D213 Task 2

Student ID: 012170282

Tyson Biegler

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. Summarize one research question that you will answer using neural network models and NLP techniques. Be sure the research question is relevant to a real-world organizational situation and sentiment analysis captured in your chosen data set(s).

Note: If you choose to use more than one data set, you must concatenate them into one data set for parts II and III.

2. Define the objectives or goals of the data analysis. Be sure the objectives or goals are reasonable within the scope of the research question and are represented in the available data.
3. Identify a type of neural network capable of performing a text classification task that can be trained to produce useful predictions on text sequences on the selected data set.

Part II: Data Preparation

B. Summarize the data cleaning process by doing the following:

1. Perform exploratory data analysis on the chosen data set, and include an explanation of each of the following elements:

According to **ss64.com (n.d.)**, the standard English characters have a ‘Dec’ value between 32-127.

```
Total non-English characters: 0
```

The vocabulary size for this tokenized data set is 5101 unique words.

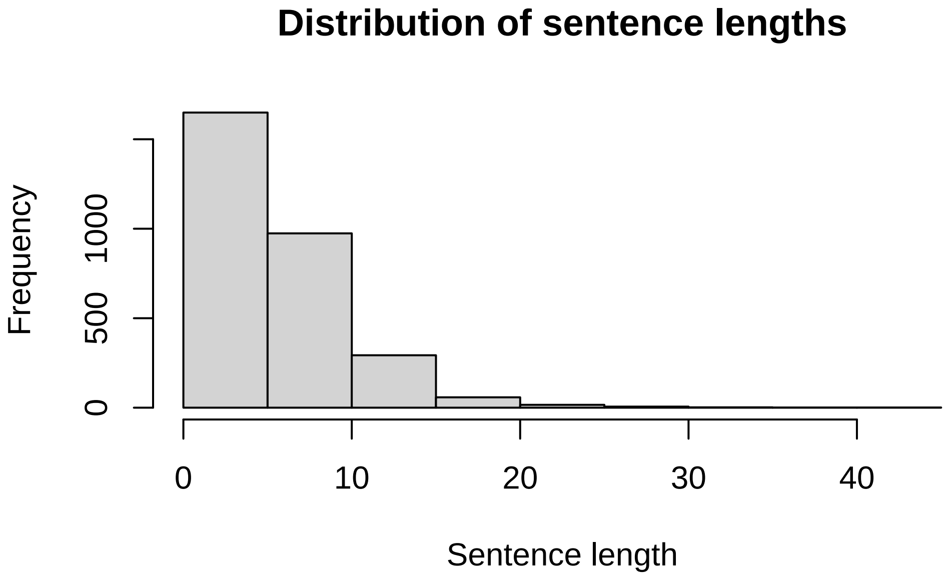
```
Vocabulary Size: 5101
```

- proposed word embedding length

The max sentence length is 41 words with a data set median of 5.

```
Longest sentence: masterful piece filmmaking many themes simmering occasionally boiling warts
Length: 41 tokens.
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.000   5.000   6.012   8.000  41.000 
```



2. Describe the goals of the tokenization process, including any code generated and packages that are used to normalize text during the tokenization process.

3. Explain the padding process used to standardize the length of sequences. Include the following in your explanation:

- if the padding occurs before or after the text sequence

- a screenshot of a single padded sequence

4. Identify how many categories of sentiment will be used and an activation function for the final dense layer of the network.

5. Explain the steps used to prepare the data for analysis, including the size of the training, validation, and test set split (based on the industry average).

6. Provide a copy of the prepared data set.

   ```
   #write.csv(data, "C:/Users/tyson/Documents/GitHub/WGU_MSDA_Portfolio/Advanced Data Analy
   ```

Part III: Network Architecture

C. Describe the type of network used by doing the following:

1. Provide the output of the model summary of the function from TensorFlow.

2. Discuss the number of layers, the type of layers, and the total number of parameters.

3. Justify the choice of hyperparameters, including the following elements:

- activation functions
- number of nodes per layer
- loss function
- optimizer
- stopping criteria
- evaluation metric

Part IV: Model Evaluation

D. Evaluate the model training process and its relevant outcomes by doing the following:

1. Discuss the impact of using stopping criteria to include defining the number of epochs, including a screenshot showing the final training epoch.

2. Assess the fitness of the model and any actions taken to address overfitting.

3. Provide visualizations of the model's training process, including a line graph of the loss and chosen evaluation metric.

4. Discuss the predictive accuracy of the trained network using the chosen evaluation metric from part D3.

Part V: Summary and Recommendations

E. Provide the code you used to save the trained network within the neural network.

F. Discuss the functionality of your neural network, including the impact of the network architecture.

G. Recommend a course of action based on your results.

Part VI: Reporting

H. Show your neural network in an industry-relevant interactive development environment (e.g., a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

I. Denote specific web sources you used to acquire segments of third-party code that was used to support the application.

J. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

- SS64. (n.d.). *ASCII table / character codes*. SS64.com. https://ss64.com/ascii.html

K. Demonstrate professional communication in the content and presentation of your submission.D213