LEVERAGING STATISTICAL ANALYSIS WITH STOCKTWITS SENTIMENTS TO CREATE ACCURATE FUTURE STOCK PRICE ESTIMATES

A paper to further enhance academic development in the world of socioeconomics

MARCH 18, 2020

SANJAY R. SWAMY
Mason, Ohio

Contents

Dedication

To my beloved grandfather, Gopal Rangaswamy, whom I look up to as a role model every day. His love for his family members transcended beyond just raising children, and he constantly serves as an inspiration for me to strive towards greatness. His passion for learning and his love for family will permanently stand firm with me wherever I go. I wholeheartedly owe this study to him.

Abstract

Guaranteeing a return in the stock market is as good as random. It is often said that a monkey is just as good a stock analyst as the top brokers on Wall Street as they essentially possess the same amount of luck as our primate friend. This study attempts to decode an immensely arbitrary practice and create something useful in order to better improve the odds of success. In this paper we review the effectiveness of multiple statistical concepts, machine learning programs, and StockTwits comments in being able to predict a future price of a security both numerically and directionally. Overall, we find that the concepts of Monte Carlo Simulations, Support Vector Regression, Linear Regression, Bayesian Regression, and Cumulative Distribution being placed into a train/test split in a machine learning program yields a cohesive future prediction for the security. We also find that these statistics are simply numbers on a screen without the all-important status quo. A StockTwits sentiment analyzing program was utilized to combat this and is used to either refute or reaffirm the analysis conducted by the machine learning program. Both these programs brought together into one create an efficacious method for both directionally and numerically predicting the future price of an underlying security.

Keywords: Sentiment, linear regression, support vector regression, Monte Carlo simulations, cumulative distribution, Bayesian regression

I: Introduction

1.1 Background

In the economy, sentiments are of principal significance, and this thesis mostly centers around the impact of conclusions from investors in combination with statistics and machine learning on the future of a financial instrument. This paper levers data from StockTwits in combination with rigorous statistical and technical analysis for a price in the future. This paper exhibits how fusing the assessments improves determining exactness of anticipating stock valuation. The Efficient Market Hypothesis (EMH) states that financial exchange costs are to a great extent driven by extra data and follow an "irregular walk design" [7, 8, 37, 39, 41]. Although this theory is broadly acknowledged by the examination network as a focal worldview, few individuals have endeavored to combine the power of statistics with StockTwits. This study in this way, asserts an immediate relationship between market price and sentiment. [42, 43, 44, 45].

The work in this paper is based on the work of Johan Bollen et al, Arijit Chatterjee et al, Arpit Goel et al, and Anshul Mittal et al. By measuring the mood of people on Twitter, they also attempted to predict future security prices. They used the tweet data of all Twitter users in 2008 and used the OpinionFinder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories: Calm, Alert, Sure, Vital, Kind and Happy. They cross-checked the resulting mood time series by applying their ability to predict the public reaction to the 2008 presidential and Thanksgiving Day elections. They also used causality analysis to test the hypothesis that public mood states are predictive of shifts in near DJIA values, as calculated by the OpinionFinder and GPOMS mood time series. The writers used self-organizing Fuzzy Neural Networks to predict DJIA values using previous values. Their findings indicate a remarkable precision of almost 87% in forecasting up and down shifts in Dow Jones Industrial Average (DJIA) closing prices. In this study, we will be attempting to create a model with qualitative and quantitative factors fueling the data to create a cohesive prediction for the future price.

1.2 Importance of the Problem

Social media has become a much more integral part of our society today than many realize. World renowned institution JPMorgan recently created an index titled the Volfefe index. This index tracks the impact that every tweet by President Trump has on the stock market. This is simply a small fraction of a society ruled by social media kings. Stock markets today are becoming harder and harder to predict, which makes the need for new research even more compelling. The number of factors which influence the prices of securities has seen a drastic uptick in recent years. Unfortunately, many techniques used by generations past to predict stock prices can simply be negated nowadays by something seemingly innocuous, like a tweet. Though the rise of social media has inevitably been detrimental for the stock market itself, the key to solving the problem may lie in the very thing that rattles markets all over the globe; Social media itself. Getting the best of both worlds as it pertains to statistics, machine learning and the rise of Twitter allows for a symbiotic bondage of the ideas to create sustainable predictions.

1.3. Investor Sentiment Measurements and Effect on the Stock Market

Barberies et al. [55] in their exploration study show speculation has a direct impact on stock prices. They demonstrated how strangely, investor sentiment under-responded to increasingly verifiable data, for example, profit declarations, share repurchases, profit inceptions, and went overboard to a drawn-out record of outrageous (positive or negative) exhibitions. In 2006 Baker and Wurgler [7, 8] inferred that "… floods of conclusion have plainly perceptible, significant, and customary impacts on singular firms and on the financial exchange all in all." In 2004, Thorp [50] likewise remarks on the slacking highlight of sentiment just as the potential feelings that drive costs of securities. He takes note that "week to week changes in part conclusion don't uncover important connections among assessment and market execution," yet he discovers that unnecessary speculator opinion in either a bullish or bearish course would flag a huge restricting reaction over the accompanying 52 weeks. A few investigations discover a few proportions of venture estimation foreseeing stock returns. In 2006, Lemmon and Portniaguina [56] found that financial specialist notion figures the profits of little stocks. In 2015, Zheng [57] records a negative prescient connection among metal prospects' profits. In 2015, Kaplanski et al. [58] assert that progressively positive feeling investors with better yield desires and higher goals to purchase stocks drove the price more so

than investors with a long-term vision. In 2015, Babu and Kumar [60] record that negative notion has a more noteworthy bearing on the NSE list return than positive slants. In 2012, Mittal and Goel [14] utilized feeling investigation and AI standards to discover the relationship between "open slant" and" market slant" to foresee open state of mind and utilize the anticipated mind-set and earlier days' DJIA esteems to foresee the securities exchange developments. In 2010, Bollen, Mao and Zeng [11] explored whether estimations of aggregate temperament states got from huge scope Twitter channels are related to the estimation of the Dow Jones Industrial Average (DJIA) after some time.

1.4 Importance of Sentiment and Hypothesis

The ongoing information blast has brought forth a fantastic increment in advancement. Assessment examination is a more current field that has as of late crossed from the scholastic domain to corporate use. A great part of the momentum distributed research regarding the matter was created by look into offices unequivocally connected with organizations, in 2009, Tang et al. [67]) with "the development of new web based life, for example, tweets, online journals, message sheets, news, and web content" significantly changing the environments of partnerships (in 2010, Cai et al. [68]). The scholarly supporters of the subject have consolidated numerous zones of etymology, software engineering, man-made consciousness, and brain science. All the more explicitly as referenced by Tang et al. [67] in 2009, it is "an order at the intersection of NLP [natural language processing] and IR [information retrieval], and thusly it imparts various qualities to different errands, for example, data extraction and content mining". AI systems, fundamental factual examination, and etymological semantic portrayal are likewise all around spoken to in the plans of the field. Similarly, as with numerous new fields, opinion examination is a mix of a couple of novel ideas reapplied to a wide scope of explicit parts of other more seasoned fields. Right now, attest the significance of opinion examination alongside utilizing AI and measurable investigation to have a repetitive calculation to reflect slant of the present day.

II: Method

The subsections of the method detailed will include the following. The sections will consist of statistical principles(a), integration of machine learning(b), and the inclusion of a StockTwits sentiment extractor (c) and the overall integration of all the parts (d). Each section will contain subsections detailing more specific subtopics. Subsections in section will include Monte Carlo Simulations, Fibonacci Sequence derivation, and cumulative distribution. Subsections in section B will include Support Vector Regression, Linear Regression, and the Train/Test split. Section C will include the details of the text extractor.

2.2 Parts of the project

(a) Monte Carlo simulations and Cumulative Distribution are both mathematics principles to create a prediction for the future simply based on past data. Monte Carlo simulations originated from the city of Monte Carlo and simulate outcomes many times over based on data and changes from the past to try and synthesize a prediction

Starting Price	$85.64						
Annual Volatility	24.32%						
Daily Volatility	1.532015997264 07%		DAYS IN THE FUTURE	SIMULATED PRICE			$88.65
			1	$84.96		1	57.89098163
Mean Price	56.24098057		2	$85.55		2	49.49676622
Median Price	56.17031878		3	$83.47		3	57.87597021
Standard Deviation	3.891858884		4	$83.20		4	57.0802456
Percentiles			5	$84.12		5	54.79961194
5%	50.06661058		6	$84.05		6	63.22151689
95%	63.21876062		7	$84.73		7	58.69350804
25%	53.47831839		8	$83.82		8	57.88300646
75%	58.67495707		9	$85.46		9	61.40545311
			10	$88.34		10	57.19543431
			11	$86.74		11	61.36746979
			12	$87.09		12	54.23232055
			13	$86.63		13	56.77448386
			14	$86.60		14	53.39311702
			15	$87.66		15	58.50291397
			16	$88.77		16	59.37187944
			17	$90.15		17	55.0863664
			18	$89.91		18	60.12371432
			19	$88.57		19	56.73561814
			20	$87.58		20	61.31491359
			21	$88.65		21	57.46222744

The cumulative distribution function (CDF) FX(x) describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x. For example, if the distribution inputs a value such as .4, the way it is being used in the stock model would indicate that the stock has a 60 percent chance of returning some value or higher the very next day based on data.

Fibonacci sequences are used to predict the price of a stock in the future through retracement levels. Fibonacci retracement is created by taking two extreme points (usually a major peak and trough) on a stock chart and dividing the vertical distance by the key Fibonacci ratios of 23.6%, 38.2%, 50%, 61.8%, and 100%. Stocks that retrace 38.2% or less of a trend will usually continue the trend. Retracements exceeding 61.8% indicate a reversal. Alerts will include ABC's up/down (multiple 38% retracements) and various reversal signals.

Time Frame	Trend	38.2%	50%	61.8%
Long	180.73 to 206.04	201.41 (18.29%)	.	.
Intermediate	191.43 to 206.04	201.41 (31.69%)	.	.
Short	200.32 to 216.16	214.18 (12.50%)	.	.

[Fibonacci retracement levels of QQQ]

(b) Using Support vector regression and Linear Regression, a program which utilized machine learning was created. Every stock has a closing price after each trading day. From Quandl, every closing price of the stock from the beginning of existence to the current moment is extracted. The program then attempts to figure out a mathematical way to predict the price of the next day given the closing prices.

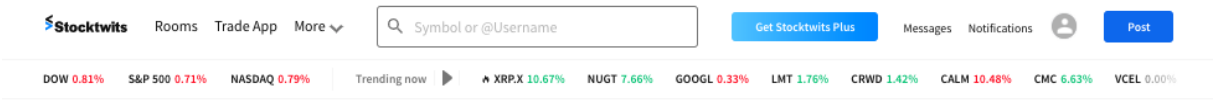
```
              Open   High    Low  ...  Adj. Low  Adj. Close  Adj. Volume
Date                              ...
2002-10-25  13.05  13.15  12.85  ...  7.611721    7.706497   13965800.0
2002-10-28  13.01  13.05  12.96  ...  7.676880    7.706497    5505900.0
2002-10-29  13.00  13.01  12.04  ...  7.131916    7.487327    2235800.0
2002-10-30  12.61  12.62  12.13  ...  7.185228    7.404398     905800.0
2002-10-31  12.56  13.02  12.50  ...  7.404398    7.469557     298900.0

[5 rows x 12 columns]
```

[example of a data set (NYSE:WYNN); Taken directly from the program]

Taking the adjusted closing price for each day, the program trains itself on the data given and attempts to come up with a method that would predict each day's adjusted closing price based on SVR and Linear Regression. A train test split is performed, and a confidence rating is spit out, which is a measure of how accurately the program was able to predict the price of the security *n* days into the future. The closer the confidence is to 100, the better. The forecast from LR and SVR are then averaged based on weighted confidence rating and a prediction is given.

(c) Statistics are simply numbers on a table without considering the status quo and the current situation. Staying up to date with the news of a stock is all important in deriving a cohesive prediction for some *x* time in the future. Using a website by the name of www.StockTwits.com, a program to stay up to date was synthesized. StockTwits is simply a huge forum where the general public can leave comments on a specific stock, where every ticker act as a sub forum, allowing comments from the public.



[area to input search query]

In the search bar, one can simply search the ticker symbol of a stock, and immediately a sub forum containing the myriad of comments from the general public is returned.

[comment example on StockTwits with flair]

This comment contains a flair, which makes it easier to attribute a sentiment to it, but some comments do not contain that flair. Even though it may come under scrutiny that random people can post comments on the stock forum, it is the general public that fuels stock markets. If more people lean one way on a stock, it will reflect that sentiment. Even though sentiment is not the only thing which determines the direction of a stock price in the future, it is an integral part of analysis. Since all comments are real time, it allows one to keep up with the current news attributed to the stock and can make a key difference in predicting the future of the stock. Each comment has a sentiment extracted and attributed to either being bullish or bearish along with percentages of each side of the sentiment.

(d) Should the sentiment extractor and the statistical analyzer disagree or have discrepancies about the directional prediction of a stock, it poses a roadblock. When this happens, the data that is fed into the statistical analysis model is altered slightly to reflect the analysis of the sentiment extractor, this is so that the status quo can be taken into consideration with the statistics. After the closing prices for the last 2 weeks are altered to reflect the sentiment, the machine learning program learns the formula and comes up with a completely new prediction. To decide how much to increase or decrease the input data by, it was decided after much testing that the amount that the sentiment leaned one way or the other is exactly how much the data would be altered that way.

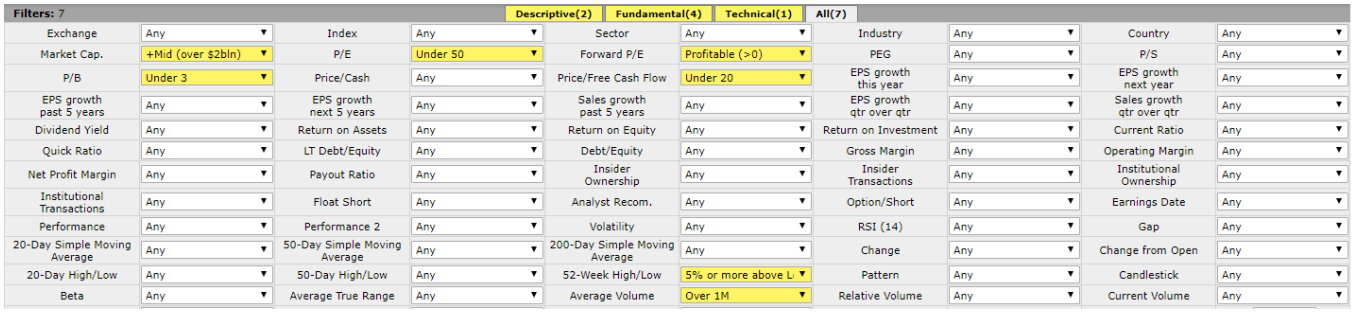
Example:

If *XYZ's* sentiment analysis indicated that **76%** of people thought that it was **Bullish** for the current situation, but the statistical analysis gave a forecast the stock would **decline** based on historical data, the changing procedure would go something like:

Take the last 2 weeks of *XYZ's* closing data and alter it by **+0.76%**. If on some given day, *XYZ*'s closing price was **$81.78**, down **0.32%** from the previous day, the new data for the closing price would instead of being down 0.32 percent, would be up **0.44%** from the previous day, with a new closing price of **$88.05**

2.3 Sampling Procedures

The sampling procedure specifically looks for stocks at 52-week lows to see if a trend reversal or continuation is in store, stocks with a volume over 1 million to eliminate the chance of low float or pump and dump stocks, and stocks which are at the median of their 52 week low and high with a market cap of over $2bn to see if a trend continuation is imminent. Fibonacci retracements (see above for explanation) were also utilized to see if on a technical side, a retracement would happen in the future.



Filters: 7					Descriptive(2)	Fundamental(4)	Technical(1)	All(7)				
Exchange	Any	Index	Any	Sector	Any	Industry	Any	Country	Any			
Market Cap.	+Mid (over $2bln)	P/E	Under 50	Forward P/E	Profitable (>0)	PEG	Any	P/S	Any			
P/B	Under 3	Price/Cash	Any	Price/Free Cash Flow	Under 20	EPS growth this year	Any	EPS growth next year	Any			
EPS growth past 5 years	Any	EPS growth next 5 years	Any	Sales growth past 5 years	Any	EPS growth qtr over qtr	Any	Sales growth qtr over qtr	Any			
Dividend Yield	Any	Return on Assets	Any	Return on Equity	Any	Return on Investment	Any	Current Ratio	Any			
Quick Ratio	Any	LT Debt/Equity	Any	Debt/Equity	Any	Gross Margin	Any	Operating Margin	Any			
Net Profit Margin	Any	Payout Ratio	Any	Insider Ownership	Any	Insider Transactions	Any	Institutional Ownership	Any			
Institutional Transactions	Any	Float Short	Any	Analyst Recom.	Any	Option/Short	Any	Earnings Date	Any			
Performance	Any	Performance 2	Any	Volatility	Any	RSI (14)	Any	Gap	Any			
20-Day Simple Moving Average	Any	50-Day Simple Moving Average	Any	200-Day Simple Moving Average	Any	Change	Any	Change from Open	Any			
20-Day High/Low	Any	50-Day High/Low	Any	52-Week High/Low	5% or more above L	Pattern	Any	Candlestick	Any			
Beta	Any	Average True Range	Any	Average Volume	Over 1M	Relative Volume	Any	Current Volume	Any			

[FinViz stock screener to narrow down basket of stocks]

2.3.1 Sample Size, Power, and Precision

The amount of stocks that were sampled vary greatly from time to time. The main idea is that the sample would come from stocks of all different backgrounds while straying from stocks which might be subject to manipulation. The aim was to take stocks from small to mid-caps as well as mid and large caps and sample a holistic figurine of stocks.

2.3.2 Research Objective

The goal of the research conducted was to find stocks which fit the category of **"Falling knife"**. A falling knife is a term used by traders to describe when one buys a security at its lowest point, in anticipation of it rising back up again. For this reason, specific stocks which were at or near their 52-week low were first screened. If stats and the sentiment behind the stock were on the rise, it would yield a good catch.

[Trend reversal example]

2.3.3 Screening Design

Every single stock subject listed on every exchange (NYSE, NASDAQ, AMEX, NYSEARCA etc.) was fed into the screener first, then the list was narrowed down to a reasonable basket of stocks. The screeners would take stocks that were 10% above their 52-week lows, and after that, the screeners explained in section 2.3 were implemented. After Fibonacci retracements were put into place did the stocks that were screened amount to around 50. In this manner, stocks that were primed for a trend reversal or continuation are returned. The next step from here was to conduct sentiment analysis on each of the stocks in an automated way.

2.3.4 Limitations of StockTwits

This and the results section explain in detail how comments can be programmatically retrieved. StockTwits has a partner API that allows users to search for tweets, users, timelines, or even post new messages. We use Selenium which is a Python based .net API which has been used to access the StockTwits API. A StockTwits developer account needs to be set up first with the necessary credentials to query the API using Selenium. The project is layered to keep the interactions, file management, application logic and algorithms separate. The following code snippet on the next page first uses. json to extract the percentage of bullish and bearish comments from StockTwits based on NLP and Machine Learning. Comments are added to make understanding the code easier.

2.3.5 Investor sentiment and comment count

In the code, I simply extract comments from every single user on StockTwits. For some stocks, this may be thousands of investors, and for some others, it may be very few. The number of users also affects the overall volatility of the data fed into the model itself, as more users implies more investment into the security yielding bigger price movements. Extracting all comments from StockTwits within a 3-week time frame provides for a general public sentiment analysis of the stock.

```
53      #adds comments that are Bullish or Bearish into corresponding counters
54      for i in json_sentiment:
55          if ((json_sentiment[i])) == 'Bearish':
56              json_sentiment[i] = 'Bearish'
57              Bearish += 1
58              totalcounter += 1
59          elif ((json_sentiment[i])) == ('Bullish'):
60              json_sentiment[i] = 'Bullish'
61              Bullish += 1
62              totalcounter += 1
63
64      total = Bearish + Bullish
65
66      #prints percentage of Bullish or Bearish comments
67      print('Percent Bullish')
68      print((Bullish/total)*100)
69      print('\nPercent Bearish')
70      print((Bearish/total)*100)
71
72      print('\n')
73      print('Total Counter')
74      print(totalcounter)
75      driver.close()
76
```

[code snippet to tally up bullish and bearish sentiment]

In order to validate whether the sentiment readings made by the program were good to use, a comment tally was utilized. The comments extracted had to be proportional to the number of watchers the stock contained.

[S&P 500 ETF on StockTwits]

With stocks like the S&P 500 ETF (NYSEARCA:SPY), the amount of comments extracted had to be a lot because comments are flying in very quickly, and there are 170,000 watchers on the stock. It is imperative that the amount of comments extracted is proportional to the number of watchers so that the program can stick to extracting comments from a certain time period and doesn't stray away from the 3-4 week timeframe.

III: Results

The results yielded by the model indicated positively about its ability to narrow stocks down from a large list of stocks fitting specific criteria. In our first study, from the 422 stocks taken, 50 stocks were narrowed down. Of those 50, the Fibonacci analysis was conducted, and 8 stocks indicated a trend continuation, and 12 indicated a reversal. Of the 8, 5 were upward trends and 3 were downward. Of the 12, 2 were bearish reversals, and 10 were bullish reversals. I have taken personal holdings in all the stocks that have been spit out by the model and reflect a bullish reading. I have also taken short positions in stocks that the model has reflected a bearish reading on.

3.2 Statistics and Data Analysis

As soon as the FinViz screener was utilized to narrow down a smaller basket of stocks, each of those securities were exported to a .csv file in Excel. Pure statistical analysis is then conducted on the data set and different predictions are given through a variety of different techniques. Another program with machine learning using Linear Regression and Support Vector Regression is also utilized in the pure statistical part of the project. The program created as a separate part from the statistics analyzes all the stocks in the original list to provide ancillary analysis for each security in the sheet. Through the power of programming and machine learning, the model takes every closing price and percentage change of the underlying security since conception. The program then runs and returns a search query from StockTwits for the ticker symbol of the stock. It then tallies up everything and gets a sentiment score by juxtaposing the percentages of Bullish/Bearish flairs and sentiment. The percentage is taken and converted into a decimal. Then, going back to the original data sets for the statistical analysis; The data of the closing prices and percentage changes from the last three weeks is altered by the sentiment score decimal to reflect a new sentiment and keep up with the status quo. After this new sheet is created with regards to the current sentiment, the algorithms and machine learning is forced to run again to create new predictions based on any sentiment score of the current day. In this manner, powerful machine learning and statistical analysis is leveraged in marriage with the status quo to create accurate future predictions.

3.5 Pulling Comments

```
23
24  website = ('https://stocktwits.com/symbol/' + str(get_ticker()))
25
26  driver = webdriver.Firefox(executable_path=r'C:\Users\Arnav.Vadnere88\Downloads\geckodriver-v0.26.0-win64\geckodriver.exe')
27  #driver = webdriver.Firefox()
28  driver.get(website)
29
30  #scroll down to allow more comments to be scrapped
31  for i in range(15):
32      driver.execute_script("window.scrollTo(0,40000)")
33      time.sleep(1)
34
35
36  comment = {}
37
38  #uses html class name to find comment and sentiment
39  el = driver.find_elements_by_class_name('st_3SL2gug')
40  le = driver.find_elements_by_class_name('st_11GoBZI')
41
42  #finds number of comment and sentiments extracted
43  length_el = len(el)
44  length_le = len(le)
45
46  #converts webelement list to regular list for easy access of data
47  for i in range(length_el):
48      comment[i] = el[i].text
49  for i in range(length_le):
50      json_sentiment[i] = le[i].text
51  data = ""
52
```

[code snippet to pull comments from stocks]

To pull comments from StockTwits, the partner API from StockTwits was utilized. With the partner API, we were able to have access to all comments from every stock ticker. The comments were then scrapped, and a sentiment score attributed to each one. The code has been annotated for better and easier understanding. This code snippet does most of the comment extraction. Using Natural Language Processing, certain words were extracted from each comment and using a variation of the bubble sorting method, the words used were compared to the words next to it. This was important because in the English language, some words on their own can mean something, but the words next to or around them could make the meaning completely different.

Examples:

Bad - Negative sentiment

Not bad - Neutral sentiment

Not bad **at all** - Positive sentiment

Good - Positive sentiment

Not good - Negative sentiment

Not **too** good - Neutral sentiment

Words in the English language that are used as modifiers to adjectives, verbs, and other adverbs can greatly alter the meaning of a sentence. These modifiers, usually adverbs, are considered in the program by looking at the comment holistically. Machine learning was also utilized in the program because it detected commonalities in the usage of those specific modifiers in the comments from the stocks. The program then learned more and more of those clauses which allowed for streamlined and efficient extraction.

Another conundrum the program had to get past was the use of curse words in comments. Since StockTwits is a crowdsourced platform with no limitations or censorship, the use of curse words is highly abundant in comments. Curse words are one of the paradoxes of the English language. For sake of this paper, I will only go into detail about one curse word, and the amount of confusion that it can cause in the program without a holistic view of the comment. It is imperative that comments with curse words are not left out in the scanning because as much as 78% of the comments extracted contained some form of a curse word. Here, we will look at the word crap which is a filler for the actual curse word which has the same meaning. Here are some clauses that were used in some comments, and after reading, it will soon be evident why machine learning was crucial to allow the program to learn the different clauses in order to make accurate sentiment predictions.

It's crap! - (it's bad) - Negative sentiment

It's THE crap! - (that's the stuff!) - Positive sentiment

Give crap to it - (telling it off) - Negative

Give a crap about it - (care about it) - Neutral/Positive/Negative depending on modifiers

Take crap - (take a beating) - Neutral/Positive ← Great falling knife indication

Piece of crap - (bad meaning) - Negative

This is only one of many curse words. There are more complex uses that curse words have in determining the meaning of the sentence. However, extracting each comment is of paramount importance in determining what the best possible sentiment score is for the stock

3.6 Code Results

```
def get_ticker():

    #gets symbol from user
    g = input("Please Enter Ticker Symbol: ")
    return g

if __name__ == "__main__":
    #declare variables
    json_sentiment = {}
    json_comment = {}
    Bullish = 0
    Bearish = 0
    totalcounter = 0
    #calls get_ticker function and creates stocktwits web address for that symbol
    website = ('https://stocktwits.com/symbol/' + str(get_ticker()))_

    #selenium requires webdrivers to perform in certain browers. In this case we are using FireFox and the webdriver
    #needed for this is geodriver.exe
    #the driver variable is set as the firefox webdriver .exe program
    driver = webdriver.Firefox(executable_path=r'C:\Users\Arnav.Vadnere88\Downloads\geckodriver-v0.26.0-win64\geckodriver.exe')

    #FireFox webdriver then open up the website that we declared in line 23
    driver.get(website)

    #scroll down to allow more comments to be scrapped
    for i in range(15):
        driver.execute_script("window.scrollTo(0,40000)")
        time.sleep(1)


    comment = {}

    #uses html class name to find all comments and sentiments
    el = driver.find_elements_by_class_name('st_3SL2gug')
    le = driver.find_elements_by_class_name('st_11GoBZI')
```

[annotated sample of code]

3.6.1 Usage of Selenium

Selenium is an intelligent web-based automation tool that allows you to manipulate a website into doing certain tasks for the programmer. This way, without using the partner API completely, we were able to extract comments from a certain time frame and still achieve the same exact results. For more details, refer to the annotated sample of code. In some cases, depending on the number of watchers a stock contains, the amount of comments may be different for each stock. Once that happens and sentiment scores are attributed, the results are given. Selenium also allows the user to input their own key words and then through machine learning, Selenium learns similar words and attaches a mood to them. During automatic annotation, any tweet with positive emoticons, like :), were assumed to bear positive sentiment, and tweets with negative emoticons, like :(, were supposed to bear negative polarity. Tweets containing both positive and negative emoticons were removed. Additional information about this data and the automatic annotation process can be found in the technical report written by Goel et al. [87].

3.6.2 Sentiment Results

```
Please Enter Ticker Symbol: TSLA
Percent Bullish
36.58536585365854

Percent Bearish
63.41463414634146


Total Counter
82
PS C:\Users\Arnav.Vadnere88>
```

[End result of sentiment analysis]

Since numerical values have been assigned to the sentiment scores of each of the comments and the stock, the next step was to alter the data that was fed into the statistical machine to reflect the current sentiment. In this case, since TSLA indicated a 63% bearish rating at the time of writing, the decimal value associated with the overall sentiment of TSLA would be -0.63. The closing price data of the stock then from the last three weeks is then altered by -0.63. If the data returns neutral, or anything below a 55% majority, the sentiment analysis will go no further, and a different stock will be analyzed.

3.6.3 Linear Regression

```
# Create and train the Linear Regression  Model
lr = LinearRegression()
# Train the model
lr.fit(x_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
# Testing Model: Score returns the coefficient of determination R^2 of the prediction.
# The best possible score is 1.0
lr_confidence = lr.score(x_test, y_test)
print("lr confidence: ", lr_confidence)
```

```
lr confidence:  0.9498228556243339
```

[Linear regression training model]

After the closing price data was inputted and the machine learning could run, a confidence rating was spit out. Refer to chapter 1 for more details on ML integration. As the model was trained further and further, a mathematical way to theoretically predict the price of the stock in the future based on trends was outputted. How close the Linear Regression model was to predicting the actual price of the security the next day is reflected in the confidence rating. The closer the confidence rating is to 1.00, the more accurate the machine learning was in predicting the price of a security n amount of days into the future. When the data is changed to reflect current sentiment, the ML program runs again and tries to find another way to predict prices in the future, but this time on a different set of data that reflects sentiment.

3.6.4 Support Vector Regression

```
# Testing Model: Score returns the coefficient of determination R^2 of the prediction.
# The best possible score is 1.0
svm_confidence = svr_rbf.score(x_test, y_test)
print("svm confidence: ", svm_confidence)
```

```
svm confidence:  0.9532119826577573
```

In the same manner as the linear regression, support vector regression was also utilized to give the model a degree of depth and complexity. To determine the best possible prediction for price in the future with sentiment, both LR and SVR had to be utilized. After this point, more analysis is conducted with different methods of extrapolation and trend analysis. Everything is then averaged out into one strong future price prediction.

3.6.5 Bayesian Regression

The Bayesian Linear Regression is a regression model based on Bayesian statistics. Bayesian technique makes use of linear regression supplemented by using additional facts in the form of a previous opportunity distribution. Prior statistics about the parameters is mixed with a likelihood function to generate estimates for the parameters. In this dissertation, we use the Bayesian Linear Regression module [17, 18] to create a regression model primarily based on Bayesian statistics. A classical treatment of regression [25] problem seeks a point estimate of the unknown parameter vector w. By contrast, in a Bayesian approach we characterize the uncertainty in w through a probability distribution p(w). Observations of facts points adjust this distribution by way of Bayes theorem, with the effect of the records being mediated through the likelihood feature. Specifically, we outline a prior distribution p(w) which expresses our uncertainty in w taking account of all records other than the information itself, and which, without lack of generality, can be written in the formα) ∝ exp - αΩ(w) where, α can once more be seemed as a hyperparameter

3.6.6 CSV Data Alteration

Close	Adj Close	Alteration	New Close	Ticker
54.73	$53.75	-0.63	$53.12	**DOX**
54.03	$53.06	-0.63	$52.43	
54.67	$53.69	-0.63	$53.06	
54.5	$53.52	-0.63	$52.89	
55.03	$54.04	-0.63	$53.41	
53.83	$52.86	-0.63	$52.23	
53.79	$52.82	-0.63	$52.19	
54.3	$53.32	-0.63	$52.69	
53.97	$53.00	-0.63	$52.37	
54.14	$53.45	-0.63	$52.82	
54.11	$53.42	-0.63	$52.79	
54.41	$53.72	-0.63	$53.09	
54.34	$53.65	-0.63	$53.02	
54.86	$54.16	-0.63	$53.53	
54.46	$53.77	-0.63	$53.14	
55.21	$54.51	-0.63	$53.88	
55.06	$54.36	-0.63	$53.73	

[Excel sheet containing original and altered closing prices]

In the *Historical Data* tab on Yahoo! Finance, any user can extract data from any time period for any stock. For the sake of this experiment, we will be using Amdocs (NASDAQ:DOX) as an example. The other forms of extrapolation for future prices take data from Yahoo! Finance. To make sure that this set of data reflects current market sentiment as well, adjusted close prices were altered by a fixed amount and the calculations were automated on Excel. This new set of data which reflected market sentiment was then imported into the other excel sheets containing Cumulative Distribution Functions and Monte Carlo Simulations. In this experiment, the original csv file from Yahoo! Finance contained data from the past 1 year and data from the last 4 weeks were changed. After the new data is fed into the new spreadsheets, new predictions are created and averaged along with the machine learning.

Most watched tickers on StockTwits

Ticker	**Watchers**
$AAPL	**360,288**
$TSLA	**293,893**
$AMZN	**251,866**
$FB	**250,002**
$NFLX	**215,007**
$AMD	**209,669**
$SPY	**172,730**
$BABA	**171,844**
$MSFT	**166,926**
$GOOG	**141,277**

For any given day, we parse the comments which contain references to the top 10 company ticker symbols from the table above and compute the average sentiment for the stock to determine validity. This data is then compared with actual closing price and extrapolated. If for a period, there are no sentiments which get generated (if there are no investors commenting) for a stock we preserve the Sentiment score which is preserved from the last day the sentiments were computed. A snapshot of the pulled data for one of the companies (Tesla) is shown in the table below.

Ticker	Trade Date	Open price	Close Price	Sentiment Category	Sentiment Score	Trade Volume	Watchers
TSLA	**2019-08-20**	**$227.62**	**$225.86**	**Bearish**	**-0.63**	**4170527**	**62,655**
TSLA	**2019-09-06**	**$227.20**	**$227.45**	**Neutral**	**0.51**	**4189372**	**71,784**
TSLA	**2019-09-17**	**$242.47**	**$244.79**	**Bullish**	**0.58**	**3946909**	**85,631**
TSLA	**2019-10-02**	**$243.13**	**$243.29**	**Bearish**	**-0.71**	**6256548**	**92,127**
TSLA	**2019-10-14**	**$247.90**	**$256.96**	**Bullish**	**0.62**	**10226860**	**101,444**
TSLA	**2019-11-04**	**$334.50**	**$337.14**	**Bullish**	**0.84**	**6074221**	**124,651**
TSLA	**2019-11-27**	**$331.12**	**$331.29**	**Bearish**	**-0.78**	**5563459**	**164,866**
TSLA	**2019-12-12**	**$354.92**	**$359.68**	**Bullish**	**0.88**	**7776211**	**195,488**
TSLA	**2019-01-21**	**$530.25**	**$547.20**	**Bullish**	**0.92**	**17803470**	**277,555**

3.6.7 Coefficient of Determination

The determination coefficient (R^2) is a measure which allows us to determine how certain predictions can be made from a given model / graph. It is the ratio of the explained variation to the total variation. It is also a measure how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain the variation. The further the line is away from the points, the less it can explain. R^2 is a statistical factor that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data. R^2 is often interpreted as the proportion of response variation "explained" by the regressors in the model. An interior value such as $R^2 = 0.4$ may be interpreted as follows - Forty percent of the variance in the response variable can be explained by the explanatory variables. The remaining sixty percent can be attributed to unknown, lurking variables or inherent variability. In case of a single regressor as in this dissertation, fitted by least squares, R^2 is the square of the Pearson product-moment correlation coefficient relating the regressor and the response variable. More generally, R^2 is the square of the correlation between the constructed predictor and the response variable. With more than one regressor, the R^2 can be referred to as the coefficient of multiple determination

3.6.8 Granger Causality

Analysis of Granger Causality [32,35,38] determines how much predictive information one signal has about another over a given time lag. The mood score for a given day is calculated over a 24 hour period The sentiment score for a given day is measured for a span of 24 hours while the close price is also determined so the sentiment score can be carried forward to affect the open price of the stock for the next day. In their study, Chatterjee and Perrizo et al [61] assert the sentiment score of the stock for a given day is seen to influence the Close Price of the stock and the Open Price of the stock for the following day.

3.6.9 Experimental Periods

In this section, we will discuss the optimal experimental periods for gathering data sets for the price of a security. For the regression analyses with machine learning, we first preprocess the daily stock data with the sentiment scores and convert it into a matrix-based format. This means that the stock price on the nth day will be affected mostly by the close price of the stock and the sentiment score for the previous n-1 days. We then move over to the different methods of cumulative distribution and Monte Carlo simulations. We consider all the stocks from the most watched stocks for the sake of this experiment and for our analysis, as dictated by Chatterjee and Perrizo [61] in their study, we perform two different forecasting experiments to test the model out – i) 7-day trading day period. ii) 14-day period including Saturday and Sunday when the markets are closed but we have sentiment data from the comments. We calculate the coefficient of determination (R^2) for each of these experiments and evaluate the optimum time frame for gathering data sets.

3.7.1 Experimental 1 - 7-day period

Ticker	Mean Absolute Error	RMSE	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
$AAPL	1.99	1.3	0.29	0.08	0.81
$TSLA	3.62	2.23	0.41	0.06	0.93
$AMZN	2.33	5.04	0.34	0.12	0.89
$FB	1.01	2.42	0.66	0.14	0.97
$NFLX	4.40	9.74	0.24	0.07	0.88
$AMD	1.22	3.88	0.19	0.16	0.96

During this time period of testing, we see that the coefficient of determination ranges from 0.81-0.96, which indicates a relatively strong line fit. We also see that the coefficients of determination are extremely close to 1. This data has been split into an 80-20 train/test split.

3.7.2 Experimental 2 - 14-day period

Ticker	Mean Absolute Error	RMSE	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
$AAPL	1.13	1.51	0.06	0.040	0.94
$TSLA	0.78	0.94	0.62	0.02	0.93
$AMZN	3.8	5.11	0.35	0.031	0.89
$FB	4.1	2.88	0.17	0.06	0.84
$NFLX	2.6	0.91	0.31	0.09	0.95
$AMD	0.35	0.57	0.78	0.052	0.97

In this fourteen-day period of testing, the coefficient of determination and the R^2 value are noticeably higher than the 7-day testing period. Our end observation is that having a longer period of testing and more sentiment heading into the trading week is extremely beneficial to being able to forecast the price for the future and contributes to a more accurate coefficient of determination. Here is a table comparing the coefficients of determination for each stock for both the 7-day experimental period and the 14-day experimental period

Ticker	Coefficient of Determination (7 days)	Coefficient of Determination (14 days)
$AAPL	0.8099	0.93681
$TSLA	0.927	0.93843
$AMZN	0.8934	0.88642
$FB	0.9722	0.8395
$NFLX	0.8766	0.948
$AMD	0.9614	0.9667

3.8 Predicting Stock Close Prices

In this subsection, we will be combining the findings from both the regression models and the extrapolation models. The forecasts will be averaged, and then given as a cohesive number. We must also determine what is the optimal time frame in which the model can forecast the closing prices accurately. We will conduct two different experiments again. The first will be attempting to predict the price daily, with the last fourteen days of sentiment being fed into the models and the data sets being altered as such. The next experiment will be conducted on a weekly basis, again with the same seven days of data being fed into the model itself. Finally, to validate the results, a RMSE test will be administered. The table below will compare the predicted closing price of a security and the actual closing price of the security. For the sake of this paper, we will be looking at the most watched stocks on StockTwits.

Daily Predictions

Date	AAPL		TSLA		AMZN	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
1/13/2020	$310.86	$311.15	$523.44	$524.86	$1,888.11	$1,880.80
1/14/2020	$311.23	$312.17	$530.65	$537.92	$1,866.63	$1,858.55
1/15/2020	$308.45	$309.55	$526.21	$518.95	$1,860.13	$1,855.09
1/16/2020	$310.76	$312.09	$520.83	$513.49	$1,864.48	$1,866.02
1/17/2020	$314.52	$315	$515.40	$510.50	$1,848.33	$1,857.25
1/21/2020	$315.77	$316	$525.66	$547.20	$1,857.80	$1,860
1/22/2020	$318.33	$317.31	$540.89	$569.56	$1,874.38	$1,883.34
1/23/2020	$316.01	$315.65	$561.42	$572.20	$1,880.51	$1,872.76
1/24/2020	$314.12	$317.52	$565.13	$564.82	$1,851.94	$1,847.44
1/27/2020	$310.54	$304.88	$555.37	$558.02	$1,824.21	$1,815.34

Date	FB		NFLX		AMD	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
1/13/2020	$220.50	$219.21	$328.06	$331.51	$47.71	$48.24
1/14/2020	$221.64	$218.63	$331.62	$335.52	$48.32	$47.91
1/15/2020	$222.02	$220.14	$334.08	$336.60	$48.90	$48.12
1/16/2020	$223.11	$220.39	$331.53	$335.85	$49.85	$48.99
1/17/2020	$222.09	$220.53	$334.48	$337.38	$50.52	$49.90
1/21/2020	$220.87	$219.12	$333.86	$332.59	$51.04	$50.70
1/22/2020	$224.16	$221.28	$328.44	$323.60	$51.38	$51.20
1/23/2020	$220.88	$219.27	$326.98	$325.01	$50.21	$50.74
1/24/2020	$218.94	$216.11	$331.63	$345.88	$50.88	$49.47
1/27/2020	$216.55	$212.50	$325.90	$341.02	$50.02	$47.90

Date	SPY		BABA		MSFT	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
1/13/2020	$326.14	$325.92	$228.33	$227.04	$160.95	$161.26
1/14/2020	$327.33	$326.84	$225.16	$224.88	$161.94	$161.72
1/15/2020	$327.85	$327.26	$222.47	$224.39	$163.20	$162.57
1/16/2020	$328.47	$329.45	$221.54	$222.73	$164.33	$164.03
1/17/2020	$324.85	$322.66	$223.56	$225.35	$166.01	$165.43
1/21/2020	$327.85	$330.82	$219.03	$220.73	$167.22	$166.43
1/22/2020	$330.62	$331.17	$221.47	$222	$166.05	$165.68
1/23/2020	$329.40	$329.41	$218.22	$216.77	$165.41	$165.27
1/24/2020	$328.51	$327.36	$214.40	$211.33	$164.13	$164.45
1/27/2020	$329.77	$330.85	$205.36	$199.50	$161.57	$160.20

Weekly Predictions

Date	AAPL		TSLA		AMZN	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
2/5/2020	$316.54	$318.95	$731.68	$734.70	$2,058.04	$2,032
2/12/2020	$327.22	$321.47	$751.31	$767.28	$2,164.96	$2,155.29
2/19/2020	$330.98	$320	$813.42	$917.42	$2,180.61	$2,161.12

Date	FB		NFLX		AMD	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
2/5/2020	$205.30	$208.71	$368.49	$362.30	$48.03	$49.31
2/12/2020	$213.05	$207.40	$380.67	$375.88	$52.51	$53.53
2/19/2020	$225.77	$216.11	$394.93	$384.90	$55.38	$57.51

Date	SPY		BABA		MSFT	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
2/5/2020	$326.86	$330.67	$220.42	$217.54	$215.31	$217.54
2/12/2020	$334.16	$336.43	$226.97	$220.21	$213.18	$181.85
2/19/2020	$339.63	$337.48	$231.15	$220.75	$204.67	$186.47

3.9 Validating the Model

The root-mean-square error (RMSE) is a frequently used degree of the differences between values (pattern and population values) predicted by using a version or an estimator and the values observed RMSE is a measure of accuracy, to examine forecasting errors of various models for a statistic. RMSE is the square root of the common of squared errors. The impact of every error on RMSE is proportional to the size of the squared errors; thus, large errors have a disproportionately massive impact on RMSE. Consequently, RMSE is touchy to outliers. We evaluate the accuracy of the models constructed for every ticker and examine it with the RMSE values of their predictions for the daily and weekly timeframe.

RMSE Values for Daily and Weekly

RMSE Values of base ticker symbol test		**RMSE Values for Daily Period 1/13/2020-1/27/2020**		**RMSE Values for Weekly Period 2/5/2020-2/19/2020**	
AAPL	1.89	AAPL	1.62	AAPL	2.43
TSLA	11.62	TSLA	10.49	TSLA	19.78
AMZN	2.74	AMZN	2.21	AMZN	3.22
FB	1.09	FB	0.92	FB	8.44
NFLX	5.62	NFLX	4.66	NFLX	6.21
AMD	1.13	AMD	1.06	AMD	0.62
SPY	0.84	SPY	0.88	SPY	0.97
BABA	4.73	BABA	4.51	BABA	5.34
MSFT	1.32	MSFT	1.34	MSFT	3.42

3.9.1 Conclusions from Testing

The RMSE test indicates that the testing values for the daily period were more in line with the testing period than the weekly period. This is indicative of the fact that the model in some parts may be underfit. However, most of the testing concludes that both the weekly and daily timeframes of forecasting seem to be in line with the original test data, which asserts that the model is for the most part, neither overfit nor underfit. The original and ending RMSEs for TSLA and NFLX are extremely high because both the stocks were at the cusp of a major unprecedented rally that increased the error percentage of the model. While some values that were predicted did not match the actual closing prices on a numerical level, the prediction from a directional standpoint was almost always correct. Even though it could not predict stocks' such as TSLA's exact values of closing, the model remained massively bullish on the stock. Overall, the model had some difficulties when it came to securities that were susceptible to volatility in its share price, or during times of general volatility.

IV: Conclusions & Discussion

In this paper, we have shown that the combination of sentiment from StockTwits together with extrapolation methods such as cumulative distribution yields an extremely accurate model for forecasting future prices of securities. Machine learning in combination with statistics is of paramount importance when considering the changes in stock prices daily. We discuss in detail how StockTwits, a crowd-sourced investment idea exchanging software, can be of the utmost help when attempting to consider the status quo and the sentiment surrounding the stock. We focused on gathering sentiment data by attributing a percentage value to bullish and bearish investors in the stock from the last two weeks. This data is converted into a percentage and then manipulated into all our extrapolation programs which then yield different numerical predictions for the future. These values are averaged into one cohesive prediction both directionally and numerically. For the sake of this paper, we focus on the most watched stocks on StockTwits.

The first portion of this paper consists of methods to screen stocks down for potential sentiment analysis and details how the screener will attempt to find stocks that fit the "Falling Knife" designation. By using Fibonacci retracements and different screening methods, stocks that were near their bottom were found. It also details the importance of sentiment analysis in determining the future of the price. We then focus on StockTwits and how it can be used to gather sentiment. By using Selenium, a web-based automation tool, we can manipulate web pages to do certain tasks. Utilizing this in combination with the Partner API of StockTwits, we were able to streamline the process of extracting comments. Natural Language Processing was utilized to determine the sentiment of words expressed in comments. Machine learning was also used to help determine the meaning of certain word modifiers and possible usages of curse words in comments. These were all tallied up to gather a percentage value of bearish and bullish.

The paper then transitions into the mathematical portion of the model. First, extrapolation methods such as cumulative distribution to yield forecasts for the future are detailed. Then, methods of regression are detailed. Three methods of regression (Bayesian, Linear, Support Vector) were all utilized as a method for extrapolation in the future. To again assert the importance of sentiment in determining the closing price of a stock, as Chatterjee and Perrizo assert in their dissertation, sentiment has a profound impact on closing price of stocks. Through a Granger Causality test, it was concluded that 14 days of sentiment should be

considered when creating a model with regression and statistical methods. Machine learning was used to create a model that utilized all three regression models with a coefficient of determination with a rating of confidence. Each method of regression took stock data from one-year past. To include sentiment in the data the regression is running on, we then explore to what degree the original data that is fed into the models is skewed. It was determined that whatever percentage the stock was bullish or bearish from the last two weeks, was the amount that each closing price for the last two weeks would be altered by. For example, if the sentiment for the last two weeks of TSLA was 73% bullish, each of the closing prices for the last two weeks of TSLA would be altered by a degree of 0.73. In this manner, the machine learning programs were forced to find a new algorithm and create a new forecast for the price in the future on skewed data that accounted for sentiment in the current day. This skewed data is fed into all the extrapolation methods and the forecasts are averaged, yielding one all-encompassing prediction.

The last part of this paper focuses on the results produced by the model and validating it. The model was tested on two different time frames. Weekly and Daily. Each time frame would have the same amount of sentiment fed into it, but they would run forecasting different time periods. We trained the regression models to predict the close price of the stocks for each day in the next two weeks. We tested these predictions and observed with the RMSE values of the trained models that the daily prediction is much more in line than the weekly forecasts. Overall, the model indicated extremely good accuracy as most of the time the model was neither underfit nor overfit from an RMSE standpoint. The model, however, was not able to account well for times of volatility on a numerical basis but despite this, the model was able to correctly predict the direction of the security. In summary, using sentiment in combination with extrapolation methods makes it possible for an investor to predict the daily closing price of stocks with relative accuracy and within a trifling margin of error.

Although the model does its best to attempt to include all actions of sentiment or FOMO, some moves are unprecedented, such as TSLA's recent skyrocket in price. Although the sentiment indicated a bullish move, an overwhelming bullish move came over the stock. This was also partly because the VIX, a volatility index, was high during TSLA's booming rally. Overall FOMO of stocks could have impacted the validity of the numbers presented but this is precisely the limitation this paper focuses on to combat, or at the very least, lessen the impact of during predictions. In addition, since so many regression methods were used to extrapolate data, the original trained model could have been slightly skewed because the three regression methods, although similar, are not the same.

4.1 Possible Changes

To better account for possible volatility in a share price, analysis of VIX may be key to creating the perfect prediction model. VIX is the ticker symbol and the popular name for the Chicago Board Options Exchange's CBOE Volatility Index, a popular measure of the stock market's expectation of volatility based on S&P 500 index options. Using the VIX values as a threshold value for future predictions could be a possible improvement on the model itself to reflect a more accurate forecast even in times of volatility.

VIX Values Throughout Jan/Feb

Date	**Value**
1/10/2020	12.1
1/21/2020	18.84
2/03/2020	17.94
2/11/2020	15.8
2/20/2020	16.54
2/28/2020	40.11

As we live in a time where nearly everything is digitalized, investors can take advantage of the internet revolution by gauging moods surrounding stocks and automating statistical analysis to make key predictions for the future. The uses of such a finding transcend beyond just the traditional investor, as an algorithm like the one presented can be used by large institutions and hedge funds to properly allocate portfolios to catch stocks that are undervalued or near their bottom. Many of the works this paper hinges on have been proven and are shown to be true by the findings presented. Algorithm traders and market makers can also utilize StockTwits as a medium for sentiment analysis, as it is a crowdsourced platform where the general public can express sentiment for the stock. Again, as stated many times, sentiment analysis from social websites is of preeminent importance in prediction of security prices in the future. In the end, even if the average investor isn't a wall street savant, the people are what drive the economy, and the people rally together as one large, ubiquitous force that is all prevailing in the current day and age; greater than any one trader on the floor.

Acknowledgements

I would like to express my utmost gratitude to both my parents, Hema and Shyam, who have continually supported me on this mission, and constantly provided me with both moral and financial support and without whose approbation I would never have achieved the things I have today. I would like to extend my sincerest thanks to Dr. Sudhakaran Prabhakaran, who motivated me to get started on such a project in the first place. I would also like to thank Arnav Vadnere, who was instrumental in creating the program that was so vital to the continuation of this project. Special acknowledgements go to Dr. Chari Ramkumar, who constantly helped me throughout the making of this paper and allowed me to get the connections to help this paper to blossom toward recognition. I would also like to extend my deepest and most sincere gratitude to all my friends who helped me in school while I was occupied undertaking this project. Meghan C, Prattyush G, Sathya P, Sai K, Neel S, and Arnav V. Again, I express my deepest gratitude to you all, and wish each one of you the best, because when a friend needed help the most, it was given without hesitation. I would also like to extend special acknowledgements to StockTwits for allowing us access to their Partner API, a valuable tool that was at the core of this project.

To all the friends who helped support my incentive, my deepest thanks extends to you even without name recognition. Without the proper motivation, I would have lost the resolve to continue this project. Additionally, to the skeptics, I am indebted to you, as it made me realize that I need to focus on a mission regardless of distractions. Skepticism motivated me to work on this more than ever and was the catalyst behind my findings, as I would stop at no length to continue my research and get a tangible solution that would work.

Last but not least, I would like to extend acknowledgements to each member of my family, I owe everything about where I am right now to the members of my family and attribute every part of me to different pieces of my family.

References

[1] Adriaans, P., & Zantinge, D. (1996). Data Mining. Harlow. England: Addison Wesley.

[2] Jiawei, H., & Kamber, M. (2001). Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann, 5.

[3] Mostafa, MM. (2013). More than words: social networks text mining for consumer brand sentiments. Expert Syst Appl. 2013;40(10):4241–51.

[4] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[5] Yang, C., Fayyad, U., & Bradley, P. S. (2001, August). Efficient discovery of error-tolerant frequent itemsets in high dimensions. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 194-203). ACM.

[6] Baker, M. and J. Wurgler (2006), "Investor Sentiment and the Cross-section of Stock Returns," Journal of Finance,(61)(4), pp. 1645-80.

[7] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[8] Baker, M. and J. Wurgler (2007), "Investor Sentiment in the Stock Market," Journal of Economic Perspectives,(21)(2), pp. 129-151.

[9] Bollen, J., Gonçalves, B., Ruan, G., & Mao, H. (2011). Happiness is assortative in online social networks. Artificial life, 17(3), 237-251.

[10] Mao, H., Counts, S., & Bollen, J. (2011, November). Computational economic and finance gauges: Polls, search, and Twitter. In Meeting of the National Bureau of Economic ResearchBehavioral Finance Meeting, Stanford, CT (Vol. 11, No. 5, p. 2011).

[11] Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. Computer, 44(10), 0091-94.

[12] Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, 450-453.

[13] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

[14] Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9(2), 181-199.

[15] Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis. Stanford University, CS229 (2011http://cs229.stanford.edu/proj2011/GoelMittalStockMarketPredictionUsingTwitterSentimentAnalysis. pdf),

[16] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[17] Hamilton, J. D. (1994). Time series analysis (Vol. 2). Princeton: Princeton university press.

[19] Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128, 1-58.

[20] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). Introduction to the Logistic Regression Model. Applied Logistic Regression.

[21] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378.

[22] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

[23] Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

[24] Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, Twitter mentions, and citations. PloS one, 7(11), e47523.

[25] Li, D., Ding, Y., Shuai, X., Bollen, J., Tang, J., Chen, S., ... & Rocha, G. (2012). Adding community and dynamic to topic models. Journal of Informetrics, 6(2), 237-253.

[26] Shuai, X., Liu, X., & Bollen, J. (2012, April). Improving news ranking by community tweets. In Proceedings of the 21st International Conference on World Wide Web (pp. 1227-1232). ACM.

[27] Mao, H., Pepe, A., & Bollen, J. (2010, July). Structure and evolution of mood contagion in the Twitter social network. In Proceedings of the International Sunbelt Social Network Conference XXX, Riva del Garda.

[28] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[29] Mao, H., Counts, S., & Bollen, J. (2015). Quantifying the effects of online bullishness on international financial markets. ECB Statistics Paper Series, 9.

[30] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

[31] Sayce, David. (2016). Number of Tweets per day. Retrieved from http://www.dsayce.com/social-media/tweets-day/

[32] Granger, C. W. (1988). Some recent development in a concept of causality. Journal of econometrics, 39(1-2), 199-211.

[33] Granger, C. W. (1988). Causality, cointegration, and control. Journal of Economic Dynamics and Control, 12(2-3), 551-559.

[34] Granger, C. W., Huangb, B. N., & Yang, C. W. (2000). A bivariate causality between stock prices and exchange rates: evidence from recent Asianflu☆. The Quarterly Review of Economics and Finance, 40(3), 337-354.

[35] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.

[36] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on (pp. 106-112). IEEE.

[37] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

[38] Arowolo, W. B. (2013). Predicting Stock Prices Returns Using Garch Model. The International Journal of Engineering and Science, 2(5), 32-37.

[39] Das, S., Poggio, T., & Lo, A. Emergent Properties of Price Processes in Artificial Markets. Ret, 10, 3.

[40] Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. The Journal of Finance, 49(3), 851-889.

[41] Dahan, E., Kim, A. J., Lo, A. W., Poggio, T., & Chan, N. (2011). Securities trading of concepts (STOC). Journal of Marketing Research, 48(3), 497-517.

[42] Chan, N. T., Dahan, E., Lo, A. W., & Poggio, T. (2001). Experimental markets for product concepts.

[43] Xu, S. Y. (2014). Stock Price Forecasting Using Information from Yahoo Finance and Google Trend. URL https://www. econ. berkeley. edu/sites/default/files/Selene% 20Yue% 20Xu. pdf.

[44] Kim, A. J., Shelton, C. R., & Poggio, T. (2002). Modeling Stock Order Flows and Learning Market-Making from Data.

[45] Lo, A., Chan, N., Lebaron, B., & Poggio, T. (1999). Information dissemination and aggregation in asset markets with simple intelligent traders (No. 653). Society for Computational Economics.

[46] Lo, A., Chan, T., & Poggio, T. (2009). U.S. Patent No. 7,599,876. Washington, DC: U.S. Patent and Trademark Office.

[47] Azure Machine Learning Group. (2016). Machine Learning. Retrieved from https://azure.microsoft.com/en-us/services/machine-learning/

[48] Parimi, Nagender. (2015). Introducing Text Analytics in the Azure ML Marketplace. http://blogs.technet.com/b/machinelearning/archive/2015/04/08/introducingtext-analytics-in-the-azure-ml-marketplace.aspx

[49] Dreman, D., S. Johnson, D. Macgregor, and P. Slovic (2001), "A Report on the March 2001 Investor Sentiment Survey," Journal of Psychology and Financial Markets, (2)(3), pp. 126-

.[50] Thorp, W. A. (2004), "Investor Sentiment as a Contrarian Indicator," The American Association of Individual Investors, Sept.-Oct. 2004.

[51] Shiller, R.J. (2003), “From Efficient Markets Theory to Behavioral Finance,” Journal of Economic Perspectives, (17)(1), pp. 83-104.

[53] Malkiel, B. G. and E. F. Fama (1970), "Efficient Capital Markets: A Review of Theory and Empirical Work," The Journal of Finance,(25)(2), pp. 383-417.

[54] Malkiel, B. G. (2003), "The Efficient Market Hypothesis and Its Critics," Journal of Economic Perspectives, (17)(1), pp. 59-82.

[55] Barberis, N., A. Shleifer, and R. Vishny (1998), “A Model of Investor Sentiment,” Journal of Financial Economics, (49), pp. 307-343

[56] Lemmon, M. and E. Portniaguina (2006), "Consumer Confidence and Asset Prices: Some Empirical Evidence," Review of Financial Studies,(19)(4), pp. 1499-529.

[57] Zheng, Y. (2015), "The Linkage between Aggregate Investor Sentiment and Metal Futures Returns: A Nonlinear Approach," The Quarterly Review of Economics and Finance, (58), pp. 128-42.

[58] Kaplanski, G., H. Levy, C. Veld, and Y. Veld-Merkoulova (2014), "Do Happy People Make Optimistic Investors?," Journal of Financial and Quantitative Analysis,(50)(1-2), pp. 145-68.

[59] Ling, D. C., A. Naranjo, and B. Scheick (2013), "Investor Sentiment, Limits to Arbitrage and Private Market Returns," Real Estate Economics,(42)(3), pp. 531-77.

[60] Babu, A. S. and R. R. Kumar (2015), "The Impact of Sentiments on Stock Market: A Fuzzy Logic Approach," The IUP Journal of Applied Finance, (21)(2), pp. 22-33.

[61] Chatterjee, A., & Perrizo, W. (2015, August). Classifying stocks using P-Trees and investor sentiment. In Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on (pp. 1362-1367). IEEE.

[62] Shefrin, H and M. Statman (1985), "The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence," The Journal of Finance,(40)(3), pp. 777-790.

[63] Shefrin, H. (1999), Beyond Greed and Fear: Understanding Behavioral Finance and the Psychology of Investing. Boston, MA: Harvard Business School Press, Revised version published 2002, New York: Oxford University Press

[64] Barber, B.M. and T. Odean (1999), "The Courage of Misguided Convictions," Financial Analysts Journal, (55)(6), pp. 41-55.

[65] Chuang, W.I. and B. S. Lee (2006), "An Empirical Evaluation of the Overconfidence Hypothesis," Journal of Banking & Finance,(30)(9), pp. 2489-515.

[66] Chatterjee, A., & Perrizo, W. (2016, August). Investor classification and sentiment analysis. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 1177-1180). IEEE.

[67] Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36(7), 10760-10773.

[68] Cai, K., Spangler, S., Chen, Y., & Zhang, L. (2010). Leveraging sentiment analysis for topic detection. Web Intelligence and Agent Systems: An International Journal, 8(3), 291-302.

[69] Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., & Chen, C. (2010). DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Systems with Applications, 37(9), 6182-6191.

[70] Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

[71] Massa, M. and V. Yadav (2015), "Investor Sentiment and Mutual Fund Strategies," Journal of Financial and Quantitative Analysis,(50)(04), pp. 699-727.

[72] Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. The American Economic Review, 1142-1161.

[73] Manski, C. F. (2004). Measuring expectations. Econometrica, 72(5), 1329-1376.

[74] Wolfers, J., & Zitzewitz, E. (2006). Prediction markets in theory and practice (No. w12083). national bureau of economic research.

[75] Berg, J., Forsythe, R., Nelson, F., & Rietz, T. (2008). Results from a dozen years of election futures markets research. Handbook of experimental economics results, 1, 742-751.

[76] Tetlock, P. (2004). How efficient are information markets? Evidence from an online exchange. Social Science Research Network.

[77] Ortner, G. (1998). Forecasting markets—An industrial application. mimeo.

[78] Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. Computers & Operations Research, 30(6), 901-923.

[79] Nagar, A., & Hahsler, M. (2012). Using text and data mining techniques to extract stock market sentiment from live news streams. In International Conference on Computer Technology and Science (ICCTS 2012), IACSIT Press, Singapore.

[80] Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. The Qualitative Report, 16(3), 730.

[81] Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015, July). Stock price prediction based on stock-specific and sub-industry-specific news articles. In Neural Networks (IJCNN), 2015 International Joint Conference on (pp. 1-8). IEEE.

[82] Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. The Journal of Economic Perspectives, 18(2), 107-126.

[83] Breen, J. (2011). R by example: Mining Twitter for consumer attitudes towards airlines. Boston Predictive Analytics Meetup Presentation.

[84] Waugh, Rob. (2012). The Tweets are paved with gold: Twitter 'predicts' stock prices more accurately than any investment tactic, say scientists. Retrieved from http://www.dailymail.co.uk/sciencetech/article-2120416/Twitter-predicts-stock-pricesaccurately-investment-tactic-say-scientists.html

Appendix

Shares of stock:

A share of stock refers to the security put out into the general public's hands by the ownership of a specific company. In return for your investment into the share of stock, you are rewarded with an extremely small percentage of ownership in the company's profits (and losses).

Bullish: Long on the stock, investors hope the price will increase

Bearish: Short on the stock, investors hope the price will decrease over time

Monte Carlo Simulations:

Monte Carlo simulations are a concept which uses the idea of randomness in order to simulate something so many times based on some number usually relating to volatility or standard deviation to find out the probability of something happening based on those simulations. For one model, I have used it to predict the stock price of a certain stock for the month based on the daily volatility of the stock. With this number I used the NORMINV function in excel to create a simulation for a given trading month. Using what-if analysis along with standard deviation, the model is able to give me an idea of what price it will be above at the end of the month and give you a general idea of where the stock is headed purely based on statistics.

FOMO: Fear of missing out. In the hopes that they won't miss another rally, there is a surge of buying in certain stocks. We saw this with TSLA's unprecedented rally around February.