Leonell Puso
ID# 302164083
CSC 180-01 Intelligent Systems (Fall 2023)

REPORT

(1) **Problem Statement:**

In this project, a prediction of how many stars a business will get on Yelp, based on what people write in their reviews, will be tackled. We want to create a program that can guess that rating just by reading what customers have said in their reviews.

The Challenge:
Imagine if you could teach a computer to read hundreds of thousands of reviews and then make an educated guess about how good a business is. That's what we're trying to do! But it's not easy because computers don't naturally understand words like we do. We have to turn those words into something the computer can work with.

The Plan:

1. First, we need to tidy up our data. Ensure that only businesses that have at least 20 reviews are in the dataset.
2. Convert all the reviews into numbers using a special trick called TF-IDF. This trick helps the computer understand which words are important in making predictions.
3. Predict star ratings of each business by training models.
4. See how the model's prediction compares with the ground truth.
5. Experiment with different settings for our model. For instance, how many layers it has and how many neurons it uses. This in turn will make our predictions as accurate as possible.

(2) Methodology

1. Data Cleaning:

- Drop unnecessary features
- Delete any empty rows
- Ensure that only businesses with at least 20 reviews are in the dataset

2. Feature Preparation:

- Aggregated the words in a review together inside each business.
- Convert the words in each review into nump arrays by using TFIDF (Term Frequency-Inverse Document Frequency) encoding.

3. Model Building:

- Create neural network models by using TensorFlow

4. Hyperparameter Tuning:

- Apply activation functions relu, sigmoid, and tanh
- Use different amounts of layers and neuron count for each layer
- Change optimizers like adam and sgd

5. Training and Validation:

- Use EarlyStopping to detect and prevent overfitting. Use this technique in five iterations but only keep the best model.

6. Evaluation and Analysis:

- Evaluate the model's performance using a metric called RMSE (Root Mean Squared Error). This score helps us understand how closely our predictions align with actual ratings.

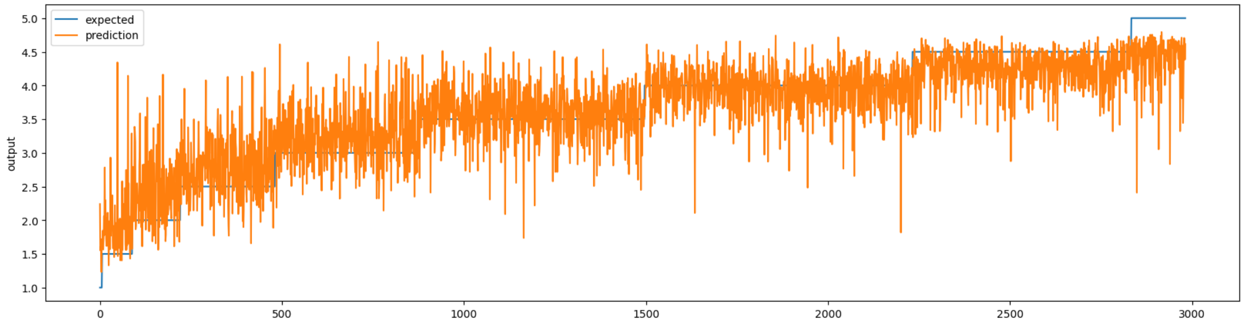
(3) Experimental Results and Analysis

For all model experimentation:

- 3 Hidden Layer
- Neuron counts: first layer has 100, second layer has 50, third layer has 20

	relu	sigmoid	tanh
adam	RMSE: 0.4456	RMSE: 0.4535	RMSE: 0.4503
sgd	RMSE: 0.4465	RMSE: 0.8388	RMSE: 0.4642

The best model is using the relu and adam with a RMSE score of 0.4456



(4) Task Division and Project Reflection

Only one member.

Project Reflection:

Throughout this project, I encountered several challenges that tested my problem-solving abilities and required me to work independently. These challenges, although demanding, enriched my learning experience and taught me valuable lessons:

- The data preprocessing has challenged me quite a bit. I had to redo most of it before training the models.

- Learning how TF-IDF works was challenging and took some time to learn. I had gained understanding of TF-IDF and its parameters. Because of overcoming this challenge, it allowed me to gain valuable insights into feature extraction techniques and their impact on model performance.

- Finding the right combination of hyperparameters for optimal model performance was a trial-and-error process. I learned to use systematic approaches and the value of patience in tuning.

- I learned that I need to improve in setting priorities and time management.

In conclusion, this project was a valuable learning experience. It showed the holes in my game, so to speak. These challenges and lessons learned will prepare me for future endeavors and should be helpful in applying these skills in upcoming projects. In hindsight, I realize the importance of starting early and managing my time effectively, especially when juggling multiple responsibilities like other classes, family commitments, and full-time work.