WGU MSDA Capstone - Twitter Sentiment Analysis on Company Stock

Executive Summary and Implications

April 25th, 2025

Jan Francis Mitchell Barlaan

Student ID: 011282929

### Problem and Context

Creating a model based on neural networks and VADER that can accurately conduct sentiment analysis on texts such as social media tweets can help 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.

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 and feelings towards investing in a company's stock, whether positive or negative, based on social media texts such as their tweets?

### Hypotheses[¶](#A3:-Hypotheses)

**Null hypothesis (H0)** - Using neural networks and VADER, we cannot create a model that can accurately predict the sentiment behind a tweet or review with at least 70% accuracy.

**Alternate Hypothesis (H1)** - Using neural networks and VADER, we can successfully create a model that can accurately predict the sentiment behind a tweet or review with at least 70% accuracy.

### Summary of Data Analysis

The Twitter data used in this analysis was obtained publicly through Kaggle and is stored in a CSV format. The raw data is composed of 80,794 different rows and four columns, with the first row being the header. Each row comprises the following columns: the Date, Tweet, Stock Name, and Company Name (Yukhymenko, 2022).

Several tools and functions were used within the Jupyter Notebooks IDE using Python kernel, but the main tools used involved VADER, Tensorflow/Keras, NLTK, pandas, matplotlib, and seaborn. Among this selection of tools, VADER, Tensorflow/Keras, and NLTK were used for preparation and later analysis. Pandas, matplotlib, and seaborn were used in data visualization tasks.

#### The data was prepared prior to its utilization in the neural network model by performing the following steps:

* Necessary packages and the beginning dataset were imported.
* EDA was performed.
* Data cleaning was performed.
* Vocabulary size was determined.
* VADER Sentiment Intensity Analyzer was used to label the tweet sentiments.
* Value counts of the sentiments were plotted to visualize the distribution.
* The proposed word embedding length was determined.
* Maximum sequence length was determined.
* The tweet data was tokenized.
* The data was padded.
* The data was split into training, validation, and test sets.
* Train, test, validation, and whole cleaned data sets were saved into individual CSV files.

Prior to data preparation, the raw format of the data is as follows.

|  | **Date** | **Tweet** | **Stock Name** | **Company Name** |
| --- | --- | --- | --- | --- |
| **0** | 2022-09-29 23:41:16+00:00 | Mainstream media has done an amazing job at br... | TSLA | Tesla, Inc. |
| **1** | 2022-09-29 23:24:43+00:00 | Tesla delivery estimates are at around 364k fr... | TSLA | Tesla, Inc. |
| **2** | 2022-09-29 23:18:08+00:00 | 3/ Even if I include 63.0M unvested RSUs as of... | TSLA | Tesla, Inc. |
| **3** | 2022-09-29 22:40:07+00:00 | @RealDanODowd @WholeMarsBlog @Tesla Hahaha why... | TSLA | Tesla, Inc. |
| **4** | 2022-09-29 22:27:05+00:00 | @RealDanODowd @Tesla Stop trying to kill kids,... | TSLA | Tesla, Inc. |
| **...** | ... | ... | ... | ... |
| **80788** | 2021-10-07 17:11:57+00:00 | Some of the fastest growing tech stocks on the... | XPEV | XPeng Inc. |
| **80789** | 2021-10-04 17:05:59+00:00 | With earnings on the horizon, here is a quick ... | XPEV | XPeng Inc. |
| **80790** | 2021-10-01 04:43:41+00:00 | Our record delivery results are a testimony of... | XPEV | XPeng Inc. |
| **80791** | 2021-10-01 00:03:32+00:00 | We delivered 10,412 Smart EVs in Sep 2021, rea... | XPEV | XPeng Inc. |
| **80792** | 2021-09-30 10:22:52+00:00 | Why can XPeng P5 deliver outstanding performan... | XPEV | XPeng Inc. |

The raw tweet data was initially unlabeled; therefore, it required labeling through a tool such as VADER’s Sentiment Intensity Analyzer. Below is the dataset after undergoing labeling of either positive or negative sentiment based on the ‘Tweet’ column. Its distribution of positive and negative sentiments was also visualized.

|  | **Date** | **Tweet** | **Stock Name** | **Company Name** | **Sentiment** |
| --- | --- | --- | --- | --- | --- |
| **0** | 2022-09-29 23:41:16+00:00 | Mainstream media has done an amazing job at br... | TSLA | Tesla, Inc. | Positive |
| **1** | 2022-09-29 23:18:08+00:00 | 3/ Even if I include 63.0M unvested RSUs as of... | TSLA | Tesla, Inc. | Positive |
| **2** | 2022-09-29 22:40:07+00:00 | @RealDanODowd @WholeMarsBlog @Tesla Hahaha why... | TSLA | Tesla, Inc. | Negative |
| **3** | 2022-09-29 22:27:05+00:00 | @RealDanODowd @Tesla Stop trying to kill kids,... | TSLA | Tesla, Inc. | negative |
| **4** | 2022-09-29 22:24:22+00:00 | For years @WholeMarsBlog viciously silenced @T... | TSLA | Tesla, Inc. | negative |
| **...** | ... | ... | ... | ... | ... |
| **55941** | 2021-10-07 17:11:57+00:00 | Some of the fastest growing tech stocks on the... | XPEV | XPeng Inc. | positive |
| **55942** | 2021-10-04 17:05:59+00:00 | With earnings on the horizon, here is a quick ... | XPEV | XPeng Inc. | positive |
| **55943** | 2021-10-01 04:43:41+00:00 | Our record delivery results are a testimony of... | XPEV | XPeng Inc. | positive |
| **55944** | 2021-10-01 00:03:32+00:00 | We delivered 10,412 Smart EVs in Sep 2021, rea... | XPEV | XPeng Inc. | positive |
| **55945** | 2021-09-30 10:22:52+00:00 | Why can XPeng P5 deliver outstanding performan... | XPEV | XPeng Inc. | positive |

A blue and orange rectangular bars

AI-generated content may be incorrect.

After the labeling process, the data needed to be lemmatized, tokenized, and padded so that the neural network can properly process it.

Lemmatization is a technique that reduces words to their dictionary form, or "lemma," based on their intended meaning and context. The process aims for more accurate and meaningful base forms.

Tokenization is used to simplify text data analysis so that a machine model can utilize the data while retaining the context. It breaks down text strings into tokens and smaller original data components. It is also an important step in preprocessing since it aids in identifying and removing both stop words and unnecessary character as well as determining vocabulary size and text sequence lengths.

Padding is the process in which variable length data is made uniformly through either the removal or addition of characters in a string.

After data preparation, including lemmatization and tokenization, the data was split into train, test, and validation sets, with an 80:20 ratio between the train and test sets and a sub-ratio of 75:25 between the train and validation sets. Each set contains its own X and y components. The ratios for the set splits were chosen since they are the most common best practices ratios.

The final steps of preparation involved padding the text sequences and changing the VADER sentiment labels to Boolean values since the neural network that processed the tweet data dealt in binary outputs.

The chosen model was based on sequential text classification. Its final structure is most accurately described as a CNN-Bidirectional LSTM, a convoluted neural network involving at least one layer of Bidirectional Long Short-Term Memory. LSTMs are a special kind of Recurrent Neural Network (RNN) capable of learning long-term dependencies in data. Bidirectional LSTM layers were employed since they account for past and future data points for every step.

The CNN portion in the final model is present due to the inclusion of a Conv1D Layer used for one-dimensional convolutional operations, typically applied to sequences like text data. The inclusion of a convolutional layer (CNNs) prior to recurrent layers (RNNs) enabled the detection of local patterns within sequences at high efficiencies. In contrast, the recurrent LSTM layers capture long-term dependencies and the data's sequential nature.

Keras/Tensorflow was used to create the model, with its output and summary listed below. Its fitness was based on the chosen criteria of accuracy. The constructed model could predict sentiment with an accuracy of about 98.6% on the train set, about 88.7% on the validation set, and about 88.7% again on the test set. Test loss was about 35.8%. The plots for validation accuracy and loss are also seen below.

**Model: “sequential”**

┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓

┃ **Layer (type)** ┃ **Output Shape** ┃ **Param #** ┃

┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩

│ embedding (Embedding) │ (None, 53, 53) │ 3,727,066 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ conv1d (Conv1D) │ (None, 51, 64) │ 10,240 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ max\_pooling1d (MaxPooling1D) │ (None, 10, 64) │ 0 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ bidirectional (Bidirectional) │ (None, 10, 200) │ 132,000 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ bidirectional\_1 (Bidirectional) │ (None, 200) │ 240,800 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ dense (Dense) │ (None, 100) │ 20,100 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ dropout (Dropout) │ (None, 100) │ 0 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ dense\_1 (Dense) │ (None, 50) │ 5,050 │

├─────────────────────────────────┼────────────────────────┼───────────────┤

│ dense\_2 (Dense) │ (None, 2) │ 102 │

└─────────────────────────────────┴────────────────────────┴───────────────┘

**Total params:** 4,135,358 (15.78 MB)

**Trainable params:** 4,135,358 (15.78 MB)

**Non-trainable params:** 0 (0.00 B)

Epoch 1/20

336/336 ━━━━━━━━━━━━━━━━━━━━ 11s 19ms/step – accuracy: 0.7329 – loss: 0.7945 – val\_accuracy: 0.8722 – val\_loss: 0.3294

Epoch 2/20

336/336 ━━━━━━━━━━━━━━━━━━━━ 6s 17ms/step – accuracy: 0.9043 – loss: 0.2777 – val\_accuracy: 0.8867 – val\_loss: 0.2973

Epoch 3/20

336/336 ━━━━━━━━━━━━━━━━━━━━ 6s 16ms/step – accuracy: 0.9414 – loss: 0.1807 – val\_accuracy: 0.8869 – val\_loss: 0.3478

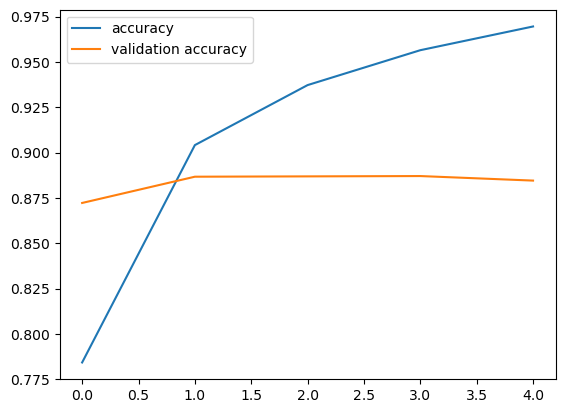
Epoch 4/20

336/336 ━━━━━━━━━━━━━━━━━━━━ 6s 17ms/step – accuracy: 0.9579 – loss: 0.1384 – val\_accuracy: 0.8870 – val\_loss: 0.3433

Epoch 5/20

336/336 ━━━━━━━━━━━━━━━━━━━━ 5s 16ms/step – accuracy: 0.9711 – loss: 0.1009 – val\_accuracy: 0.8845 – val\_loss: 0.3641

A graph with lines and numbers

AI-generated content may be incorrect.

This data frame showcases the prediction capabilities of the constructed model in comparison to the actual sentiment, the one labeled by VADER at the beginning of the analysis.

|  | **Date** | **Tweet** | **Predicted\_Sentiment** | **Actual\_Sentiment** | **Stock Name** | **Company Name** |
| --- | --- | --- | --- | --- | --- | --- |
| **40981** | 2022-05-11 18:40:19+00:00 | Published new plan thoughts on CPI, $AAPL and ... | positive | positive | AAPL | Apple Inc. |
| **6936** | 2022-06-02 16:28:52+00:00 | I know many are excited today for this rally i... | negative | negative | TSLA | Tesla, Inc. |
| **55089** | 2021-12-22 23:34:40+00:00 | BREAKING: NIO ET5 vehicle found to cure Covid-... | positive | positive | NIO | NIO Inc. |
| **22678** | 2021-11-09 12:14:28+00:00 | EV space and stocks are heating up. I hope I g... | positive | positive | TSLA | Tesla, Inc. |
| **54494** | 2022-03-25 08:45:10+00:00 | $NIO - Aims to achieve break-even in Q4 2023 w... | positive | positive | NIO | NIO Inc. |
| **...** | ... | ... | ... | ... | ... | ... |
| **21155** | 2021-11-28 13:21:16+00:00 | GREAT NEWS: Tesla Model S Plaid led the way wi... | positive | positive | TSLA | Tesla, Inc. |
| **6433** | 2022-06-09 08:29:44+00:00 | $TSLA China May deliveries slightly better tha... | positive | positive | TSLA | Tesla, Inc. |
| **53971** | 2022-08-26 12:08:28+00:00 | China Securities Regulatory Commission: The Ch... | positive | positive | NIO | NIO Inc. |
| **17231** | 2022-01-16 00:45:33+00:00 | @realMeetKevin @Tesla Thank god my kid just wa... | positive | positive | TSLA | Tesla, Inc. |
| **1616** | 2022-08-29 23:08:17+00:00 | @fraggelcurris @latestinspace @All\_That\_Ash601... | negative | negative | TSLA | Tesla, Inc. |

#### Findings

The dataset used to create this model was initially composed of 80,793 rows of tweets regarding several publicly traded companies and their stock. After initial positive/negative sentiment labeling with VADER's Sentiment Intensity Analyzer, the model was trained on 80% of the overall dataset and tested on the remaining 20% (half of this was used for validation). The constructed model could predict sentiment with an accuracy of about 98.6% on the train set, about 88.7% on the validation set, and about 88.7% again on the test set. Test loss was about 35.8%. The overall neural network architecture and the chosen set of hyperparameters contributed to these scores. Further optimization may increase the accuracy. However, the accuracy of all three split components of the data is already relatively high. Therefore, an even larger dataset to train on may yield more gains in comparison to hyperparameter optimization.

In the context of the original research question and the initial model built to answer it, using neural networks and VADER, we can successfully create a model that can accurately predict the sentiment behind a tweet or review with at least 70% accuracy. Therefore, we can reject the null and accept the alternative hypothesis since our accuracies between the train, test, and validation sets all score significantly above 70% accuracy.

#### Limitation of Chosen Analysis[¶](#E2:-Limitation-of-Chosen-Analysis)

One major limitation of the chosen analysis stems from the initial sentiment labeling performed by VADER's Sentiment Intensity Analyzer. This step was unavoidable and necessary since the tweets in the raw data set had no prior sentiment labeling applied to them. VADER was one of the best choices for this task since its strength lies in handling social media texts and their associated nuances. In a worst-case scenario, improper/inaccurate labeling of VADER could lead to bigger trickle-down effects as the data is fed into more complicated models, such as the neural networks we employed. However, this inherent weakness was also the reason for selecting a combination of convoluted and recurrent neural network features. Hybrid solutions between VADER and these neural network architectures were chosen to use the strength of one to cover the inherent weakness of another in a complementary fashion.

#### Recommended Course of Action [¶](#E3:-Recommended-Course-of-Action-Based-)

According to the results of the analyses, the hybrid approach using lexicon-based tools such as VADER in combination with neural networks is quite capable of creating a model with high accuracy for sentiment analysis. We will treat this as proof-of-concept verification and recommend exploring other similar lexicon-neural network solutions. One neural network architecture we may be interested in attempting this approach with is Transformers. One intriguing key feature of this specific architecture is that it can parallel processes and uses an attention mechanism that focuses on different parts of the input sequence when making predictions. For example, this allows it to learn relationships between words that might be far apart in a sentence.

#### Expected Benefits and Further Applications[¶](#E4:-Future-Study-and-Further-Applicatio)

Further study and application of a model based on this type of sentiment analysis can be combined with forecasting methods. One example would be time series analysis. It would also be interesting to see the results of combining the results of a sentiment analysis such as this one with basic linear regression. Regardless of the method used, it would prove insightful to determine whether a correlation exists between the current public opinion of a company/stock and whether that stock's closing price increased or decreased relative to its opening price.

Another implementation building upon the model we created could come in the form of a churn analysis and its relationship with the public perception of a company based on the sentiment analysis results.

Whether the further implementation of a sentiment analysis model such as the one we created is used for finance/company stock applications or used in conducting studies into the state of a company's PR, rate of churn, and many other applications, it proves itself as a handy tool in business and data analysis.

### Sources[¶](#F:-Sources)

Adam, F. M. (2023, July 1). Sentiment analysis with vader: Analyzing twitter sentiment. Medium. <https://medium.com/@vafaisalm93/sentiment-analysis-with-vader-analyzing-twitter-sentiment-86ae8078dacb>

Luay, M. (2023, September 22). Sentiment analysis using recurrent neural network(rnn),long short term memory(lstm) and... Medium. <https://medium.com/@muhammadluay45/sentiment-analysis-using-recurrent-neural-network-rnn-long-short-term-memory-lstm-and-38d6e670173f>

Yukhymenko, H. (2022, 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>