News-based Machine Learning and Deep Learning Methods for Stock Prediction

Junjie Guo¹, * and Bradford Tuckfield²

¹International education college, Henan University, Kaifeng, People's Republic of China
²The Wharton School, University of Pennsylvania, Philadelphia, United States of America

*Email: swlee143@gmail.com

Abstract. Stocks occupy a vital position in the financial market. Over the years, scholars have made unremitting efforts in forecasting the stock market. Because the more accurate the prediction, the more people will profit from the stock market. Machine learning has achieved excellent results in stock prediction. Nowadays, with the rise of deep learning, the stock prediction methods used by people are beginning to lean towards deep learning, and many results have been achieved. This paper will use news rather than traditional stock structured data for stock prediction, and we will use machine learning and deep learning methods in contrast. Moreover, we use natural language processing to process the news. The objects of prediction are stock indexes (DJIA, S & P500) and individual stocks (IBM, JPM). We find that deep learning performs at least 4.5% better than machine learning on prediction tasks related to stock indexes, and at least 3% better in the prediction of individual stocks. We discuss the implications of this result.

1. Introduction

Stock prediction has been considered an ideal way to obtain more stable and generous returns in the stock market. Stock prediction has important significance and unique appeal both in theoretical research and in actual financial markets. The earliest studies on stock predictions date back about 50 years ago. However, today, stock prediction is still an important subject in financial research.

From previous research, the methods of stock prediction are roughly divided into two categories--machine learning and deep learning. Machine learning methods are relatively mature, and previous studies have proposed a large number of different machine methods, and the accuracy of predictions is continuously improving. Deep learning has risen in recent years, and its rapid development is closely related to the development of artificial intelligence. The models involved in deep learning are widely used in speech recognition, computer vision, and other fields, and have achieved remarkable achievements in these fields[1]. Applying deep learning to stock prediction is a novel method, but how to apply deep learning models to stocks was difficult at first, gradually with the continuous research in recent years, people have gradually explored the use of various deep learning models or deep neural networks, such as LSTM, CNN[2][3]. Although the prediction accuracy of different models is different, the overall improvement of machine learning methods is significant. The trend of stock indexes and individual stocks is the main target of the forecast. From a research perspective, there is not much difference between the prediction of stock indexes and individual stocks.
At present, most research uses structured data, such as opening and closing prices, volume. The difference in prediction accuracy is mainly reflected in the model. However, the direct use of structured data also has disadvantages[4]. Structured data often ignores other relevant factors that can affect the stock market, such as national economic trends, foreign exchange, international politics, and ideally, if people can take these factors into consideration to the greatest extent. The accuracy of the prediction will be very high. However, it is impossible for people to take all factors into consideration because the methods often limit people, and some factors are difficult to quantify, such as political factors. News has a significant influence on the trend of stocks, which has been proven in previous studies[5]. The use of news to predict stocks is not widely used in the era of popular machine learning, because how to use news as input is a problem, which hinders the use of news makes predictions this way. With the development of technologies such as natural language processing (NLP), this problem will be improved[6]. Word embedding is a new method for processing news text, and its application will make news prediction stocks more accurate and reasonable[7].

The research in this paper is based on news. In the prediction method, this paper uses machine learning (Naive Bayes, Random Forest, Gradient Boosting Machine) and deep learning methods (LSTM, CNN). This paper mainly explores the performance and differences of machine learning and deep learning in stock index and individual stock prediction under the premise of news-based and makes a rational analysis of the differences.

2. Literature review
The stock prediction has always been considered complicated. However, people have always desired to predict the direction of stocks with higher accuracy. The more accurate the prediction, the more people will profit from the stock market[8,9]. EMH (Efficient-market hypothesis) is an important theory about the stock market. It holds that in a stock market with sound laws, proper functions, high transparency, and sufficient competition, all valuable information has been timely, accurately, and fully reflected in the stock price. Among the trends, including the current and future value of the enterprise, unless there is market manipulation, investors cannot obtain excess profits higher than the market average by analyzing past prices[10]. Although this theory is an important theory of the securities market, there is considerable controversy, and both supporters and opponents have provided a lot of evidence[11].

In the early years, people were keen to use traditional machine learning methods to predict stock movements [12][13][14]. These studies have tried different models. Although they have made good results, the gap between the various models is not significant. Some studies use the news to predict stocks. Because related researches show that news has a high impact on the stock market, it is reasonable to use the news to predict stocks[15][16]. The use of news has improved prediction accuracy compared to previous methods[17][18][19]. Later, with the popularity of neural networks, people began to use various neural network models for stock prediction[20][21]. Prediction accuracy is also much improved compared to previous machine learning methods.

With the development of deep learning in recent years, deep learning has provided more prediction methods for people. CNN is one of the models for deep learning. The strong performance of CNN is in its image processing method, which has been used in various areas. Studies using CNN models in stock prediction are rare. Existing researches show that CNN also shows excellent performance in stock prediction[22][23]. However, CNN is often applied directly to prediction, and few studies have improved the model to make it more suitable for prediction. Based on previous studies, we found that whether it is a widely used machine learning algorithm or the latest deep learning model, these studies mainly use structured stock data to make predictions, such as opening price, closing price, and trading volume. However, they rarely consider other factors that can affect stocks and how they affect stocks, such as news[24]. Hence, there are still many things that can be improved in stock prediction.
3. Method

3.1. Data description and preprocessing

Our stock data and news text are from Kaggle. We have selected stock indexes (DJIA, S & P500) and individual stocks data (IBM, JPM). The time period of these four sets of data is selected from 2008-8-8 to 2015-7-22, for a total of 1750 trading days. The data includes the opening price, closing price, high price, low price, and volume. News data is from 2008-08-08 to 2016-07-01. The data includes the 25 headlines of the day and the rise and fall of the Dow Jones Index that day. The news comes from major media. News data includes news for 1989 trading days. To ensure the consistency of the data period, the period of stock data and news data is from 2008-08-08 to 2015-07-22, for a total of 1750 trading days. A summary of these four sets of data is shown in Table 1.

Table 1. A summary of data

| Stock&Index | Mean price(Index) | Mean change | Days of price(Index) increase | Days of price(Index) drop |
|-------------|------------------|-------------|-------------------------------|--------------------------|
| DJIA        | 12947.63529      | 0.036%      | 943                           | 807                      |
| S&P500      | 1420.34278       | 0.038%      | 961                           | 789                      |
| IBM         | 160.30966        | 0.023%      | 877                           | 873                      |
| JPM         | 45.15599         | 0.076%      | 888                           | 862                      |

Before the experiment, we performed data preprocessing. The news text was preprocessed at the time of retrieval, so we did not repeat the processing. We created labels for the closing prices of two sets of stock index data and two sets of individual stock data -- 0 or 1. 0 means that the index or price on the D + 1 trading day is lower than the D trading day, and 1 means that the index or price on the D + 1 trading day is higher than the D trading day or equal to D trading day. The formula is shown in (1).

\[
Label = \begin{cases} 
0, & D > D + 1 \\
1, & D \leq D + 1 
\end{cases}
\] (1)

3.2. Naive Bayes, Random Forest, Gradient Boosting Machine

According to relevant researches experience, for these three machine learning methods, we classify 80% of the data as the training set and 20% of the data as the test set. That is, the data of 1400 trading days is used as the training set, and the data of 350 trading days is used as the test set.

We use the N-Gram model[25] and TF-IDF (term frequency-inverse document frequency) to process news text. First, we use TF-IDF to convert the text into a vector that the model can process. In TF-IDF, the importance of a word is proportional to its frequency in the text (TF) and inversely proportional to its frequency in the corpus (IDF )[26]. The formula is shown in (2)(3)(4).

\[
tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} 
\] (2)

\[
idf = \log \frac{|d|}{|\{j : t \in d_j\}|} 
\] (3)

\[
tfidf_{i,j} = tf_{i,j} \times idf_i 
\] (4)

At the same time, we should strictly ignore entries with document frequencies below or above the given threshold, and stop words specified by the corpus. Because if too many features are retained, too much noise will be included during training. Furthermore it will reduce the generalization ability of the model. On the contrary, too few features will cause insufficient training and weaken the generalization ability of the model. In the Naive Bayesian model, we set the highest frequency of words to 0.7 and the minimum value to 0.1. The highest frequency of words to appear in the random
The forest model was set to 0.99, and the minimum value was 0.01. In the gradient lifter model, the maximum frequency is set to 0.9, and the minimum value is 0.1. The maximum capacity of the vocabulary in the three models is 200,000, and \( n = 1 \) in N-Gram, that is, the number of occurrences of a single word is counted. The unary model formula of N-Gram is shown in (5). For the significance level in Naive Bayes, we set it to 0.01 i.e., \( \alpha = 0.01 \).

\[
P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} p(w_i) \tag{5}
\]

After setting the parameters, the process is divided into three parts. First, training the vector of news text conversion, testing with a test set, and finally calculating the accuracy. We predict the rise and fall of stock indexes and individual stocks, so the accuracy is calculated as the number of days that the correct rise and fall are predicted divided by the total number of days in the test set. The formula is shown in (6).

\[
Accuracy = \frac{RightDay}{TestSetDay} \quad (TestSetDay = 350)
\tag{6}
\]

3.3. LSTM, CNN

Long Short-Term Memory neural network (LSTM) is a kind of time loop network. LSTM is suitable for processing and predicting important events with very long intervals and delays in time series. The LSTM contains three important gates - Forget gate, Input gate, Output gate. Forget gate[27]. The LSTM model is shown in Figure 1. First, we preset the parameters. The number of words in the dictionary is 10,000, and each sentence The maximum length is 200, and the embedded dimension is 100. In order to optimize the processing speed, we set the batch size to 32. Next, the text samples are vectorized into 2D integer tensors so that they can be input into the LSTM. The output dimension of the embedding layer is 128-dimensional, dropout is 0.2. The output dimension of LSTM is 128-dimensional, and dropout is 0.2. The fully connected layer uses the softmax function as the activation. The final step is the output accuracy.

**Figure 1. LSTM model**

Convolutional Neural Network (CNN) has achieved outstanding achievements in image processing and language recognition. CNN includes input layer, convolutional layer, pooling layer, fully connected layer, and output layer[27].CNN model is shown in Figure 2 Applying CNN to stock prediction is in nearly several years. It was proposed that the main difficulty of this method is input. News text cannot be directly used as input. Like LSTM, we use embedding method to solve this problem. In this paper, we use Convolution1D to make predictions. The number of filters is 120, and the length is 2. The activation function is ReLU, and the formula is shown in (7).
Figure 2. CNN model

\[ f(x) = \max(0, x) \] (7)

4. Result

The accuracy of the prediction is shown in table 2. We will analyze the final result from three aspects, the accuracy of machine learning, the accuracy of deep learning, and the comparison of the two. A reasonable explanation will be given based on the results. In the table, boldface is the highest accuracy of machine learning, and bold italics is the highest accuracy of all methods.

| Method                | DJIA   | S&P     | IBM     | JPM     |
|-----------------------|--------|---------|---------|---------|
| Random Forest         | 0.508951 | 0.485933 | **0.496163** | **0.514066** |
| Naive Bayes           | **0.514066** | **0.529411** | 0.457812 | 0.473145 |
| Gradient Boosting Machine | 0.501278 | 0.498721 | 0.445012 | 0.503836 |
| LSTM                  | **0.537084** | 0.506393 | **0.514066** | **0.547314** |
| CNN                   | 0.534526 | **0.539641** | 0.468032 | 0.460358 |

Among the three machine learning methods, it can be seen from the table that Naive Bayes has an advantage in the prediction accuracy of stock indexes, and Random Forest has an advantage in the prediction accuracy of individual stocks. The prediction accuracy of the three machine learning methods is generally high when it comes to prediction accuracy for individual stocks.

For the two deep learning methods, LSTM is higher than CNN in the prediction accuracy of DJIA, IBM, JPM, and CNN is higher than LSTM in S & P prediction accuracy. The accuracy of CNN in individual stock prediction is significantly lower than its prediction Performance in stock index. Although Naive Bayes, Random Forest, Gradient Boosting Machine, and CNN performed poorly in individual stocks, LSTM still performed well.

From the overall results, both in terms of stock index prediction and individual stock prediction, the accuracy of deep learning is higher than machine learning. However, in terms of predicting individual stocks, both machine learning and deep learning perform worse than predicting stock indexes. It is reasonable to predict better because deep learning models have better generalization ability when dealing with complex problems. However, this does not mean that people will use deep learning methods when predicting stocks. Some machine learning methods have formed stable structures and methods in a long-term exploration, which may be more stable in predicting stocks. Deep learning models are more complicated and changeable. Different methods may perform considerable differences in results by changing the structure of the model and various parameters.

Regarding why these methods generally perform poorly in predicting individual stocks. We infer three reasons from the results. The first reason could be that because the stock index is more sensitive
to the news, or the correlation is higher than individual stocks. Secondly, the two stocks we selected (IBM, JPM) have very long listing time, and it is no longer a highly active company in the market. Maybe the news has a limited impact on these two stocks. Finally, in order to make the comparison more reasonable, we have performed similar treatments on various methods. We have not optimized too much for one method. Because this may cause the viability of the result to decrease. So we can see that some methods perform well in predicting stock indexes, but they do not perform well in predicting individual stocks.

5. Conclusion
In this paper, we use the news to predict the stock market. In terms of methods, we use machine learning and deep learning methods and compare the differences between machine learning and deep learning. We use stock indexes (DJIA, S & P500) and individual stocks as prediction targets (IBM, JPM). In terms of news processing, we used the TF-IDF, N-Gram methods in Natural Language Processing, and the two neural networks LSTM and CNN in deep learning. The final results show that although the overall method of deep learning is accurate, the degree is slightly higher than the machine learning method, but the deep learning method has a significant difference in predicting stock indexes and individual stocks. We give our analysis in the results section.

Future Study
(1). In addition to using structured data and news, we may use other data to make stock predictions, such as user comments in financial news. However, this data is difficult to obtain and requires complex data cleaning.
(2). People can not only predict the ups and downs or trends of stocks but also can consider from a practical point of view, for example, assuming that a certain amount of capital is invested, how much people can profit or lose under the predicted results.
(3). Pay more attention to the optimization of the model. By improving the results and parameters of the model, to make the generalization ability of the model for stock prediction stronger.

References
[1] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2012, March). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In 2012 IEEE international conference on Acoustics, speech, and signal processing (ICASSP) (pp. 4277-4280). IEEE.
[2] Siripurapu, A. (2014). Convolutional networks for stock trading.
[3] Wang, X., Liu, Y., Sun, C. J., Wang, B., & Wang, X. (2015, July). Predicting polarities of tweets by composing word embeddings with long short-term memory. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1343-1353).
[4] Mizuno, H., Kosaka, M., Yajima, H., & Komoda, N. (1998). Application of neural network to technical analysis of stock market prediction. Studies in Informatic and control, 7(3), 111-120.
[5] Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics, 70(2), 223-260.
[6] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).
[7] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2016, December). Knowledge-driven event embedding for stock prediction. In Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers (pp. 2133-2142).
[8] Fung, G. P. C., Yu, J. X., & Lam, W. (2002, May). News sensitive stock trend prediction. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 481-493). Springer, Berlin, Heidelberg.
[9] De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? The Journal of finance, 40(3), 793-805.
[10] Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.
[11] Malkiel, B. G. (2003). The efficient market hypothesis and its critics. Journal of economic perspectives, 17(1), 59-82.
[12] Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications, 40(14), 5501-5506.
[13] Li, F., & Liu, C. (2009, August). Application study of BP neural network on stock market prediction. In 2009 Ninth International Conference on Hybrid Intelligent Systems (Vol. 3, pp. 174-178). IEEE.
[14] Shah, V. H. (2007). Machine learning techniques for stock prediction. Foundations of Machine Learning| Spring, 1(1), 6-12.
[15] Howe, J. S. (1986). Evidence on stock market overreaction. Financial Analysts Journal, 42(4), 74-77.
[16] Pearce, D. K., & Roley, V. V. (1984). Stock prices and economic news.
[17] Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics, 70(2), 223-260.
[18] Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. Information Processing & Management, 45(5), 571-583.
[19] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In Twenty-fourth international joint conference on artificial intelligence.
[20] Schöneburg, E. (1990). Stock price prediction using neural networks: A project report. Neurocomputing, 2(1), 17-27.
[21] Refenes, A. P., Burgess, A. N., & Bentz, Y. (1997). Neural networks in financial engineering: A study in methodology. IEEE transactions on Neural networks, 8(6), 1222-1267.
[22] Tsantekidis, A., Passalis, N., Tefas, A., Kannaiainen, J., Gabbouj, M., & Iosifidis, A. (2017, July). Forecasting stock prices from the limit order book using convolutional neural networks. In 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 1, pp. 7-12). IEEE.
[23] Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187-205.
[24] Fung, G. P. C., Yu, J. X., & Lam, W. (2002, May). News sensitive stock trend prediction. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 481-493). Springer, Berlin, Heidelberg.
[25] Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. Computational linguistics, 18(4), 467-479.
[26] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).
[27] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.