

D213 Task 2 - Sentiment Analysis

Western Governors University

Danielle Van Sant

Dr. William Sewell

## Part I: Research Question

Can a neural network, NN, model coupled with natural language processing, NLP, be used to predict the sentiment of new customer reviews?

The objectives of this data analysis are to build a NN model and to train, test and validate the model so that it can be used with confidence to run sentiment analysis on incoming text based customer reviews. Another objective of the analysis is to preprocess the data and prepare it to build the model. A final objective of this data analysis is to take data in the form of text-based customer reviews from three major corporations, Amazon, Yelp, and IMDB, and transform that data so that it can be used to build an NN model.

The website Geeks for Geeks defines LSTM **as “Long Short-Term Memory** is an improved version of recurrent neural network designed by Hochreiter & Schmidhuber.

A traditional[RNN](https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/) has a single hidden state that is passed through time, which can make it difficult for the network to learn long-term dependencies.**LSTMs model** address this problem by introducing a memory cell, which is a container that can hold information for an extended period.

LSTM architectures are capable of learning long-term dependencies in sequential data, which makes them well-suited for tasks such as [language translation](https://www.geeksforgeeks.org/language-translator-using-google-api-in-python/), speech recognition, and [time series forecasting](https://www.geeksforgeeks.org/time-series-forecasting-using-recurrent-neural-networks-rnn-in-tensorflow/).” (GeeksforGeeks, 2024)

## Part II: Data Preparation

Data cleaning is one of the most important steps in any data analysis process, and machine learning processes are no exception. The following steps were taken to prepare the data for processing.

Unusual characters, such as but not limited to emojis, are often found in customer reviews. Unusual characters must be removed from the text in the data cleaning step. Below is the code that was used to remove these characters.

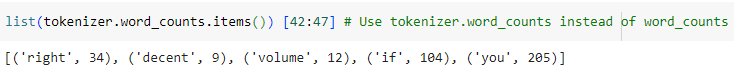
A screenshot of a computer code

Description automatically generated

The vocabulary size is the number of unique or tokenized words featured in the dataset or across all comments. Below is a screenshot of the code used to determine the vocabulary size.

A close up of a computer code

Description automatically generated



“Word embedding is the position of the word in the learned vector space. It is a mapped vector in a number format that allows a NN to determine the probability distribution of all words and the occurrence of the next word in the vector” (Jason Brownlee, Oct.11, 2017).

A screenshot of a computer code

Description automatically generated

The maximum sequence length is not something that is set in stone. Instead, there are several factors that need to be considered. The below excerpt from LinkedIn explains the factors that must be considered before selecting a sequence length. I arbitrarily decided to start low with a sequence length of 25.

* **Incremental adjustments:**

  Begin with a short sequence and increase it bit by bit, keeping an eye on performance metrics. This trial-and-error method helps you strike a balance between learning complex patterns and computational efficiency.

* **Utilize mixed methods:**

  Combine quantitative metrics like accuracy with qualitative insights from visualizations to gauge how well your RNN handles data. This holistic approach can lead to a more informed decision on the ideal sequence length. (Machine Learning Machine Learning , 2023)

25

A screenshot of a computer code

Description automatically generated

Tokenization is a process that breaks words down into smaller chunks. The chunks are easier for a machine-learning algorithm to understand.

A screenshot of a computer program

Description automatically generated

The padding occurs at the end of the text sequence. An example of this is provided in the screenshot below.

A screenshot of a computer screen

Description automatically generated

My model has two categories of sentiment: positive and negative. The activation function is sigmoid.

A computer screen shot of a program

Description automatically generated

To prepare the data for analysis several steps were taken.

1. Read the data from the dataset or from all three sources.

2. Check on unusual or any abnormal characters

3. Clean the data

4. Explore the data

5. Create a two-dimensional NumPy array with our ratings by encoding

6. Split data into training and testing sets with an 80/20 split

* **Training Data:** 64% of the original dataset
* **Validation Data:** 16% of the original dataset
* **Testing Data:** 20% of the original dataset

7. On the training set, apply the tokenizer using the fit\_on texts() method

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

I elected to use post-padding on my data. Padding individual samples is necessary because input data must be a single tensor in deep learning models. Therefore, a value, in this case 0, must be added to the individual reviews.

A screenshot of a computer program

Description automatically generated

A copy of the prepared dataset is included with this submission.

## Part III: Network Architecture

A screenshot of a computer

Description automatically generated

**Layer 1 = Embedding**

**Layer Type = Input**

**Number of parameters = 275,400**

**Layer 2 = LSTM**

**Layer Type = LSTM**

**Number of parameters = 120,400**

**Layer 3 = Dense**

**Layer Type = Output**

**Number of parameters = 101**

Geeks for Geeks explains the role of activation functions like this, “An activation function in the context of neural networks is a mathematical function applied to the output of a neuron. The purpose of an activation function is to introduce non-linearity into the model, allowing the network to learn and represent complex patterns in the data. Without non-linearity, a neural network would essentially behave like a linear regression model, regardless of the number of layers it has.

The activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias to it. The purpose of the activation function is to introduce non-linearity into the output of a neuron.” (GeeksforGeeks, 2024)

•   activation functions = Softmax

The Softmax activation function activates the final dense output layer of the NN. Logistically softmax generalizes the inputs to output results as probabilities close to 0 and 1.

•   number of nodes per layer = 200, 100, 1

* Raw text data (sentences\_train, sentences\_test) is the initial input.
* This raw data is processed into numerical sequences using Tokenizer and pad\_sequences.
* These numerical sequences are the direct input to the embedding layer, which acts as the first layer.
* The input layer takes in the entire vocabulary so that it has a number of nodes equal to the vocabulary size.
* The embedding layer has 200 nodes.
* The output layer has 1 node because it has two possible outcomes: positive and negative.

•   loss function = binary crossentropy

Binary crossentropy was selected as the loss function due to the binary classification of the sentiments.

•   optimizer = adam

The adam optimizer is a solid choice due to its ease of use and quickness to implement, making it more computationally efficient and using less memory.

•   stopping criteria = early stopping

Early stopping criteria was selected to avoid overfitting the model. The model runs a set number of epochs but will stop before reaching the final epoch if validation has not improved over a set number. I selected a patience of 3 for my early stopping criteria.

•   evaluation metric = accuracy

The evaluation metric I selected is accuracy. Model accuracy determines the ability of the model to classify consumer comments based on sentiment correctly.

## Part IV: Model Evaluation

Early stopping criteria were employed not to run additional epochs of the model. This is done because running the model is computationally intensive, and early stopping stops the process after val loss has not improved for a set number of epochs, in this case, 3.

A screenshot of a computer program

Description automatically generated

My original model was overfitted, so test loss was close to 0. To avoid overfitting the model I changed the output dim to 200. The model is more accurate with a greater loss.

A screenshot of a computer code

Description automatically generatedA graph of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated

The predictive accuracy of the model was determined by scoring on the model. The final accuracy of the model and final validation accuracy are below. These values indicate that the model can accurately predict positive and negative reviews.

Accuracy = 0.8197

Validation Accuracy = 0.7455

## Part V: Summary and Recommendations

A screenshot of a computer program

Description automatically generated

Amazon describes the functionality of a neural network in this manner: “A neural network is a method in [artificial intelligence](https://aws.amazon.com/what-is/artificial-intelligence/) that teaches computers to process data in a way that is inspired by the human brain. It is a type of [machine learning](https://aws.amazon.com/what-is/machine-learning/) process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and improve continuously. Thus, artificial neural networks attempt to solve complicated problems, like summarizing documents or recognizing faces, with greater accuracy.” (*What is a neural network? - artificial neural network explained - AWS*)  
 The results indicate that the company should implement the model and use it to predict customer reviews.

Resources

GeeksforGeeks. (2024, June 10). *What is LSTM - long short term memory?* <https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/>

Machine Learning                                                                                                          Machine Learning. (2023, November 6). *What is the optimal sequence length for an RNN?*. How to Optimize Sequence Length for RNN Models. <https://www.linkedin.com/advice/3/what-optimal-sequence-length-rnn-skills-machine-learning> nu4gc#:~:text=The%20optimal%20sequence%20length%20for%20an%20RNN%20depends%20on%20the,the%20training%20too%20computationally%20expensive.

GeeksforGeeks. (2024b, July 16). *Activation functions in neural networks*. https://www.geeksforgeeks.org/activation-functions-neural-networks/

What is a neural network? - artificial neural network explained - AWS. (n.d.). https://aws.amazon.com/what-is/neural-network/