**Data Analytics Capstone Topic Approval Form**

**Student Name:** Jan Francis Mitchell Barlaan

**Student ID:** 011282929

**Capstone Project Name:** Twitter Sentiment Analysis on Company Stock

**Project Topic**: Sentiment Analysis with Neural Networks and VADER

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

**Research Question:** Using neural networks such as LSTM combined with other sentiment analysis techniques such as VADER, can we create a model that can accurately predict people's current sentiments/feelings towards investing in a company’s stock, whether positive or negative based on their tweets?

**Hypothesis**: **Null hypothesis** - Using neural networks and VADER, we cannot create a model that can accurately predict sentiment behind a tweet or review with at least 70% accuracy. **Alternate Hypothesis** - Using neural networks and VADER, we can successfully create a model that can accurately predict sentiment behind a tweet or review with at least 70% accuracy.

**Context:** Being able to create a model based on neural networks and VADER that can accurately conduct sentiment analysis on texts such as social media tweets can be a very useful tool to gauge overall public opinion on company stocks. Those with positive sentiments may be worth investing in at that moment, while those carrying more negative public sentiments should be treated more cautiously when investing. Further study and application of a model based on this type of sentiment analysis can be combined with forecasting methods such as time series analysis or even basic linear regression to determine whether a correlation of some sort exists between the current public opinion of a company/stock and whether the closing price of that stock increased or decreased relative to its opening price. However, for the purposes of this capstone project. I will only conduct the sentiment analysis portion.

**Data:** Date, Tweet, Stock Name, Company Name

I will be using the “stock\_tweets” csv file obtained from Kaggle. This dataset contains Date, Tweet, Stock Name, and Company Name columns. It contains 80,793 rows and 4 columns.

This dataset is publicly accessible on Kaggle and is unrestricted for use in academic and research applications. Additionally, tweets are obtained from X/Twitter, which is a publicly traded company, thereby further lending to its unrestricted use.

**Data Gathering:** The data used for this project will be collected via downloading the “stock\_tweets” dataset from Kaggle, which I accessible through the following URL:

<https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction/data?select=stock_tweets.csv>

Additionally, data cleaning and preprocessing techniques will be used to prepare the dataset prior to feeding it into the models used for sentiment analysis.

**Data Analytics Tools and Techniques**: The technique used will be sentiment analysis conducted through combined usage of VADER and neural networks such as LSTMs. VADER will be initially used to label the sentiments for the entire dataset. Next, the data will be split into train-validation-test sets with the train portion used to train the model and the validation set used to validate the training of the model. Finally, the trained model will be fitted on the test set in order to observe predicted vs actual sentiments from the tweets. Jupyter Notebook will be used as the IDE for this project

**Justification of Tools/Techniques:** Sentiment analysis is an appropriate technique for this dataset since tweets are bodies of text in nature that carry positive or negative sentiments. VADER is a very powerful tool that can be used to label unlabeled texts such as tweets. LSTMs can then be trained on prelabeled texts to create a model that can accurately predict sentiments from similar bodies of text/tweets.

**Project Outcomes**: The key anticipated outcome is to build a trained LSTM model that can predict sentiment from a text with at least 70% accuracy as the model evaluation metric.

**Projected Project End Date**: On or before 04/31/2025

**Sources**:

Yukhymenko, H. (2022b, December 5). *Stock tweets for sentiment analysis and prediction*. Kaggle.

https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction/data?select=stock\_tweets.csv

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

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

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.



Reviewed by:



Comments: Click here to enter text.