Report for Assignment 0100
Introduction to Deep Learning

Fida Rahman
Ricardo Frumento
John Ha
Alvaro Montoya Ruiz
Department of Computer Science
College of Engineering
University of South Florida
Tampa, Florida

1. Introduction

Genre classification based on the lyrics of music is a challenging task even for humans. It is a current problem in the natural language processing field and the work done to tackle it has room to improve. The current project tries to find a simple model to reach acceptable accuracy compared to the novel solutions proposed by other authors. The intent is to give a good starting point for future work.

2. Related Work

Anna Boonyanit and Andrea Dahl [2] tried the classification of three genres using lyrics and got 68% accuracy using LSTM, a balanced dataset, and multiple correct genres. Alexandros Tsaptsinos [6] tried the same problem but with 20 genres and reached 49.5% accuracy using a hierarchical attention network. As shown using only lyrics is a hard problem so other sources are also considered. Rajanna et al [5] classification of nine genres tried using audio and got 38% of accuracy using a two-layer deep neural network. Finally, the best result was achieved by Li, Y. et al [3] for five genres using an input being a combination of audio and lyrics, the authors got 0.87 for the F1 score using their own model.

3. Model

3.1. Implementation

For the implementation discussed in this report, it was used Python and Google Colab to design, write, and test code for four different models: SimpleRNN, LSTM, GRU, and CNN. The architecture for all four is simple, all methods are composed of an embedding layer, a specific layer, a dropout layer, and a dense layer. The difference is on the specific layer and CNN has instead of a specific layer one 1-dimensional convolutional layer, one pooling layer, and one layer to flatten.

Genre	Country	Rock	Reggae
1	Know	Know	Like
2	Time	Time	Know
3	Heart	Come	Come
4	Come	Like	Time
5	Like	Away	Yeah

TABLE 1: Most Frequent Words per Genre

3.2. Dataset

The dataset used [4] is a compilation of song lyrics, genres, and annotations for sentiment analysis. It has 28372 lyrics distributed in 7 genres but for this project to achieve a balanced dataset it was decided to use country, rock, and reggae, and 2500 lyrics of each were used. It was also tested using the unbalanced dataset that used all the available lyrics for each genre. The average number of words per lyrics is 72.99 and the max length of words is 199. A bag of words was produced to understand the dataset better. It shows, once again, the size of the challenge. Table 1 with each genre and its five most frequent words.

Table 1 shows that the most frequent words are not what the models should be looking for as features because they show up in all the genres.

The preprocessing of the data consisted of dropping the unused genres and the unnecessary sentiment analysis columns. The embedding was performed using GloVe [1].

For the test, train, and validation split it was used 10% for testing and 20% of the remaining data for validation.

4. Description of Analysis and Results

4.1. Hyperparameter Optimisation

The multiple runs to find the optimal parameters used four different models and several configurations. The following table shows the results. L1 (0.01) and L2 (0.01) regularization were used in all the trials and the dropout rate

was kept constant at 60% because overfitting was a problem from the beginning. The trials were run for 20 epochs, the batch size used was 32, the RMSProp optimizer was used also.

As a baseline, a neural network using an LSTM layer and a dense layer was used with all the layers frozen. The baseline used a balanced dataset, with a learning rate of 0.001.

Finding Best Hyperparameters				
Model	Dataset	Learning Rate	Training Accuracy	Validation Accuracy
Baseline	Balanced	0.001	40.51%	44.13%
SimpleRNN	Balanced	0.001	77.81%	56.33%
LSTM	Balanced	0.001	53.62%	49.47%
GRU	Balanced	0.001	32.84%	32.60%
CNN	Balanced	0.001	33.16%	32.67%
SimpleRNN	Unbalanced	0.001	74.76%	61.10%
LSTM	Unbalanced	0.001	59.40%	60.77%
GRU	Unbalanced	0.001	45.06%	47.08%
CNN	Unbalanced	0.001	45.06%	47.08%
SimpleRNN	Balanced	0.0001	86.05%	36.80%
LSTM	Balanced	0.0001	51.40%	52.07%
GRU	Balanced	0.0001	32.66%	32.60%
CNN	Balanced	0.0001	33.41%	32.13%
SimpleRNN	Unbalanced	0.0001	50.45%	51.67%
LSTM	Unbalanced	0.0001	48.66%	49.87%
GRU	Unbalanced	0.0001	80.08%	53.55%
CNN	Unbalanced	0.0001	57.51%	47.54%

TABLE 2: Results for the hyperparameter tuning

4.2. Quantitative Results

After testing the best combination of hyperparameters was to use LSTM with a learning rate of 0.001, and the unbalanced dataset. With these parameters, five full runs were made and the results are shown below. 50 epochs were used for the trials.

Trial	Training Accuracy	Validation Accuracy	Testing Accuracy
1	62.14%	62.94%	63.10%
2	60.30%	58.10%	58.18%
3	60.77%	63.15%	63.18%
4	58.16%	52.50%	53.00%
5	60.37%	56.89%	55.25%
Average	60.34%	58.71%	58.54%
Standard Error	1.429	4.468	4.581

TABLE 3: Average results for the best hyperparameters setting

4.3. Analyses

The plot for training and validation accuracy and loss for the first run is shown below with the confusion matrix for the same run.

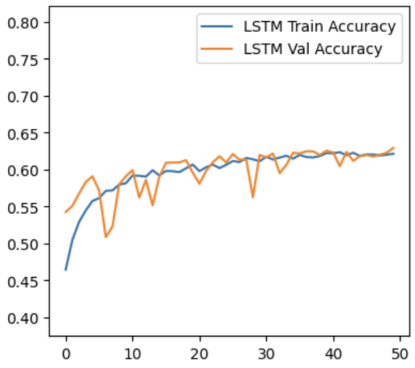

Figure 1: Accuracy for training and validation

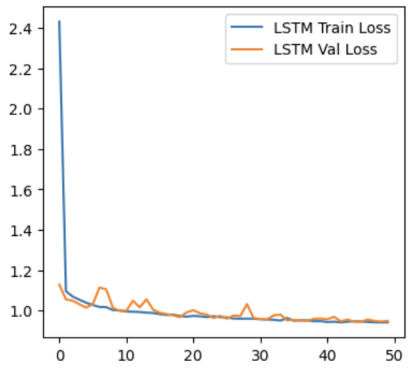

Figure 2: Loss for training and validatiom

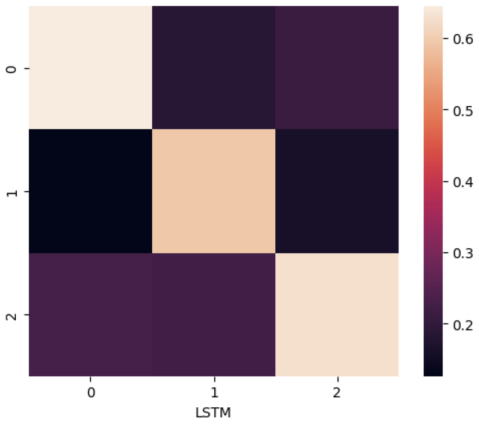

Figure 3: Confusion matrix

One other metric used was the F1 score. As the score was between 0.5 and 0.8 it shows a decent model, showing there is still work to be done to improve the efficiency of the model but the current state is statistically acceptable.

$$F_1 = 0.631$$

5. Conclusions

This project was able to meet the current results showed by the best model presented that only used lyrics. One difference was that the proposed architecture uses only one correct genre instead of multiple like [6]. The results presented show how the problem is hard and perhaps requires much more attention from the community.

References

[1] Jeffrey Pennington et al. "GloVe: Global Vectors for Word Representation". In: (2014).

[2] Andrea Dahl Anna Boonyanit. "Music Genre Classification using Song Lyrics". In: *Stanford CS224N* (2022). DOI: https://web.stanford.edu/class/cs224n/reports/final_reports/report003.pdf.

[3] Ding et al. Li Zhang. "Music genre classification based on fusing audio and lyric information." In: *Multimed Tools Appl 82, 20157–20176* (2023). DOI: https://doi.org/10.1007/s11042-022-14252-6.

[4] Luan et al. Moura. "Music Dataset: Lyrics and Metadata from 1950 to 2019." In: *Mendeley Data* (2020). DOI: 10.17632/3t9vbwxgr5.2.

[5] Arjun Raj Rajanna et al. "Deep Neural Networks: A Case Study for Music Genre Classification". In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. 2015, pp. 655–660. DOI: 10.1109/ICMLA.2015.160.

[6] Alexandros Tsaptsinos. "LYRICS-BASED MUSIC GENRE CLASSIFICATION USING A HIERARCHICAL ATTENTION NETWORK". In: *Stanford* (2017). DOI: https://arxiv.org/pdf/1707.04678.pdf.