Statistical and Sentiment Analysis of Reddit to predict stock market movements

Hwa Chong Institution
Project 8-22

Low Jit Yong Ernest
Ho Cheng Xun Bennett
Koh Jing Wei Jae

1 Introduction

1.1 Rationale

The Efficient Market Hypothesis(EMH) implies that it is not possible to make an accurate prediction of the stock price consistently such that profit can always be earned from trading a stock. However, as hypothesised in behavioural economics, it is possible to predict stock market movements using trader sentiments. Thus, we aim to assess the validity of the EMH via mathematical means by analysing comments on specific stocks on online economic forums to craft a model to accurately predict the change in stock prices.

1.2 Research Questions

1. Is there a relationship between traders' sentiments on online forums and stock market movements?
2. Can a model that quantifies this relationship be crafted with reasonable accuracy to predict future stock market movements?
3. Is the Efficient Market Hypothesis valid?

1.3 Objectives

1. Develop a classifier to determine average daily sentiments about the market(positive/negative) from forum comments.
2. Identify if a relationship exists between changes in stock price and sentiment scores and construct a regression model that accurately regresses stock price movements against sentiment scores.
3. Quantify the strength of the relationship between sentiment scores and changes in stock price

1.4 Scope of Study

1. Forum: Reddit (subreddit: r/wallstreetbets)

   The subreddit is focused on trading so sentiments expressed are focused on short-term changes in stock price

2. Stock: SPY[1] (SPDR S&P 500 ETF Trust)

   SPY has a large trading volume so it will respond to sentiment changes quickly and it has a strong following on Reddit, providing us with a large sample group

[1] Technically, SPY is not an individual stock, but an exchange-traded fund. For most purposes, we can treat it as a stock.

2 Literature Review

2.1 Previous work

Fuzzy Neural Networks were used to map twitter sentiments against the Dow Jones Industrial Average(DJIA) closing values with an 86.7% accuracy and a reduction in Mean Average Percentage Error(MAPE) by more than 6% (Bollen et al, 2011). A Granger-causal relationship was also found between changes in the Anxiety Index characterised by sentiments showing worry and changes in returns of the stock market (Gilbert & Karahalios, 2010).

Previous studies show that there is the possibility of a cause-and-effect relationship between investors' sentiments and market movements that can be investigated.

Current programs that predict stock prices utilise machine learning that analyses historical pricing and technical indicators(patterns that signal projected change of stock prices).

Instead, our novel approach analyses traders' sentiments, predicting stock prices from a different perspective.

2.2 Sentiment Analysis

Sentiment Analysis is the qualitative and quantitative classification of emotions in text data by ranking words along a polarity spectrum of emotions:

Negative Neutral Positive

It is used to analyse comments and assign a sentiment score to comments based on the sentiment scores tagged to the individual words in the text.

The sentiment analysis program Azure Machine Learning, a Naive Bayes Classifier functions by:

1. Cleaning up the text

   Grammar errors are corrected and stop words(words denoting grammar) are removed.

2. Generating sentiment scores

Words denoting emotion are compared with the lexicon(library of words with sentiment scores tagged to them). Then, an overall sentiment score of the text is calculated.

Sentiment scores are determined by logistic regression which models the probability that sentiment scores are positive. Positive sentiments belong to the interval (0.5,1] and negative sentiments belong to the interval [0,0.5) where $|p(x) - 0.5|$ is the extent of deviation from a neutral sentiment.

3 Methodology

3.1 Data Collection

1. Comments with the keyword “SPY” from the subreddit r/wallstreetbets from 1/1/2019-6/30/2019 are scraped from the Reddit API using a Python Program and compiled in a .csv file, recording the publish date of the comment and text.
2. Historical data of SPY stock is downloaded from Yahoo Finance and the closing price of SPY stock from 1/1/2019-6/30/2019 is retrieved.

3.2 Data Processing

1. Comments are run through the Azure Machine Learning Sentiment Analysis Program and average sentiment scores are calculated for each day using a pivot table.
2. The percentage change in SPY price stock per day is calculated using the formula:

$$\%increase = \frac{p_1 - p_0}{p_0} \times 100\%$$

p_1 = closing price of current day

p_0 = closing price of previous day

3.3 Linear regression

1. A scatter plot of percentage change of SPY closing price/% against sentiment score from the previous day is plotted and a best fit line is drawn between the data points. The one day gap

allows time for traders' sentiments to take effect.

3.4 Quantifying the relationship

1. The p-value is calculated from the t-statistic for hypothesis testing. If the p-value<0.05, null hypothesis is rejected and the relationship is validated.
2. The R^2 value is calculated. The model is viable if the R^2 value>0.70
3. The Mean Squared Error of the model is calculated.

4 Results

4.1 Data processing

Publish dates and comment text were compiled in a .csv file

Publish Date	body
6/1/2019 8:04	I was in
6/1/2019 8:06	TSLA not on spy
6/1/2019 8:09	Cool, now I feel less bad
6/1/2019 8:14	I would say
6/1/2019 8:21	You are assuming our Pr
6/1/2019 9:54	Only been playing SPY th
6/1/2019 11:40	IWM =
6/1/2019 11:42	Riding 170
6/1/2019 12:08	Cashing out at 290+ wou

Figure 1: Sample dataset of publish dates and comment text [2]

Historical closing price of SPY stock is downloaded

[2] Some text is cut off due to the size of the cell

date	close
4/1/2019	285.83
4/2/2019	285.97
4/3/2019	286.42
4/4/2019	287.18
4/5/2019	288.57
4/8/2019	288.79
4/9/2019	287.31
4/10/2019	288.29
4/11/2019	288.21

Figure 2: Sample dataset of closing price of SPY stock

Comments are run through Azure Machine Learning Sentiment Analysis Program and average sentiment score for each day is calculated in a pivot table(Objective 1 achieved).

Sentiment	Score
positive	0.618285
negative	0.12281
positive	0.821467
positive	0.627812
negative	0.08364
negative	0.045017
negative	0.238759
negative	0.060741

Figure 3: Sample dataset of sentiment score output

Row Labels	Average of Score
Jan	0.471925778
1-Jan	0.390255926
2-Jan	0.483915369
3-Jan	0.47761601
4-Jan	0.482967048
5-Jan	0.444207909
6-Jan	0.40609475
7-Jan	0.506098987
8-Jan	0.509816355

Figure 4: Sample dataset of average daily sentiment score

Percentage change in closing price of SPY stock is calculated

date	percentage change
1/3/2019	-5.97
1/4/2019	8.18
1/7/2019	1.99
1/8/2019	2.39
1/9/2019	1.2
1/10/2019	0.91
1/11/2019	0.1
1/14/2019	-1.58
1/15/2019	2.95

Figure 5:Sample dataset of percentage change in SPY closing price

4.2 Measure of fit

Regression model of percentage change in SPY stock/% against the sentiment score is plotted(Research Question 2 answered). However, due to a weak R^2 value of 0.037 of the regression

model, we extended the number of days between the comments and the percentage change to 2 and 3 days. The R^2 values were also quite poor at 0.069 and 0.029 respectively.

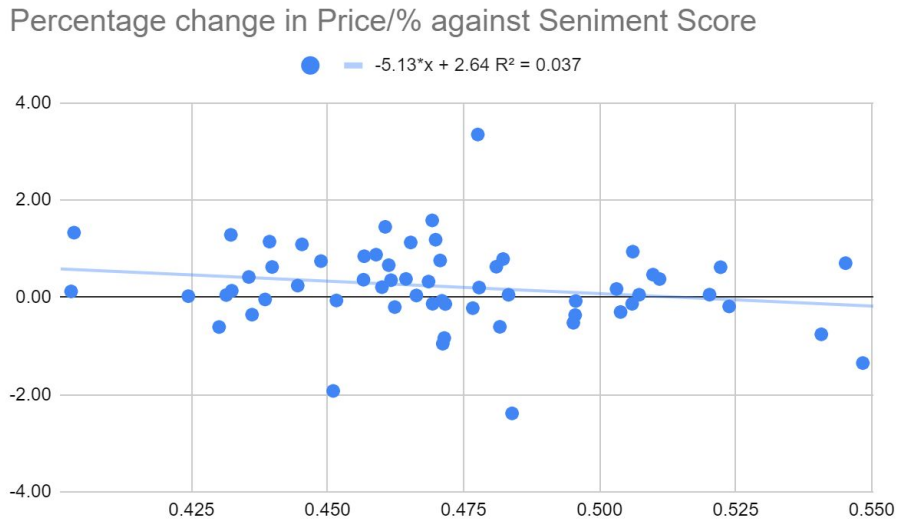
Since $R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$

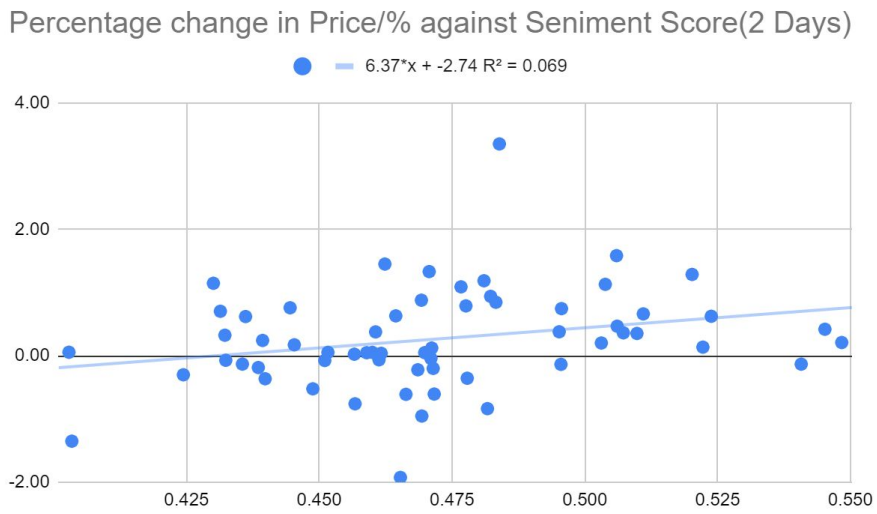
where TSS=Total Sum of Squares

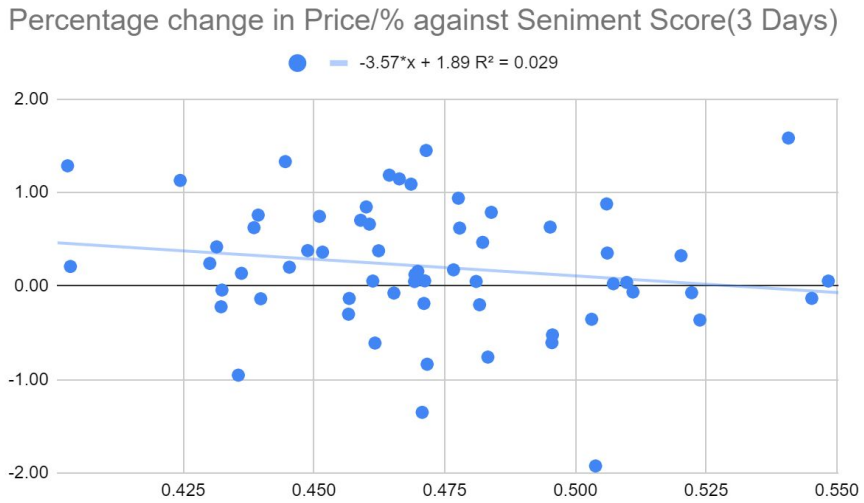
RSS=Residual Sum of Squares

TSS≈RSS, the prediction model is not any better than if the average value was taken as the output regardless of the x-value.

The low R^2 value shows that there is no relationship between sentiment scores and the percentage change in price of SPY stock. After varying the days, we also concluded that there is no specific time difference between the day the sentiments are made and the day that the sentiments are reflected in the stock market, reinforcing that there is no relationship between sentiments and change in price of SPY(Research Question 1 answered).







The t-statistic that we calculated was equal to 0.0467 while the critical value for a 20% significance level was 1.318. Since the t-statistic<critical value, null hypothesis is confirmed.

The p-value was equal to 0.998, much higher than 0.05 so there is a high probability that the t-statistic distribution was achieved by chance and there is no significant correlation between sentiment scores and percentage change. The null hypothesis is confirmed(Objective 2 achieved).

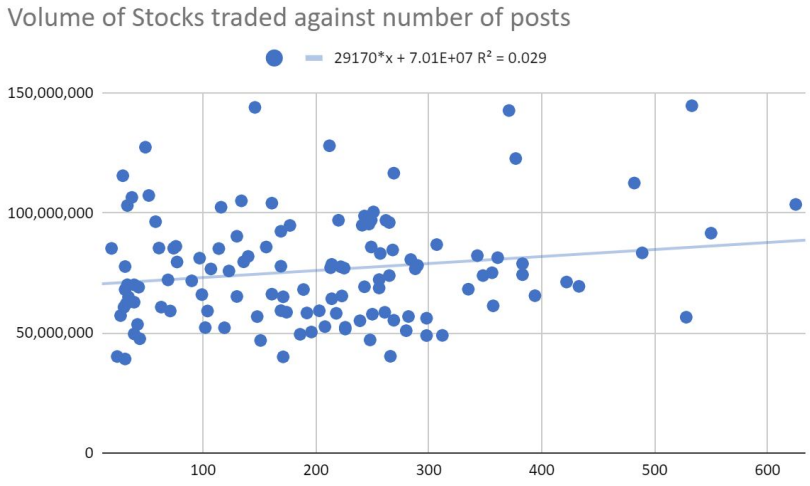
The Mean-Squared Error was equal to 0.579, indicating that on average each prediction is ±0.761 from its actual value. Since the mean percentage change of the stock market prices is 0.21, there is a 360% error of predictions. Therefore, our model cannot predict stock market movements accurately (Research Question 3 answered/Objective 3 achieved).

4.3 Extension

We tried creating a sentiment analysis program using Python which targeted keywords relating to the direction of change of stock price(e.g. increase/decrease).However, Azure Machine Learning was still used for the final results as it was able to allocate more appropriate sentiment scores to the comments with its larger library of words.

Furthermore, the relationship between the number of comments and the trading volume of the next day was tested as it was hypothesised that an increased amount of discussion would indicate an increased number of shares traded.

The number of comments represented the x-value while the trading volume represented the y-value and a best fit line was obtained.



Due to the weak R^2 value of 0.029, the conclusion was that there is no correlation between number of posts and volume of stocks traded on the market.

5 Conclusion

5.1 Analysis

The model is not viable in modelling the relationship between sentiment scores of comments and the percentage change in stock price as the null hypothesis was not rejected, there was a poor R^2 value and high Mean Squared Error.

5.2 Error discussion

A few reasons could have contributed to the inaccuracy of the model:

1. External factors

Changes in stock prices may be related to trader's sentiments but there could be other factors affecting it too such as major crises(Covid-19), government stimulus and the pre-existing trend and they may not be reflected in forums, causing a deviation of prices independent of traders' sentiments.

2. Differences in demographic

A research by the Pew Research Centre from 2011-2014 found that Reddits' target audience are young to middle-aged adults(Reddit Demographics, 2014). However, the Federal Reserve System's Survey of Consumer Finances in 2017 showed that Americans 75+ years old are most likely to directly own stocks among all age groups. The difference in user age groups highlights that the views on Reddit may not be representative of the trading community.

3. Limitations of Sentiment Analysis

The sentiment analysis program is unable to capture nuances of sarcasm and decipher double negatives, an inherent flaw of sentiment analysis. The library provided by the program is also not suited for specific stock terms(“bulls”/”bears”).

4. Nature of Forum

Forums like Reddit allow open discussion, resulting in sentiments not being solely directed at perceived changes in stock prices the next day. Some comments contain statements or facts and convey feelings from past trading, instead of expressing sentiments for the future. To solve this problem, we tried to better select posts with keywords indicating time(“tomorrow”) but the relationship was as weak as the one above.

5.3 Improvements/Extension

1. Sentiments could be analysed from platforms with a wider demographic, or one that is more representative of traders of SPY stock such as Facebook or Twitter.
2. Simpler charts could be analysed instead that show market sentiments like polls or votes.
3. Machine Learning with more complex algorithms could be used to get a better fit.
4. The library of the sentiment analysis software could be updated to detect nuances and identify stock jargon.

References

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. doi: 10.1016/j.jocs.2010.12.007

Gilbert.E., & Karahalios.K. Widespread worry and the stock market. Artificial Intelligence, pages 58–65, 2010

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Introduction. *Springer Texts in Statistics An Introduction to Statistical Learning*, 1–14. Doi: 10.1007/978-1-4614-7138-7_1

Kiger, P. (2017, December 05). Adults Aged 75+ Tend to Invest More and Own More Stocks. Retrieved August 07, 2020, from https://www.aarp.org/money/investing/info-2017/stock-ownership-fd.html

Response. (2015, November 25). Reddit Demographics: Primarily Young Adult Males. Retrieved August 07, 2020, from https://response.agency/blog/2014/02/reddit-demographics-and-user-surveys/