Stock Price Prediction using Natural Language Processing

Justin Huynh

7745112

jhuyn017@uottawa.ca

CSI4900 Honours Project

Winter 2019

Supervisor: Dr. Diana Inkpen

TABLE OF CONTENTS

1 Introduction 1
2 Review of Related Literature 1
2.1 The Nature of the Stock Market 1
2.2 Machine Learning 2
3 Description of Data 3
3.1 Financial Data 3
3.1.1 Financial Features 3
3.2 Reddit Posts 4
3.2.1 Sentiment Features 5
4 Methodology 6
4.1 Parameters 6
4.2 Baseline Model 6
4.3 Sentiment Model 7
5 Description of Evaluation Methodology 7
6 Description of Results 7
7 Discussion of Results 9
8 Conclusions 9
9 Future Work 9
9.1 International Social Media 9
9.2 Granularity 10
9.3 Feature Selection 10
10 Bibliography 11

1 INTRODUCTION

Stock price prediction is finding the future value of a company stock traded on a financial exchange. Successfully predicting a stock's future price can maximize investor gains. Ideally, investors should invest in stocks that are reliably predicted to rise in value. The stock market is volatile and prone to wild and unpredictable swings. The field of stock forecasting is influenced by a high degree of uncertainty, data intensity, noise, and hidden relationships. Factors such as political events, general economic conditions, and traders' expectations all influence the financial markets. Thus, predicting future stock price movements and developing forecasting models that consistently achieve high returns is hard [1]. Still, the problem of stock prediction remains enticing. Small improvements to current forecasting models can increase investment returns and profits by millions.

Recently, researchers have consulted computer science techniques for stock forecasting. There is an immense ocean of financial, economic, and text-based information produced every day. To identify patterns from all this data, researchers are using the power of machine learning. Machine learning algorithms are trained using given data to produce a statistical and mathematical model. They learn from patterns in the training data to come up with a solution to a problem.

The main goal of this study is to explore the predictability of stock market movement directions with Support Vector Machines (SVM) and social media sentiment analysis.

I focused on two machine learning models. First, I explored a baseline model developed using SVM and traditional financial indicators. In comparison, I added sentiment analysis as a feature to the baseline model and contrasted the classification accuracy between both models. I explored the technology sector with four stocks in particular—Apple, Amazon, Google, and Microsoft. In addition, I experimented on 30 stocks from the Dow Jones Industrial Average (DJIA). These stocks were chosen due to their popularity. It is much easier to find social media posts about large companies for analysis. In summary, the study finds that Reddit sentiment analysis in SVM can improve prediction accuracy for stock market movements between 1 and 20 days.

The repository of code and results can be found here: https://github.com/Ehsap/StockPricePredictor.

2 REVIEW OF RELATED LITERATURE

2.1 THE NATURE OF THE STOCK MARKET

The stock market is the collection of markets and exchanges where the buying, selling, and issuance of shares of publicly-held companies take place. Someone that trades on the stock market buys and sells shares of a company listed on a stock exchange. Techniques for forecasting stock price movements are essential for making well-informed buying and selling decisions.

Fundamental and technical analyses were the first methods used to predict stock prices. Fundamental analysis is a method of evaluating a stock's intrinsic value by examining related economic, financial, and other qualitative and quantitative factors. Technical analysis is the analysis of statistical trends gathered from trading activity, such as price movement and volume, to predict the value of a stock. In summary,

technical analysis is based on the belief that past trading activity and price changes of a stock can indicate the stock's future price movements [2].

A prominent theory in economic research is the Efficient Markets Hypothesis (EMH) theory, which assumes that stock prices are set from all available information and are therefore unpredictable and follow a "random walk" [3]. The random walk idea states that if the flow of information is constant and information is efficiently reflected in stock prices, then tomorrow's price changes will only reflect tomorrow's news and will be independent of the price changes today. Since news is unpredictable by its definition, and, thus, future price changes must be unpredictable and random, which implies that technical analysis is unreliable.

However, recent research has outlined that stock prices don't always follow a random walk, and that technical analysis and machine learning can help identify trends for predicting stock prices [4].

2.2 Machine Learning

Research in stock forecasting has identified Support Vector Machines (SVM) as a promising machine learning model for predicting future stock price movements. For example, in a study focused on predicting price movements of 500 companies on the Tokyo Stock Exchange, SVM outperformed linear regression and backpropagation neural network models [4]. SVM is a supervised learning model that analyzes data for classification. Given a set of training examples, each marked with the category they belong to, the SVM training algorithm builds a model that assigns new examples to one category or the other. The SVM model plots examples as points in space, mapped such that the examples of different categories are clearly separated by a hyperplane that is as wide as possible. New examples mapped into that same space would then be classified as belonging to the category in which side of the hyperplane they fall [5].

In the modern era, social media use has reached unprecedented levels, which has led to the hypothesis that the displayed public sentiment could be correlated with fluctuations in stock price. For example, public sentiment measured by tweets is correlated or even predictive of stock values for 16 of the most popular tech companies on Yahoo! Finance [6].

The thoughts and emotions of people are always an important piece of information. On social media platforms such as Twitter, Facebook, and Reddit, people can express and share their opinions. Processing text to identify and extract its subjective information and emotions is known as sentiment analysis. The goal of sentiment analysis is to identify the polarity of a given text: positive, negative, or neutral [7]. Given the vast amount of posts made on social media platforms everyday, it is possible to determine the public mood by analyzing the sentiment of all the posts. Recent research in this area has suggested that mining sentiment data from Twitter posts can help predict stock price movements [6].

Research has outlined SVM and sentiment analysis as promising methods for stock forecasting. I combined the benefits of SVM and sentiment analysis into a machine learning model for predicting stock price movements. I built upon the findings of Bollen [6] by selecting a different social media platform for analysis: Reddit. Further, I differentiate from Bollen's testing methodology by predicting stock price movements for specific stocks along with the DJIA.

3 DESCRIPTION OF DATA

I chose two dimensions of data to use as input features for the SVM models: financial data and sentiment from Reddit posts. The timeline of data collected is from January 1st, 2010 to December 31st, 2018. I chose this timeline because it has a wide variety of periods where stock prices rise and fall. For instance, this timeline has periods of market decline, such as the 2010 European sovereign debt crisis and the 2015 Chinese stock market crash.

3.1 FINANCIAL DATA

I collected daily stock price data from 2010-2018 for four key technology stocks: Apple, Amazon, Google, and Microsoft. In addition, I collected the daily prices of the Dow Jones Industrial Average (DJIA), which is a summary of the stock values of 30 of the largest American companies. This data includes the daily open, close, high, and low prices. I extracted this data from Yahoo Finance [8].

3.1.1 Financial Features

Stock price prediction is a time-series problem, in which we aim to predict future values based on data from the past. To do this, you need to look at historical trends that cover days, weeks, months, and years. I used four traditional indicators for interpreting stock price trends: stock momentum, stock volatility, index momentum, and index volatility.

Stock momentum measures how quickly a stock has changed in value within the last *n* days. For example, stocks that rise quickly in value have high momentum. Stocks that stay around the same price have low momentum. Similarly, index momentum measures the momentum of the DJIA and is calculated the same way.

Stock volatility refers to the amount of uncertainty or risk related to the size of changes in a stock's value within the last *n* days. For example, stocks that fluctuate frequently between highs and lows have high volatility. It is a measure of how risky and unstable the stock is. Likewise, index volatility measures the volatility of the DJIA and is calculated the same way.

I calculated the momentum and volatility for the specific stocks and DJIA using the formulas in Table 1 below. These formulas originate from papers by Achelis and Madge [9] and [10].

Table 1: Financial Features

Feature Name	Description	Formula
Stock Momentum	Average of the given stock's momentum over the past *n* days. Each day is labeled 1 if the closing price that day is higher than the previous day, and -1 if the price is lower than the previous day.	$\frac{\sum_{i=t-n+1}^{t} y}{n}$
Stock Volatility	Average over the past *n* days of percent change in a given stock's price per day.	$\frac{\sum_{i=t-n+1}^{t} \frac{C_i - C_{i-1}}{C_{i-1}}}{n}$
Index Momentum	Average of the DJIA index momentum over the past *n* days. Each day is labeled 1 if the closing price that day is higher than the previous day, and -1 if the price is lower than the previous day.	$\frac{\sum_{i=t-n+1}^{t} d}{n}$
Index Volatility	Average over the past *n* days of percent change in the DJIA index price per day.	$\frac{\sum_{i=t-n+1}^{t} \frac{I_i - I_{i-1}}{I_{i-1}}}{n}$

Let Ct be the stock's closing price at day t, where t is the current day, and define Ii as the DJIA index's closing price that day. A stock's directional change on a given day is labeled as $y \in \{-1,1\}$, and the index's directional change on a given day is labeled as $d \in \{-1,1\}$.

3.2 Reddit Posts

Reddit is an online forum home to thousands of communities. Reddit is free to use for the public, and every day millions of people around the world post, vote, and comment on community forums organized around their interests. For example, there are communities dedicated to breaking news, TV, sports, and cooking [11]. I chose Reddit as the source for social media posts due to the website's large community and frequent engagement. Reddit is the third largest social media platform in North America, with an average of 330 million monthly active users—making its monthly usership on par with Twitter. Furthermore, the average Reddit user spends 15 minutes 47 seconds on Reddit.com each day, compared to just over 11 minutes for visitors to Facebook.com, and 6 minutes 23 seconds on Twitter.com [12]. The frequency and volume of Reddit posts make Reddit a strong candidate for sentiment analysis.

Reddit has a ranking system, where users can vote for their favourite posts. Posts are then sorted in descending order, with the most popular posts displayed first. Posts with the highest number of votes represent the most important news of the day as ranked by the community.

To develop features for the sentiment model, I created a Python script to scrape the daily top posts from the Reddit World News forum. This forum consolidates top news articles from across the world. I

aggregated the top five daily posts specific to a company and the top ten daily posts in general. For example, posts specific to Microsoft would contain 'Microsoft' in the post title or within the post's content. I extracted the top daily posts in general by taking the 10 posts with the highest number of votes for that day. The extraction timeline is consistent with the financial data timeline—2010-2018.

3.2.1 Sentiment Features

After collecting the top daily posts from Reddit, I aggregated the daily sentiment using the Natural Language Toolkit (NLTK) and Valence Aware Dictionary and sEntiment Reasoner (VADER) library for Python. The sentiment features focus on two dimensions: company specific sentiment and general sentiment about world events. The motivation behind these two dimensions is that stock prices fluctuate based on news that targets a company directly, and indirectly due to world economic and political events.

NLTK is a library that performs natural language processing tasks [13]. From this library, I used the VADER library for sentiment analysis. I chose VADER because it is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains [14]. VADER is trained to interpret common internet slang and emoticons. For example, VADER incorporates a list of numerous lexical features common to sentiment expression on social media, including: a full list of Western-style emoticons, sentiment-related acronyms (e.g., LOL and WTF), and commonly used slang (e.g., nah, meh, and giggly). This is essential for parsing Reddit posts, which frequently contain slang and emoticons. In addition, VADER was empirically validated by multiple independent human judges, which helps to ensure sentiment classification accuracy.

VADER can process a piece of text and give it a sentiment score. It does this by summing the valence scores of each word in the lexicon, and then normalizing the score to be between -1 (most extreme negative) and +1 (most extreme positive). I used VADER to calculate the sentiment score for every Reddit post. I then aggregated the daily sentiment scores and calculated daily averages for company specific sentiment and general sentiment for the day.

I calculated the historical sentiment trends using four metrics: stock sentiment momentum, stock sentiment volatility, general sentiment momentum, and general sentiment volatility. The inspiration for these metrics came from the financial momentum and volatility indicators.

Stock sentiment momentum measures how quickly sentiment towards a company has changed in the last *n* days. A company that has received increasingly positive news in the past *n* days will have high sentiment momentum. Similarly, general sentiment momentum is a measure of the momentum of general world news.

Stock sentiment volatility refers to the amount of uncertain sentiment directed towards a company. It is a measure of the size of changes in a company's sentiment in the past *n* days. A company that constantly fluctuates between positive and negative news would have high sentiment volatility. Similarly, general sentiment volatility measures the volatility of general world news.

The calculations for each of these sentiment features are listed in Table 2:

Table 2: Sentiment Features

Feature Name	Description	Formula
Stock Sentiment Momentum	Average of the given stock's sentiment momentum over the past *n* days. Each day is labeled 1 if the sentiment that day is higher than the previous day, and -1 if the sentiment is lower than the previous day.	$\frac{\sum_{i=t-n+1}^{t} y}{n}$
Stock Sentiment Volatility	Average over the past *n* days of percent change in a given stock's sentiment per day.	$\frac{\sum_{i=t-n+1}^{t} \frac{C_i - C_{i-1}}{C_{i-1}}}{n}$
General Sentiment Momentum	Average of the general sentiment momentum over the past *n* days. Each day is labeled 1 if the sentiment that day is higher than the previous day, and -1 if the sentiment is lower than the previous day.	$\frac{\sum_{i=t-n+1}^{t} d}{n}$
General Sentiment Volatility	Average over the past *n* days of percent change in the general sentiment per day.	$\frac{\sum_{i=t-n+1}^{t} \frac{I_i - I_{i-1}}{I_{i-1}}}{n}$

Let Ct be the stock's average sentiment at day t, where t is the current day, and define Ii as the average general sentiment for day t.

4 METHODOLOGY

4.1 PARAMETERS

There are two main parameters in this experiment: *forecastPeriod* and *lookbackPeriod*. The *forecastPeriod* is the number of days into the future that we want to predict stock prices. The *lookbackPeriod* is the number of days into the past that we use to calculate historical trends such as momentum and volatility. Both the *forecastPeriod* and *lookbackPeriod* are in the range of {1, 5, 20, 90, 180, 270} days. Stock markets are closed on weekends, so each day range represents one day, one week, one month, one quarter, two quarters, and one year respectively. For each combination of {*company*, *forecastPeriod*, and *lookbackPeriod*} I created SVM classifiers for both the baseline model and sentiment model. In total, I created 245 baseline models and 245 sentiment models for comparison.

4.2 BASELINE MODEL

I used 66% of the data for training and reserved 33% for testing. This is equivalent to training on data from January 1st, 2010 to December 31st, 2015 and testing on data from January 1st, 2015 to December 31st, 2018. For each day, I calculated the stock momentum, stock volatility, index momentum, and index volatility for the past *lookbackPeriod*. These are the input features. It is important to note that the training data will start at *lookbackPeriod* days. This is because we can't calculate trends that exist prior

to our data's timeline. For example, if the *lookbackPeriod* is 1 day, the training data will start on January 2nd, 2010 instead of January 1st, 2010. Furthermore, I created a vector *Y* that holds class labels for each day *d*. The *Y* vector holds whether a stock will rise or fall in *d* + *forecastPeriod* days. If a stock rises in *d* + *forecastPeriod* days, the class for day *d* is labeled 1. If the stock falls, it is labeled -1.

After calculating the input features and class labels, I induced a classifier using the SVM classifier from the Scikit-learn library [15]. I used the default Radial Basis Function (RBF) kernel. The RBF kernel is useful for classifying data that is not linearly separable, which is often the case for stock price prediction. Linearly non-separable features often become linearly separable after they are mapped to a higher dimension feature space. In summary, the RBF kernel helps to create this high dimension feature mapping which makes it easier to classify non-linearly separable data [16].

After inducing the classifier, I tested the classifier's accuracy using the test set from January 1st, 2015 to December 31st, 2018. For each day in the test set, I used the classifier to predict stock price movements within the forecast period. I then compared the predictions with the actual classification to get an accuracy score. The accuracy score measures how many predictions are the same as the actual classification.

4.3 Sentiment Model

The sentiment model uses the same train-test paradigm as the baseline model. The only difference is that the sentiment model incorporates sentiment features into training and prediction.

5 Description of Evaluation Methodology

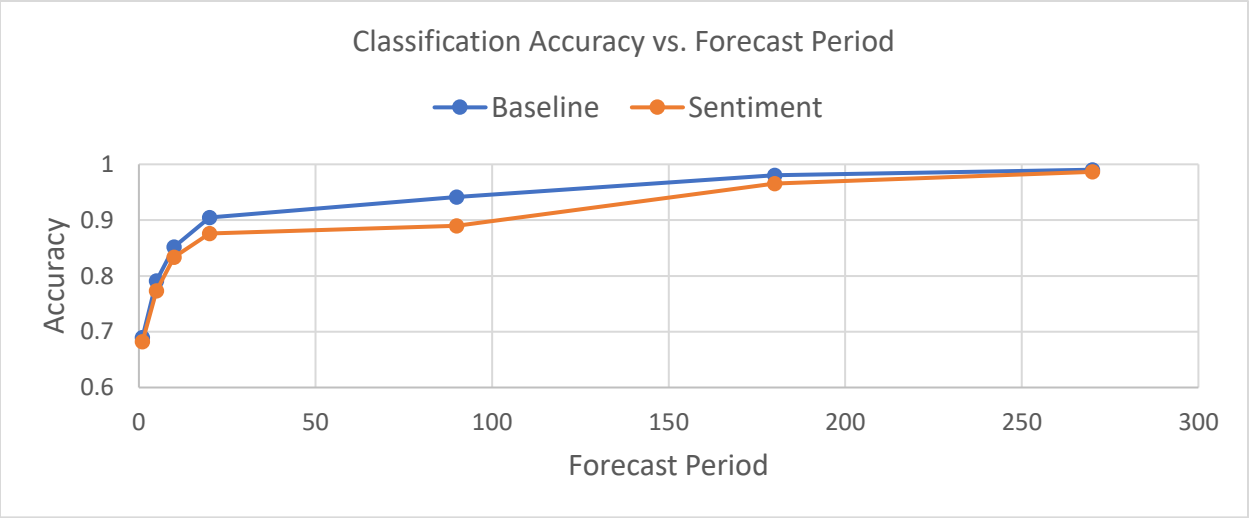
To evaluate the performance of the models, I compared the accuracy scores of each test combination for each stock (Apple, Amazon, Google, DJIA). These metrics will help indicate the average performance of each model. I identified the test combinations from the sentiment model that outperformed the baseline model and vice-versa. The usefulness of the sentiment model will be shown in the cases where it outperforms the baseline model, which indicates that sentiment analysis can be a useful input feature for classifying stock price movements.

In addition, I calculated the mean, median, min, and max accuracies for each combination of {*forecastPeriod, lookbackPeriod*} by aggregating the accuracy scores across each stock. For example, to get these metrics for the baseline model with a *forecastPeriod* of 5 days and a *lookbackPeriod* of 5 days, I averaged the accuracy scores from all baseline models with *forecastPeriod* = 5 and *lookbackPeriod* = 5.

6 Description of Results

The chart below represents the average classification accuracy for the models in relation to the forecast period. The blue line is the baseline model and the orange line is the sentiment model. For both models, the classification accuracy is lower for forecast periods between 1 and 20 days, with a classification accuracy ranging from 69% to 90%. However, for forecast periods between 90 and 270 days, the accuracy starts to get very high, with accuracies ranging from 95% to 99%. Furthermore, the average classification accuracy for the baseline model outperforms the sentiment model for all forecast periods.

Chart 1: Average Classification Accuracy vs Forecast Period



However, when examining all configurations of {*company, forecastPeriod, lookbackPeriod*}, the sentiment model outperforms the baseline model in 81/245 configurations, which is roughly 33% of configurations. When the sentiment model does outperform the baseline model, it does it with a mean improvement of 5%, minimum improvement of 0.15%, and a maximum improvement of 23%. The maximum improvement of 23% came from predicting Google stock prices 90 days into the future with a lookback period of 90 days. Given this configuration, the baseline model had a classification accuracy of 68% and the sentiment model had an accuracy of 91%. Further, the ability for the sentiment model to beat the baseline model is limited to forecast periods between 1 and 20 days. The baseline model consistently outperforms the sentiment model for longer term forecasts—180 and 270 days into the future.

Furthermore, the experimentation revealed several optimal lookback periods—the lookback periods that give the highest classification accuracy for each forecast period. For example, the best lookback period to use for predicting 1 day into the future is a lookback period of 5 days. For the rest of the forecast periods {5, 10, 20, 90, 180, 270} the optimal lookback period was symmetrical to the forecast period. In other words, if you want to predict stock prices X days into the future, you should lookback X days into the past to get the best accuracy.

Table 3: Optimal Lookback Periods

Forecast Period	Lookback Period
1	5
5	5
10	10
20	20
90	90
180	180
270	270

The accuracy scores for all configurations can be accessed at:
https://github.com/Ehsap/StockPricePredictor/tree/master/Results.

7 Discussion of Results

Initially, I expected the sentiment model to consistently beat the baseline model. This expectation would be consistent with research on Twitter mood improving the accuracy of stock price prediction for DJIA closing values [6]. However, on average, the baseline model using only financial features outperformed the sentiment model. This indicates that financial trends are more reliable than sentiment analysis for predicting stock price movements using an SVM model.

Classifier performance is limited by the quality of the input features. If the input features do not generalize the data well, then the classification accuracy will be low. Thus, the sentiment model's lower accuracy could be explained by the poor ability of Reddit Posts to correlate with stock price movements. There is a lot of noise in social media posts that may not accurately represent global sentiment towards a company. Further, I only extracted the top 10 general news posts from Reddit to calculate the daily general sentiment. The predictive accuracy of the sentiment model could be improved if I extracted a higher number of general news posts. The top 10 general news posts may not consistently represent news that impacts economic and commercial indicators, which influences the stocks in our study.

Furthermore, the ability of the sentiment models to beat the baseline models in 33% of configurations could indicate that Reddit sentiment analysis can reveal predictive trends in various economic and commercial indicators—which help predict stock price movements. This makes sense, as stock prices fluctuate based on emotional reaction to news. However, the sentiment model is unreliable when it comes to outperforming the baseline model, as it only beats it in 33% of configurations.

8 Conclusions

There are three key takeaways from these results. First, on average, a baseline model with just financial features is better for predicting stock price movements than a sentiment model using Reddit sentiment analysis. Second, a sentiment model can outperform the baseline model, but this capability is limited to short and medium term predictions between 1 and 20 days. Third, to get the best accuracy when predicting stock price movements X days into the future, you should use a lookback period of X days.

9 Future Work

9.1 International Social Media

Its worth mentioning that this experiment ignores many factors that contribute to stock price fluctuations. Firstly, the text dataset comes from Reddit—a predominately English speaking online forum. The sentiments expressed from Reddit reflect only the English-speaking part of the world that use this forum. It would be interesting to see the results of this study if they included social media sentiment from Chinese, French, and Spanish speaking social networks etc., as this would more

accurately represent global sentiment. All these indicators are features that should be explored in future experiments.

9.2 Granularity

The stock data consists of daily prices only. At many financial institutions, they build their forecasting models based on price data from a minute and second granularity. It would be interesting to see how adding this level of depth to our model would affect classification accuracy. I imagine that accuracy for one day forecasts would be improved by incorporating price data on a minute/second granularity.

9.3 Feature Selection

The accuracy for a machine learning model's accuracy is heavily influenced by feature selection. In this project, I only focused on momentum, volatility, and sentiment for a specific stock and the DJIA. Selecting input features with high correlation to the output will improve classification accuracy. The stock market has many hidden relationships and factors that can be further explored. For example, fundamental business indicators from quarterly reports such as: revenue, operating income, assets, liabilities etc.; all play a role in evaluating the health of a company, which influences the company's value on the market. In addition, we should also use market trends from international markets such as China, London, and Japan. The world's economy is heavily globalized and market fluctuations in one country can influence markets across the globe. Future work should study the incorporation of these features.

10 BIBLIOGRAPHY

[1] Y. N. S.-Y. W. Wei Huang, "Forecasting Stock Market Movement Direction with Support Vector Machine," Elsevier, 2004.

[2] J. Kuepper, "Technical Analysis: Indicators and Oscillators," 2019. [Online]. Available: https://www.investopedia.com/university/technical/techanalysis10.asp.

[3] B. G. Malkiel, "The Efficient Market Hypothesis and Its Critics," Journal of Economic Perspectives, 2003.

[4] K.-j. Kim, "Financial Time Series Forecasting Using Support Vector Machines," Dongguk University, Seoul, 2003.

[5] OpenCV, "Introduction to Support Vector Machines," 2019. [Online]. Available: https://docs.opencv.org/2.4.13.7/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html.

[6] H. M. X.-J. Z. Johan Bollen, "Twitter Mood Predicts the Stock Market," Cornell University, 2010.

[7] A. F. Diana Inkpen, Natural Language Processing for Social Media, Morgan & Claypool Publishers, 2018.

[8] Yahoo, "Yahoo Finance," 2019. [Online].

[9] S. B. Achelis, Technical Analysis from A to Z, McGraw-Hill Education, 2001.

[10] S. Madge, "Predicting Stock Price Direction using Support Vector Machines," Princeton, 2015.

[11] Redditinc, "How does Reddit work?," 2019. [Online]. Available: https://www.redditinc.com/.

[12] A. Hutchinson, "Reddit Now Has as Many Users as Twitter, and Far Higher Engagement Rates," 20 April 2018. [Online].Available: https://www.socialmediatoday.com/news/reddit-now-has-as-many-users-as-twitter-and-far-higher-engagement-rates/521789/.

[13] S. E. L. a. E. K. Bird, Natural Language Processing with Python, O'Reilly Media Inc., 2009.

[14] C. &. G. E. Hutto, "VADER: A Parsimonious Rule-based Moel for Sentiment Analysis of Social Media Text," Eigth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, 2014.

[15] Scikit-learn, "sklearn.svm.svc," 2019. [Online]. Available: https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC.

[16] A. Ng, "Non-linear SVM Classification with Kernels," 2012. [Online]. Available: http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex8/ex8.html.