Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/71/20241645

Review of machine learning with sentimental analysis method for cross-model stock price prediction

Zhiwei Li

Software Engineering and Management Program of McMaster University, 1280 Main Street W, Hamilton, Ontario, L8S 4L8, Canada

lucas.lizhiwei@gmail.com

Abstract. Stock market prediction has been a popular area of research for years. Machine learning, as a fast-developing popular algorithm, is applied to stock market prediction by many previous researchers to better solve the time series involving problems over the changing prices. Different from traditional machine learning algorithms that focus on stock prices only, this paper gives a brief description and review of stock price predictions models that contain sentimental analysis over the recent paper works. Different from the prices in number format, sentimental analysis is more based on textual information extraction and mining, converting them into usable input to pass to a prediction model. This extends the prediction model's takeable input domain and strengthens the accuracy. To better classify the differences between the models, discussion and introduction are given based on different model types about whether they are traditional type or deep neural network embedded. Even though traditional types of models are more popular for sentimental analysis and neural networks perform better in prediction tasks, traditional methods are relatively easy to build or train with more explainability, compared to deep learning models suitable for larger data sets.

Keywords: Machine learning, stock price prediction, sentiment analysis, deep neural network.

1. Introduction

Stock market predictions have been a hot field of study for decades. Even though two famous financial theories, Efficient Market Hypothesis (EMH) and Random Walk Theory (RWT), have argued that financial markets cannot be predicted well using the information people can access [1], researchers around the world have shown that the prediction results can help people to make financial decisions [2]. The most generally used element for prediction is a series of past stock prices over a period. And on the other hand, deep learning, as a method of machine learning algorithms, has a neural structure to store or process data within its multi-layer hierarchy. This makes it able to feed forward the valuable information in the neural network, and thus it can easily help to extract useful information over a given time. This is one greatly recognized advantage of applying machine learning to conduct prediction models. Apart from stock prices, context information is also an important source of data mining where people can extract sentimental information for stock predictions. One of the oldest examples of stock predictions based on context information is from B. Wuthrich and his partner, 1998 [3]. They developed a technique which extract the keywords from the five selected stock indices, and these keywords are associated with a weight based on their occurrent frequency. Then these data are combined with daily closing values,

DOI: 10.54254/2755-2721/71/20241645

which later on develop the probabilistic rules for prediction. This example marks an early success to make good use of text information as a supplement to prices values as prediction input, and the result showed to be satisfactory. This text mining procedure is improved and summarized to be a sentimental analysis process in later works.

Since text mining can bring positive supplements to machine learning predictions, this paper summarized and compared 20 past papers that combine the textual sentimental analysis model and prediction model into a single cross-model for stock prediction, taking the machine learning's advantage to analysis the time-series-involved price data, but also gaining the valuable information from textual data mining. This multidimensional cross-model structure has generally good prediction result [4]. There are two big section to discuss the sentiment analysis model part and the prediction model part separately. Comparison is made based on different types of method within each section. The most used model suggestion will also be given after the discussion.

2. Sentiment analysis model

Sentiment analysis, which is also known as opinion mining, is aiming to determine and extract the emotional tendency over the given context. Text input can be in the format of words or sentences, which are extracted from selected articles or paragraphs. After being gathered, these input texts are passed to the cleansing phase before getting into the sentiment analysis model, which includes removing unwanted characters, tokenization, stop words and punctuation remove. Then, in the sentiment analysis model, a score is produced as the output to represent the sentiment tendency of the input text. This score can be notified with a choice of negative, zero and positive as a usual example, or a selection from a richer pre-decided result set. For stock price prediction, this score works as the technical indicator to summarize and simplify the initial information, and feed into the following prediction model as supplement data with price data. By this way, it is possible to make the relative news title or article information about stock usable in the prediction phase.

Common methods used in this model can be classified into two types based on whether they use deep neural networks or not: one is the traditional category, including support vector machines, lexicon methods, and some other alternative algorithms. The other type is the NN applied method, which mostly includes long short-term memory structures or transformer structures.

2.1. The conventional category

In terms of traditional types of method, lexicon is one of the most common ways to implement. To implement the lexicon method, a pre-determined lexicon dictionary is needed, which collects some commonly used keywords and assigns them into some preset classification with different sentimental scores. Then the word matching will perform across the input text to generate an overall score labeled to this text, which can be used in next phase. In some papers, these keywords are selected to be more impactful ones [5] and used with valence aware dictionary and sentiment reasoner (VADER) [6], which checks polarity, the intensity of the emotion by checking how intensely the statement is positive, negative, or neutral [7]. Another common method is SVM, a supervised machine learning algorithm. To use SVM in text mining, a SVM model needs to be trained based on a pre-labeled training set. SVM's main idea is to find a hyperplane as a boundary that can separate different classifications, and here is to find the mapping between sentiment vectors and stock trends [8]. Then, after training, input texts are passed to the SVM to determine the sentiment result, as we did with the lexicon. SVM can also be usually used with other methods together, like VADER, to calculate the sentiment score for the training set [8].

Lexicon and SVM are relatively easy in structure and implementation. Due to Lexicon's word-score mapping algorithm, this obviousness brings better interpretability, and there is no training phase needed as long as the developer set the mapping correctly, but it has more limited coverage since it and only use preset mapping between words, and it is insensitive to context actual meaning when sarcasm and negation are used. SVM, as a machine learning algorithm, it is robust to overfitting problem due to its algorithm quality, and it is suitable for text analysis these kinds of large numbers of feature involved

DOI: 10.54254/2755-2721/71/20241645

situation. But SVM need a training phase before doing classification, and this training process might take long time and effort when datasets are large.

There are also other kinds of methods or algorithms that can be used in sentiment analysis models, like N-grams and Naive Bayes [9,10]. There are also some useful libraries and tools for this procedure, like the Natural Language Toolkit (NLTK) [11], Text Blob [12] and SSWN [13].

2.2. The Neural network model

Neural networks are another type of model generally used in sentiment analysis models. Deep neural networks (DNN) usually have complex structures, and they need a large amount of training with a big, labeled training dataset as a sample and adjusting over the arguments before implementing the analysis. But DNN generally has better performance over complex sentiment analysis cases, better generalization capability, and fitting ability for different tasks.

The most commonly used types of DNN for sentiment analysis are recurrent neural networks (RNN), convolutional neural networks (CNN), and transformers. RNN is the type of DNN that uses feedforward neural structure to deal with sequential input data, and RNN is mostly implemented in Long Short-Term Memory (LSTM) structures in recent years to make the network find and remember some important information which can be used in later works. Common cases are LSTM is used after word embedding, which is a technique of representing words as vectors of real numbers. The words with similar meanings will have similar representation [8]. CNN is special for its convolutional computing neural networks, and it is effective for capturing local patterns and spatial hierarchies in text data. CNN is used in some sentence level sentiment analysis, and in two of the papers, it is used with max-pooling to extract feature maps that represent the most important events during specific periods [14,15].

So in comparison to the traditional model and the DNN applied model, the traditional model is generally easier to build and implement. These traditional methods have a simple training data set, and the result is also better interpretation and explanation. DNN-applied models, on the other hand, cost more effort and data for training and are complex in structure. But DNN has better coverage over the various general cases and performs better on more general data sets.

3. Prediction model

Prediction model is usually takes in the input data combined with output from previous phase to produce prediction of the future stock price. For input used in this phase, the most general type is of past stock price, including open-close, high-low price with associated volume data. Some technical indicators are also applied as supplement in some of the examples. Traditionally, a large amount of model takes stock price as input only, but in select reviewed papers, they cooperate with the sentiment analysis result from former model and strengthen the prediction precision. Depending on whether they used DNN or not, this paper can also be separated into two categories, traditional method and DNN applied method. Following is the discussion separately for these two types.

3.1. The conventional prediction model

Conventional prediction models can contain one or multiple algorithms inside. Typical model used is SVM. In one of the papers, SVM takes the merged data of sentiment data and stock price, then make predict suggestion whether a buy or a sell is recommended [16]. SVM can also used with a realistic rolling window to eliminate the bias [5]. Instead of SVM, which is aiming for classification, a derivative algorithm Support Vector Regression (SVR) is also used in some of the examples to do regression elimination. It builds a hyperplane using the article text with stock price, and then makes a prediction of the price in 20 minutes future [17]. Decision tree, or sometimes composed in random forest algorithm, is another non-DNN algorithm can be used for prediction. Decision tree is implemented by separating the data into subset based on their feature differences, aiming to make the new input fall into specific categories based on feature classification. Random forest, on the other hand, contains several decision trees trained on different bagging result, and the input will be past to different decision trees to get gathered prediction result. Decision trees have a good explainability, since they are white box model

DOI: 10.54254/2755-2721/71/20241645

[18] and their prediction result can be expressed in Boolean logic. Random forest can resolve the overfitting issue in decision tree by averaging, which is used in many other paper [9] [19]. Some other machine learning algorithm like naïve bayes, logic regression is also considered and implemented in some of the paper [19,20].

3.2. The Deep learning prediction model

In terms of Deep learning prediction model, RNN in format of LSTM is generally popular in selected papers because RNN is suitable for stock price this time series type of input data. LSTM as a type of RNN, can avoid the problem of exploding and vanishing gradients of sequential data, and filter out the meaningful long-term information from marketing price movement [21]. This is thanks to LSTM's three types of gates: forget gate to decide the information to discard, input gate to decide the information to add in cell state, and output gate to decide the output based on cell state and input. A mentioned improvements over the original LSTM model is using the hard sigmoid function instead of the original sigmoid function as an activator of each gate layer in LSTM neural network. Hard sigmoid function is a linear approximation of sigmoid and it can accelerate the calculation of network [22]. LSTM can also combine with RCNN, which is a neural network taking the advantages of CNN and RNN. CNN has a good ability to extract semantic information from texts and RNN is better to catch the context information and to simulate complex temporal characteristics. This allows the model to take information form news title, relating them to temporal features, and driven to directional movements used as prediction guidance [23]. In one paper, LSTM's output is passed to a layer that is connected with two perceptrons that use a SoftMax activation function, and this layer, which summarizes the output feature from price data and sentimental analysis results, is responsible for the final prediction over the prices [22].

There are also some other DNN model proposed for prediction structure, like ISCA algorithm with Back propagation neural networks. Sine cosine algorithm (SCA) uses a mathematical model based on sine and cosine functions to produce random solutions make them fluctuate outwards or towards the best solution parameters for BPNN.

Like mentioned in the sentimental analysis model, the prediction models involved in DNN are complex to build and take more effort and dataset to train, but DNN models tend to have a better coverage of the large dataset. There are varies algorithm options for non-DNN traditional prediction models, but in terms of DNN models, LSTM is still the main stream to implement.

4. Conclusion

There are plenty of examples where the stock price prediction model can cooperate with the sentimental analysis phase to improve the result. Based on our comparison of the selected works, the SVM-based sentimental analysis model is agreed to be relatively easy to build and train. This convenience and good extracting results make the SVM-based model popularly used in many works. In terms of prediction model, a NN-involved structure is preferred, and LSTM is the most popular structure. With some adjustments to improve the feature, LSTM can carry meaningful information to future reuse, which is very suitable for time series-involved prediction work.

References

[1] Cootner, P. H. (1964). The random character of stock market prices. (No Title).

[2] De Fortuny, E. J., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. Information Processing & Management, 50(2), 426-441.

[3] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998, October). Daily stock market forecast from textual web data. In SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218) (Vol. 3, pp. 2720-2725). IEEE.

[4] Usmani, S., & Shamsi, J. A. (2021). News sensitive stock market prediction: literature review and suggestions. PeerJ Computer Science, 7, e490.
[5] Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Systems Journal, 13(1), 760-770.
[6] Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision support systems, 55(3), 685-697.
[7] Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A., & Ganaie, I. A. (2023). Stock prediction by integrating sentiment scores of financial news and MLP-regressor: A machine learning approach. Procedia Computer Science, 218, 1067-1078.
[8] Tipirisetty, A. (2018). Stock price prediction using deep learning.
[9] Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.
[10] Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. PeerJ Computer Science, 7, e476.
[11] Bouktif, S., Fiaz, A., & Awad, M. (2019, October). Stock market movement prediction using disparate text features with machine learning. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1-6). IEEE.
[12] Sirimevan, N., Mamalgaha, I. G. U. H., Jayasekara, C., Mayuran, Y. S., & Jayawardena, C. (2019, December). Stock market prediction using machine learning techniques. In 2019 International Conference on Advancements in Computing (ICAC) (pp. 192-197). IEEE.
[13] Albahli, S., Irtaza, A., Nazir, T., Mehmood, A., Alkhalifah, A., & Albattah, W. (2022). A machine learning method for prediction of stock market using real-time twitter data. Electronics, 11(20), 3414.
[14] Oncharoen, P., & Vateekul, P. (2018, August). Deep learning for stock market prediction using event embedding and technical indicators. In 2018 5th international conference on advanced informatics: concept theory and applications (ICAICTA) (pp. 19-24). IEEE.
[15] Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications, 178, 115019.
[16] Batra, R., & Daudpota, S. M. (2018, March). Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In 2018 international conference on computing, mathematics and engineering technologies (ICoMET) (pp. 1-5). IEEE.
[17] Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. Information Processing & Management, 45(5), 571-583.
[18] Carta, S. M., Consoli, S., Piras, L., Podda, A. S., & Recupero, D. R. (2021). Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. IEEE Access, 9, 30193-30205.
[19] Bouktif, S., Fiaz, A., & Awad, M. (2020). Augmented textual features-based stock market prediction. IEEE Access, 8, 40269-40282.
[20] Gupta, R., & Chen, M. (2020, August). Sentiment analysis for stock price prediction. In 2020 IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 213-218). IEEE.
[21] Fataliyev, K., & Liu, W. (2023, November). MCASP: Multi-Modal Cross Attention Network for Stock Market Prediction. In Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association (pp. 67-77).
[22] Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 57(5), 102212.

DOI: 10.54254/2755-2721/71/20241645

[23] Vargas, M. R., Dos Anjos, C. E., Bichara, G. L., & Evsukoff, A. G. (2018, July). Deep leaming for stock market prediction using technical indicators and financial news articles. In 2018 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.