**D214 Task 2**

**Data Analytics Graduate Capstone**

**Kurt Harris**

**College of Information Technology Information, Western Governors University**

**William Sewell**

**November 29, 2021**

**Research Question**

**Research Question:** Can the sentiment of a short statement be accurately predicted as positive, neutral or negative?

**Hypothesis**: H0-. A predictive sentiment model cannot be created. H1-. A predictive sentiment model can be predicted with an accuracy better than 70%, the threshold a company is willing to make decisions on.

**Context:** The sentiment from an interaction could provide valuable information for a company trying to improve customer experience and has been requested providing justification for this study. This study will provide a model that can make predictions of sentiment on short text. The value of this information is that it can be acted upon to reduce churn by proactive outreach for unresolved service issues or extremely poor experiences.

Virahonda (2020) explains how various architectures of a Sequential Keras models could be used to predict sentiment on text in a three-way classification, published on towardsdatascience.com. The article hypothesizes that by hyper tuning the model’s parameters and good data, a highly accurate model could be constructed.

**Data Collection**

To build the sentiment model, short text from SMS messages, survey reviews, chat support or Twitter will be required. Publicly available on Kaggle, Twitter airline review dataset will be used that contains a three-way categorical airline\_sentiment target feature.

A disadvantage to public data is the quality, volume and time it takes to find a suitable dataset compared to private data a company has. This challenge was overcome by diligently researching and locating a good dataset to train the model on.

One advantage to using publicly available data is that it avoids many legal issues that one would face if trying to use private data from a company and easily repeatable.

**Data Extraction and Preparation**

**Data Gathering:** A dataset of airline reviews and sentiment classification is available from Kaggle.com in a csv format and is publicly available.This data contains Twitter text of customer reviews about airlines and positive, neutral or negative sentiment.

Graphical user interface, website

Description automatically generated

The text will need to have some cleaning including spelling, trimming, padding and tokenization to be then transformed into a numeric array for the tensor Keras model (Virahonda, 2020).

Text

Description automatically generated

Text

Description automatically generated

Basic exploratory data analysis will first be performed on this dataset to check for nuances within the data like; skew or distribution of classification, outliers, null volume and central tendencies around number of words and lengths of text (Malik, 2019).

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Sentiment will be converted into three integer binary features by creating dummy values.

Graphical user interface, text, application

Description automatically generated

**Text

Description automatically generated**

**Data Analytics Tools and Techniques**: Design: The length of words within each row will be checked for normality with a quantile-quantile and a density plot. The distribution of sentiment with a histogram. Neither of these are necessarily needed but provide insights that could help tune the model (Malik, 2019).

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generatedA TensorFlow Keras model will be built for the sentiment analysis on a test portion of the dataset. This will be constructed of one or more hidden layers by way of trial and error with accuracy on a held-out sample as the metric.

Text

Description automatically generated



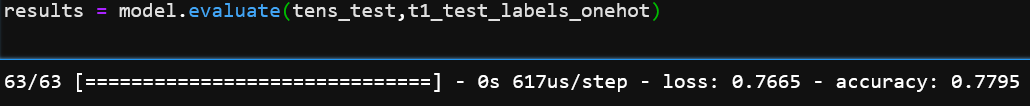
Python will be used for both data processing and model creation being an open-source tool with a wide audience who contribute often. Johnson (2021) explains how Python is better at replication and growth at scale making it easier to go from the lab to field.

A disadvantage for this model is the use of tensors which are required to be of the same shape and length. Understanding how much padding or trimming is needed can impact the model accuracy (TensorFlow, 2021).

**Analysis**

A confusion matrix analysis will be used for additional insights into the accuracy of the model and a one-way ANOVA to test if there is a statistical difference between the predicted and actual results. With ANOVA, if the test p value is less than .05 then the two sets are different, and the null hypothesis would be accepted. Otherwise, the test shows no different between the groups validating the model prediction accuracy.

ANOVA is used since it is an omnibus test for 3 or more groups, it tests for a difference overall, i.e. at least one of the groups is statistically significantly different than the others. However, if the ANOVA is significant, one cannot tell which group is different (Python for Data Science, 2020).



Chart, line chart

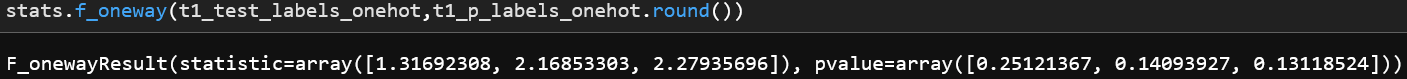
Description automatically generated

Confusion Matrix shows 78% accuracy:

A screenshot of a computer

Description automatically generated with medium confidence

ANOVA Results: p-value for all three classifications are greater than 0.05 indicating no statistical difference between the predictions and the actual labels. This confirms the accuracy results are valid as the predictions are on par with the actual labels (Python for Data Science, 2020).



**Data Summary and Implications**

The key objective and deliverable for this project is a three-way neural net sentiment model that can be used on short text of 30 characters or less and maintain an accuracy greater than 70%.

Results of the confusion matrix and ANOVA analysis indicate accuracy at a statistically significant level have been achieved and we accept the alternative hypothesis.

One limitation is that the dataset is all Twitter data and only on one subject. Model accuracy might not be consistent with other sources of short text data (Virahonda, 2020).

Recommend further testing on other sources of short text data to understand how the model accuracy varies and possible retraining on these additional sources. Two directions for future study:

1. Model trained on a combination of short text like, Twitter, SMS text messages and survey reviews.
2. Tweaks to the model architecture with new data could improve accuracy like adding additional hidden layers or nodes to the model.

**Reference**

Virahonda, Sergio (2020, October 8). Sentiment Analysis with Deep Learning and Keras Retrieved November 17, 2021, from

<https://towardsdatascience.com/an-easy-tutorial-about-sentiment-analysis-with-deep-learning-and-keras-2bf52b9cba91>

TensorFlow (2021, November 12) Masking and Padding with Keras. Retrieved November 20, 2021, from

<https://www.tensorflow.org/guide/keras/masking_and_padding>

Malik, Usman (2019, August 7). Python for NLP: Movie Sentiment Analysis Using Deep Learning in Keras. Retrieved from <https://stackabuse.com/python-for-nlp-movie-sentiment-analysis-using-deep-learning-in-keras/>

Python for Data Science, LLC (2020) Retrieved November 23rd, 2021, from

<https://www.pythonfordatascience.org/anova-python/>

Twitter US Airline Sentiment (2019, October 19). Retrieved November 5, 2021 from

<https://www.kaggle.com/crowdflower/twitter-airline-sentiment>

Johnson, Daniel (2021, October 7) R Vs Python: What’s the Difference? Retrieved November 19, 2021, from

<https://www.guru99.com/r-vs-python.html>