SDSC4016 Fundamental of Machine Learning II

City University of Hong Kong



SDSC4016 Final Project

Instructor:

Xiangyu ZHAO

**Amazon Digital Music Product Review Sentiment Classification and Time Series Prediction Using LSTM Deep Neural Network and ARIMA**

Garifullin Aidar 56443506

Pattaravarodom T-touch 56325772

Law Sau Ching 56656131

Date: 20-04-2023

**Part 1: Motivation**

Recommender systems have become increasingly prevalent in recent years, with sentiment analysis being crucial for such systems. However, the mismatch between a customer's review and rating poses a significant challenge to recommendation systems, as individual ratings compute the average rating. As such, predicting a user's rating for those items that were not rated yet by the Product Rating Predicting system is essential to improve the accuracy of the recommendation system.

Sentiment classification can help businesses understand customer opinions about their products, including digital music products. By analyzing customer reviews and comments, businesses can gain insights into customer satisfaction levels, identify areas for improvement, and adjust their marketing strategies accordingly. Gaining insights into customer satisfaction and identifying areas for improvement in their products or services can lead to better product development and increased customer loyalty, resulting in higher profits for the business. From the customers' viewpoint, it can lead to better customer retention and loyalty and increased sales through positive word-of-mouth marketing.

Furthermore, identifying the demand season for a product is critical in the market, as it can help businesses develop effective strategies to address seasonal fluctuations in demand. Using the rating data to identify the season of high ratings, we can forecast the high rate season and summarize which season has the best rating, implying the seasonal demanding market movements.

Time series analysis can provide valuable information for businesses better to manage their inventory, production, and pricing strategies. For example, businesses can use time series analysis to forecast product demand, identify seasonal trends, and optimize production and inventory management accordingly. This can lead to better cost management, improved customer satisfaction, and increased revenue for the business.

Additionally, time series analysis can provide valuable insights into market trends and customer preferences, helping businesses to adjust their product offerings to meet customer needs better. It can also be helpful for customers in predicting future trends and making informed purchasing decisions. Time series analysis can help them to make more informed decisions when shopping online, leading to increased satisfaction and overall customer experience.

Therefore, this project focuses on developing a deep learning approach for sentiment classification and time series prediction of Amazon digital music product reviews. The project aims to predict a user's rating for items not yet rated by the Product Rating Predicting system and forecast the demand for the product by identifying the season of high ratings in the year. The problem is crucial in recommender systems and market analysis and can significantly affect businesses regarding revenue and market share.

**Part 2: Background**

Sentiment analysis is a common task in natural language processing that involves computationally identifying and categorizing the sentiment expressed in a text. In recent years, sentiment analysis has become increasingly crucial for identifying customer satisfaction and dissatisfaction from reviews and social media posts. One application of sentiment analysis in the industry is forecasting market movements based on sentiment expressed in news and blogs.

Previous research on sentiment analysis in Amazon product reviews has focused primarily on predicting the rating of a product/venue based on the sentiment expressed in user reviews. In addition, researchers have also explored ways to overcome the mismatch between a customer's review and rating, as well as visualizing customer opinions in the form of charts based on sentiment analysis. Machine learning techniques have been applied to detect unfair comments left by users through sentiment analysis.

For instance, Choi and Cardie (2009) applied a hierarchical method to categorize adjectives, adverbs, and noun phrases with scores for domain-specific sentiment classification. Raut and Londhe (2014) focused on overcoming the mismatch between the review and the rating and visualizing customer opinions in charts. Elmurngi and Gherbi (2018) used machine learning techniques for unfair review detection on Amazon reviews using sentiment analysis.

In brief, sentiment analysis in Amazon product reviews has garnered significant attention from researchers, as it can potentially provide valuable insights for both businesses and consumers.

**Part 3: Description**

Sentiment analysis of text review data and time series analysis using the ARIMA method was the project's two main responsibilities;

For the sentiment classification challenge, we used the *Bidirectional-additive-attention-LSTM* model, an improved variant of the conventional LSTM technique that can handle sequential input in both directions. The Bidirectional additive attention LSTM is based on a Bidirectional LSTM (BiLSTM) architecture that can process sequential data in both directions. The network can capture past and future context for each element in the input sequence. It consists of two LSTM layers that process the input sequence in forward and backward directions, making the model capture more comprehensive and accurate information about the sequence than a traditional LSTM network. The two LSTM layers that make up the Bidirectional-Additive-Attention-LSTM process the input sequence in both forward and backward directions, allowing it to collect both past and future context for each element in the input sequence.

Additionally, the model employed an attention mechanism that helped the network to focus on essential parts of the input sequence when making predictions. More than that, the additive attention mechanism helps the network to focus on the most critical parts of the input sequence when making predictions. It computes a weighted sum of the output of each BiLSTM layer based on a learned attention vector.

On the other hand, we utilized the *ARIMA* model to analyze time series data. ARIMA analyzes data with known underlying structures, such as seasonal variations or trends. The ARIMA model used various methods, such as the Augmented Dickey Fuller Test (ADF Test), to identify trends in time series data, analyze autocorrelation and partial autocorrelation functions, and predict future values based on historical data. By employing the ARIMA method, we were able to identify patterns and trends in the customer review data, which can help businesses to optimize their resource allocation and minimize waste.

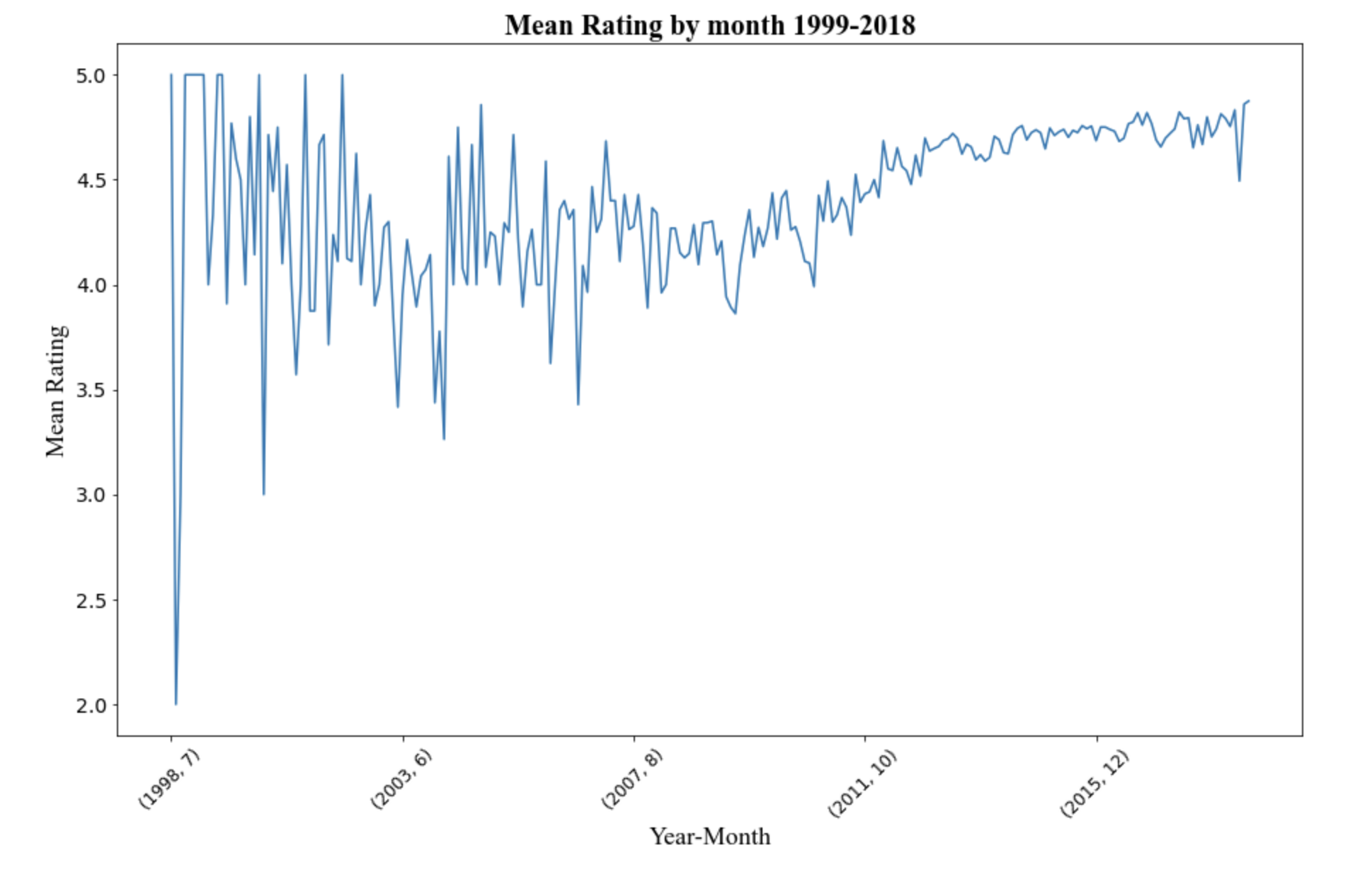
**Part 4: Data**

In conducting the data analysis for our selected task, we explored several key statistics to gain insights into the nature of our data. Our dataset comprises 146,149 unique reviews, with only 146 reviews containing NaN values in the *‘review text’* column. Therefore, we had 146,003 valid reviews that we could work on.

We calculated the overall mean rating of the reviews to be 4.6934, indicating that most reviews in our dataset were positive. The overall standard deviation of the ratings was 0.7212, suggesting considerable variation in the ratings provided by users.

To gain insights into the time series pattern of our data, we further examined the monthly mean ratings and standard deviation. The mean monthly ratings were found to be 4.3962, with a standard deviation of 0.4027. We also observed that the monthly standard deviation for the first 10 years (1998-2008) was 0.4632, which was higher than the monthly standard deviation for the last 10 years (2009-2018), which was found to be 0.2315.

After plotting the time plot of mean rating by month, it became clear that the data is suitable for time-series analysis since it shows significant fluctuations with time (See Fig. 1).



*Fig. 1: Mean Rating by month 1999-2018.*

Figure 1 shows frequent upward and downward trends, with an overall increasing pattern in the mean rating from 1999 to 2018.

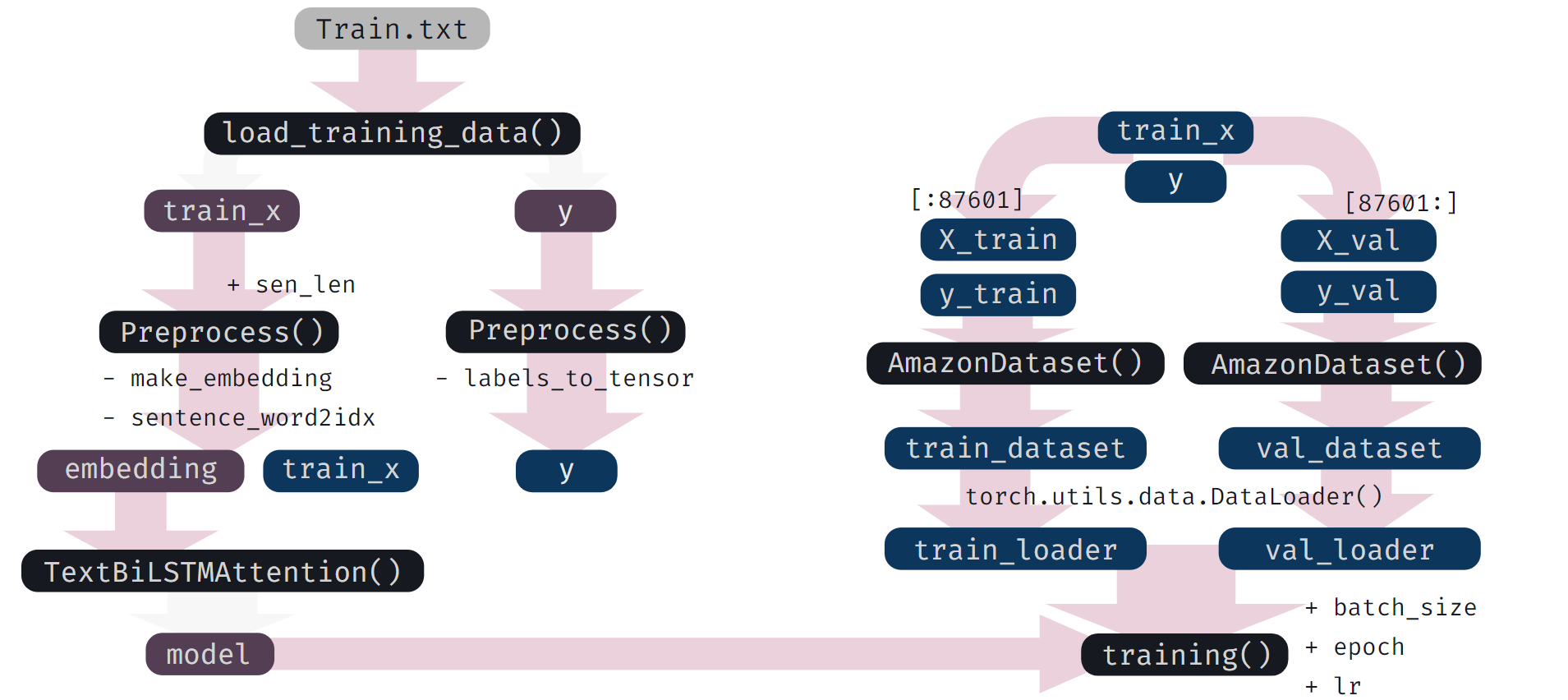
Furthermore, we discovered that only 4 months in 1998 and 5 months in 1999 had valid reviews, while all others had reviews available. This finding suggested that our dataset was more complete in more recent years. We also observed that the mean ratings for the reviews were more fluctuating for the first 10 years compared to the last 10 years, indicating a possible change in user behavior or the product itself over time.

**Part 5: Implementation**

For the first task, after we preprocessed the data, we modeled the data using *Bidirectional-additive-attention-LSTM* for the sentiment analysis; the key variables and functions to represent the main workflow and structure of the codes are utilized, as seen in Table 1. For the activation function in the hidden layer, which consists of the hidden attention layer, we use Tanh and Softmax for the output attention layer. But we decided to use the Sigmoid function for the output classifier layer.

**Table 1: Parameters and Activation function value set up**

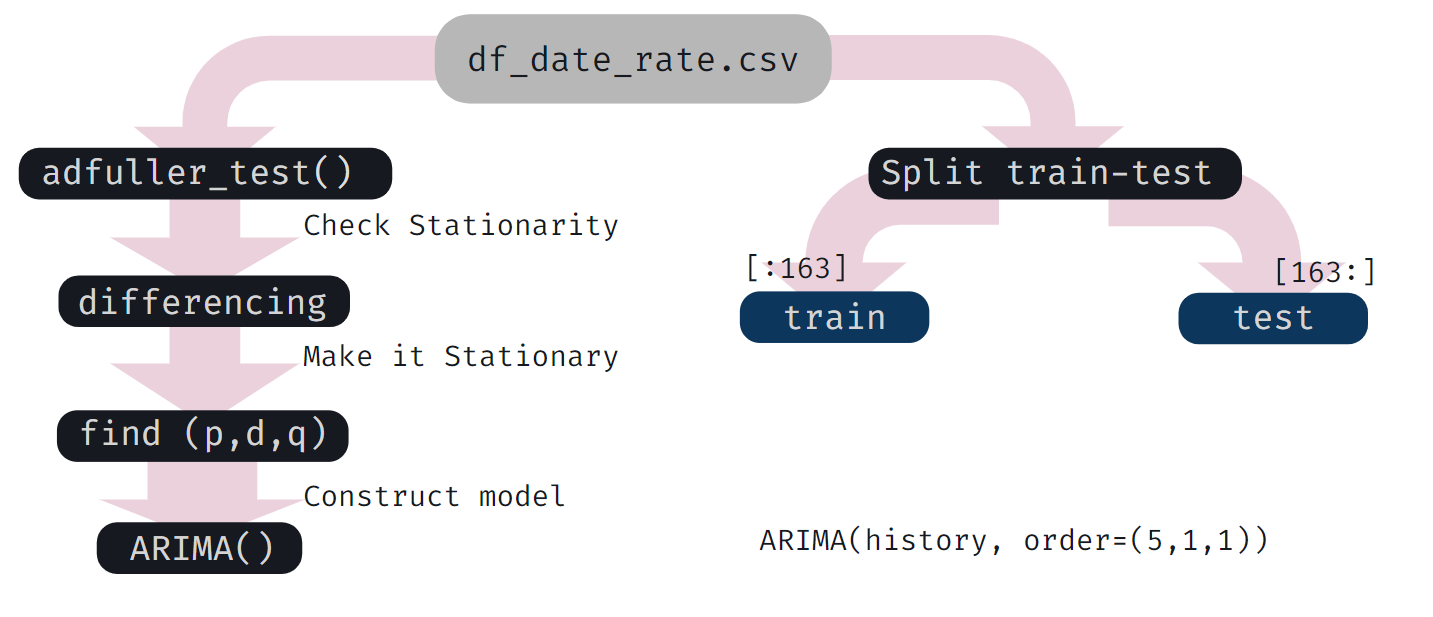
| **Parameters** | **Values** |  | **Activation layer** | **Value** |
| --- | --- | --- | --- | --- |
| **sen\_len** | 60 |  | Hidden Attention Layer (Hidden) | Tanh |
| **fix\_embedding** | True |  | Hidden Attention Layer (Output) | Softmax |
| **batch\_size** | 64 |  | Output Classifier | Sigmoid |
| **epoch** | 50 |  |  |  |
| **Learning rate** | 0.001 |  |  |  |
| **optimizers** | Adam |  |  |  |
| **weighted avg** | 0.79 |  |  |  |



*Fig. 2: Workflow of Task 1 implementation*

As can be seen in the workflow in figure 2, the train text data is initially plugged-in into the loading function to output the X-input-attribute and y-target-attribute. After that, the function preprocess makes a word embedding, keeps in the variable embedding, converts the sentence of text into a sequence of word indices, keeps in the train\_x variable, and turns the y attribute into a tensor. The embedding matrix will then be used to construct the Neural Network function Bidirectional-attention-LSTM as a model. Before training, we split the dataset into train and validation sets for X and y and passed them through the AmazonDataset function to return a trainset and validation set.

Eventually, the model, training set, validation set, and all the hyperparameters are used in a training function to start the training process. And repeat the same workflow for the testing dataset, just changing the name and the function.



*Fig. 3: Workflow of task 2 implementation*

The ARIMA model was implemented for the second task. Figure 3 shows the general workflow. Once the mean rating was grouped by month for every year and the corresponding data frame was created, the average rating column was put through an augmented Dickey-Fuller test (ADF) (*short - Adfuller test*) to check for stationarity. Since the data was found to be non-stationary, this led to the next step - differencing, which aimed to ensure the stationarity of the data for further analysis. The first differences seemed sufficient for this purpose.

The next step concerned finding the right tuning of parameters for our ARIMA model, that is, finding the right p - the lag order, d - the degree of differencing, and q - the order of the moving average. Autocorrelation and partial autocorrelation were the means of identifying those. Having dealt with the necessary data preprocessing for time series, the latter was ready to be put through the ARIMA model. This meant splitting the data into training and testing sets, running the ARIMA model with its optimal parameters on the training set, and assessing the accuracy of the resulting model by finding the mean squared error between the predicted and the actual values of mean rating in the testing set.

**Part 6: Results and Observations**

**Table 2: Classification Report and task 1**

| **Label** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **0** | 0.48 | 0.55 | 0.51 | 6117 |
| **1** | 0.88 | 0.84 | 0.86 | 23084 |
| **macro avg** | **0.68** | **0.70** | **0.69** | **29201** |
| **weighted avg** | **0.79** | **0.78** | **0.79** | **29201** |

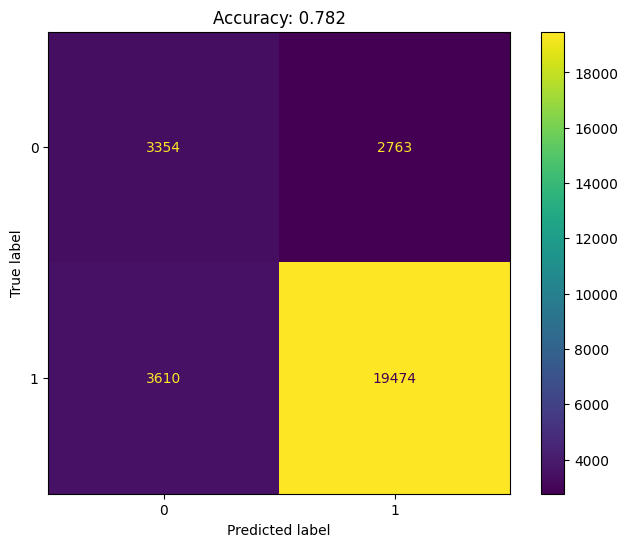
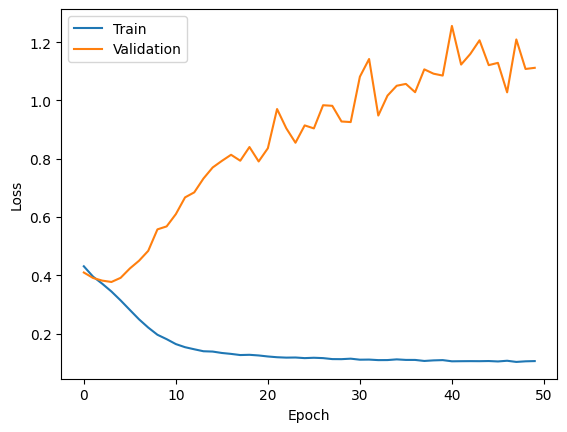
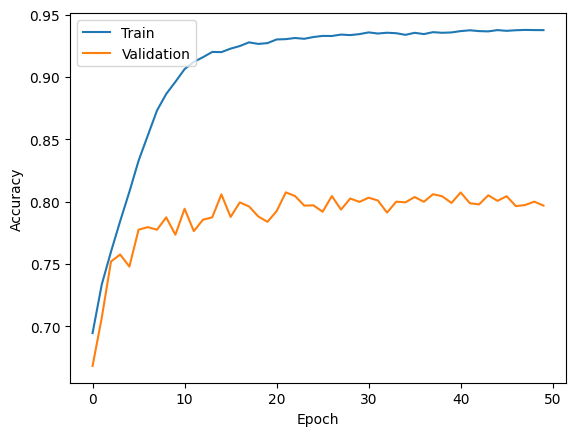
The result from classification in the sentient analysis task was quite promising at the accuracy rate around 78 percent accurately predicted, but after having a look at the confusion matrix and classification report in table 2 and figure 4, we found that the classification is considerably biased toward a single class which is the *‘good rating’* class since the number of data point in the class is significantly imbalanced.

Figure 5 revealed that the validation accuracy reached a stable value of around 80 percent after 10 epochs. While the training accuracy reached a stable value of around 93 percent, indicating a significant overfitting for the model. The loss of validation also gradually increases by epoch. While the loss of training set plummeted below 0.2 since the first 10 epochs

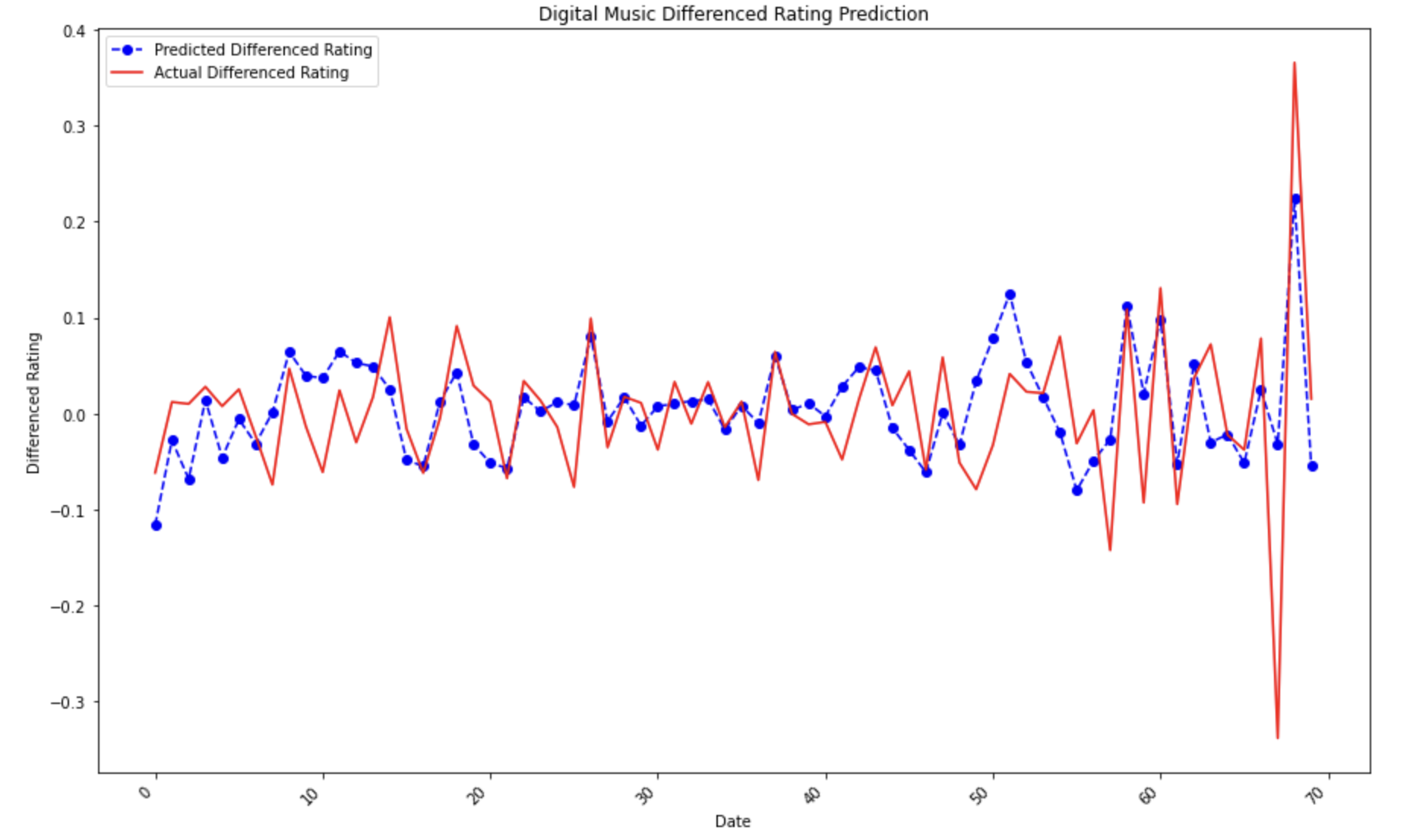


*Fig. 5: Accuracy Plot and Loss Values Plot of task 1*

For the time series analysis, analyzing trends over the two decades can predict how certain digital music products will likely be rated in the future. The results of our time-series analysis are summarized here.

The autocorrelation and partial autocorrelation plots showed that the significant lags are 1 and 5. Hence, these values were put into our model for the first trial. Later the auto\_arima function was utilized to find the optimal order based on BIC and AIC criteria. The resulting values for p, d, and q are (5, 1, 1). Therefore, the model was trained using this order.

The training results are promising, considering the low MSE of 0.00436 and low MPE of 1,892. The graph of predicted values for the mean rating was plotted together with the actual values in the testing set (Fig. 6). As it follows from Fig. 6, there are cycles that repeat around every 2-3 months with characteristic growth and fall patterns.



*Fig. 6: Digital Music Rating Prediction vs. Actual Rating.*

**Part 7: Discussions**

The results and observations from the sentiment analysis can be discussed from various perspectives, considering the need for the development of the recommender system with sentiment analysis;

From the *Deep Learning Architecture perspective*, the Bidirectional additive-attention LSTM has already outperformed this classification task despite a very imbalanced dataset, without balancing the data, the model is likely to accurately predict class 1 or ‘good rating’ class. But developers can consider a more advanced architecture like Transformer or Network Pruning to remove the computational cost and improve efficiency.

From the *Data Collection Perspective*, increasing the number of data can also improve the data imbalance and accuracy, especially the imbalanced class. Another way is to balance the data by the undersampling method. However, each data balancing method required a trade-off consideration.

After knowing this model result, Amazon can also apply the *future business strategy* of adjusting the purchasing demand of some Digital Music to put on sale. For example, Amazon uses the sentiment predicting score to improve the efficiency of customers’ purchase decisions and fill the unrated review in the future without requiring the customer to rate.

The limitation of the model is mainly the vulnerability to overfitting since the model becomes too closely adapted to the training data and fails to generalize well to new data.

For the time series analysis, the model seems to work well on the rating data of digital music from 1998 to 2018. The visualization in Fig. 6 shows that the trained model closely follows the actual values of the mean rating in the testing set, proving that the ARIMA time-series model is well-fit for modeling the average rating of the digital products by the users. The results and observations can also be discussed from various perspectives;

From the *Music Industry Perspective*, ARIMA might as well be utilized by the music industry to know when to release their songs to increase the probability of receiving a high rating on their product. The music production industry can better tailor its marketing or product release to meet customers’ demands.

From the *Technology Development Perspective,* we extracted the insights that from 1998-2008, the technology for online shopping and review platforms were still in its early stages. The development of music streaming services has made it easier for consumers, as seen from the results.

From the *E-commerce Perspective,* for now, and in the future, more people will begin to use the Internet for shopping and sharing their opinions, increasing the development of the E-commerce industry. Still, they may, in turn, exploit the traditional commercial industry. Therefore, humans must be aware and adaptive to the new technology in the shopping and commercial industry.

The limitation of this task is that a simple ARIMA model might not be accurate months and especially years ahead. Better models, such as Time-Series LSTM neural networks, could be used to improve prediction accuracy further. However, we must consider the computational capability that may be a trade-off with the time series prediction using Times-Series LSTM.

**Part 8: References**

Choi, Y., & Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*. https://doi.org/10.3115/1699571.1699590

Elmurngi, E. I., & Gherbi, A. (2018). Unfair reviews detection on Amazon reviews using sentiment analysis with supervised learning techniques. *Journal of Computer Science*, *14*(5), 714–726. https://doi.org/10.3844/jcssp.2018.714.726

McAuley, J. (n.d.). *Digital\_Music\_5.json.gz* [Data set]. Retrieved from

https://jmcauley.ucsd.edu/data/amazon\_v2/categoryFilesSmall/Digital\_Music\_5.json.gz

Ni, J., Li, J., & McAuley, J. (2019). *Justifying recommendations using distantly-labeled reviews*

*and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 188–197). Association for Computational Linguistics.

Raut, V. B., & Londhe, D. D. (2014). Opinion mining and summarization of Hotel Reviews. *2014 International Conference on Computational Intelligence and Communication Networks*. https://doi.org/10.1109/cicn.2014.126

Shrestha, N., & Nasoz, F. (2019). *A Comparative Study of Deep Learning Models for Sentiment*

*Analysis of Short Texts* [PDF file]. Retrieved from https://arxiv.org/ftp/arxiv/papers/1904/1904.03705.pdf