Taylor Cina

D213 – Advanced Data Analytics

***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).*

Can a customer review be analyzed through the use of NLP and neural networks to predict a positive or negative sentiment?

*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.*

The goal of this analysis is to produce a deep learning model capable of processing user reviews and predicting a positive or negative sentiment value.

*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.*

An artificial neural network takes inputs, passes them through a series of hidden layers and produces an output. Artificial neural networks excel at determining binary outputs from large amounts of input data, which is suitable for a sentiment analysis.

***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:*

*•   presence of unusual characters (e.g., emojis, non-English characters)*

The dataset did not appear to have any unusual characters.

*•   vocabulary size*

The vocabulary size of the original dataset was 1879. After tokenization and lemmatization, the vocabulary size was reduced to 1568.

*•   proposed word embedding length*

A common embedding length is the fourth root of the vocabulary size rounded up. Therefore, the first tested embedding length will be 1568^0.25 = 7

*•   statistical justification for the chosen maximum sequence length*

The maximum sequence length chosen was 17, which matches the maximum length of a list of tokens in the dataset after preprocessing.

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

The goal of the tokenization process was the convert a list of sentences into integer tokens representative of the original vocabulary and lemmatize them to reduce the total vocabulary whilst keeping a similar amount of information. Nltk was used to tokenize and lemmatize the review list. All code is included in the attached file “TCina D213 T2.ipynb”

*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*

The padding process consisted of finding the largest sequence length, 17 and padding all sequences using 0’s after the data. Below is 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.*

Two categories of sentiment will be used. 0 is a negative review, 1 is a positive review. The softmax activation function will be used for the final dense layer

*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).*

The steps to prepare the data for analysis are as follows:

- Import the Amazon dataset as a csv with tab separation. Label the columns “Review” and “Sentiment”

- Check for missing values.

-Print a list of unique characters to check for special or unusual characters.

-Normalize the text by eliminating punctuation and special characters using regex.

-Tokenize the reviews, then lemmatize the tokens.

-Remove stopwords from the reviews.

-Get the vocabulary size and max length of a sequence.

-Train test split the data 80/20.

-Sequence and pad the training and test data.

-Convert the dataframes to numpy arrays for use in the model.

*6.  Provide a copy of the prepared data set.*

The train and test datasets are provided as 4 separate files labelled appropriately.

***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.*

The model summary output is pictured below:

A screenshot of a computer program

Description automatically generated

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

The model has 4 layers: an embedding layer, a 1D pooling layer to flatten the embedding layer output, and 2 dense layers. Across all layers, the model has 11,478 parameters, all of which are trainable.

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

*•   activation functions*

The ReLu activation function in the hidden layers is a commonly used activation function for hidden layers. For the final layer, Softmax is used because the output is binary.

*•   number of nodes per layer*

Nodes and layers were added via experimentation. Nodes were added until adding complexity stopped improving accuracy.

*•   loss function*

Sparse categorical cross entropy functions on integer outputs, which the sentiment value is.

*•   optimizer*

The adam optimizer is an improvement to stochastic gradient descent and is a robust optimizer suitable for this analysis.

*•   stopping criteria*

A stopping monitor was implemented with a patience of 6 to ensure that the model would stop training when improvements were no longer evident.

*•   evaluation metric*

Accuracy is an acceptable evaluation metric for a binary decision.

***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.*

Using a stopping criterion prevents the model from overfitting. It also prevents the model from training itself to a less accurate version on the training set.

The final training epoch of a run of the model is pictured below:



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

The model fitness was addressed by using a stopping parameter during training and tweaking the balance of nodes and layers in the model definition. The model appears to be well fitted as the accuracy and loss of the training and validation data form roughly similar curves.

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

Below are visualizations of the model loss and accuracy metrics during the training process:

A graph of blue and orange lines

Description automatically generatedA graph with blue and orange lines

Description automatically generated

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

The accuracy of the model against the testing set is 81%, which is very good. A printout of the model accuracy and loss is shown below.

A computer code with text

Description automatically generated with medium confidence

***Part V: Summary and Recommendations***

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

The code used to save and load the model is shown below:

**A screen shot of a computer program

Description automatically generated**

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

The neural network took in a training sample of 800 customer reviews and positive and negative sentiment values and managed to achieve a testing accuracy of 81%. The model architecture didn’t require many nodes in its single hidden dense layer likely due to the use of embedding and vectorization in the model building process which allowed it to learn the meanings of positive and negative words more effectively.

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

Using this model, customer reviews can be analyzed for positive and negative sentiment to find what consumers like or dislike about a product without any built-in rating on the part of the consumer. It would be beneficial to find what customers dislike about a product in order to rectify any issues.

***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.*

HTML file included as “TCina D213 T2.html”

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

*Kotzias, D. (2015). Sentiment Labelled Sentences [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C57604.*

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