A DEMONSTRATION ON BIG DATA USAGE AND ITS APPLICATIONS IN ALGORITHMIC TRADING AND SENTIMENT ANALYSIS: A CASE STUDY OF S&P 500 STOCKS

Prepared by Amin ILYAS, Nizar AQACHMAR, Pritam RITU RAJ, Zahi SAMAHA

INTRODUCTION – Algorithmic trading refers to the use of computer programs for entering trading orders, in which computer programs decide on almost aspect of the order, including the timing, price, and quantity of the order. The core to algorithmic trading is to produce trading strategies which are built upon technical analysis rules, statistical methods, and machine learning techniques. Fundamentals of **big data** (i.e., volume, variety, and velocity), play various important roles in algorithmic trading analysis, real time analysis and machine learning. Moreover, the big data analytics can provide real-time picture giving the potential to improve investment opportunities for individual traders as well as trading firms. At the same time, the advancement in computer natural language processing and understanding capabilities has facilitated the inclusion and deployment of additional factors in trading prediction models, such as the introduction of a news sentiment score. To leverage **sentiment analysis** in forecasting stock market movements, one can utilize the sentiment data as a directional indicator to determine the appropriate long or short positions of specific stocks in a portfolio. Several quantitative hedge funds have integrated sentiment analysis as a key component of their trading strategies. ***Keywords:*** *big data value chain, algorithmic trading, sentiment analysis*

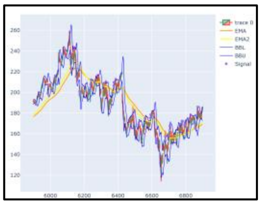
AIMS AND OBJECTIVES – This project aims at: (1) providing an example case on how big data is used in algorithmic trading and sentiment analysis; (2) demonstrating how outputs of the two techniques can be used to better understand and then forecast the markets, and (3) providing insights by capturing business opportunities made available by results of the project. Outputs of this project will be useful for both individual traders and also trading firms.

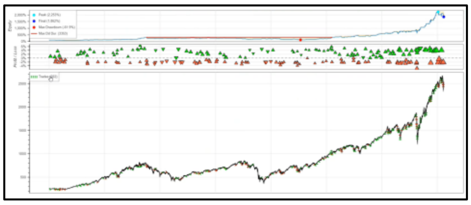
METHODOLOGY – To fulfill the objectives, we write and deploy two separate codes. The first code is developed as algorithmic trading tool to produce signals that that enable us estimate outcomes and returns, deliver accurate predictions, and back testing strategies. The second code serves as a complement strategy to the first code in which we run sentiment analysis on twitter conversations that outputs allow us to keep track of positive and negative sentiment towards a certain stock.

DATA ACQUISITION – Sources of data used in this project are summarized by in the following table. All data acquisitioned are automatized using codes we develop for in this project.

NO.	DATA	SOURCE	SIZE	USAGE	CLASSIFICATION
1.	Ticker symbols of S&P 500	Wikipedia	503 common stocks by 500 large-cap companies	Algorithmic trading	Structured data
2.	Stock data for S&P 500	Yahoo Finance	3,055,083 entries (combined for all S&P 500 stocks from 1993-01-05 to 2022-02-05	Algorithmic trading	Structured
3.	Twitter conversation	Twitter.com	More than 2,000 entries (varies with stock ticker used)	Sentiment analysis	Unstructured

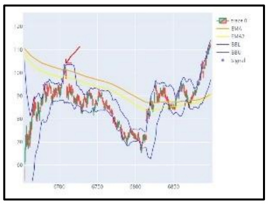
DATA TECHNICAL ANALYSIS – In this stage, we focus on adding more structure to the stock data by first selecting a ticker. Example is shown by the following diagram where a technical analysis graph is drawn for the ticker 'MMM'. From left to right:, the pictures informs is about signal (purple dots), trading strategy (200 trades as shown), and summary statistics from back testing. From this data we can see that return for the trading strategy proposed by our algorithm is about 1761 returns.





```
Start                        0.0
End                       7119.0
Duration                  7119.0
Exposure Time [%]      21.966292
Equity Final [$]    55858.968654
Equity Peak [$]     67586.549054
Return [%]           1761.965622
Buy & Hold Return [%]  887.056149
Return (Ann.) [%]            0.0
Volatility (Ann.) [%]        NaN
```

In addition, our data may also produced negative results as shown below.



```
Start                       0.0
End                      7127.0
Duration                 7127.0
Exposure Time [%]     19.304153
Equity Final [$]     527.388117
Equity Peak [$]         8977.33
Return [%]           -82.420396
Buy & Hold Return [%] 463.963704
Return (Ann.) [%]           0.0
Volatility (Ann.) [%]       NaN
```

DATA SENTIMENT ANALYSIS - After finding the desired signals we chose one dot and run our sentiment analysis around that date. The following is shown for the same stock, MMM, sentiment analysis run for the date 2020-12-03 where one of the signal produced positive returns. On the data there are more negative sentiment than the positive sentiment.

```
   ...:
   ...: neg_des = neg.describe()
   ...: neu_des = neu.describe()
   ...: pos_des = pos.describe()
   ...:
   ...: # Neutral words
   ...: x = round(neu_des.iloc[1,0]*100/2,2)
   ...:
   ...: print("Negative:", str(round(neg_des.iloc[1,0]*100+x,2))+'%')
   ...: #print("Neutral:", str(round(neu_des.iloc[1,0]*100,2))+'%')
   ...: print("Positive:", str(round(pos_des.iloc[1,0]*100+x,2))+'%')
Negative: 50.22%
Positive: 49.77%
```

DATA USAGE - For the particular signal we used in our example above, there is a conflict found between returned suggested by algorithmic trading and sentiment analysis. This can be explained by the fact that around the date conversation of the stock MMM could have not been too frequent in favor of twitter users. However, other results of us shows some consistency between the two techniques.

CONCLUSION – The algorithmic trading we develop for this project has successfully forecasted trading patterns that results in positive returns such as shown by trades tested for stock MMM used in this project. The outcome of this strategy also reveals that, given accuracy of signals produced, trading gives better returns compared to buy-and-hold strategy. On the other hand, our data sentiment analysis has also successfully processed twitter conversations on the stock and then classified the sentiment into to positive and negatives. Overall results show some consistencies between our algorithmic trading signal produced and the positive sentiment occurs around the signal. However, some inconsistencies between the two techniques are observed which then provides areas of improvement for future projects on the same or higher big data scale than one used in this project.

RECOMMENDATION – An integration of sentiment analysis to algorithmic trading will improve the quality of forecast. The algorithmic trading, we deploy in this project can be improved to accommodate sentiments by modifying the technical variables EMA1, EMA2, RSI and BB. The approach will generate another algorithmic trading based on sentimental analysis as opposed to pure technical analysis such as the algorithm used in this project.

BUSINESS INSIGHTS – We propose an establishment of a trading company that make use of big data to analyze the market and applies algorithmic trading technical analysis and sentimental analysis simultaneously to trade the market. Our project results provide a preliminary result to support the proposed business idea. The integrated approaches will provide various competitive advantages for the firm, including: (1) accurate forecasting of market trends, (2) quick and efficient trade execution which can result in increased profitability and reduced transaction costs; (3) better risk management by identification and avoidance of trades with high probability of negative outcomes. Business model we propose (i.e. with the integration of sentiment analysis to algorithmic trading) is estimated to provide an **average annualized return** between 7.5% to 9% and an **increase in trading performance** up to 15%. In order to maximize profitability of the firm, we propose that the use of integrated tool will be supported by quality of big data used, availability of alternatives trading strategies employed and favorable market conditions.

***