Lora Milam

Western Govenors University

D213 Advanced Data Analytics

22 September 2023

D213 Performance Assessment Task 2

**1 Introduction**

Lora Milam Masters Data Analytics (8/11/2023) Program Mentor:d213@wgu.edu

* 1. **Research Question**

This analysis will utilize neural network models and NLP techniques to examine datasets from the paper ‘From Group to Individual Labels using Deep Features’(Kotzias). These datasets “contain sentences labelled with positive and negative sentiment, extracted from reviews of products, movies, and restaurants” (WGU). The reviews were obtained from Amazon, IMDB, and Yelp. Each dataset consists of 500 positive and 500 negative sentences (WGU). The primary objective is to address the research question:

Is it possible to create a reliable model to predict the sentiment, dictated as positive (1) and negative (0), of a review?

* 1. **Objective and Goals**

The main objective is to build a neural network model for text classification of positive and negative reviews through the train/test process. The goal is to have the model accurately classify positive and negative reviews.

* 1. **Prescribed Network**

The type of neural network to be used to perform this task is a recurrent neural network (RNN). Within the python language, TensorFlow and Kera libraries can perform text classification and will be used within this task.

**2 Data Preparation**

**2.1 Data Exploration**

After performing the exploratory data analysis, the following was determined:

* A presence of non-English unusual characters

After performing the statistical analysis, the following was determined:

* Vocabulary size: 5155
* Maximum length of words: 479
* Minimum length of words: 7
* Vocabulary sized of cleaned dataset: 3000
* Maximum length of cleaned words: 42
* Minimum length of cleaned words: 2

A proposed word embedding length of the cleaned dataset could be 45 words based on the statistical analysis.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A white screen with red text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**2.2 Tokenization**

The goal of the tokenization process is to prepare the text data for the machine learning process. In this task, tokenization cleans the data by removing stop words, punctuation and split the words into individualized ‘tokens’.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**2.3 Padding Process**

The padding process used to standardize the length of sequences by adding zeros to the length of sequences until the predetermined max length is met. If the sequence is greater than the preset max length, then the remaining sequence length is cutoff. Within this task, the max length of the padding process was set to the max length of cleaned words in a sequence, 42.

A screenshot of a computer code

Description automatically generated

**2.4 Categories of Sentiment**

In previous sections, it was determined that this is a total number of 3000 unique words after cleaning the data which will be used for categories of sentiment. The final dense layer of the network will be a sigmoid activation function.

A screenshot of a computer program

Description automatically generated

**2.5 Steps to Prepare the Data**

The data preparation steps include:

1. Load libraries
2. Import dataset
3. Explore descriptive features and perform statistical analysis
4. Clean dataset by removing punctuations, numbers, emojis, and stopwords
5. Perform tokenization
6. Normalize words by performing padding process
7. Perform Train/Test on cleaned dataset
8. Build model
9. Perform validation

**2.6 Prepared Dataset**



A screenshot of a computer

Description automatically generated

**3 Network Architecture**

**3.1 Model Summary**

A screenshot of a computer

Description automatically generated

**3.2 Network Architecture**

The model uses a four layer neural network: an embedding layer for input, two flatten type hidden layers, and a dense layer. The total parameters used for this model is 52,333.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**3.3 Hyperparameters**

The hyperparameters for this model are:

Activation functions: ReLu and Sigmoid

Number of nodes per layer:

A screenshot of a computer

Description automatically generated

Loss function: Binary\_crossentropy

Optimizer: Adam optimizer

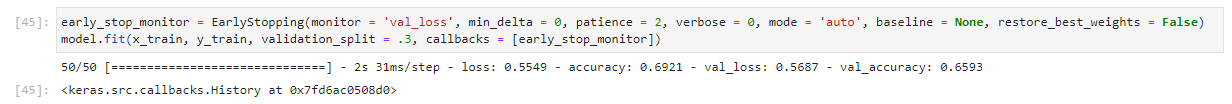
Stopping criteria: EalryStopping

Evaluation metric: Classification accuracy

**4 Model Evaluation**

**4.1 Stopping Criteria**

Defining the exact number of epochs can lead to underfitting or overfitting, while utilizing stopping criteria aids in counteracting these outcomes by providing the stopping monitor with a large arbitrary number and having the stopping criteria determine the best performing epoch number.



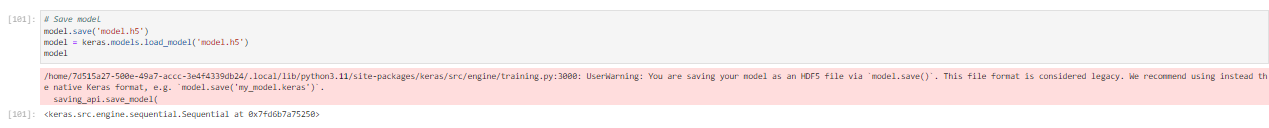
A screenshot of a computer code

Description automatically generated

**4.2 Training Process**

A screenshot of a graph

Description automatically generated



**4.3 Fit**

According to the graph in section 4.2, the training and validation accuracy are very similar. This means that the model is decently fit. The stopping criteria aided in preventing overfitting.

**4.4 Predictive Accuracy**

According to the graph in section 4.2, the training and validation accuracy are very similar. This means that the model is decently fit. There would be over a 50% accuracy if the model was used to predict new data.

**5 Summary and Recommendations**

**5.1 Code**

Code is provided in previous sections.

**5.2 Functionality**

The functionality of this neural network is to predict sentiment of new data. After preprocessing the data, the model can predict the sentiment label of the data, negative (0) or positive (1). However, the model can only accurately predict about 67% of new data.

**5.3 Recommendations**

This model can be used in its current state; however, it only has an accuracy of about 67%. To improve the accuracy, it is suggested to continue to adjust the hyperparameters until a desired accuracy percentage is reached.

**6 Reporting**

**6.1 Reporting**

Report is provided as “D213 Part 2 Performance Assessment Lora Milam Report.pdf”.

**7 Supporting Documentation**

**7.1 Sources**

Kotzias, D., Denil, M., de Freitas, N., & Smyth, P. (2015). From group to individual labels using Deep Features. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2783380

Western Governors University. (n.d.). D213 Advanced Data Analytics. Salt Lake City.