Finding Wisdom in the Crowd: How Stock Market Returns can be Effectively Predicted Using Twitter Sentiment and Past Return Data

By:

Connor Collins

Honors Thesis

Presented to:

Santa Clara University Honors Program

Mentor: Dr. David Zimbra

Reader: Dr. Wendy Ku

June 2021

1. Introduction

Twitter is one of the most widely used social media platforms, boasting over 330 million monthly users[1]. One of the popular features of tweets is the "cashtag," which is used to talk about a particular stock by using its stock ticker preceded by a '$' (i.e., for Apple, a tweet would reference $AAPL). Numerous studies have been conducted on the correlation between tweet sentiment and returns for the mentioned security. For example, Mao et al. found in 2012 that tweets related to the S&P 500 had a correlation with market returns at a significance level of 1%, and Ranko et al. found a similar correlation in 2015 between tweets and the following 10 days of returns[2].

In this study, the utility of using tweet sentiment in tandem with prior return data to inform a trading strategy will be investigated. This is useful to conduct against the context of prior research for the purpose of reaffirming past studies and comparing results from these studies with an environment of increased retail investment interest. Largely because of the COVID-19 pandemic, with many Americans stuck at home without work and with stimulus checks at their disposal, trading volume in American stock exchanges increased by over 50% from 2019 to 2020[3]. Coined the "retail wave" by Deutsche Bank, 53% of respondents in a survey put out by the investment bank planned to use half of their stimulus check for investing, representing a "sizable inflow" of capital into the market[4]. Additionally, use of cashtags has increased in recent years, with over 20 million tweets in 2020 containing the syntax as a reference to stock[5]. Thus, it is useful to analyze ever changing trends in the market and how they may be predicted using Twitter data.

With the implications on the economy of the stock market, numerous theories have been posited about the returns an investor can expect. One of the most popular is the Efficient Market Hypothesis, which comes in three forms: weak, semi-strong, and strong. The weak form suggests that all past information is priced into securities, but fundamental analysis of a security can provide investors an advantage in the short term. The semi-strong form states that all public information is instantly priced into securities, only allowing investors an advantage if they are privy to private information. Lastly, the strong form states that all information, private and public, is priced into securities, and there is no ability for an individual investor to gain an advantage over the market. This study will evaluate the accuracy of these forms. Another popular theory is that of behavioral finance, which states that individual investors do not behave rationally, letting their emotions influence their trading decisions. Issues like overconfidence and the fear of regret negatively impact investors' ability to rationalize decisions and inhibit them from consistently achieving significant returns. This study's results will indirectly comment on the effects behavioral finance has on retail investors.

2. Dataset

A dataset of approximately 465,000 tweets with cashtags referencing a stock in the Dow Jones Industrial Average (DJIA) from July 1, 2020 to January 30, 2021 will be used in this analysis. To scrape these tweets, the Python module *Tweepy* was used, a common interface for interacting with Twitter's API. The DJIA was used because of its size and makeup; with 30 stocks that cover every sector except utilities, and its significance as one of the main American stock indexes, the DJIA is a well-known, diversified index to conduct this study on. Of note, three companies were replaced in the DJIA in August of 2020. Because this event occurred close to the beginning of the time range, the added companies were used for

this study and removed companies were ignored. A period of seven months was utilized to capture an adequate sample size of varying returns.

Tweepy's parameter for filtering retweets was used, meaning that only the original tweet and not retweeted copies were scraped, as gathering the same tweet multiple times could lead to redundant data collection. Instead, the number of retweets was pulled for each record, in addition to the username, date, tweet ID, and text of the Tweet.

Generating a time frame of this length required a creative approach. The free Twitter API only allows searching for a keyword's occurrence in public tweets within the last week. However, Twitter allows searching of individual users for up to their most recent 3,200 tweets. To acquire seven months of tweets related to the DJIA, Twitter was searched for the 30 DJIA cashtags multiple times over the course of December and January, and then each user's account was searched for additional mentions of DJIA tickers dating back to July 1, 2020. This approach is imperfect but was the best option available to locate several months of tweets. As over 400,000 tweets were obtained related to the DJIA through this method, an adequate sample was sufficiently compiled.

For the stock price information, data was downloaded from Yahoo Finance, using the very user-friendly Python module *yfinance.* Extracting the closing price from the data for each company within the same date range as the tweets, the return day over day was calculated for each stock.

3. Methodology

3.1 Preprocessing

As a first step in the classification process, the text in each tweet must be filtered down to the words that are most valuable in sentiment analysis. To begin, each tweet's text was tokenized into a list of words split at spaces, to allow for easier text cleaning and later use with the sentiment classifier.

In a typical tweet, there is a small proportion of words that may signal the ultimate sentiment. Thus, it is very important to refine the words in each tweet down to words of value and standardize these across the corpus so that the classifier can generalize about tweets of equal sentiment. Examples of words worth removing are stop words (words like "the", "and", etc. that do not add value), usernames in tweets, and numbers. Using the Natural Language Toolkit's (NLTK) list of stop words, any token that matched a word on the list was removed from the tweet's tokens. Once stop words were removed, all tokens were set to lowercase, and stemmed using Porter Stemmer. This reduces all words to their stem (i.e., "buying" and "bought" are stemmed to "buy") so that positive and negative words in different forms can be generalized as the same for the classifier. Additionally, using the *words* package from the NLTK that provides a list of nearly every word in the English language, any words that did not match a word in this lexicon or its stemmed equivalent was removed as these do not aid in English sentiment classification.

Lastly, there were several words removed by the various NLTK techniques that were added back to the token list for tweets because of their relevance to this study. NLTK's stop words list is general and not specific to financial analysis, so words such as "high" and "low" that may be pertinent to stock discussion were added back as tokens to any tweets that mentioned them, after being stemmed. Further, there are non-English tokens that may aid in sentiment analysis that were added back in, such as shorthand references to financial terms like "$vix" (using the cashtag for the VIX index) or "eps".

3.2 TF-IDF

As is consistent in natural language processing literature, a Term Frequency-Inverse Document Frequency (TF-IDF) matrix was utilized to represent the importance and location of tokens. To construct this matrix, sklearn's *TfidfVectorizer* is applied to the tokens for each tweet. This creates a matrix of TF-IDF values for later use in classification. The formula for TF-IDF can be seen below, where $w_{i,j}$ is the TF-IDF score for word i in tweet j, $tf_{i,j}$ is the proportion of words in tweet j that word i makes up, N is the number of tweets, and df_i is the number of tweets that the word appears in.

$$w_{i,j} = tf_{i,j} * log(\frac{N}{df_i})$$

The TF-IDF matrix that is created is a two-dimension array that contains each array of TF-IDF scores for each tweet inside an outer array. The length of each inner array is equal to the total number of unique tokens, and each tweet contains either a 0 (representing that the word is not contained in the tweet's tokens) or a non-zero TF-IDF score calculated using the previously explained process for the length of its array. As English words mean nothing to Python, this matrix allows for classification algorithms to interpret the location and importance of each token in the corpus. The specific application of the TF-IDF matrix will be explained in the following section.

3.3 Classification

To generate the sentiment for each tweet, a rating of either bearish, neutral, or bullish for 1,000 tweets from 2020 was manually classified. In the sample set, 70 bearish tweets, 578 neutral tweets, and 352 bullish tweets were identified. To conduct classification of these tweets, a TF-IDF matrix was created and utilized as the independent variable and the sentiment value of -1, 0, or 1 was the variable to predict. When applying classification algorithms to this subset, performance was low as bias to the majority neutral class was very high. The most effective classifier, a Logistic Regression algorithm, failed to perform above a macro F1-score of 0.57 in 10-fold Cross Validation (CV).

To solve this problem, the minority classes were upsampled to better balance the dataset, with the optimal performance coming from a distribution of 550, 578, and 552 for the bearish, neutral, and bullish tweets, respectively. After testing various classifiers, a Support Vector Classifier (SVC) was selected due its superior performance. By tuning hyperparameters through 10-fold CV techniques, a linear kernel and regularization parameter of 17 were identified as the optimal parameters. Using macro F1-score as the metric of choice for its summarization of both precision and recall, tuning of the SVC led to a 10-fold CV macro F1-score of 0.87. Once the SVC was tuned to satisfaction, the *TfidfVectorizer* was refit to the sample tweets and used to transform a TF-IDF matrix for the rest of the tweets in the sample. Lastly, the SVC was refit to the sample set of tweets and used to predict the sentiment for the remaining tweets in the study using the TF-IDF matrix.

3.4 Bullishness Factor

With sentiment classification now applied for each tweet, a bullishness factor was calculated for each stock, each day. In previous work done by Mao, H., et al., bullishness has been defined as below, where B_t represents the bullishness score for a company's stock on a particular day, $M_t Buy$ is the number of bullish tweets for the day, and $M_t Sell$ is the number of bearish tweets for the day.

$$B_t = \frac{(1+M_t Buy)}{(1+M_t Sell)}$$

This method of scoring bullishness was originally used in the analysis but failed to lead to a model with satisfactory returns. Building off work done by Nisar, M. & Yeung, M., a modified bullishness formula was created to reflect the influence of retweets on a tweet's importance[6]. Fundamentally, the more a post has been retweeted, the more that sentiment is likely to be held throughout the Twittersphere. In accordance with this, a weighted sentiment for each tweet was recalculated, as shown below where WS_t represents the weighted sentiment for tweet t, S_t is the original sentiment of tweet t, and RT_t is the number of retweets for tweet t.

$$WS_t = |1 + (S_t * RT_t)|$$

Fed back into the formula for the bullishness factor, with M_t *Buy* now representing the summation of the weighted sentiments for bullish tweets and M_t *Sell* now representing the summation of the weighted sentiments for bearish tweets, a new bullishness index was created. This now represented the bullishness factor for each stock, each day throughout the duration of the study.

3.5 Feature Engineering

Utilizing the closing price data from Yahoo! Finance, a data frame of returns for each day was constructed and merged with the bullishness index to create a unified set of data. Dates falling on weekends and holidays were dropped, so that previous day sentiment and return data could be analyzed on the same scale. For sentiment and return data, the previous four entries were compiled as additional potential variables in the data frame. The data frame was then split into two, where data frame one (DF1) comprised the 110 trading days in 2020 to be used for model selection and data frame two (DF2) comprised the 19 trading days in January to test the model on.

4. Building a Predictive Model

The strategy utilized for this paper's strategy was a buy or sell indicator based on the model's predicted return for the next day. Importantly, there were two assumptions used. First, the investor using the model would allocate an equal amount of money for each stock purchased per day, and the amount invested each day is equal. This is a realistic, attainable feat given today's brokerages offering fractional shares, and makes interpretation of the results much more straightforward. Second, the investor would buy or sell stock each day at that day's closing price. While in theory this may not be exactly possible, assuming that a trader would utilize the model in the closing minutes of each day is effective enough for the purposes of this study.

The general process for building the model was to train the algorithm of choice on each company's independent variables individually, and then evaluate the results by using the average score. The first approach was to use regression models like sklearn's *Linear Regression* and *Ridge Regression*. However, these models interestingly proved ineffective at predicting returns for both DF1 and DF2 and failed to beat the market for either date range. One likely reason is the low correlation between any of the variables and the returns for the following day, shown below in Table 1. These numbers represent the average of each stock's Pearson correlation with each variable in DF1. With a large timeframe, companies in multiple sectors, and the overall nature of the market, it is not necessarily surprising that the correlation coefficients were low.

Table 1. Correlations by Variable

Variable	Sentiment same day	Sentiment yesterday	Sentiment 2 days ago	Sentiment 3 days ago	Sentiment 4 days ago	Return yesterday	Return 2 days ago	Return 3 days ago	Return 4 days ago
Correlation Coefficient (R^2)	0.0162	0.0009	0.0188	0.0182	-0.0502	-0.0270	0.0376	0.0546	-0.0793

With the knowledge that linear regression was ineffective at predicting returns, the next step was to investigate classification algorithms. To do so, dummy variables were created to represent the next day's returns either being positive or negative and represented as a 1 or 0, respectively. Classification proved to be more effective at generating alpha returns, and the best model found was sklearn's *Logistic Regression* (LR). Utilizing 10-fold CV, a regularization parameter of 0.025, and class weights of 0.2145 for 0 and 0.53 for 1 were found to generate the highest alpha. Additionally, the LR model performed best when only training on sentiment from one and two days ago, and returns from two and three days ago.

5. Findings

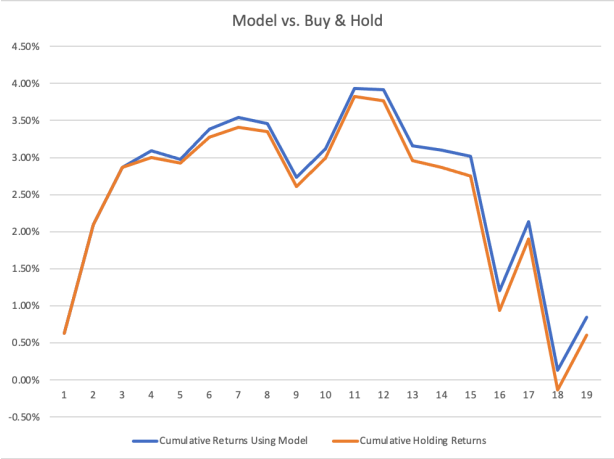
As a metric of comparison, the model was compared against a market portfolio that bought each DJIA company's stock at an equal weight on the final trading day of 2020 and held until the final trading day of January 2021. For the month of January, the model returned a cumulative gain of 0.84% versus the market's 0.61%. Utilizing a paired t-test, the model's daily returns versus the market's generated a t-score of 1.55, implying a p value of 0.0695 for the one-tailed test. Interestingly, the model only beats the market on nine of the nineteen trading days in January, while it ties on four and loses on six. While this margin of victory seems small, it is the magnitude of the wins versus losses that accounts for the model's success. The average margin of loss to the market is -0.025%, compared to the average margin of victory of 0.042%. The results by day can be seen in Table 2, and cumulative gains of the model versus the equally weighted market portfolio can be seen in table 3.

Looking at the table of gains, the model's highest cumulative success occurs at day 14 and 15, aided by two of the four biggest gains happening in days 11 and 12. One possible explanation for this is the significance of the presidential inauguration in the days leading up to this, and the relationship this had to highly polarized tweets regarding the DJIA. The changing of powers has obvious implications to the stock market, and some of the highest weighted sentiment scores occurred leading up to Biden's inauguration. The relationship of highly weighted sentiment scores and high gains further supports the relationship between tweet sentiment and returns that this study investigated.

Table 2. Gains and Losses by Day

Day	Model Gain	Holding Gain	Difference	Cumulative Difference
0	0.0062984	0.00631361	-1.5E-05	-1.51666E-05
1	0.0145991	0.01459915	0	-1.51666E-05
2	0.0077646	0.00776462	0	-1.51666E-05
3	0.0022121	0.00131504	0.000897	0.000881844
4	-0.001128	-0.0007308	-0.0004	0.000484778
5	0.0040512	0.00347684	0.000574	0.001059141
6	0.0016109	0.00129792	0.000313	0.001372163
7	-0.000854	-0.000547	-0.00031	0.001065169
8	-0.007243	-0.0074154	0.000172	0.001237437
9	0.0038537	0.00385368	0	0.001237437
10	0.0081617	0.00827881	-0.00012	0.001120374
11	-0.000136	-0.0005263	0.00039	0.001510737
12	-0.007628	-0.0081345	0.000507	0.002017708
13	-0.000565	-0.0008966	0.000332	0.002349339
14	-0.000854	-0.0011443	0.00029	0.002639311
15	-0.018112	-0.0181117	0	0.002639311
16	0.0092929	0.00966052	-0.00037	0.002271682
17	-0.020034	-0.020398	0.000364	0.002635301
18	0.0071296	0.00741283	-0.00028	0.002352074

Table 3. Cumulative Gains versus Market



Also of interest is the model's success in predicting excessive, abnormal returns. When looking at all predictions, the model only correctly predicts the next day's direction 50.5% of the time. However, when looking at returns whose absolute value is greater than 1%, the model is right 58.5% of the time; for 2% and 3%, this value increases to 70.2% and 96.7%, respectively. This likely serves as the explanation for the lackluster accuracy rate, but significant results: the model isn't always correct, but it rarely misses an abnormal return.

6. Conclusion and Future Work

Given the statistically significant results based on six months of publicly available data, this study rejects the Efficient Market Hypothesis. Even in its weak form, the Hypothesis states that future stock prices are not influenced by past events. With a return for January that is 37.7% higher than an evenly weighted market portfolio, this paper validates those investors that believe they can beat the market with an informed, fact-based approach. On the other hand, this study validates the psychological imperfections that humans bring to investing through. No model was able to beat the market using Twitter sentiment alone; it was only through combining sentiment with historical return data that a superior trading strategy was able to be constructed.

For future work, there is room for both improvement in the design of this analysis and other potential applications. First, a more comprehensive dataset would do a keyword search for all tweets throughout the entirety of the time frame, as opposed to the two-step approach that was taken here. While this was all that was possible due to constraints of Twitter's free API, it is certainly worth mentioning. Second, a better representation of the investing world would include all languages, and not just English. There are many bilingual Americans that presumably are involved in the stock market, in addition to all the international investors who have money in American exchanges. A better study design could be to scrape relevant tweets from all languages, and then use a translation tool to determine sentiment. Lastly, given the Logistic Regression's success at predicting outsized returns, another interesting application of a study like this could be to focus on projecting this phenomenon, perhaps using a different strategy such as options to exploit this success.

7. Acknowledgments

I would like to thank my mentor, Dr. David Zimbra, for his assistance in developing my methodology and providing guidance throughout my study. His knowledge in the field of data science is vast, and this thesis would not have been possible without him. I also would like to thank my reader, Dr. Wendy Ku, for her aid in writing my paper and identifying relevant financial conclusions in my work. She is an expert in the financial field, and this also would not have been possible without her.

8. References

[1] Ellyatt, Holly. "Young Retail Investors Plan to Spend Almost Half of Their Stimulus Checks on Stocks, Deutsche Survey Claims." CNBC. March 09, 2021. https://www.cnbc.com/2021/03/08/how-the-young-plan-to-spend-stimulus-checks-deutsche-bank.html.

[2] Hentschel, Martin, and Omar Alonso. "Follow the Money: A Study of Cashtags on Twitter." *First Monday*, 2014. doi:10.5210/fm.v19i8.5385.;
Mao, Yuexin, Wei Wei, Bing Wang, and Benyuan Liu. "Correlating S&P 500 Stocks with Twitter Data." *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, August 2012. doi:10.1145/2392622.2392634.

[3] Nisar, Tahir M., and Man Yeung. "Twitter as a Tool for Forecasting Stock Market Movements: A short-window Event Study." *The Journal of Finance and Data Science*4, no. 2 (2018): 101-19. doi:10.1016/j.jfds.2017.11.002.

[4] Pisani, Bob. "Trading Volume Is up from 2020's Breakneck pace as Retail Investors Jump in." CNBC. January 22, 2021. https://www.cnbc.com/2021/01/22/trading-volume-is-up-so-far-from-2020s-breakneck-pace-as-retail-investors-get-even-more-active.html.

[5] Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. "The Effects of Twitter Sentiment on Stock Price Returns." *PLoS ONE*10, no. 9 (September 21, 2015). doi:10.1371/journal.pone.0138441.

[6] Tankovska, H. "Twitter: Monthly Active Users Worldwide." Statista. January 27, 2021. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

9. Appendix

Appendix A. Number of Tweets Per Company

Ticker	Number of Tweets
$VZ	5613
$MSFT	47131
$MMM	4205
$TRV	2011
$WBA	4956
$AXP	2839
$UNH	3891
$AAPL	104253
$IBM	7437
$HON	2032
$CSCO	6784
$JPM	16033
$KO	19607
$CVX	4254
$WMT	19755
$PG	7100
$DOW	3606
$MCD	7149
$GS	21228
$CAT	8106
$NKE	6582
$INTC	15266
$JNJ	13581
$V	101738
$DIS	23730
$BA	76985
$HD	10387
$MRK	5660
$CRM	15808
$AMGN	3181
Total	570908