Lyrics Matter: Exploiting the Power of Learnt Representations for Music Popularity Prediction

ARR Cycles: December ARR 2024 (Long Paper), February ARR 2025 (Long Paper), Commit to ACL 2025 (Long Paper)

ACL 2025 Metareview:

The work sounds promising and is a well-executed application, but sounds like it is better suited for a short paper with a focused contribution, unless substantially revised to broaden empirical scope, deepen insights, and demonstrate methodological innovation.

Past ARR Scores

- Dec ARR (Meta Reviewer: 3 ; Reviewers Soundness: 4, 3, 3.5 ; Reviewer Overall: 4, 2.5, 2.5)
- Feb ARR (Meta Reviewer: 2 ; Reviewers Soundness: 3, 4, 1.5; Reviewer Overall: 2, 3.5, 1.5)

Summary Of Changes:

1. The final meta-reviewer comment on our ACL 2025 submission acknowledged the strength and applicability of the work, and recommended its conversion to a short paper. The paper has been revised accordingly.

2. Feedback from the December ARR cycle highlighted the need for additional error analysis, ablation studies, and interpretability analysis. These elements are incorporated in the experiments and appendix.

3. The dataset used is no longer English-only raised as an issue in Feb ARR; it now includes five languages, allowing us to evaluate the model's generalizability across multilingual inputs. Model performance remained consistent, confirming robustness across languages.

4. The novelty of our work lies in identifying key limitations in existing state-of-the-art approaches—substantiated through prior literature—and addressing them through targeted architectural and representational improvements. These enhancements are validated by consistent improvements in regression-based metrics, demonstrating the effectiveness of our proposed techniques.

Official Review of Submission6827 by Reviewer Qhmc

Summary Of Strengths:

- This paper is relevant to the ACL fields of interest.
- Experimental results demonstrate significant improvements over the baseline.
- The use of interpretability tools (SHAP and LIME) helps the reader better understand the logic behind the model.

Summary Of Weaknesses:

- Overall, the method proposed in this paper appears to be a collection of sub-methods that collectively contribute to the state-of-the-art results. It is hard to find any real novelty in the paper.
- Experiments were conducted on one dataset only, which raises doubts about the framework's generalizability.
- The authors did not provide a detailed analysis comparing the computational complexity of HitMusicLyricNet against baselines. It remains unclear whether the performance improvement is justified for the extra computational cost.

Comments Suggestions And Typos:

- The SPD dataset used in the paper is English-only and limited to specific genres, which may limit generalizability. Future work could explore multilingual datasets or cross-lingual transfer.
- It would be beneficial to include qualitative examples or case studies showing how specific lyric features (e.g., sentiment, thematic content) contribute to the prediction.

Author Comment and Changes Incorporated:

1. Every architecture's foundation is built upon small building blocks, i.e., components developed separately in different domains for different purposes. One of the strong contributions of the paper is in identifying key shortcomings of the baseline architecture, HitMusicNet, with reference to the domain of music popularity prediction and its complexity. Building upon insights from previous architectures and literature studies, the authors propose two major architectural contributions: building an automated lyric feature extraction pipeline using LLMs, and carefully designing an auto-encoder to preserve directional information in embeddings.
   A detailed experimental study is presented on how different LLMs capture the semantic meaning of lyrics, highlighting the need for lyrics-understanding LLMs for further improvements. An in-depth experimental study is also proposed on various techniques of auto-encoders for reliable embedding compression. The second contribution involves redesigning the multi-modal deep learning architecture to properly encapsulate each modality to aid in music success prediction. These improvements and additions in architecture, motivated by domain knowledge and past literature, lead to significant baseline improvements. A detailed ablation study is also provided by the authors to highlight the contribution of various components of HitMusicLyricNet and data modalities in music success. The authors also present an in-depth study of model performance, error analysis, and interpretation analysis, showcasing HitMusicLyricNet's robustness while also

highlighting areas of sub-optimal performance and their reasoning, thereby opening doors for a clear direction toward further research.

2. In Lines 277–288, we discuss in detail our consideration of multiple open-source datasets to provide scores on them and highlight the issues in using them to train HitMusicLyricNet. The current dataset used, SPD, was built by highlighting various issues with previous open-source datasets and ensures diversity across genres, cultures, etc.
3. The major addition in computational complexity comes with the usage of LLMs and the lyrics feature extraction pipeline, whose computation details and cost of running are discussed in Section 4, Lines 398–418. Please let us know if further details are required.

Official Review of Submission6827 by Reviewer mLXF

Official Reviewby Reviewer mLXF20 Mar 2025, 20:38 (modified: 28 Mar 2025, 15:12)Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer mLXFRevisions

Summary Of Weaknesses:

1. While I understand that other datasets do not provide access to all modalities required for the proposed model I still think that it's worthy to look into assessing the model with less modalities on these datasets. It would strengthen the study by considering popularity beyond the Spotify's definition
2. Some other baselines should be included. A simple one using the metadata, but also other baselines from the literature which do not necessarily need all the modalities.

Comments Suggestions And Typos:

Author Comment and Changes Incorporated:

1. Thank you for your detailed feedback. As we discussed the various shortcomings of multiple open-source datasets for training HitMusicLyricNet in Lines 277–288, a point worth noting is that the majority of these datasets did not provide song lyrics, which directly removes one of our major contributions in the architecture, leading to ineffectiveness in showcasing results when incorporating them.
2. Secondly, all of these datasets provided the Spotify popularity score as the only metric for assessing music popularity success. There is strong reasoning associated with this, owing to the unavailability of other metrics publicly/open-source, which makes it difficult to obtain a robust popularity score.
3. The particular reason for not including other baselines was primarily that they do not combine modalities, which play a major role in music popularity prediction, as shown in the results and modality analysis. Hence, such comparisons would further raise concerns about their reliability. HitMusicNet, on the other hand, includes all modalities in predicting song success, making it ideal for comparison. Additionally, the HitMusicNet paper highlighted the effectiveness of the approach over previous baselines, helping us verify our improvements over the state of the art.

Official Review of Submission6827 by Reviewer QWsd

Summary Of Weaknesses:

1. The proposed method offers limited novelty, as its differences from the baseline approach are minimal. The results in Table 4 show only a marginal improvement over the baseline, and the absence of test set MSE results raises concerns about the impact of the approach.
2. The baseline method is evaluated on the original SPD dataset, whereas the proposed method is tested on a cleaned version (SPD*), making the comparison potentially unfair.
3. The overall presentation of the paper lacks clarity and structure. For example, in Table 3, both the maximum and minimum values in each column are highlighted in bold, leading to inconsistency in the presentation.
4. The meaning of the phrase "LLMs to extract MATHEMATICAL representations from lyrics" is unclear and requires further clarification.
5. This research does not appear to fall within the scope of natural language processing or computational linguistics.

Comments Suggestions And Typos:

1. Line 191: "Large Language Models" should be written in lowercase.
2. Line 338: "tranpose" should be corrected to "transpose."
3. Line 403: The phrase "for compute requirements" is grammatically unclear and needs revision.
4. Sections 3.2 (Baseline Methodology) and 3.3 (Dataset) describe background information rather than the proposed method. They would be more appropriately placed under Section 4 (Experiments and Results) for better organization.
5. Table 4 presents the main experimental results but appears later than Tables 2 and 3. To improve readability, it should be placed before them.

Author Comment and Changes Incorporated:

Starting with highlighting a few of the paper's main contributions and work areas:
The paper's main objective is to demonstrate the power of lyrics in music success prediction in comparison to other modalities (audio and social metadata). The work is motivated by the HitMusicNet architecture proposed in IEEE 2020 for a multi-modal deep learning-based architecture for music popularity prediction. The authors carefully identify bottlenecks in the HitMusicNet architecture, supported by literature review, and present HitMusicLyricNet with significant improvements over the baseline (compared to improvements proposed by previous studies in this domain). An automated lyric feature extraction pipeline along with a newly designed auto-encoder is presented that carefully preserves the dimensional information encoded in embeddings.
Authors carry out a detailed study on how different LLMs/LMs encode the semantic structure of lyrics, highlighting the need for domain-specific LLMs for carrying out lyrics feature extraction. The work also solves a major bottleneck in previous studies where multi-modality data was compressed through the same auto-encoder, leading to significant information loss for underrepresented modalities like metadata and lyrics features (statistical features).
A detailed experimental study is conducted on auto-encoder architecture for compression, LLMs for lyrics embeddings, and ablation study on the contribution of various modalities (presented in Appendix A). Detailed error analysis and interpretation analysis helped in assessing model strengths and weaknesses in music success prediction, providing a clear direction for the research community on future work and improvements.

We would also like to address the weaknesses mentioned:

1. Our work identifies major bottlenecks in previous approaches supported by literature review and provides improvements on those bottlenecks, along with the addition of newly designed components. The improvements of 9% and 20% are seen as significant in the field of MIR and the domain of music popularity prediction, as seen in the improvements posed by previous studies. The MSE and MAE were shown both for train and validation sets to demonstrate model robustness, and over that, only MAE Test was reported. We can surely help provide MSE Test scores for you.
2. Our experiment indicated that the results of training HitMusicNet on SPD vs. that of the SPD* dataset were almost similar, and in Section 3.3 we have highlighted our approach to obtaining the SPD* dataset, which shows no bias in selecting the set (~60%).
3. We understand your concern about the reporting of results in Table 3. But that was done deliberately to showcase overall which two configurations of LLMs we are taking further for our study. The bold was made across the row to demonstrate the high weightage given to test set values in comparison to train, given the high dimensionality nature of data.
4. The line "LLMs to extract mathematical representations ..." conveys that we have used LLMs encoder model forward pass to extract embeddings (mathematical vector representation), which is the standard process of extracting embeddings out of text data.
5. The research objective lies in the domain of NLP application and the sub-domain of multimodal application, which is of interest to ACL. This work showcases how we can carefully use NLP models to extract the deep semantic structure from lyrics and use it in music success prediction as compared to using statistical NLP techniques. Previous works in a similar domain submitted to ACL showcase the interest in such areas of work.
6. Section 3.2, as named Baseline Methodology, discusses the HitMusicNet architecture in detail, whereas the literature review on architectures used in music popularity is constrained to the section of Related Work. Section 3.3 discusses the SPD dataset in detail and data diversity and also the methodology in cleaning the SPD dataset.