Review Score Prediction Project: Scientific Report

Aksa Fatima
Faculty of Electronics and Information Technology
Warsaw University of Technology
Email: aksa.fatima.stud@pw.edu.pl

***Abstract*—This paper aims to develop a machine learning model to predict Amazon Music review scores based on the textual content of the reviews. We frame the task as a regression problem, with scores ranging from 1 to 5. This report provides an overview of the problem, discusses the dataset, describes preprocessing, justifies the evaluation strategies, and explains the selected classification algorithms.**

I. Introduction

Review score prediction is important for e-commerce platforms, as it allows businesses to gauge customer sentiment and predict product ratings. This project focuses on predicting Amazon Music review scores based on the content of the reviews.

II. Problem Overview and Existing Methods

Predicting review scores from text involves various machine-learning methods:

- Linear Regression: A simple baseline approach to predict scores.

- Random Forests: An ensemble method that handles complex relationships and assesses feature importance.

- Neural Networks: Advanced models like Long Short-Term Memory (LSTM) networks or transformers, which can process text data effectively.

These methods vary in complexity and performance. In this project, we compare some of these approaches to determine the best method for review score prediction.

III. Dataset Description

The dataset used is the "Amazon Music Reviews" from Kaggle. It contains text-based reviews and corresponding review scores. Given the raw nature of the data, preprocessing is required.

A. Data Preprocessing

To prepare the data for modeling, several preprocessing steps are needed:

- Text cleaning: This involves removing special characters, converting to lowercase, and possibly removing stopwords.
- Tokenization and vectorization: Converting text to numerical representations using techniques like TF-IDF or word embeddings (e.g., Word2Vec or GloVe).
- Handling missing data: Check for and handle NaN values.
- Consistent sequence length: For neural network models, padding or truncating is used to ensure a uniform sequence length.

B. Data Balance

An important aspect of data analysis is checking for data imbalance. If the review scores are not evenly distributed, techniques like resampling or weighted loss functions may be required.

C. Data Splitting

To train and evaluate the model, the dataset is typically split into training, validation, and test sets. A common split is 70% training, 15% validation, and 15% test.

IV. Performance Evaluation

The following metrics are used to evaluate the performance of the regression models:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

These metrics help gauge the accuracy of the model's predictions. Performance evaluation should be conducted on a test set to ensure generalization.

V. Solution Description

Given the problem of predicting review scores from text, several machine-learning methods can be employed:

- Linear Regression: A baseline model that provides a simple approach to predicting scores.

- Random Forests: An ensemble method that can capture complex relationships and feature importance.

- LSTMs (Long Short-Term Memory): A type of Recurrent Neural Network (RNN) that is useful for sequential text data.

These methods require different preprocessing steps, particularly in tokenization and sequence length. This section describes the advantages and disadvantages of each method and explains the adjustments made to optimize the dataset for each approach.

VI. Conclusion

In this report, we presented a comprehensive approach to predicting Amazon Music review scores based on text content. We discussed the problem, existing methods, dataset preprocessing, performance evaluation, and solution description. The results from different methods can guide future improvements and refinements in the review score prediction process.