FINEDEC Project

Overview

Generate and test a stock trading signal using unstructured, non-numerical data. Sources may include social media, news outlets, or company filings. We hope to create signal which can boost the risk-weighted performance of a systematic trading strategy.

Natural Language Processing

Data: Fintwit

We found a dataset about top companies in the S&P 500 index from 2015-2020.
https://www.kaggle.com/omermetinn/tweets-about-the-top-companies-from-2015-to-2020

On Twitter, stocks are normally tagged using a '$' prefix (eg. $AAPL), which makes it easier for us to find out the stock being referred to in each tweet. For further filtering for relevance, we used a subset of Twitter users which are verified to be part of financial media or retweet information along those lines.

```
fintwit_users =
['CNBC','FT','BBC','WSJ','DeItaone','CNN','ReutersBiz','Quicktake','Market
Watch','TheStreet','YahooFinance','markets','FirstSquawk',]
```

Preprocessing

For each tweet, we needed to match it to a certain stock return. We are able to extract the mentioned stock(s) using the "$" tag and the tweet date from the Kaggle dataset.

The features would be the remaining body of text in the tweet excluding:

1. Stock tags ("$%")
2. Hash tags ("#%")
3. URLs ("http:%")
4. User tags ("@%")

The remaining text is processed to remove punctuation and English stop words.

Sentiment Analysis

Creating a Sentiment Model using Word2Vec

First, we must label the dataset with a dependent variable. By knowing the stock(s) mentioned and the date of each tweet, we decided to label the dataset with the residual open-to-close return of the next trading day.

$$r_{outright} = \frac{p_{close_t}}{p_{open_t}} - 1$$

Then, we must calculate each stock's beta to a benchmark index to remove the market return. As the dataset uses the top stocks mentioned in the S&P 500, we chose to use the ETF SPY as the benchmark index.

$$\beta_{stock} = \rho \frac{\sigma_{stock}}{\sigma_{index}}$$

The residual return of the stock is defined as:

$$r_{resid} = r_{stock} - (\beta_{stock} * r_{index})$$

We hypothesise that the impact of Twitter sentiment when mentioning a stock is idiosyncratic to the company, therefore, the noise of overall market movements shall be removed.

The labels for the dependent variable are defined below:
If r_resid > 0 -> label = 1
Else label = 0

This is to set up our data to be used to train using the word2vec deep learning algorithm.

	body	date	stk	ret	beta	bench_return	resid_return
0	jeff worst bezos billion year lost since	2015-01-01	AMZN	-0.012989	0.784360	-0.004603	-0.009378
1	earlier suddenly caused via month drop glitch ...	2015-01-01	AAPL	-0.018494	1.043523	-0.004603	-0.013690
2	company term tech diversity rank like see	2015-01-01	FB	-0.001654	1.114565	-0.004603	0.003476
3	company term tech diversity rank like see	2015-01-01	AAPL	-0.018494	1.043523	-0.004603	-0.013690
4	hint street pick wall top	2015-01-01	GOOGL	-0.005727	0.907207	-0.004603	-0.001551
5	hint street pick wall top	2015-01-01	AAPL	-0.018494	1.043523	-0.004603	-0.013690
6	market underperformed likely street year wall ...	2015-01-02	AAPL	-0.018838	1.043523	-0.012000	-0.006316
7	boon free investor cost holiday high delivery ...	2015-01-02	AMZN	-0.015700	0.784360	-0.012000	-0.006288
8	boon free investor cost holiday high delivery ...	2015-01-02	WMT	-0.000817	0.508070	-0.012000	0.005280
9	boon free investor cost holiday high delivery ...	2015-01-02	TGT	-0.015831	0.696950	-0.012000	-0.007467

Indexing each word to the frequency of usage in the dataset, we can turn the tweets into vectors of numerical values. Afterwards we can use the Sequential model to train the word2vec network.

Model Description:

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 40, 100)           1000000
_________________________________________________________________
flatten (Flatten)            (None, 4000)              0
_________________________________________________________________
dense (Dense)                (None, 32)                128032
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 33
=================================================================
Total params: 1,128,065
Trainable params: 128,065
Non-trainable params: 1,000,000
_________________________________________________________________
```

However, it only has a test accuracy slightly above random.

```
Test loss: 0.6924309134483337
Test accuracy: 0.5193687081336975
```

.

Findings base on our testing model

Technical analysis based on historical data to overview the stock performance and test our trading model. We doubt if any news or new regulations published will impact the stock price,

whether add in sentiment analysis can give a signal to the investors to make buy or sell decision which able to react faster than simply using technical analysis.

We use technical analysis + sentiment analysis to generate Amazon (stock: AMZN) stock price performance from 1-1-2015 to 31-12-2019.

Technical analysis base on RSI+Trading volume(MFI) over the past 60days to generate the MFI line in below table. We buy when MFI over 50 and sell once it reaches 70.

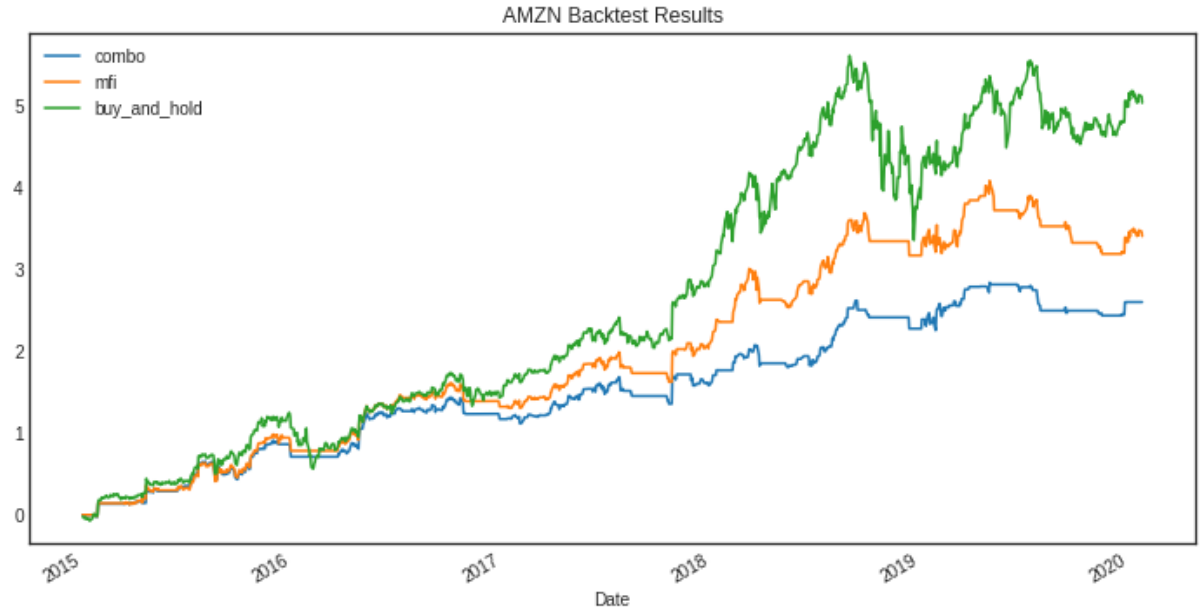
The blue line (combo) combines the technical analysis + sentiment analysis.

Green line is buy and hold strategy.

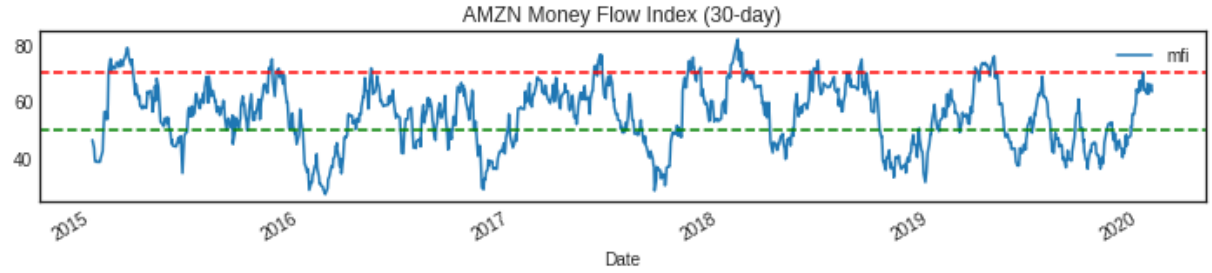
We can see the return of buy and hold strategy is the highest, an investor has around 5 times gain.

If we use MFI strategy, around 3 times gain from investment.

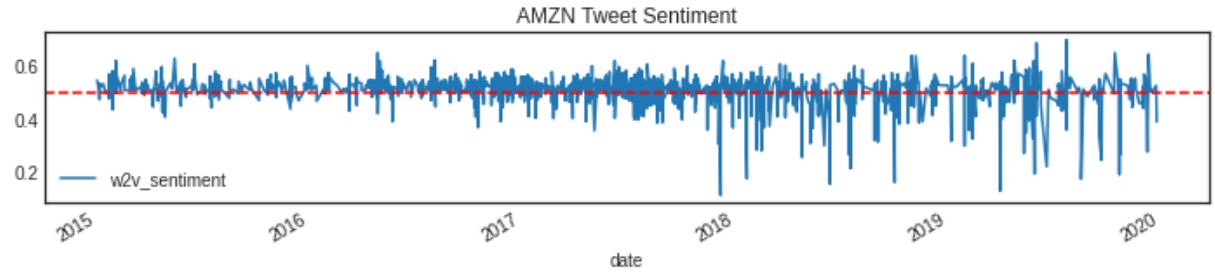
Using combine strategy, the return performance is worse than the above strategy, approximately 2.5 times gain from the initial investment.




```
<Figure size 432x288 with 0 Axes>
```




```
<Figure size 432x288 with 0 Axes>
```



If we look at the drawdown in 2019. Combine strategy reacts to sell the stock faster than the other 2 strategies.

The drawdown due to the Macro-environment factor in 2019 which harms the stock market.

Base on this finding, we expect if stock price changes due to Micro-environment factor eg. news that harms the company's reputation, combine strategy can give a signal to investors to react faster than only based on technical analysis.

Other than investment return's performance, we look at risk management.

We use Sharpe ratio [(Average return on investment – risk free)/volatility of risk] and Pain vs gain ratio (Average monthly return in positive / absolute value of Average monthly return in negative) to measure the risk against return.

Combine strategy has a steady uptrend of return performance, it considers as conservative approach in our model. It has around 40% return on average per year.

In retail investor's point of view, the return of 40% is not bad with a strategy taken into account of the risk management.

For a financial institution which helps their investors to invest. Having a steady return could benefit both-side of financial institution and their investors. For example, if an investor has a mortgage to finance their Real Estate investment, the investor uses the gain from the stock market to pay the mortgage monthly. At the same time, the investor leveraged to invest in the stock market. When stock price in the portfolio drops suddenly, the investor requires to pay Margin call. The cashflow of the investor may have problem that not able to pay both-side.

Compare to buy and hold strategy, MFI and combine strategy has some cash on hold time when hit sell signal. We consider the opportunity cost that the investors can seek for another opportunity to invest compare to buy and hold strategy. We could add in the correlation analysis to seek for the potential investment.

Our thoughts of improvement of our model are add in correlation analysis in our model. We could compare the combine strategy signal with different single stock, overview industry performance. If investors can get notice of the stocks or stock market may affect by some Macro or Micro-environment, investor can shift the investment in other market eg. Bond market. Base on the projection we do for Amazon stock, if we expect the stock price in uptrend, the investor may consider to boost their profit by using derivatives.

Limitation of our model:

Twitter is the source we use for sentiment analysis. It includes US stock only, sentiment analysis here not applied for Hong Kong stock market. Besides, not all of the stocks do have comments in twitter that are not able to do the sentiment analysis. We use Amazon to test our model due to comments volume is big enough.

Transaction cost does not include in our analysis. Additionally, we use single stock for testing. If we use other stock for testing, the result of our analysis may not be the same as Amazon analysis we did.