**D214 Task 3**

**Data Analytics Graduate Capstone**

**Kurt Harris**

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

**William Sewell**

**November 30, 2021**

**Executive Summary and Implications**

Understanding the quality of a customer interaction with a service department is historically based on a post survey response, leaving little room for proactive outreach and prevention of churn. The lack of insight into customer interactions leads to undue churn that might have been prevented.

The hypothesis for this analysis is that a predictive sentiment model can be created with accuracy above 70%. The null hypothesis is that one cannot be created above said threshold.

Twitter sentiment data set was gathered from Kaggle.com and contains over 14,000 records (Twitter US Airline Sentiment, 2019). After removing nulls, review of distribution and density helps identify a good max length for the text. Next, the text has lower case and spelling standardizations applied then converted into numeric arrays. Last, these are padded or trimmed to the max length identified earlier which is 30 words.

Chart, histogram

Description automatically generated

After training the model it was tested on the held-out sample of 2000 records. Confusion matrix analysis shows accuracy of 78% and ANOVA analysis p-value confirms the predictions and actual labels are not statistically different. The alternative hypothesis is thus accepted (Python for Data Science, 2020).

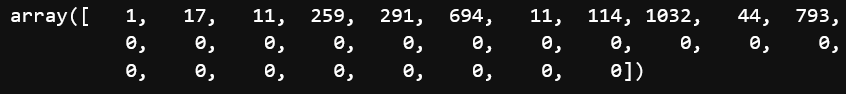
A screenshot of a computer

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

A limitation to this model is that it requires the use of tensors, arrays of data that must be the same size and shape. This model is specific to short text and might not be as accurate on longer text lengths (Virahonda, 2020).



Further testing on other types of short text like SMS or online chat is necessary to understand the model’s accuracy generally on short text as the model was trained only on Twitter data.

Tweaks to the model architecture with new data could improve accuracy like adding additional hidden layers or nodes to the model (Python for Data Science, 2020).

Main benefit from this model is the ability for proactive customer outreach when negative sentiment is predicted. Possible reduction in churn and improvement within the customer experience could lead to more profit. Additional analyses could be conducted using the model’s predictions to cluster the most painful topics based on negative sentiment. This would help identify some of the biggest customer experience opportunities.

**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>

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>