**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 dataset.

A. Can the NLP and Neural Network Model be used on customer feedback to analyse their sentiments? This will help the company understand area to improve.

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.

A. The goal is to create a data model that will predict customer’s sentiments by learning off the words from the previous feedback.

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.

A. Deep learning neural network and natural language processing both can perform text classification and identify sentiments. NLP model ingests unstructured data and converts it to structured data for insight. Neural networks include layers that are connected by nodes. I will be using the Recurrent Neural Networks (RNN) for classification. These nodes form a network to answer simple question like is it black or white and then build upon these decisions. Rowe, W. (n.d.)

**Part II:  Data Preparation**

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

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

•    Here is a list of all the special character and the function to remove them.

Graphical user interface, text, application, email

Description automatically generated

•   vocabulary size

Graphical user interface, text, application

Description automatically generated

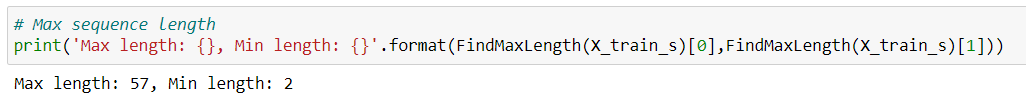
•  proposed word embedding length

Graphical user interface, text, application

Description automatically generated

•   statistical justification for the chosen maximum sequence length

Maximum sequence length is needed to preserve the length of the input data, it ensures that the generated model doesn’t yield results that are not likely to generalize. Therefore, inputs that are shorter than the maximum length are used for padding.



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

Tokenization simply separates the given text into small chunks or tokens. Then, a unique index ‘word\_index’ is assigned to each word that help understand the context, lemmantize words, transform text into sequences, interpret the meaning by analysing the sequence of words and provide padding information for the sequences. (Natural Language API Basics  |, n.d.)

Text

Description automatically generated

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

The padding process is used to improve and preserve the shape of the tensor dimensions. It requires the input to be of the same size and shape. Since sentences can have different lengths, padding is used to make all the sentences of the same size.

•   both pre and post sentence padding were tested, and pre-sentence padding had the highest level of accuracy. It is also the default setting.

•   A picture containing text

Description automatically generated

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

* There are two categories in the dataset, 1 (positive) and 0 (negative). It will be using ‘sigmoid’ activation as the ‘softmax’ did not provide good accuracy number.
* Text

  Description automatically generated

5.  Explain the steps used to prepare the data for analysis, including the size of the training, validation, and test set split.

1. Import the ‘imdb\_labelled.txt’ into Pandas dataframe using Python in Jupyter.
2. Perform EDA then visualize the data using countplot and boxplots.
3. Check the data for null values
4. Identify and remove whether there reviews with longer than usual length that can skew the padding.
   1. Check if any outlier exists in the review length
      1. If the length of the review is > 0.99 and < .01
      2. It was removed from the dataset.
5. Create the two-dimensional numpy array with our ratting by encoding
6. Split the data into training and test by using the ‘train\_test\_split’ method from sklearn (80/20 rule)
7. Apply tokenizer using the fit\_on\_text() method from
8. Create the vocabulary data
9. Retrieve word\_index
10. Apply pre-padding to the sentence to get the maximum length.
11. Fit the model, convert the training, and test set to Numpy array.

6.  Provide a copy of the prepared dataset.

Attached to the submission

Text, letter

Description automatically generated

**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.

Text

Description automatically generated with medium confidence

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

1st layer: Core Type= Embedding

* Layer type = Input
  + Number of parameters= 99360

2nd layer: Core Type= LTSM

* Layer type = Input
  + Number of parameters= 19400

3rd layer: Core Type= Dense

* Layer type = Input
  + Number of parameters= 51

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

•   activation functions

‘Sigmoid’ function is ideal for binary classification problems as their output values are either 0 or 1, it allows them to predict the probability of corresponding binary class. (Navlani, 2019)

•   number of nodes per layer

Since the prediction contains only 2 outputs, 0 for negative and 1 for positive, the output layer has only 2 nodes.

•   loss function

‘binary\_crossentropy’ is being used as it is ideal for classifying the ratings as they 0 and 1.

•   optimizer

‘adam’ optimizer is being used as it is well known for deep learning. It is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. (Team, n.d.)

•   stopping criteria

‘Early stopping criteria’ is being used to stop the overfitting of the training data and improve the generalization. It stops the training when the metric stop improving. In the model, it is set to 10 epochs with parameters ‘patience’ set to 2, which states the number of no improvements before stopping (Team, n.d.),

Text

Description automatically generated

•   evaluation metric

Text

Description automatically generated with low confidence

**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 instead of defining the number of epochs, including a screenshot showing the final training epoch.

When an epoch 10 is provided to the model, it will get trained from the beginning to the last epoch [1 -10]. In this situation, even if the validation score drops the model will continue till the end. This requires the observer to manually investigate and identify the optimum epoch where the validation score was lowest. This is where the stopping criteria like ‘EarlyStopping’ comes in handy. Once enabled, a higher epoch number with ‘patience’ parameter of 2 will monitor the validation score. This will ensure training is stopped if there has been no improvement in 2 epochs. ‘EarlyStopping’ needs to be set during the fit with callbacks. Example is available in C3e.

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

A picture containing text

Description automatically generatedIn the visual below, the epoch is set to 10, with the patience at 2 (patience setting are visible in C3e). Since, the accuracy did not improve after 8, the model stopped the fitting process.

Chart, line chart

Description automatically generated

3.  Assess the fitness of the model and any measures taken to address overfitting.

To lower the overfitting of the model, I took the model with highest validation accuracy and then test it with the testing set. The model can be further improved by providing additional data for training and optimizing the network. The deep network performs best with a very large dataset. (Real Python, 2021)

4.  Discuss the predictive accuracy of the trained network.

The model accuracy is at .73 and loss at .58 per batch size. That means, model was able to accurately predict 73% of the time and of the predict 73% it was off .58%. This can be improved by providing additional data and testing out various variables.

Text

Description automatically generated with low confidence

Graphical user interface, text, application, email

Description automatically generated

**Part V:  Summary and Recommendations**

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

Graphical user interface, text, application

Description automatically generated

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

The model was trained using approx. 500 customer feedback and it was validated and tested. NLP was used to analyse customer mood and/or sentiments from the review. The model is trained using the customer reviews and recording their sentiment as labels. It was also tuned to perform predictions and when new feedback is left by the customer, it can be used to identify whether their feedback was negative or positive.

Impact:

Using the NLP network architecture helped detect customer’s sentiments with an acceptable degree of accuracy.

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

Since the model was trained with an acceptable degree of accuracy, I would recommend the business to use this model to analyse customers sentiments. This will help the company understand area to improve.

**Part VI: Reporting**

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

File attached D213T2.pdf

I.  List the web sources used to acquire data or segments of third-party code to support the application.

Rowe, W. (n.d.). What Is a Neural Network? An Introduction with Examples. BMC Blogs. Retrieved 27 September 2022, from https://www.bmc.com/blogs/neural-network-introduction/

Text Classification: What it is And Why it Matters. (n.d.). MonkeyLearn. Retrieved 26 September 2022, from <https://monkeylearn.com/text-classification/#:%7E:text=The%20two%20main%20deep%20learning,Recurrent%20Neural%20Networks%20(RNN)>.

Natural Language API Basics  |. (n.d.). Google Cloud. Retrieved 27 September 2022, from <https://cloud.google.com/natural-language/docs/basics#tokenization>

Team, K. (n.d.). Keras documentation: Adam. Retrieved 28 September 2022, from <https://keras.io/api/optimizers/adam/>

Team, K. (n.d.-b). Keras documentation: EarlyStopping. https://keras.io/. Retrieved 28 September 2022, from https://keras.io/api/callbacks/early\_stopping/

Real Python. (2021, August 6). Practical Text Classification With Python and Keras. Retrieved 29 September 2022, from https://realpython.com/python-keras-text-classification/

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

Navlani, A. (2019, December). Neural Network Models in R. https://www.datacamp.com/. Retrieved 28 September 2022, from https://www.datacamp.com/tutorial/neural-network-models-r