Universe
Scientific Publishing

Crude Oil Futures Price Prediction Based on CNN-LSTM incorporating News Headlines Feature Extraction

Jiaying Li

Southeast University, Nanjing 211102, China.

Abstract: In this study, a framework for predicting crude oil futures price movements is constructed based on topic modeling of news texts and a combined model of convolutional neural networks and long and short-term memory for feature mining for sentiment analysis. Based on the dual significance of reality and statistics, this paper takes the logarithmic return of crude oil futures settlement price as the prediction target, and uses news headlines to identify the intrinsic influences affecting crude oil futures prices using topic modeling and sentiment analysis. In this study, by selecting market data and text data of WTI crude oil futures from 2011 to 2022 for empirical analysis, the CNN-LSTM prediction model is significantly better than other models.

Keywords: Topic Mining; Financial Time Series Forecasting; Sentiment Analysis; Influencing Factors

1. Introduction

Accurate prediction of crude oil futures market price can help investors to obtain considerable returns and formulate reasonable investment strategies from the theory and practice of urgent and important significance. The key to accurate forecasting of crude oil futures prices lies in identifying the factors that influence the price of crude oil futures and choosing a more appropriate forecasting scheme.

Initially, Kailian[1] studied the role of supply and demand factors in driving oil prices from the structure of supply and demand equilibrium in 2009 and concluded that oil prices are mainly dominated by global demand. Pejić et al.[2] showed that news headlines are the most effective textual data, and the most commonly used methods are topic mining and sentiment analysis. Topic models are able to infer the topic probability distribution of a document set given the document set to identify the topic and thus identify the influencing factors of crude oil futures price, LDA model is the most commonly used topic model, it is chose in this paper. In addition, investor sentiment is increasingly considered in asset price studies. Tissaoui et al.[3] showed that uncertainty factors such as investor panic are the main factors leading to oil price volatility. However, The financial domain has a large number of jargons and lacks an appropriate corpus, and Araci [4] proposed the FinBERT model based on the BERT pre-training model to deal with the NLP task in the financial domain. In preprocessing financial text can be speeded up by this pre-training approach.

Recurrent neural networks have better ability to memorize sequence information, Siami-Namini et al.[5] showed that LSTM model and BiLSTM fit better compared to ARIMA model in financial time series. CNN can extract more deep features from complex time series both horizontally between features and vertically in time space. Vidal et al.[6]recognized that images associated with time series have static and dynamic information of the data and added LSTM to CNN network model to predict gold volatility using a combined model.

The rest of the paper is organized as follows. Section 2 introduces the related theories, including the prediction framework, theme mining and sentiment analysis. Section 3 is empirical analysis. Section 4 shows the experimental results comparing the models, prediction assessment metrics. Finally, Section 5 provides conclusions, and recommendations for further research.

2. Methodology

2.1 Structure

The forecasting structure of this thesis consists of five parts. First, this study obtains market data and media news headlines of crude oil futures. Second, the two parts of the data are preprocessed, the data are cleaned in the market data, and several time series of price movements are analyzed statistically after removing outliers, and the forecast target series with practical significance and satisfying the characteristics of smoothness and non-white noise are selected. Segmentation of text data, removal of deactivated words, stemming extraction and so on. Besides, do the construction of technical indicators on the market data, thematic modeling on the processed text data to identify new influencing factors and do sentiment analysis to get the sentiment score.Furthmore all the factors are integrated and data alignment is done and finally this thesis predicts the movement of crude oil futures price using some models as the baseline model and evaluates its prediction results.

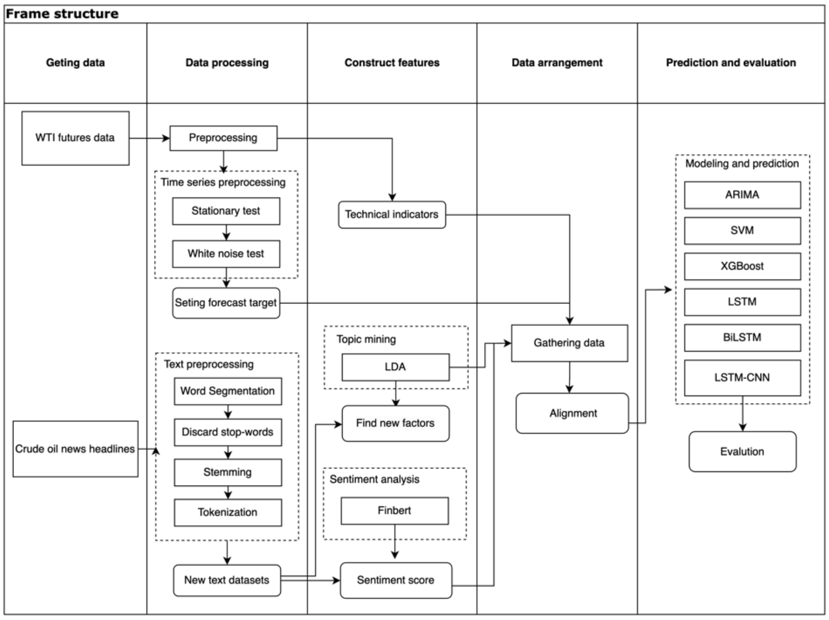

Fig. 1 Frame structure in this paper

2.2 Latent Dirichlet Allocation

LDA (Latent Dirichlet Allocation) is an unsupervised Bayesian model, a topic model that gives the topic of each document in the document set as a probability distribution. It is also an unsupervised learning algorithm that does not require a manually labeled training set for training, but simply gives the document set and a specified number k of topics. A document can contain multiple topics, and each word in the document is generated by one of the topics. Each word in a document is composed by selecting a certain topic with a certain probability, and selecting a certain word from that topic with a certain probability. One of the important formulas is:

$$p(word|text) = p(word|topic) \times p(topic|text) \qquad (4)$$

Where, $p(word|text), p(word|topic), p(topic|text)$ are the conditional probability of word in text, conditional probability of word in topic, and conditional probability of topic in text, respectively.

3. Empirical analysis

3.1 Data processing and target sequence analysis and selection

In this paper, daily frequency data of WTI crude oil futures main continuous contract and news headlines and hot comments on oilprice.com are obtained from wind, and the text and market data are selected from July 1, 2011 to June 30, 2022, with a total of 2,750 samples.

In this study, four alternative prediction targets were initially selected: Y^1,Y^2,Y^3,Y^4,where Y^1 is the closing price of crude oil future，Y^2 is the one-day log return at the settlement price，Y^3,Y^4 are the first-order and second-order difference of the settlement price，Y^1,Y^2,Y^3,Y^4 is computed in Equation (1)- Equation (4):

$$Y_t^1 = s_t \qquad (1)$$

$$Y_t^2 = log\left(s_t \div s_{(t-1)}\right) \qquad (2)$$

$$Y_t^3 = \Delta s_t = s_t - s_{(t-1)} \qquad (3)$$

$$Y_t^4 = \Delta s_t - \Delta s_{(t-1)} \qquad (4)$$

The variation curves of Y^1,Y^2,Y^3,Y^4 over the intercepted time horizon are presented as follows:

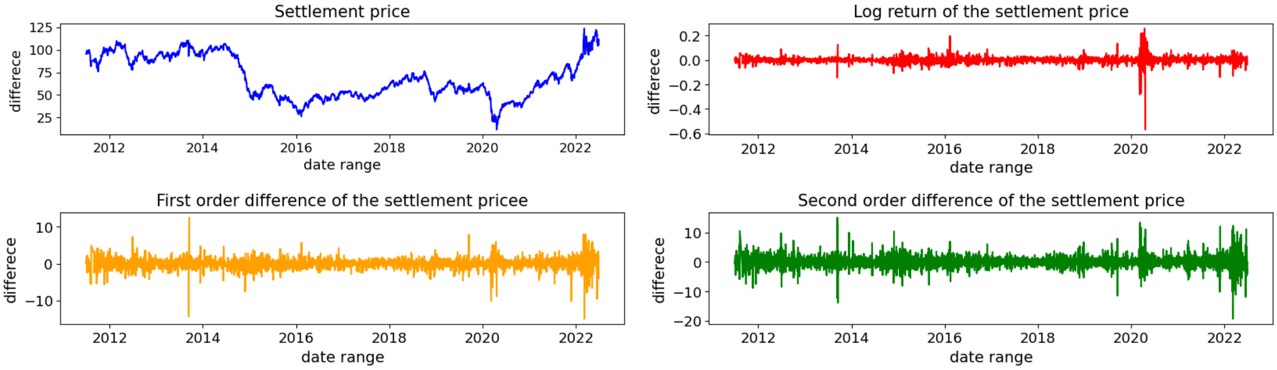

Fig. 2 Change curves of four time series from 2011 to 2022

Perform the smoothness test and white noise test for Y^1,Y^2,Y^3,Y^4 in turn, the ADF of Y^1,Y^2,Y^3,Y^4 is shown in the Table 1, and all but Y^1 are smooth sequences, And Y^2,Y^3,Y^4 are not white noise sequences:

Table 1 ADF test of Y^2,Y^3,Y^4

	Y^1	Y^2	Y^3	Y^4
Adf Test Statistic	-1.820595171	-8.947899781	-11.294228937	-11.294228937
pvalue	0.370285758	8.885366722e-15	1.356142462e-20	3.682854027e-29
Critical Value(1%)	-3.432744970	-3.43275377	-3.432744093	-3.432755542
Critical Value(10%)	-2.567333274	-2.567335344	-2.567333067	-2.567335760

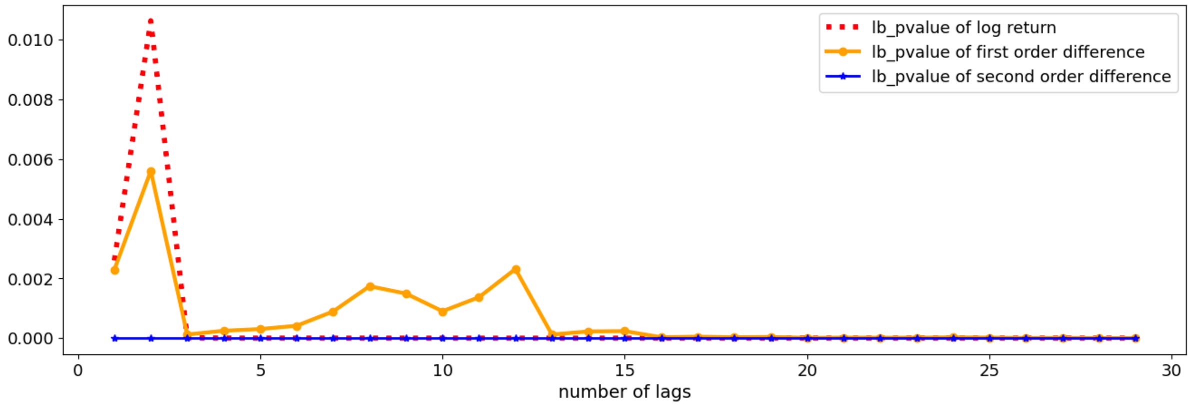

Fig. 3 White noise test of settlement price's log return

In terms of the practical significance of the three target series, Y^2 is the logarithmic return of one day's based on the settlement price in the historical period, which embodies the staring profit and loss under the daily no-liability system of futures trading and determines the change of the investors' settlement reserve.

3.2 Potential factors and topic modeling

In addition to the initial data, Firstly, in this study, the LDA model was used and other relevant features were condensed from it. The first thing that is done is text preprocessing. Second, In preforming LDA , it is a crucial question to choose a reasonable topic number. In this paper, perplexity was used to confirm this question. Perplexity can also be used to compare two probability distributions or probability models. The perplexity is calculated in the interval from 10 to 50, as shown in Fig. 4. And the smallest point of perplexity to the number of topics being 36, and therefore select the number of topics of 36 for the LDA model.

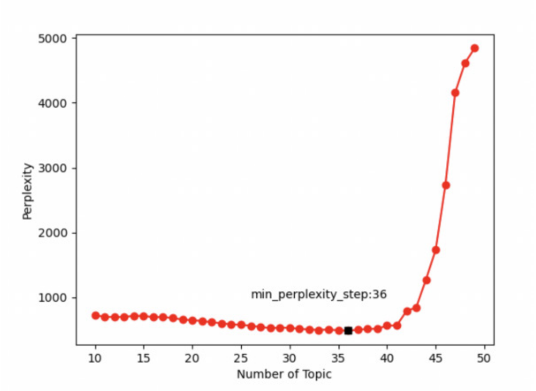

Figure 4 Perplexity of LDA in the number of topics in (10, 50)

The first four keywords were selected for each of the 36 themes, and the final results were manually categorized by experts to extract the five major theme categories,[7]Furthermore, we summarized 8 factors that affect the price of WTI crude oil futures :Geopolitical, Supply and demand, Exploitation, Storage, Transportation, Global economy, Monetary policy, season. Then, we selected the WTI crude oil spot price, USDX and DJIA[8]. Besides, according to the study of Zhang et al. [9] ,we chose geopolitical risk index as a new factor.

3.3 News headlines' sentiment analysis

For hard-to-quantify factors that cannot be directly queried, they can be represented by adding investor sentiment together with sentiment labels. We used the K-means algorithm of clustering to divide the target value into five classes, with the label set of , and this is used as the label for the training and prediction of the textual sentiment analysis, with the distribution of the labels as shown in the Fig 6.

The data selected in this paper includes 2750 trading days, in order to get the text labels corresponding to the text data, we splited the data into a training set including 1500 samples and a test set including 1250 samples.

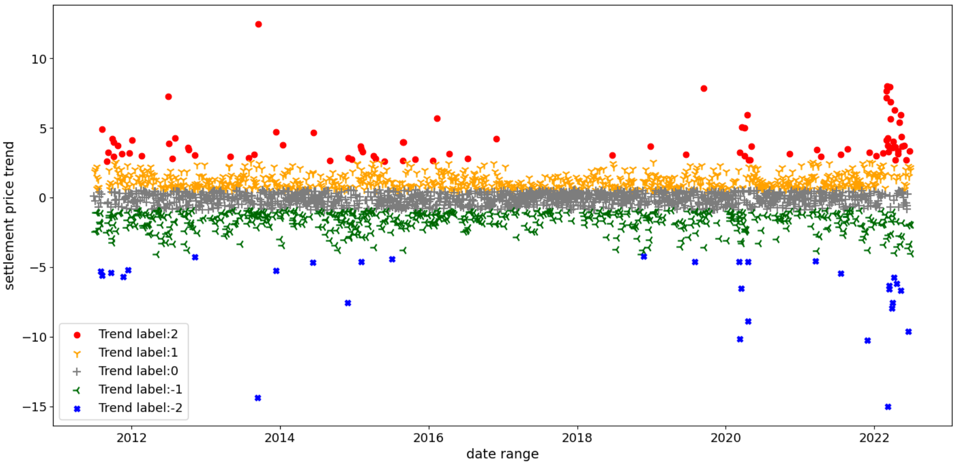

Fig. 6 Trend classification graph by k-means

In sentiment analysis, an output layer is added on top of Finbert to accomplish the sentiment classification task, by comparing it with Finbert's native sentiment analysis task, and the causal analysis of its prediction results shows that the latter has no relevance at all to the final prediction goal. While the other not, so Finbert is fine-tuned to achieve the sentiment analysis of the original news headlines.

4. Result and discussion

4.1 Model configuration

In order to obtain the optimal effect on each model, we adopted the grid search method to determine the optimal parameters of each model, taking SVR as an example, the optimization of the selected kernel function, penalty coefficients and error thresholds are needed. To prevent model overfitting, simple parameter values are used and dropout layers are added in the middle of the neural network. And the CNN-LSTM framework constructed in this paper is based on two Cov2d and two LSTM in series.

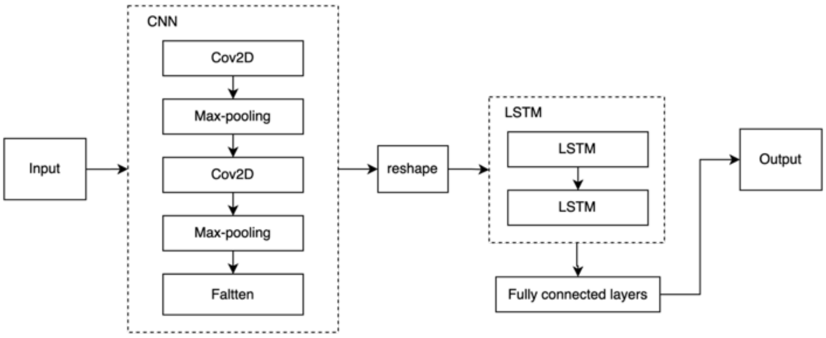

Figure 7 the structure of CNN-LSTM

4.2 Futures log return forecasting results

MAE and RMSE were used in this paper, the two criteria could evaluate the prediction performance of the models during the testing period.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_{i-1}| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i-1})^2} \tag{10}$$

$$IR_{mae} = \frac{MAE_O - MAE_F}{MAE_O} \tag{11}$$

$$IR_{rmse} = \frac{RMSE_O - RMSE_F}{RMSE_O} \tag{12}$$

Where n is the number of observations in the test period, y_i is the actual value of the log return on day i, and $y\hat{}_{i-1}$ is the log return forecast obtained using the forecasting models. O,F denote raw data and complete data, IR denotes the magnitude of change between the two states, respectively.

Table 2 Demonstration of evaluation metrics for different situations

Predictior	**ARMA**	**XGBoost**	**LSTM**	**BiLSTM**	**CNN-LSTM**
Without new factors or sentiment analysis					
MAE	0.05140	0.05189	0.04834	0.04923	0.02431
RMSE	0.05739	0.05987	0.05493	0.05905	0.03132
With new factors and sentiment analysis					
MAE	0.05092	0.04571	0.04227	0.04562	0.01542
IR_{mae}	0.9339%	11.9098%	12.5569%	7.333%	36.5844%
RMSE	0.05723	0.05800	0.04810	0.05231	0.02091
IR_{rmse}	0.2788%	3.1234%	12.4340%	11.4140%	33.2375%

In Table 2, it can be seen that among the models, the overall effect of CNN-LSTM is better than others, the ARMA model performs better compared to the machine learning model in terms of prediction. The difference between LSTM and BiLSTM is small. Whereas, the prediction accuracy of CNN-LSTM with labeled sequences of sentiment analysis and other new factors under a uniform training set is the highest among all the comparison experiments. From Table 2, it can be seen that CNN-LSTM reduces MAE and RMSE by about 36.6% and 33.24%, respectively.

5. Conclusion

In this paper, we develop a framework for predicting crude oil futures price movements by introducing news texts to identify intrinsic influences and sentiment labels. And as it shows, the prediction effect of the model proposed in this paper is higher than other comparative models.

A deeper exploration of the futures price prediction problem can also be carried out on the basis of this paper. Although some of the potential influences found by the thematic model with proxy variables have been applied in the study of this paper, such as weather, etc., for which no proxies can be found directly, that are attributed to investor sentiment in a generalized way. Therefore, the realizability of this part of influencing factors can be fully explored in future research.

References

[1] Kilian L. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market [J]. American Economic Review, 2009, 99(3): 1053-69.

[2] Pejić Bach M, Krstić Ž, Seljan S, et al. Text mining for big data analysis in financial sector: A literature review [J]. Sustainability, 2019, 11(5): 1277.

[3] Tissaoui K, Zaghdoudi T, Hakimi A, et al. Do Gas Price and Uncertainty Indices Forecast Crude Oil Prices? Fresh Evidence Through XGBoost Modeling [J]. Computational Economics, 2022: 1-25.

[4] Araci D. Finbert: Financial sentiment analysis with pre-trained language models [J]. arXiv preprint arXiv:190810063, 2019.

[5] Siami-Namini S, Tavakoli N, Namin AS. A comparative analysis of forecasting financial time series using arima, lstm, and bilstm [J]. arXiv preprint arXiv:191109512, 2019.

[6] He Z, Zhou J, Dai HN, et al. Gold price forecast based on LSTM-CNN model; proceedings of the 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), F, 2019 [C]. IEEE.

[7] Li J, Li G, Liu M, et al. A novel text-based framework for forecasting agricultural futures using massive online news headlines, [F].

[8] Li X, Shang W, Wang S. Text-based crude oil price forecasting: A deep learning approach, [F].

[9] Zhang Z, He M, Zhang Y, et al. Geopolitical risk trends and crude oil price predictability, [F].