Natural Language Processing in Trading

Yufei Jing, Xinyi Zhu, Yuefei Chen

August 14, 2022

1 Abstract

The paper uses financial news posts in a three-year period (2017-2019) to generate a sentiment-based market-neutral trading strategy. The strategy is mediocre in terms of the 3-year performance though generates profits. Besides, the paper introduces a deep learning implementation of the sentiment-based strategy. This model is based on Long Short Term Memory Networks, which is pervasive in predicting time-series of data. Applied in this case, the improved strategy result has over 3% 1-year returns. The paper also incorporates NLP technology when doing the analysis on the fundamental side, specifically, when analyzing the similarity between companies' two consecutive financial statements from SEC EDGAR.

2 Introduction

According to the efficient market hypothesis (EMH), the stock prices fully reflect the publicly available news such as those revealed by financial news in an efficient market. Given the theory, and based on a view of technical analysis, the project assumes that the current market is largely inefficient (inefficient market hypothesis), which means that some investors can gain excess returns with their active strategies. This project largely employs the sentiment-based information in the financial news, and following the above assumption, the strategy is logically solid since we try to capture the under-reaction of the price with respect to recent financial news such that active investors could benefit from price changes in a short future (in the project, we assume a 5-trading-day investment horizon).

In this project, we have the following main contributions:

1. *Studies of the text data* - We conducted preprocessing including cleaning the text, removing stop words, stemming and lemmatization to news data and media data (the Kaggle news posts, details revealed in 3.1) and fundamental data (details revealed in 3.3). Then we conducted sentiment analysis on the news posts (for both models) and calculated the similarity indices of the tf-idf vector of two consecutive years for each stock (for the enhanced model).

2. *Sentiment-based stock trading strategy* - We implemented a sentiment-based strategy that relies solely on the sentiment analysis of the financial news posts. The probability of the sentiment strategy being accurate is over 50%, and the investor conduct the sentiment-based

trading on a 14-trading-day basis and last for 3 years, he or she will get 3% portfolio return given this stock universe.

3. *Enhanced DL-based model incorporating sentiment and fundamental data* - The paper also proposes using LSTM networks to predict stock return based on sentiment information. The model uses stock return data from 2017 to 2019 as the training set and we backtest the model using the 2019 stock return data.

4. *Validation of two models* - We conduct backtest for the two models. For the sentiment-based strategy, the model implementation is also a backtest, so we use the data from 2017 to 2019 to test the strategy. For the LSTM networks model, we test the strategy using the data in 2019.

3 Data

3.1 Text Data

Text data downloaded from Kaggle (`https://www.kaggle.com/datasets/gennadiyr/us-equities-news-data`) named Historical Financial News Archive. Data shown here represents the historical news archive covering the last 12 years(2008-2020) of US equities that still have a price greater than 10 dollars per share on NYSE/NASDAQ. In our project, we focused on a selected group of 20 companies for 2017-2019. The reason we chose this period that's consecutive and where most of the news for the 12 years has

been. We conducted a sentiment analysis of each company's news and calculated the sentiment score of each company's daily news in preparation for the following analysis.

3.2 Stock Data

Our stock data is obtained from Yahoo finance, a comprehensive database with time series on including stock open price, adjusted close price, and financial reports, that provides financial data we used to calculate stock return and portfolio profit. Here we only consider 20 stocks listed of the most recent active in the market on July 24. These include 10 stocks with the highest positive change in return and 10 stocks with the highest negative change in return.

3.3 Fundamental Data

We downloaded 2016-2019 financial statement datasets from SEC website(`https://www.sec.gov/edgar/searchedgar/companysearch.html`) for 20 selected stocks. The TF-IDF calculation is applied to the financial statements to see the Jaccard and Cosine similarity between each stock for two consecutive years. The results of TF-IDF will be one of the indicators of our second model, because we believe that the large difference in TF-IDF scores between financial statements of two consecutive years will have a great impact on the sentiment analysis of investors. This project also

collects basic pieces of information of company such as company age, company size and calculates company P/E ratio that are obtained from macrotrends website (www.macrotrends.net), which contains quarterly P/E ratio for selected firms.

4 Methodology

4.1 Algorithm of the Sentiment-based Model

4.1.1 Sentiment Analysis

For the first model, we propose that based on the sentiment analysis of the financial news over a 14-trading-day period, an active investor can possibly generate alpha. First, we analyze the sentiment for every financial news that contains the information of only one stock, then generate a trade signal by grouping the sentiment over fourteen trading days. By generating the sentiment, we use the compound score from vaderSentiment api to record an overall sentiment:

$$\text{Sentiment} = \begin{cases} 1 & \text{if coumpound score} > 0.05 \\ -1 & \text{if coumpound score} < -0.05 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

When the sentiment is 1, this is a long signal of a certain stock, and -1 is a short signal of a certain stock. If the signal is 0, we assume a neutral

position of a stock. The resulting sentiment matrix is of the following form:

$$S_{m,n} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,n} \end{pmatrix}$$

where m is the number of 14-trading-day periods within 2017-2019 (i.e. 56) and n is the number of stocks (i.e. 20 stocks).

4.1.2 Returns Matrix in 5-day

Accordingly, since we assume the under-reaction in terms of financial news of the investment population, we believe the price change with respect to news spans in a short future (i.e. 5 trading days). The resulting future 5-day return matrix from 2017/02/09 to 2019/12/12 is:

$$R_{i,j} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,j} \\ \vdots & \vdots & \ddots & \vdots \\ r_{i,1} & r_{i,2} & \cdots & r_{i,j} \end{pmatrix}$$

where i is the number of 14-trading-day news collection and processing periods within 2017-2019 (i.e. 56 key trading day) and j is the number of stocks (i.e. 20). The matrix begins at 2017/02/09 since we begin our sentiment analysis using the news from 2017/01/23 to 2017/02/09 and we assume by initiating the investment on 2017/02/09 and hold the portfolio

by 5 trading days, an investor will capture the market reaction to the last two week's news during his/her horizon.

4.1.3 Portfolio Return

Based on the sentiment signal matrix $S_{m,n}$, and the future 5-day return matrix $R_{i,j}$, we have enough information to calculate the portfolio return for each key trading day:

$$R_{p,m} = \frac{M}{n_1} * \sum_{i=1}^{n_1} r_{long,i} - \frac{M}{n_2} * \sum_{j=1}^{n_2} r_{short,j} \tag{2}$$

where M is the initial amount distributed to long and short sub-portfolios on a single trading day, n_1 is the number of stock that has buy signal (sentiment = 1 of a row vector of S), n_2 is the number of stock that has buy signal (sentiment = -1 of a row vector of S), $r_{long,i}$ is the 5-day future stock return of the i^{th} stock that has buy signal, $r_{short,j}$ is the 5-day future stock return of the j^{th} stock that has sell signal, and m denotes the operation is implemented on m^{th} trading day of the selected 56 trading days.

Notably, we assign M initial amounts to long and short sub-portfolios on each day in order to construct a market-neutral strategy that avoids possible market risks. In this project, we also assume initially the investor has a long position of the j^{th} stock that he/she is going to short based on the news sentiment signals. Finally, we obtain a column vector that contains

56 portfolio returns in 56 key trading days:

$$R_{p,m} = \begin{pmatrix} r_{p,1} & r_{p,2} & \cdots & r_{p,56} \end{pmatrix}^T$$

where m denotes the operation is implemented on m^{th} trading day of the selected 56 trading days.

4.1.4 Model Implementation and Validation

Since the model bases only on the sentiment of financial news, we combine the model implementation and model back-test. We use the media data and historical returns from 2017/01/23 to 2019/12/12. For each of 56 key trading days, we calculate a portfolio return with an investment horizon being 5 trading days based on the previous 14-(trading)-day financial news. And based on the back-test result, we can see that sentiment-based trading strategy does give insights on the portfolio construction. For detailed back-test result, please check 5.1.

4.2 Algorithm of the Enhanced DL-based Model

In the second model, we apply the preprocessing result (i.e. the sentiment matrix) from the first model and build a deep learning model to predict the 5-day return of each stock as the basis of analyzing portfolio return. We also consider fundamental information when we construct the portfolio.

4.2.1 Prediction of 5-day Return with LSTM Networks

Long short-term memory (LSTM) networks have been successfully applied in time series data [3]. LSTM is a neural network model that has feedback connections. Its connection weights and biases will change with the epoch of training. The activation patterns in the network will change in every time-step. The LSTM architecture aims to provide a short-term memory for RNN that can last thousands of timesteps [4]. In details, there are two inputs in the LSTM in contrast to RNN, h and c. In the definition, h is a hidden state and means a short memory input and it will change fast with the state going because it has a complicated function in each state related to h. The c is a cell state which can be regarded as a long term memory input because it is just the previous state value added by something. So it will not change very fast and it can store some long term information. The selection of reason is based on the fact that short terms of sentiments variation can only extract the relationship between each state or very close states. When adding a cell state in the model, features and relationships in a long period will be extracted. It is pragmatic to analyze previous weeks sentiments and returns, then predict next 5 days returns.

In this model, we utilize LTSM to generate the 5-day return prediction and then calculate the portfolio return. The set-up of the model is: we first set a look-back-period h, where h in this case is 14 trading days. This means the historical sentiments of stock i of period $S_i = (s_{i,1}, s_{i,2}, ..., s_{i,14})$ affects the final prediction of 5-day return of stock i. At each day t, we use the sentiment value $S_{i,t}$ as the input and the model is expected to output a

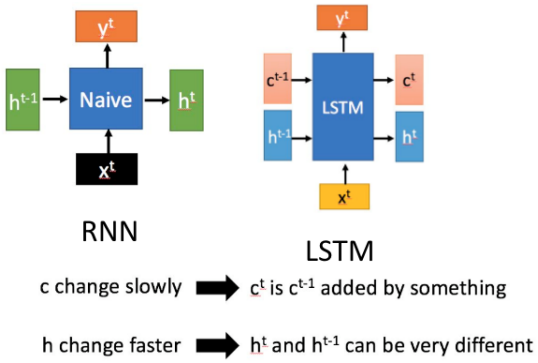

Figure 1: The Structure of Long Short Term Memory and Recurrent Neural Network

value r_{5-day}. For the model fitting, we use the data from 2017 to 2018 as the training set and data in 2019 as the test set.

4.2.2 Thinking of Portfolio Construction: Weight Matrix W and Fundamental Information

We collect information of companies' sizes, 4-year mean P/E ratio (2013-2016) and the financial statements from 2016 to 2018 of companies from SEC EDGAR. For the latter, we first preprocess the data and then turn the text data to tf-idf. Our idea is to check if the company has encountered substantial changes that is possibly affecting the stock investment decisions. We analyze the idea by calculating the Jaccard similarity of the consecutive tf-idf of the financial statements:

$$J_{f_t,f_{t+1}} = \frac{f_t \bigcap f_{t+1}}{f_t \bigcup f_{t+1}} \tag{3}$$

where f is the tf-idf vector of the financial statement and the integer $t \in [2016, 2018]$.

Based on the results of the fundamental information, we rule out the possible effects on the portfolio construction from company size and the F/S similarity (please check 5.2). The only factor we consider to construct the weight matrix W is P/E ratio. Since different sectors have different P/E ratio, it is hard to analyze the P/E ratio's effect on W in a short time; besides, there are other fundamental information could be considered, thus we decide to use the market-neutral construction methods as we do in the first model.

5 Results

5.1 Test Result for Sentiment-based Model

Based on Figure 2, in the period of Jan 2017 to Jan 2018, Our portfolio returns fluctuated over the months but up trending in overall. Between the period Jan 2018 and Jan 2019, there have been large fluctuation at the third quarter of 2018, a large rise around September of 2018. There are 29 trading periods have positive return and 27 trading periods have negative return over selected three years. The largest return is over 1210.901 dollars at November 9, 2018 and lowest return is -545.03 dollars at July 23,2018.The portfolio return over three years is 0.0294 that means the overall profit for our portfolio is approximately 3% which is mediocre but

profit. Surveying the news over the period, September is a time when stocks enter their usual turmoil and are near record highs, making this part of 2018 more likely than ever for a major event to hit the market. A series of global disputes erupted that fall, including tariffs on Chinese imports, the possibility of auto tariffs on other countries, the start of sanctions on Iranian oil, and the November election. All of this contributed to financial market volatility in 2018.

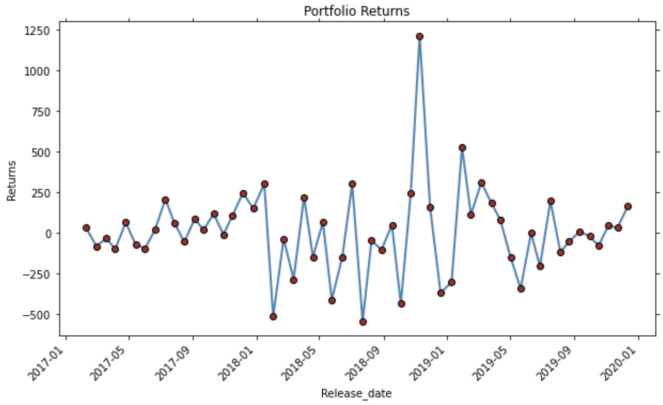

Figure 2: The Portfolio Return over three years

5.2 Test Result for Fundamental Analysis

5.2.1 PE Ratio

Figure 3 represents that the each company's PE ratio which is a pair of company's share price to the company's earnings per share. The ratio is used for valuing companies and to find out whether they are overvalued or undervalued. It is obviously identified each Google, Amazon and Meta has

the highest PE ratio over the selected stocks. Rest of stocks' PE ratio are similar. We want to do more detailed analysis based on different industry if we have more time.

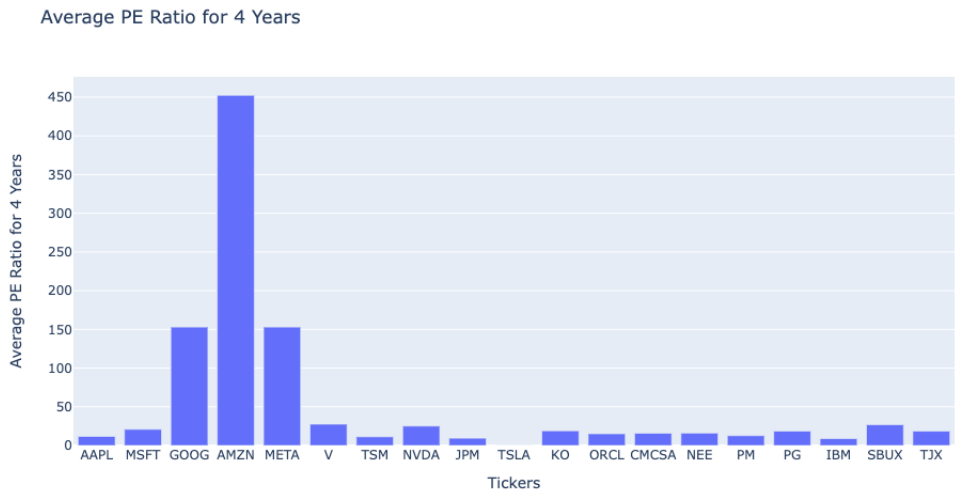

Figure 3: P/E Ratio for 20 stocks

5.2.2 Jaccard Similarity

Figure 4 shows that jaccard similarity between 2016 VS 2017, 2017 VS 2018, 2018 VS 2019 for selected 20 companies financial statements. There is a slightly difference of jaccard similarity within these stocks that means no significant decision been made within three years. Since we assume that a major change or decision of a company will affect investor sentiment and their subsequent investments.

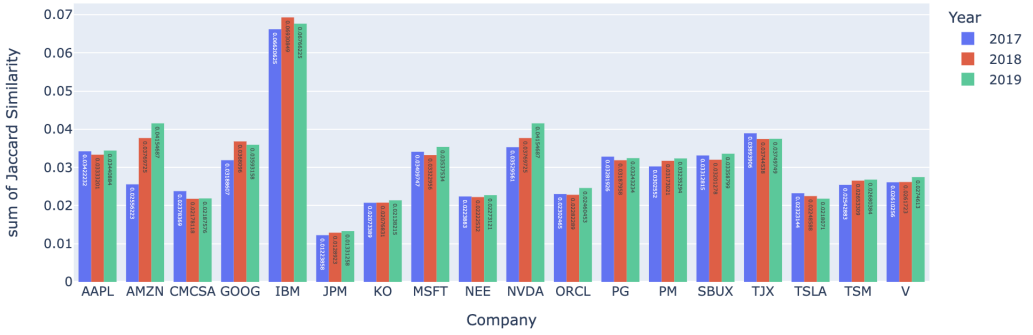

Figure 4: Jaccard Similarity for 20 Stocks over Three Years

5.3 Test Result for DL-based Model

In order to compare with the first strategy, the dataset used is identical to the previous simulation, Applied into the dataset, the LSTM will train each stock's sentiments and returns, and extract the features from each of them. In practice, each stock will train one model. These 19 models trained by 2017-2018 data will be combined and predict stock return in 2019. In order to make a decision, we add an classifier to divide Long or Short operations based on predicted return. If the result is positive, which the return is positive, we can long them. And if the return is negative, we need to short them. Additionally, The portfolio management strategy is the same as the sentiment based model–market neutral portfolio. The model is simulated using 2019's data and the result is shown in Figure 5. This chart is the return when we fund $10000 in a trade day. The X-axis is trading period and Y-axis is 5-days' return. It is obvious that most of them are profitable. And the 1-year return is over than 3% in 2019, which is better than the baseline model (model 1) in that year.

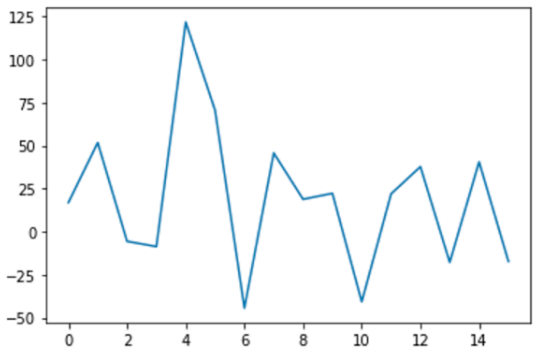

Figure 5: LSTM model Returns in Every Trading Periods

6 Conclusion

6.1 Project Conclusion

In conclusion, the paper, at first, applied financial news releases over a three-year period (2017-2019) to generate sentiment based market-neutral trading strategies. The strategy produced a mediocre profits. In addition, this paper also introduces the deep learning implementation of the emotion-based strategy. The improved strategy result has over 3% 1-year returns, which is a better compared with the first method. In conducting fundamental analysis, this paper also introduces natural language processing techniques, specifically, to analyze the similarity of two consecutive financial statements of a company from SEC EDGAR. Moreover, the deep learning model preliminary shows that LSTM Networks can be used in the stock market. Using the LSTM model can significantly improve the retracting of Stockranker stock selection model in the backtest stage.

6.2 Discussion and Future Work

There are three points in this project that we can improve in future. First, we can involve more methods in the stock picking process. In this project, we choose the most active stocks in July 2022, and this may arise the problem of survivorship bias. We can incorporate fundamental and technical analysis during the stock picking. Second, we collect the fundamental information from the companies: P/E ratio, similarity of two consecutive 10-K statements etc. We can incorporate the fundamental information in the model to decide the final stock holdings. Third, the data we collected is outdated and prepandemic, plus we do not have the detailed transaction data from each investor, so the assumption that we use the adjusted closed price of each stock to calculate 5-day return is questionable in some sense. The holding period, 5 days, is also a hyper-parameter. It is expected to collect more information to decide the relationship between sentiment generation from news and the portfolio holding period.

References

[1] Zhang, W., Skiena, S. (2010). Trading Strategies to Exploit Blog and News Sentiment. Proceedings of the International AAAI Conference on Web and Social Media, 4(1), 375-378. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14075

[2] Aboody, D., Even-Tov, O., Lehavy, R., Trueman, B. (2015). Overnight Returns and Firm-Specific Investor Sentiment. SSRN Electronic Journal. Retrieved from https://doi.org/10.2139/ssrn.2554010

[3] Wang, J., Zhang, Y., Tang, K., Wu, J., amp; Xiong, Z. (2019). Alphastock. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining. Retrieved from https://doi.org/10.1145/3292500.3330647

[4] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. Retrieved from https://ieeexplore.ieee.org/abstract/document/6795963/