Individual Final Report: Lyrics Genie Project

[Sanjana Godolkar]

1 Introduction

This report presents an overview of my contributions to the ”Lyrics Genie” project, an innovative endeavor employing NLP and machine learning to understand and predict song popularity. The project encompassed various aspects, from data preprocessing to model implementation and evaluation.

2 Description of Individual Work

My role primarily involved algorithm development and implementation. I focused on preprocessing data, extracting features, and applying machine learning models, particularly XGBoost for classification and BERT for NLP tasks. Additionally, I developed code for a Streamlit application, integrating our models for real-time song popularity prediction.

2.1 Failed Models and Challenges

- **Binary Classification Attempt**: Initially experimented with a binary classification approach for song popularity prediction.
- **Transformer Models**: Explored advanced Transformer-based models for NLP analysis.
- **Challenges with Feature Engineering**: Faced complexities in determining impactful features for song popularity.
- **Data Sparsity and Quality Issues**: Encountered issues with data sparsity, particularly in less popular songs.
- **Overfitting in Complex Models**: Struggled with overfitting in more complex models.

2.2 Streamlit Application Development

Integration of models into a Streamlit application for comparative analysis.

3 Detailed Contribution

- **Data Preprocessing and Feature Extraction**: Responsible for data cleaning, feature selection, and handling missing values.
- **Model Development and Evaluation**: Developed and assessed an XGBoost classifier for song popularity prediction.
- **Streamlit Application Development**: Contributed to a Streamlit application for real-time popularity prediction of lyrics.

4 Results

- The XGBoost model achieved a test accuracy of 36.72%.
- Successfully integrated predictive models into the Streamlit application.

5 Summary and Conclusions

This project highlighted the complexities of predicting song popularity using NLP and machine learning. The moderate accuracy of the model suggests future improvements, such as refining feature selection and exploring advanced NLP techniques. The Streamlit application demonstrated practical application in a user-friendly format.

6 Code Attribution

Approximately 73% of the code used in my contributions was adapted from online resources, with significant modifications and additions.

7 References

1. Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

3. Online resources and documentation for Python libraries (Pandas, Scikit-Learn, XGBoost, PyTorch, Transformers), and Streamlit.