Comparison of Different Machine Learning Scenarios for Stock Price Prediction of Netflix

Haoxin Chen[1, *]

[1]Department of Jinan University, University of Birmingham Joint Institute, Jinan University, Guangzhou, China.

*Corresponding author: conniechen@stu2019.jnu.edu.cn

Abstract. Buying stocks based on relatively accurate predictions to make a profit is what investors have been yearning for. However, due to the volatility and stochastic intrinsic of the stock, the price is also full of uncertainty, which is difficult to predict. With the improvement of computer performance and the popularization of machine learning methods nowadays, effective stock prediction methods emerge one after another. In this paper, random forest, XGBoost, and LSTM techniques are utilized for predicting the closing price of Netflix, which is one of the major stocks in Nasdaq and has been fluctuating this year. The closing price of the previous five days and the last day's volume are used as inputs. The models are evaluated by MSE and R-square. According to the analysis, similar model evaluation results in random forest and LSTM indicate that the models are efficient. Furthermore, the prediction could be improved by considering the fluctuations of stocks simultaneously and introducing more variables in various dimensions. These results shed light on guiding further exploration of predicting stock price movements based on advanced machine learning methods.

Keywords: random forest; XGBoost; LSTM; Stock price prediction.

1. Introduction

The information era witnessed great changes in society. Data permeates almost every industry. Traditional economic and financial theories believe that stock prices are mainly affected by market and company characteristics. Under the same regulatory, policy environment, and similar macroeconomic conditions, companies in the same industry have highly similar operating conditions when facing changes in the economic environment. The impact of COVID-19 on traditional industries, (e.g., transportation, mining, and electric power) is quite serious, but it also creates opportunities for the development of high-tech industries. The manufacturing and health industries responded strongly to the epidemic in a positive manner, boosting the confidence of the stock market [1].

In economics and finance, from both short-term and long-term perspectives, there is a positive correlation between the effective stock market and economic growth, and an indirect transmission mechanism for the impact of stock market development on investment [2]. Therefore, whether for the whole economic environment or ordinary investors, the direction of the stock market is extremely critical. Contemporarily, most investors make decisions based more on data and analysis than experience and intuition now. In the securities market, ordinary investors account for the majority of all investors, and how to choose stocks and profit becomes the focus of their attention. Thus, a reasonable stock prediction method is conducive to individual investors to discover risks and opportunities in time and optimize investment decisions.

The subject of this paper is Netflix, one of the five major stocks in the Nasdaq index (FAANG). With large scale, good liquidity, and strong industry representation, the constituent stocks of NASDAQ are more likely to attract the attention of ordinary investors. The NASDAQ index has maintained continuous growth in the past ten years, but it has declined in the past six months due to the decreasing liquidity in the market and other reasons and has recently shown an upward trend. As a members-only Internet streaming platform, Netflix provides users with rich and high-quality movies and TV series, with 222 million registered users worldwide, accounting for 80% of the market. Thanks to its exclusive membership and excellent resources, its stock price rose from $15 per share when it went public to nearly $700 per share by the end of 2021. However, Netflix's results for the first two quarters of 2022 show that the company's business model is being tested by massive

subscriber losses and intensifying competition. In the future, the development of Netflix measured by its stock prospects also become a point of concern.

Machine learning and deep learning algorithms can process a large amount of data and quickly analyze rules from it. Hence, they are favored by many researchers and widely used in the financial field. Stock data has the characteristics of nonlinear, high noise, and non-stationary, so it is of practical significance to establish a machine learning model to explore the rules of stock rise and fall. At present, the mainstream research models include regression, neural networks, SVM, LSTM, random forest, etc. Sharma et al. surveyed several efficient regression approaches, e.g., linear, polynomial, RBF, and sigmoid regression to predict the stock market [3]. Mobin Akhtar et al. mainly used a Support Vector Machine to improve the overall accuracy of stock price prediction [4]. Jayanth Balaji et al. mainly focused on deep learning methods and found that LSTM, GRU, CNN, and ELM performed well in terms of RMSE, DA, and MdAPE [5]. Chhajer et al. additionally compared the performance of ANN and BP applied to different situations [6]. Vijh et al. created new variables such as stock price's seven days moving average and its standard deviation from the original data to improve the model [7]. Based on these models, further optimization has been made. Huang et al. defined a data set unit by week and estimated the unit number dynamically in the different economic cycles through Bayesian model to improve the effectiveness of traditional LSTM. The new B-LSTM increased over 25% prediction [8]. Kim et al. analyzed a sentimental dictionary with news articles and obtain the positive index for each date, which is correlated with the stock return value [9]. Hogenboom et al. proposed an advanced natural language processing pipeline containing word sense disambiguation for event-based stock price prediction [10].

Based on the above research results, this paper intends to focus on the performance of tree algorithms and neural network models in predicting prices measured by day. The rest part of the paper is organized as follows. Sec. 2 will describe the data source in detail and explain the principles of the models and metrics. Netflix is selected as the representative stock, its closing price of the current day as the target, the closing price of previous days and the volume as variables, which are slightly transformed from the original data. Three methods, random forest, XGBoost, and LSTM, are utilized. Sec. 3 will present the predicted results, including the time series plot of predicted prices, and make a comparative analysis of the performance of the three models. The effectiveness of the models is tested and compared by two metrics, i.e., MSE and R-square.

2. Data & Method

2.1 Data

The daily data of Netflix stock in this article is obtained from Yahoo Finance. The data starts on 2017/01/03, the first trading day of 2017, and ends on 2022/04/22. The variables of the original data include open price, close price, high price, low price, and volume. In this article, the stock price of the current day is considered as the dependent variable, that is, the target of prediction. The independent variables are the stock price of the past five days and the natural logarithm of the volume of the previous day as given in Table 1.

Table 1. Variables

	Notations	Definition
Dependent variable (y)	$Price_t$	Stock price today
Independent variables (X)	$Price_{t-1}$	Stock price 1 day ago
	$Price_{t-2}$	Stock price 2 days ago
	$Price_{t-3}$	Stock price 3 days ago
	$Price_{t-4}$	Stock price 4 days ago
	$Price_{t-5}$	Stock price 5 days ago
	$ln(Volume_{t-1})$	Natural logarithm of Volume 1 day ago

2.2 Models and metrics

This paper uses OLS regression, random forest, XGBoost, and LSTM four models to analyze the stock price. Among them, the OLS model is an important benchmark to compare the predictive ability of machine learning methods. However, for multivariable models, this method may fail to identify valid variables or cannot be accurately estimated due to multicollinearity or too high dimensionality. A tree algorithm is a simple nonlinear model and an ideal high-dimensional prediction model. Random forest is a traditional tree algorithm, but it has the problem of insufficient fitting ability. Ensemble learning tree algorithms, such as XGBoost, take regularization into account in the idea of gradient increase, reduce the complexity of the model by ensuring small deviation, preventing over-fitting, and thus improve the out-of-sample prediction results. The advantages of the neural network model are the nonlinear mapping ability, adaptive ability, and generalization ability, while the long short-term memory model further considers the time series relationship of predictor variables and can transmit the prediction information of lagged data, which is very important for the stock market.

The simple linear predictive regression model is implemented based on the least square method,

$$f(X) = y = \beta_0 + \sum_{k=1}^{p} \beta_k X_k \tag{1}$$

where β is the coefficient.

$$\hat{y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} \tag{2}$$

$$e_i = y_i - \hat{y}_i \tag{3}$$

$$Q = \sum {e_i}^2 = \sum (y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \cdots - \beta_k X_{ik})^2 \tag{4}$$

The residual is minimized by taking partial derivatives, then the coefficient is

$$\beta = (X'X)^{-1} X'y \tag{5}$$

Unlike regression, tree algorithms are completely nonparametric, divided according to groups of observations with similar characteristics, and make predictions based on the average of the results in each region. For a regression tree, the model is in the form of

$$f(X) = \sum_{m=1}^{M} c_m * 1_{X \in R_m} \tag{6}$$

where $R_1, \ldots, R_m$ represents a partition of feature space. For random forest, a subset of predictors is randomly selected and divided at each branch to combine the prediction results of many different tree structures.

XGBoost is an ensemble learning method based on tree regression. Firstly, the objective function is transformed through the second-order Taylor expansion, then the regularization term is introduced, including the number of leaf nodes and the L2 norm of weight vector. Finally, the final objective function is obtained by merging the coefficients. For a tree with T leaves,

$$L^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2] + \gamma T \tag{7}$$

where G_j is the sum of the first-order partial derivatives of samples contained in leaf node, H_j is the sum of second-order partial derivatives,w_j is the weight vector of leaf node j in the tth tree, $G_j, H_j, \lambda, \gamma$ are constant.

Long short-term memory network model is a special kind of RNN, which solves the problem that when the time line is long, the residual will decrease exponentially, which leads to the slow updating of network weight. It adds a storage unit on the basis of RNN, and does linear processing for long time information. At time t, the output is h_{t-1} and the input is x_t. LSTM includes forget gate (f_t), input gate (i_t), and output gate (o_t), all of which are controlled by Sigmoid function (σ):

$$f_t = \sigma \left(W_f [h_{t-1}, x_t] + b_f \right) \tag{8}$$

$$i_t = \sigma \left(W_i [h_{t-1}, x] + b_i \right) \tag{9}$$

$$o_t = \sigma \left(W_o [h_{t-1}, x_t] + b_o \right) \tag{10}$$

where $\sigma(\cdot) \in (0,1)$. Further, new C_t is computed by the sum of the product of C_{t-1} and f_t, and the product of $\widetilde{C_t}$ and i_t :

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{11}$$

Here, $\widetilde{C_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_C)$, C is the memory cell. Finally, through output gate:

$$h_t = o_t \cdot \tanh(C_t) \tag{12}$$

In this paper, R^2 and MSE are utilized as the metrics, where the formulae can be given as:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2} \tag{13}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{14}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the sample mean. $R^2 \leq 1$, the model is more precise when R^2 is closer to 1. MSE directly reflects the prediction error, but is squared. When the model is trained as a loss function, the model is greatly affected by outliers.

2.3 Procedure

After importing Netflix stock data, this research divides the data set into the training and the testing set. To make the period clearer, the training set in this paper is from 2017 to 2020 and the test set is from 2021 to April 22 of 2022, as shown in Fig. 1, and the ratio is approximately 7:3. Subsequently, the data is applied to the four models: OLS regression, random forest, XGBoost, and LSTM. The parameters of each model are adjusted to improve the accuracy. By comparing the prediction results of the four models and the real price, the advantages and disadvantages of the models can be obtained.

Fig. 1. Dataset separation.

3. Results & Discussion

3.1 Correlation analysis

The correlation coefficients between the variables are given in Fig. 2. The price today has a high correlation with the price of each of the previous days, almost 1, but has a low negative correlation of 0.38 with the natural logarithm of yesterday's volume. Similarly, the correlations are also quite high between the prices of previous days, while lower with the volume at about 0.37.

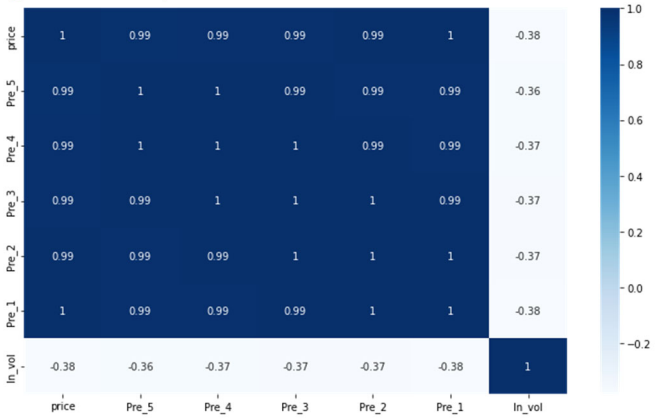

Fig. 2. Correlation matrix.

3.2 Comparison of regression

After constructing the models with the training data, test data is applied to the models and make predictions. The parameters of each model have been optimized. The predictive price timelines of the four models are shown in Figs. 3-6. The importance of the features in random forest is ranked below in Table 2. The price of the previous day is the top important, at 0.92, followed by the price five days ago. In addition, the mean square error (MSE) and R-square (R^2) of the four models is calculated. As

shown in Table 3, OLS has the lowest MSE of 220.2621 and highest R^2 of 0.9732. LSTM has MSE of 1843.5983 and R^2 of 0.7758, followed by random forest (MSE=2723.1944, R^2 =0.6689) and XGBoost (MSE=3510.0922, R^2 =0.5732).

Table. 2 Feature importance of the random forest

Variable	Importance
Pre_1	0.92
Pre_5	0.04
Pre_4	0.01
Pre_3	0.01
Pre_2	0.01
ln_vol	0.0

Table 3. Effectiveness of the models

Model	MSE	R^2
OLS	220.2621	0.9732
Random Forest	2723.1944	0.6689
XGBoost	3510.0922	0.5732
LSTM	1843.5983	0.7758

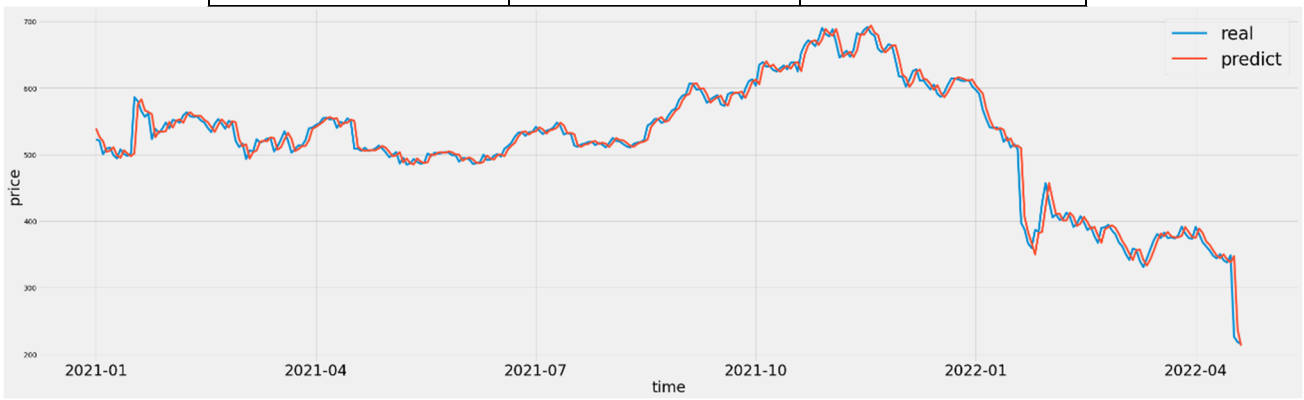

Fig. 3 Real price VS linear regression predicted price.

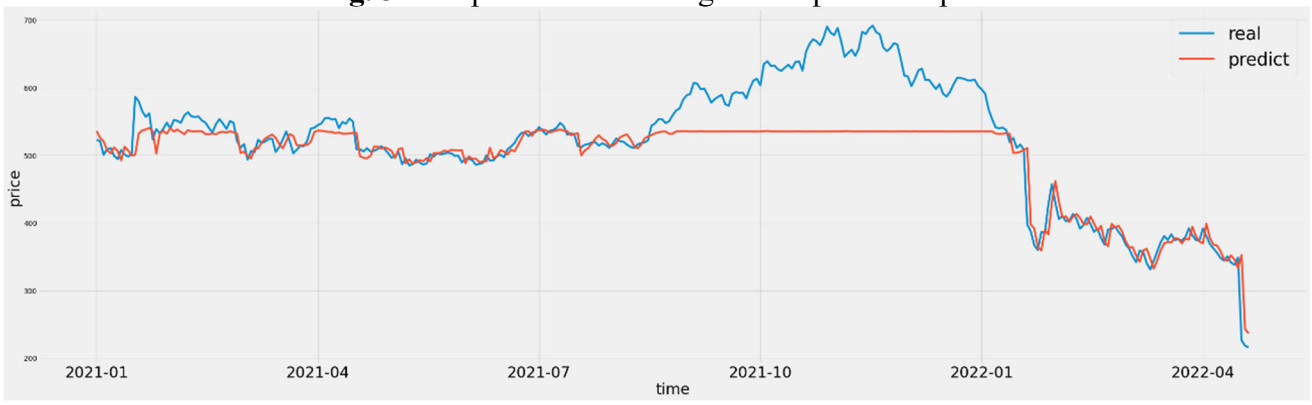

Fig. 4 Real price VS random forest predicted price.

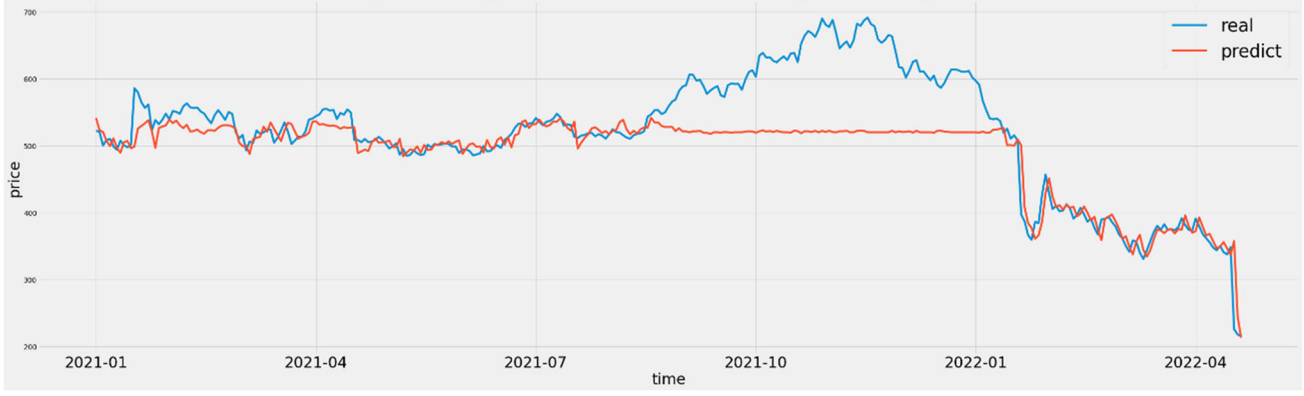

Fig. 5 Real price VS XGBoost predicted price.

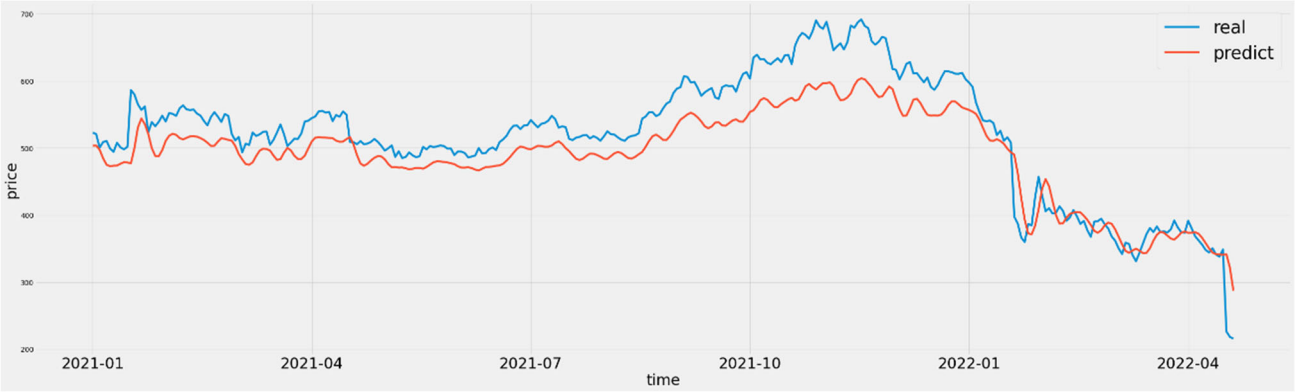

Fig. 6 Real price VS LSTM predicted price.

3.3 Explanation

According to the results of the metrics, OLS has the best prediction performance, with R^2 reaching 97.32%, followed by LSTM, random forest, and Xgboost, all exceeding 50% and ideally reaching 77%. However, this is unreasonable. In the correlation analysis, except for the volume, the other variables are the stock prices of the previous days, which are highly correlated with the prediction target. Nevertheless, the correlation between the price of previous days is high, which may lead to multicollinearity. In this case, the OLS model is not suitable, and the results indicate that the prediction lag is high. For the two tree algorithms, random forest and XGBoost, the random forest seems to perform better in terms of metrics, but the price predicted by the two methods is relatively stable, which in reality cannot deliver future price information to investors in time. The metrics of LSTM are higher than that of random forest, the predicted price change is more time-sensitive and the trend is close to the real situation, but the range of price change is not that large.

3.4 Limitations

The limitation of this research is that the number of dependent variables selected is not enough, and the stock price change is not measured from more dimensions. The correlation coefficients between the independent variables are high, which may lead to multicollinearity. Additionally, it is more practical to evaluate the prediction effect of each model not only by the gap between the predicted price and the real price but also by considering whether the price rises and falls are consistent with the real situation. Apart from the comparison and analysis of mainstream stock prediction models, the existing models can be improved or new models can be created to upgrade the prediction ability of machine learning models for stock prices in the face of different situations.

4. Conclusion

In summary, this paper investigates the stock price prediction based on several frequently used machine learning methods. Due to the difficulty of price forecasting, scholars have been optimizing models to predict stock market returns more accurately. By obtaining the basic historical data of the stock Netflix, this paper defines the input as the closing price of the past five days and the trading volume of the previous day. Three machine learning models, random forest, XGBoost, and LSTM, are constructed to predict the closing price of the stock on the current day and are compared and analyzed. The comparative analysis based on MSE and R-square shows that LSTM and RF have a relatively better prediction effect on stock price compared with XGBoost. For LSTM, the value of MSE is 1843.5983 and R-square is 0.7758, while for RF, the value of MSE is 2723.1944 and R-square is 0.6689. However, the variables considered in this work may have a single dimension and multicollinearity. At the same time, there is a lack of qualitative prediction of the rise and fall. Overall, the two models effectively predict the price of Netflix stock. These results offer a guideline for selecting more appropriate models in stock price prediction. In the future, one can further improve

the model and optimize the prediction effectiveness by adding variables like stock price's moving average or emotion analysis.

References

[1] He Pinglin, et al. COVID–19's Impact on Stock Prices Across Different Sectors—An Event Study Based on the Chinese Stock Market, Emerging Markets Finance and Trade, 2020, 56:10, 2198-2212.

[2] Masoud N. M. The Impact of Stock Market Performance upon Economic Growth. International Journal of Economics and Financial Issues 2013, 3: 788-798.

[3] Sharma A. et al., Survey of stock market prediction using machine learning approach, 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 506-509.

[4] Md. Mobin Akhtar, et al. Stock market prediction based on statistical data using machine learning algorithms. Journal of King Saud University - Science, Volume 34, Issue 4, 2022, 101940.

[5] Balaji A. Jayanth, et al. Applicability of Deep Learning Models for Stock Price Forecasting An Empirical Study on BANKEX Data, Procedia Computer Science, Volume 143, 2018, Pages 947-953.

[6] Chhajer Parshv, et al. The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction, Decision Analytics Journal, Volume 2, 2022, 100015.

[7] Vijh Mehar, et al. Stock Closing Price prediction using Machine Learning Techniques, Procedia Computer Science, Volume 167, 2020, Pages 599-606.

[8] Huang Biao, et al. Stock Prediction based on Bayesian-LSTM. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018). Association for Computing Machinery, New York, NY, USA, 2018, 128–133.

[9] Kim J. et al. Stock Price Prediction Through the Sentimental Analysis of News Articles. 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), 2019, pp. 700-702.

[10] Alexander Hogenboom, et al. The impact of word sense disambiguation on stock price prediction, Expert Systems with Applications, Volume 184, 2021, 115568.