Data Analytics Capstone Topic Approval Form

Student Name: Kurt Harris

Student ID: 000178150

Capstone Project Name: Three-way Neural Net Sentiment Model on Short Text

Project Topic Predictive three-way classification sentiment model on short text, less than 30 words like survey reviews, twitter or chat messages.

X This project does not involve human subjects research and is exempt from WGU IRB review.

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

Hypothesis: H_0-. A predictive sentiment model cannot be created. H_1-. 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. 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.

This study will contribute to the field of data and analytics by exploring the predictive accuracy of a neural net three-way classification sentiment model on short text less than 30 words. This additional layer of sentiment could provide a useful feature for further models or analysis.

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.

Allibhai (2018) touches on using multiple classifications by passing the target as an array and the configuration for output in a matching array. This will be useful for the three possible categories.

Data: 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. No personable identifiable features will be used nor any business names. Row count before cleaning is 14,640.
The data set includes the following features:

https://www.kaggle.com/crowdflower/twitter-airline-sentiment?select=Tweets.csv

Field	Type
Tweet_id	Integer
Airline_sentiment	Categorical
Airline_sentiment_confidence	Float
Negativereason	Text
Negativereason_confidence	Float
Airline	Categorical
Airline_sentiment_gold	Text
Name	Text
Negativereason_gold	Text
Retweet_count	Integer
Text	Text
Tweet_coord	Object List of float x, y cords
Tweet_created	DateTime
Tweet_location	Text
User_timezone	Categorical

Twitter and Kaggle have made this information publicly available, the only features used are the airline_sentiment and text. Limitations: The dataset is limited by the accuracy of the sentiment labels, the target feature which was classified by a group of researchers. Delimitations: Longer text containing more than 30 words will be trimmed and less padded to fit the model and any null or one word text will be removed (Malik, 2019).

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. 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).

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). Sentiment will be converted into three integer binary features by creating dummy values. Data sparsity is less than 5%.

Data Analytics Tools and Techniques: Design: 1. 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). For example, since this model uses tensors which are required to be of the same shape and length, understanding how much padding or trimming is needed can impact the model (TensorFlow, 2021). 2. A 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. 3. 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 (Python for Data Science, 2020).

Justification of Tools/Techniques: 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.

SAS is not considered for this project as a business decision citing cost and is slow to get updates both points supported by Dutta (2021) in the article titled '8 differences Between SAS and Python'.

Project Outcomes: 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%. Support for the alternative hypothesis comes from Virahonda (2020) that a successful three-way sentiment model could be built with relatively high accuracy using a Tensorflow Keras Sequential model and enough good data.

Projected Project End Date: 2021-12-31

Sources:

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

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

TensorFlow (2021, November 12) Masking and Padding with Keras. Retrieved November 20, 2021, from https://www.tensorflow.org/guide/keras/masking_and_padding

Dutta, Bhumika (2021, August 15) 8 Differences Between SAS and Python. Retrieved November 19, 2021, from https://www.analyticssteps.com/blogs/8-differences-between-sas-and-python

Allibhai, Eijaz(2018) Building a Deep Learning Model using Keras. Retrieved November 1, 2021, from https://towardsdatascience.com/building-a-deep-learning-model-using-keras-1548ca149d37

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

Institutional Review Board Quiz and Approval

Have you read and understood the "Human Subjects FAQ" page and completed the "Human Subjects FAQ Quiz" at the WGU Institutional Review Board (IRB) website? (https://irb.wgu.edu/info/Pages/Home.aspx)

☒ Yes, I have read and understood the "Human Subjects FAQ" and have provided email proof of my completed quiz in appendix A. (https://irb.wgu.edu/info/Pages/Human-Subjects-FAQ-Quiz.aspx)

□ No, I have not completed the Human Subjects FAQ quiz.

Assess whether your capstone proposal complies with WGU's IRB standards for exemption status. Explain why you believe the proposed project complies with the standards for exemption status. If it does not, make arrangements with a course mentor and the IRB for approval.

☒ The research complies with WGU's IRB exemption status because:

- Research involving the collection or study of freely available de-identified existing data

□ The research requires approval from WGU's IRB because:

□ Yes, I would like to schedule a conference to discuss my project.

To be filled out by a course mentor:

☒ The research is exempt from an IRB Review.

□ An IRB approval is in place (provide proof in appendix B).

Course Mentor's Approval Status: Approved

Date: 11/24/2021

Reviewed by:

Comments: Click here to enter text.