Review Text to Review Number Prediction

Shruti Ghargi
Department of Computational Linguistics
Heidelberg University, Germany
shruti.ghargi@stud.uni-heidelberg.de

1 Introduction

In the realm of music album reviews, understanding the sentiments and opinions expressed by users is of paramount importance. Predicting numerical ratings from textual reviews can provide valuable insights into the quality and reception of music albums. A novel neural network method is introduced for review number prediction by taking user information into account. The report explores the process of leveraging Natural Language Processing (NLP) techniques and deep learning models to predict review numbers from review text.

2 Dataset

The dataset used for this project is "music_album_reviews.csv." The dataset comprises 80,000 music album reviews authored by users of rateyourmusic.com. It was obtained through web scraping in May 2022, and it encompasses a total of 79,922 album reviews, including associated numerical ratings where available. For reference, the album chart containing the albums from which these reviews were gathered can be accessed through the "provided".

Split	#Text
Train	60000
Test	20000

3 Pretraining

Data Pre-processing:

3.1 Handling Missing Values:

The dataset contained a few review rows without their corresponding numerical ratings and certain rating rows with missing review text. Using the

3.2 Text Cleaning:

The process of text cleaning played a pivotal role in refining the textual content of the dataset. It involved the following key actions:

- Removing Punctuation: Punctuation marks, often extraneous to the core text, were scrupulously expunged to streamline the content.
- Removing Numbers and Links: Numeric values and hyperlinks present in the reviews were systematically eradicated to ensure a focus on the textual essence.
- Removing URLs: Any embedded URLs, that could introduce noise or artifacts, were meticulously eliminated.
- Resetting Index: To ensure consistent and accurate data handling, reset the index of the DataFrame. This maneuver was performed to mitigate any potential index-related inconsistencies. As a result, the DataFrame dimensions were

uniformly set to [78,162 rows x 2 columns].

3.3 Tokenization and Stopword Removal:

Harnessing the capabilities of the Natural Language Toolkit (NLTK), embarked on tokenization – a process of breaking down the reviews into individual words. Concurrently, common stopwords, which contribute little to the essence of the text, were systematically pruned from the tokenized words. Furthermore, all tokens were harmoniously converted to lowercase to facilitate uniformity and avoid case-related disparities.

3.4 Lemmatization:

To standardize word forms and reduce words to their base or lemma forms, NLTK's WordNetLemmatizer is employed. This lemmatization process was instrumental in simplifying the dataset's vocabulary, ensuring greater consistency in word representations.

3.5 Word Clouds:

Visualizing the dataset's content is often illuminating. To this end, word clouds are generated – graphic representations of the most frequent terms in the lemmatized reviews. These word clouds provided an insightful overview of the dataset's textual composition. Additionally, distinct word clouds for each rating value were produced, ranging from 0.5 to 5, allowing us to visually explore sentiment distributions within the reviews.

3.6 Custom Stopwords Removal:

Recognizing the specific nature of music reviews, custom stopwords that were unique to this domain were removed. Words like "album," "song," "CD," and others of similar like were systematically excluded from our analysis. Following this, the top 10
most frequent tokens that emerged after the custom stopwords were removed. This analysis provided valuable insights into the dataset's underlying content, shedding light on prevalent themes and expressions within the reviews.

In summary, our journey through meticulous data preprocessing encompassed a comprehensive suite of actions, meticulously executed to prepare the dataset for subsequent model training. These steps collectively enhanced the data's quality, usability, and relevance, setting the stage for robust and insightful analyses.

- To enhance the model's understanding of the review text, pre-trained GloVe word embeddings were incorporated.
- GloVe embeddings were loaded into the project to map words to vectors.
- A tokenizer was initialized to convert text data into sequences.
- The review text was tokenized into sequences and padded to a fixed length to prepare it for the model.
- Each word in the tokenized text was mapped to its corresponding GloVe vector, enabling the use of pre-trained embeddings as features.

4 Algorithms:

The heart of the project lies in the application of deep learning algorithms, specifically Convolutional Neural Networks (CNNs).
enhance accuracy and predictive capabilities. This project underscores the significance of NLP and deep learning in extracting valuable

4.1 CNN Model Architecture:

- The model consisted of an embedding layer, a convolutional layer, max-pooling, dense layers, dropout, and an output layer.
- Hyperparameters such as embedding dimension, batch size, and number of epochs were adjusted during model training.

5 Results and Decision:

After training the CNN model, the results were analyzed.

5.1 Model Performance:

- Train Loss: 0.8431
- Train Accuracy: 0.7147
- Test Loss: 1.5387
- Test Accuracy: 0.4148

The model demonstrated reasonable training and validation accuracy as shown in figure 1 and 2; however, there is room for improvement in the test accuracy. Further experimentation with hyperparameters, model architecture, and data preprocessing techniques may yield better results.

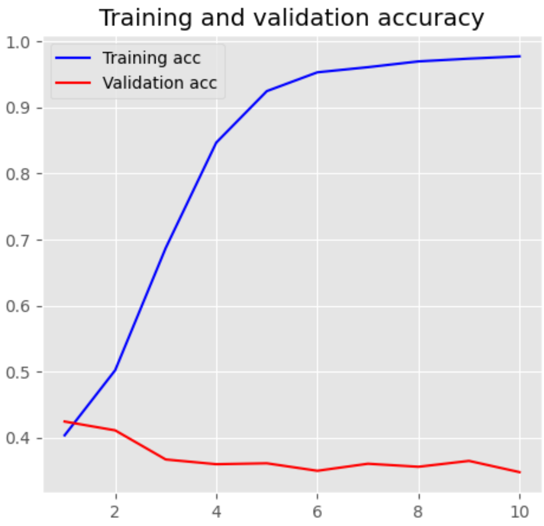

Figure 1

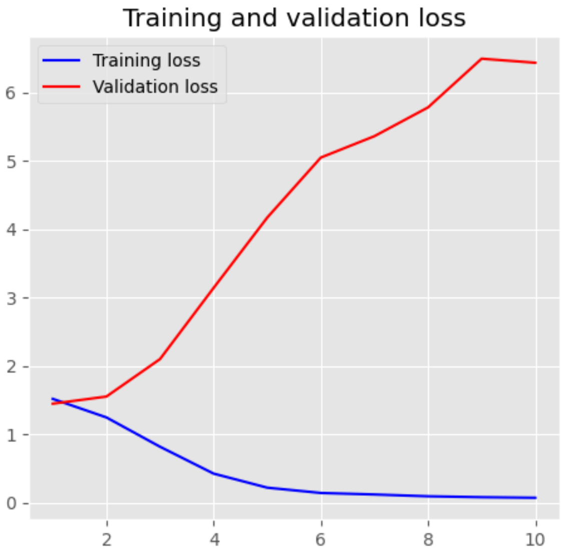

Figure 2

6 Conclusion:

This report delved into the task of predicting numerical ratings from textual music album reviews. By following a systematic data preprocessing pipeline, leveraging pretrained GloVe embeddings, and utilizing a CNN model, gained insights into the challenging problem of review rating prediction. While the initial results show promise, ongoing refinement and optimization of the model are essential to

insights from textual data in the context of music album reviews.

7 References:

[1] Shapira, Oren. "Using Neural Networks and NLP Word Embedding to Predict Amazon User Review Ratings." *Medium*, 23 October 2022, https://medium.com/mlearning-ai/using-neural-networks-and-nlp-word-embedding-to-predict-amazon-user-review-sentiment-28156f69e1e1.

[2] Jane, Ryon. *YouTube*, 30 August 2022, https://github.com/omshapira/Amazon_star_rating_predictions/blob/master/Project_Amazoz_Reviews.ipynb.

[3] Brownlee, Jason. "How to Use Word Embedding Layers for Deep Learning with Keras - MachineLearningMastery.com." *Machine Learning Mastery*, 2 February 2021, https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/.

[4] Janakiev, Nikolai. "Practical Text Classification With Python and Keras – Real Python." *Real Python*, https://realpython.com/python-keras-text-classification/#using-pretrained-word-embeddings.

[5] "GloVe: Global Vectors for Word Representation." Stanford NLP Group, https://nlp.stanford.edu/projects/glove/.