illustrate the model's performance, the study present two sample outputs generated from different thematic prompts.

Example 1: Successful Generation

Input Themes: <love> <pain>

Generated Verse:

I gave you all my soul, still I'm left in the rain

Love turned cold, now I'm stuck with the pain

Every night I write, just to numb what remains

Heartbreak rhymes flowing deep in my veins

Thematic coherence is easily gleaned from references to emotional loss and suffering, giving adequate coverage to both love and pain.

Example 2: Weak Generation

Input Themes: <politics> <social>

Generated Verse:

We ride with the team, living the dream

Money and cars, we shining so clean

Pop the champagne, we on the scene

Ain't nothing to lose when you the king

Here, the content reflects themes of party or wealth rather than social or political commentary. This indicates that the model may default to dominant patterns and thus, requires stronger structural or semantic framing, which is a gap for future work on targeted thematic control.

5. Conclusion

This research proposes a conditional rap lyric generation model using a thematically controlled multi-label annotated BART-based framework. The model's incorporation of content word extraction and topic prompting in the input format enhances the coherence and thematic alignment of the generated lyrics. The results of the automatic evaluation indicate strong performance as well, achieving a micro-F1 of 0.96 and macro-F1 of 0.86 on the test set, which confirms persuasive thematic control.

Despite these contributions, the study is limited in a number of ways. The current step in the process of annotation is based on keyword-matching rules which are far too rigid and overly simplistic, thus missing more complex and subtle themes going beyond textual representation. Furthermore, the model neglects essential rhythmic elements such as rhyme and beat found in rap music. Results can be improved in future work by extending and enhancing the evaluation framework professionally.Adding semantic clustering or weakly supervised techniques for topic annotation augmentation, rhyme density modeling, beat alignment, phonetic feature incorporation, multi-lingual regional rap stylistic data augmentation for expanded generalization, and dataset extension improves lyrical quality. This work aids in the domain of text generation with a high level of control and sets the stage for advanced systems for lyric generation that require intricate style considerations.

References

[1] Bradley A. (2017). Book of Rhymes: The Poetics of Hip Hop. Basic Civitas Books.

[2] Mayer R., Neumeyer D., et al. (2008). Popular Music and Society, 31(4): 443-460.

DOI: 10.54254/2755-2721/2025.BJ25483

[3] Potash P., Romanov A., Rumshisky A (2015). Ghostwriter: Using an LSTM for Automatic Rap Lyric Generation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[4] Malmi E., Takala P., Toivonen H. (2016). DopeLearning: A Computational Approach to Rap Lyrics Generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[5] Dathathri S., Madotto A., Lan W., et al. (2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In: International Conference on Learning Representations (ICLR), .
[6] Nikolov N. I., Malmi E., Northcutt C. G., Parisi L. (2020). Rapformer: Conditional Rap Lyrics Generation with Denoising Autoencoders. In: Proceedings of the 13th International Conference on Natural Language Generation.
[7] Lewis M., Liu Y., Goyal N., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).
[8] Vaswani A., Shazeer N., Parmar N., et al. (2017). Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS).
[9] Wolf, T., Debut, L., Sanh, V., Chaumond, J., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations(pp. 38-45).