Crude oil price prediction using CEEMDAN and LSTM-attention with news sentiment index

Zhenda Hu

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles **76**, 28 (2021)
© Z. Hu, published by IFP Energies nouvelles, 2021
https://doi.org/10.2516/ogst/2021010

Available online at:
ogst.ifpenergiesnouvelles.fr

REGULAR ARTICLE OPEN ACCESS

Crude oil price prediction using CEEMDAN and LSTM-attention with news sentiment index

Zhenda Hu*

School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, PR China

Received: 27 September 2020 / Accepted: 8 March 2021

Abstract. Crude oil is one of the most powerful types of energy and the fluctuation of its price influences the global economy. Therefore, building a scientific model to accurately predict the price of crude oil is significant for investors, governments and researchers. However, the nonlinearity and nonstationarity of crude oil prices make it a challenging task for forecasting time series accurately. To handle the issue, this paper proposed a novel forecasting approach for crude oil prices that combines Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Long Short-Term Memory (LSTM) with attention mechanism and addition, following the well-known "decomposition and ensemble" framework. In addition, a news sentiment index based on Chinese crude oil news texts was constructed and added to the prediction of crude oil prices. And we made full use of attention mechanism to better integrate price series and sentiment series according to the characteristics of each component. To validate the performance of the proposed CEEMDAN-LSTM_att-ADD, we selected the Mean Absolute Percent Error (MAPE), the Root Mean Squared Error (RMSE) and the Diebold-Mariano (DM) statistic as evaluation criterias. Abundant experiments were conducted on West Texas Intermediate (WTI) spot crude oil prices. The proposed approach outperformed several state-of-the-art methods for forecasting crude oil prices, which proved the effectiveness of the CEEMDAN-LSTM_att-ADD with the news sentiment index.

1 Introduction

Crude oil is one of the most powerful resources in the world. The fluctuation of crude oil price plays an important role in the development of bulk commodity and global economy [1]. Under the comprehensive effects of market supply and demand game, US dollar exchange rate, speculative trading, geographical conflicts, natural disasters and other factors, the international crude oil price fluctuates sharply, which increases the difficulty of crude oil price prediction. Therefore, to build a scientific and reasonable model to accurately predict the trend of international crude oil price has become a hot and difficult issue in academic circles, investment circles and political circles.

However, due to the comprehensive effects of factors mentioned above, the fluctuation of crude oil price presents nonstationarity and nonlinearity [2], making the prediction of crude oil price a challenging task. The research of crude oil price forecasting mainly includes two directions. The first direction is choosing effective models or improving the algorithm to better extract the features of price series and then predict. The second direction is to find the external indicators that affect the crude oil prices series, including financial policy, the price of related financial products, news sentiment and public opinions, to better predict the future trend of the original series.

For the first direction, a lot of models have been proposed [3–11]. And in recent years, a novel "decomposition and ensemble" framework has been widely used in time series prediction, which can significantly improve the forecasting accuracy [12]. In this framework, the original sequence is first decomposed into several components, and then each component is predicted by a single model. Finally, the several prediction results are integrated to get the final prediction result. For the second direction, some researchers [13, 14] found that news articles and social media data were pretty beneficial in financial prediction. And other research methods proved that crude oil price had a significant relationship with different economic indicators. Oladosu [15] used Empirical Mode Decomposition (EMD) method to study the relationship between Gross Domestic Product (GDP) of US and crude oil prices. King *et al.* [16] found that political events and economic news, the same as oil supply and demand, played an important role of oil prices.

In the financial market, information sentiment contained in news articles and social media data is an important index reflecting the sentiment and viewpoint of

* Corresponding author: huzhenda2020@gmail.com

investors and traders. The text contents of news include not only the report of facts, but also the intonation of language and emotion. Therefore, the news describing the fluctuation of crude oil price reflects the crude oil market situation through texts and influences the investor sentiment through the way of network communication. However, the consideration of these text data makes the analysis of financial market even more complex [17, 18]. Inspired by this correlation, we quantify crude oil news as a sentiment index and introduce it into crude oil price prediction model.

Overall, we combine these two directions mentioned above for crude oil price prediction. This paper leads a news sentiment index to predict crude oil price more precisely in the framework of "decomposition and ensemble". Concretely, we first construct a crude oil news sentiment index, which quantifies news text as a numerical index. Then we decompose the raw crude oil prices series and crude oil news sentiment index into several components, respectively, via Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). Next, Long Short-Term Memory (LSTM) with attention mechanism is applied to forecasting each prices component with corresponding sentiment component. Finally, the prediction values of each component are summed to get the final price prediction value.

The main contributions of this paper are as follows: 1) A news sentiment index based on crude oil news texts was constructed and added to the prediction of crude oil prices. 2) This paper proposed a novel forecasting approach for crude oil prices that combines CEEMDAN, LSTM with attention mechanism and addition, following the well-known "decomposition and ensemble" framework. And we made full use of attention mechanism to better integrate price series and sentiment series according to the characteristics of each component. 3) Abundant experiments were conducted on West Texas Intermediate (WTI) spot crude oil prices from US Energy Information Administration (EIA). The proposed approach outperformed several state-of-the-art methods for forecasting crude oil prices, which proved the effectiveness of the news sentiment index and attention mechanism.

2 Related work

In recent years, a large number of prediction models have been proposed. It can be divided into three categories: time series models, Artificial Intelligence (AI) models and hybrid prediction models. For the first categories, Abledu and Agbodah [3] established Autoregressive Integrated Moving Average (ARIMA) model to quantitatively predict the international oil price. Baumeister and Kilian [4] proved that Vector AutoRegression (VAR) models could have good performance when forecasting short-term crude oil prices. Although the time series model can better describe the linear characteristics of crude oil price series, it is hard to fit the nonlinear characteristics of crude oil price series. Therefore, many researchers introduced AI models into oil price prediction. Shin *et al.* [5] used neural network model to forecast the monthly price of WTI crude oil from January 1992 to June 2008. Yu *et al.* [6] utilized Least Squares Support Vector Regression (LSSVR) model to predict US WTI crude oil price. And taking gold prices into consideration in the forecasting, Tang and Zhang [7] built a multiple wavelet Recurrent Neural Network (RNN) model for crude oil price forecasting. The experimental results showed the effectiveness of the model. Zhao *et al.* [8] applied deep neural network model to the price prediction of WTI crude oil and achieved good results. Because of the multiple characteristics of the crude oil price series, the mixed model with different models became an effective choice. Kristjanpoller and Minutolo [9] showed that the proposed network approach had ability to improve the prediction results for both spot oil prices and future oil prices. Safari and Davallou [10] proposed a hybrid forecasting model which integrated exponential smoothing model, ARIMA model and nonlinear autoregressive neural network.

Owing to nonstationarity and nonlinearity of the original price series, the family of EMD provides a new method for processing time series data. It starts from the characteristics of data itself and reveals the internal fluctuation characteristics of data by decomposing the fluctuation information of original signal on different scales. Some researches [19–23] have demonstrated that it is an effective time series analysis tool and applied it to price forecasting.

For the consideration of text data, many studies have proved the correlation between investor sentiment and stock market volatility. Devitt and Ahmad [24] used the emotional tendency of financial reviews to predict future financial trends. Das and Chen [25] proved that there was a high positive correlation between stock index and online sentiment analysis using linear regression. Bollen *et al.* [26] tracked the public mood state from the content of huge amount of micro-blog feeds by simple text processing techniques. However, in the field of crude oil market, there are few researches on news sentiment analysis and crude oil price fluctuation. Lechthaler and Leinert [27] utilized the VAR model to study the price fluctuation of global crude oil market, which showed that news sentiment has an important influence on the fluctuation of crude oil price. Alfano *et al.* [28] demonstrated news sentiment not only has an impact on the noise residual of oil, but also on the basic price trend through the regression analysis of news sentiment and oil time series decomposition components.

3 Approach

3.1 The construction of news sentiment index

3.1.1 The construction of sentiment dictionary

Basic sentiment dictionary

This paper uses HowNet Sentiment Dictionary as basic dictionary and adds NTUSD Chinese Sentiment Dictionary of Taiwan University of China in order to enrich the vocabulary. To adapt to the field of the news texts, we also add common financial sentiment words and related sentiment vocabulary in the field of crude oil. After deleting duplicate words, the final basic sentiment dictionary can be obtained.

Modified dictionary

Next, this paper constructs the modified dictionary which includes degree level dictionary and the dictionary of negator. The degree level dictionary consists of adverbs of degree, transition words and so on with different weights. According to [29], the weights of adverbs of degree are directly from HowNet Dictionary and shown in Table 1. The weights are determined according to the intensity of degree adverbs. The weights defined in HowNet Dictionary are widely used in Chinese sentiment analysis [30, 31]. Level One and Level Two represent emotional decline so we set the value less than 1. And from Level Three to Level Six, the words represent emotional enhancement so we set the value more than 1. The weight difference of each level is set to 0.2. The detailed weight distribution is shown in Table 1.

In addition, the dictionary of negator includes all negative words, such as never, no, shouldn't, don't, won't and so on.

3.1.2 The calculation of weighted sentiment series

Combined with the characteristics of international crude oil news and financial news, this paper designs the relevant news sentiment analysis rules. The contents of news sentiment analysis are as follows:

The score and weights of sentiment words

After text segmentation, we utilize the basic sentiment dictionary to conduct sentiment analysis. We set the score of positive sentiment words as 1 ($W_p = 1$) and the score of negative sentiment words as -1 ($W_n = -1$).

In addition, in order to distinguish the importance of different sentiment words, we use TFIDF (Term Frequency–Inverse Document Frequency) value as the weight of each sentiment word. TFIDF is used to evaluate the importance of a word in a document. If a word or phrase appears frequently in one article and rarely appears in other articles, it is considered that the word or phrase has good classification ability and should be given a higher degree of importance.

The calculation of modified words and position

After setting the score and weights of sentiment words, we need to calculate the influence of modified words and the position between modified words and sentiment words to optimize the weights. We develop weight optimization rules based on semantic rules. The rules are as follows:

(a) If there are no modifiers or negatives in front of semantic words, the score of the emotional word does not change.
(b) If there is any negative in front of semantic words, the score of the semantic word needs to be multiplied by -1.
(c) If there is any modifier in front of semantic words, the score of the semantic word needs to be multiplied by the weight of this modifier.

We can obtain the sentiment scores optimized by the above three semantic rules.

Table 1. Adverbs of degree and the weights.

Level	Word (English translation)	Weight
Six	Extreme, Extraordinary, Doubly and so on	1.7
Five	Over, Excessive, More than and so on	1.5
Four	Especially, Particularly and so on	1.3
Three	Further, More and so on	1.1
Two	A little, slightly and so on	0.9
One	Relatively, Not much and so on	0.7

The calculation of daily news sentiment index

We can get the final emotional value by summing all the sentiment words in the news. The summing formula is represented as equation (1):

$$\text{article\_sentiment} = \sum_{i=1}^{n_1} w_i (w_i > 0), \tag{1}$$

where article_sentiment represents the sentiment score of each news, n_1 represents the number of sentiment words in each news. And then, we can obtain the sentiment score of crude oil news in each trading day by averaging the sentiment scores of all crude oil news in the day. The formula is represented as equation (2):

$$\text{daily\_sentiment} = \sum_{i=1}^{Z} \text{article\_sentiment}_i / Z, \tag{2}$$

where daily_sentiment represents the sentiment score of each trading day, Z represents the number of articles in that day.

Through the above calculation of sentiment scores for crude oil news, we get the Chinese news sentiment index.

3.2 The proposed CEEMDAN-LSTM_att-ADD with news sentiment index

3.2.1 Ensemble empirical mode decomposition

Huang *et al.* [32] first proposed EMD in 1999. It is a signal processing method, which can be used to process nonlinear and nonstationary signals. However, EMD has some shortcomings, which may lead to the problem of "mode mixing". In order to solve this problem, Wu and Huang [33] proposed ensemble EMD based on EMD in 2009. The steps of EEMD are as follows:

Step 1: Determine the standard Gaussian white noises $g_i(t) \sim N(0, \sigma^2)$, (the standard deviation σ is usually set as 0.1 or 0.2), the ensemble number E and a loop variable $i = 1$.

Step 2: Add a Gaussian white noise $g_i(t)$ to the raw series $Y(t)$ to obtain the following new series:

$$Y_i(t) = Y(t) + g_i(t). \tag{3}$$

Step 3: Conduct EMD on $Y_i(t)$ to get m intrinsic mode functions (IMFs) and one residue series $r_i(t)$:

$$Y_i(t) = \sum_{j=1}^{m} c_{ij}(t) + r_i(t), \tag{4}$$

where $m = \lfloor \log_2 T \rfloor - 1$ [33], determined by the length of raw series T.

Step 4: Add 1 to the loop variable i. If $i > m$, execute Step 5; otherwise, go back Step 2.

Step 5: Calculate the jth final IMF $C_j(t)$ in E trials as shown in equation (5):

$$C_j(t) = \frac{1}{E}\sum_{j=1}^{I} c_{ij}(t). \tag{5}$$

Step 6: Obtain the residue series as shown in equation (6):

$$r(t) = Y(t) - \sum_{j=1}^{m} C_j(t). \tag{6}$$

Finally, the raw series can be divided into m IMFs and one residue.

However, due to the difference of the chosen white noise, the mode functions obtained by decomposition are different, which makes EEMD unstable. And EEMD method is difficult to completely eliminate the reconstruction error caused by Gaussian white noise. In order to further solve these problems, Torres *et al.* [34] proposed CEEMDAN, on the basis of EEMD in 2011, which can better obtain the intrinsic mode functions and accurately reconstruct the original signal, with lower operation cost than EEMD algorithm.

3.2.2 LSTM with attention mechanism

Hochreiter and Schmidhuber [35] proposed LSTM in 1997 to overcome the problem of gradient disappearing in RNN. The difference between LSTM and RNN is that three gates including input gates, output gates and forget gates are added in LSTM. LSTM is well suited for dealing with long-term dependency issues. The specific structure of the model is shown in Figure 1, where the cell state C is used to record the long-term status of the sequence and hidden state h is used to record the current status of the sequence.

The forget gate decides what information needs to be retained. It is a sigmoid function whose value is from 0 to 1 and determines the degree of forgetting the input information. And forget gate's input includes output of last sequence and input of this sequence. The formula is represented as equation (7):

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right). \tag{7}$$

The input gate determines what information will be added to the current input information in the long-term state. And then we building a new vector. The formulas are defined as follows in equations (8) and (9):

$$i_t = \sigma\left(W_i h_{t-1} + U_f x_t + b_i\right), \tag{8}$$

$$a_t = \tan h(W_a h_{t-1} + U_a x_t + b_i). \tag{9}$$

Passing through forgetting gate and input gate, old cell state C_{t-1} is updated to C_t. The formula is represented as equation (10):

$$C_t = C_{t-1} \odot f_t + i_t \odot a_t. \tag{10}$$

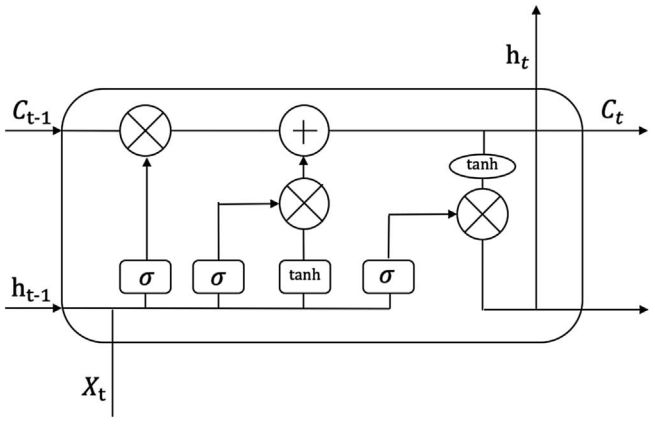

Fig. 1. The unit structure of LSTM.

The output gate controls what input data and long-term status should be output at the current time. Output gate determines the output information. This gives the output cell state which is obtained through sigmoid layer and then through a tan h function. The formulas are given in equations (11) and (12):

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o), \tag{11}$$

$$h_t = o_t \odot \tan h(C_t). \tag{12}$$

LSTM method has been widely used in many fields, such as Natural Language Processing (NLP), image processing, speech recognition and so on. In recent years, LSTM has also been applied in time series analysis and achieved good performances.

And the attention mechanism [36] was proposed by Google Deep Learning team in 2014. The mechanism of attention originates from the study of vision, which is proposed by referring to the operation of human brain. It can help us to find more important information in the deep learning task. Bahdanau *et al.* [37] used the similar attention mechanism to translate and align in the machine translation task, which is the first work to propose the attention mechanism to apply to the NLP field. Recently, attention mechanism has become a hot research field of machine learning.

The principle of attention mechanism is to change the hidden state h_t which should be directly transferred to the neural network of the previous layer to the weight of all hidden states of the previous all layers. The formula is expressed as equation (13):

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i. \tag{13}$$

The attention mechanism makes that the final output is not the previous value of all hidden layers, but the weighted sum of hidden layers values obtained by the similarity of input vector and goal vector so as to focus on the important information. The attention mechanism can better integrate the historical price series and news sentiment series to forecast future price.

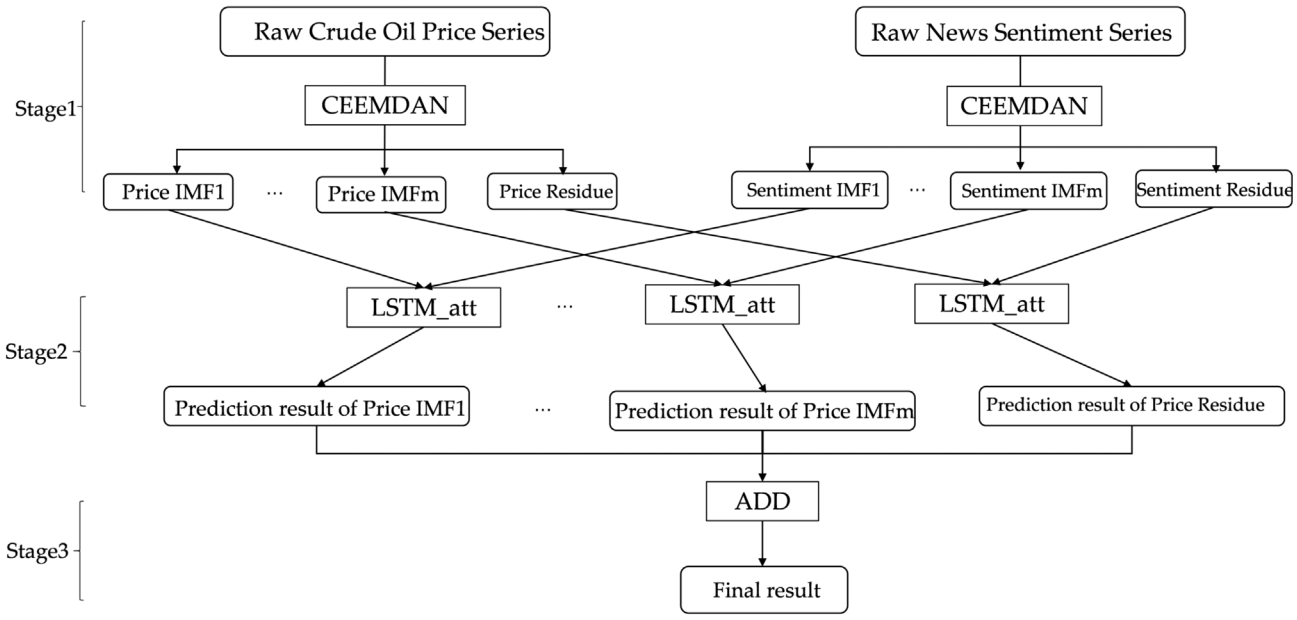

Fig. 2. The flowchart for the proposed approach.

3.2.3 The proposed CEEMDAN-LSTM_ att-ADD with news sentiment index

Considering the powerful abilities of CEEMDAN on signal preprocessing and the advantages of LSTM on time series prediction, this paper puts forward a novel approach that integrates CEEMDAN, LSTM with attention mechanism and addition, termed CEEMDAN-LSTM_att-ADD, in addition with news sentiment index, for forecasting crude oil prices, which follows the "decomposition and ensemble" framework. The approach is shown in Figure 2, consisting of three stages:

Stage 1: Decomposition. CEEMDAN is applied to decompose the raw series of crude oil prices $Y(t)$ ($t = 1, 2,\ldots, T$) into two parts: (1) m components $hj(t)$ ($j = 1, 2,\ldots, m$); (2) one residue component $r(t)$. The same operation is applied to the raw news sentiment index $X(t)$ ($t = 1, 2,\ldots, T$).

Stage 2: Individual prediction. For each component from Stage 1, the sentiment index component is attached with corresponding price component. The LSTM with attention model is fit on the training set, and then, the model is applied to the test set.

Stage 3: Ensemble prediction. The test results of all components from Stage 2 are aggregated by simple addition as the final forecasting results.

The CEEMDAN-LSTM_att-ADD with news sentiment index first used CEEMDAN to decompose the original crude oil price series and the sentiment index into several components, which transformed the complex series into several relatively simple components. And each component reflects some characteristics of raw series. Specifically, the first several IMFs reflect high-frequency parts, while the last several IMFs and the residue reflect the low-frequency parts of raw series. Secondly, the CEEMDAN-LSTM_att-ADD builds a forecasting model for each single price component along with corresponding sentiment index component, and then, the forecasting model can predict the test data of each single component. Finally, the predicted values from each component can be aggregated as the final forecasting results of crude oil prices. All these steps contribute to improve the performance for crude oil prices prediction.

4 Experiment

4.1 Data description

4.1.1 Crude oil Chinese news texts

According to the authority and integrity of news, we selected two websites including the website of ZhongYou (http://www.cnoil.com/) and the website of international oil (http://oil.in-en.com/) as our data sources. The news contents of these two websites mainly focus on the field of crude oil. We crawl crude oil news texts from the two websites using crawler technology, totally 31 525 articles from 28 February 2006 to 9 March 2020. Due to the lack of news data in some time periods, we choose the news data from 4 December 2014 to 25 February 2020 to ensure data continuity. Finally, after deleting the part of news, we get totally 23 033 news texts and an average of 17 news samples per day.

The specific steps are as follows: 1) get the URL list of the page to be grabbed; 2) grab the page and 3) parse the page. First of all, by analyzing the source code of the two websites, the URL lists of relevant news can be obtained. And then we grab the pages. In the final step of page parsing, we use HTML tag rules and regular expression to match and crawl the content of the news we need.

After obtaining the news texts, we utilized them to construct the news sentiment index via the approach mentioned above.

4.1.2 Crude oil price series

In order to test the validity of our approach CEEMDAN-LSTM_att-ADD with news sentiment index, this paper used an open crude oil price series, WTI crude oil daily spot prices, from US EIA for its authority and importance in the international crude oil markets. Considering the start and end time of crude oil news in the two websites, we used the daily close prices from 4 December 2014 to 25 February 2020, with a total 1311 samples, as experimental data. Among the samples, the first 80% ones were used as training samples, and the remaining 20% ones were used as testing samples.

The news sentiment index and the original crude oil prices with corresponding components decomposed by CEEMDAN of WTI are shown in Figure 3 and Figure 4, respectively. The top series in the two figures are the original data, followed by the IMFs and the residual series.

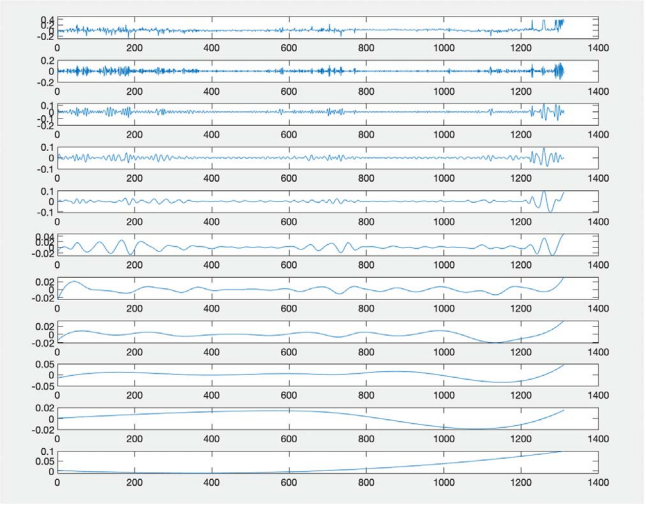

Fig. 3. The daily news sentiment index decomposed by CEEMDAN.

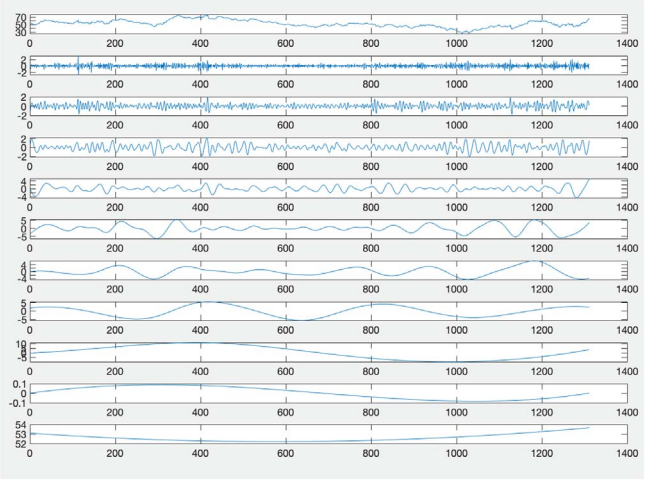

Fig. 4. WTI crude oil daily prices decomposed by CEEMDAN.

4.2 Evaluation criteria

As for the evaluation criteria, we chose two most frequently used evaluation indicators for time series prediction: the Mean Absolute Percent Error (MAPE) and the Root Mean Squared Error (RMSE). MAPE and RMSE are defined as equation (14) and equation (9), respectively:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - y_i}{Y_i}\right|, \tag{14}$$

$$\mathrm{RMSE} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(Y_i - y_i)^2}, \tag{15}$$

where Y_i is the actual value, y_i is the predicted value at time i and N is the total number of test samples.

In addition, we utilized the Diebold-Mariano (DM) statistic to compare the prediction accuracy of two models, which is defined as equation (16):

$$S = \frac{\frac{1}{n}\sum_{t=1}^{n} d_t}{\sqrt{\frac{1}{n}\left(\gamma_0 + 2\sum_{k=1}^{n-1}\mathrm{cov}(d_t, d_{t-k})\right)}}, \tag{16}$$

where $d_t = (y_t - \hat{y}_{1t})^2 - (y_t - \hat{y}_{2t})^2$, $\hat{y}_{1t}$ represents the values predicted by the first model at time t, while $\hat{y}_{2t}$ is predicted by the other model. If the value is negative, it means the first model is statistically better than the other one.

4.3 Experimental settings

As for crawling crude oil news, we used the package Beautiful Soup 4 because it can provide users with different resolution strategies and strong speed flexibly. And for word segmentation, we used the tool called jieba.

To show the prediction ability of LSTM with attention mechanism on crude oil price series, firstly, we conducted experiments on raw series using single models. Three popular models of time series prediction including SVR, AdaBoost and Random Forest were used to compare with it. Then we did experiments on the decomposed series via CEEMDAN under the framework of "decomposition and ensemble". And the prediction values of each component were added to get the final price prediction value owing to the ADDition (ADD) operation's simplicity and effectiveness. We call this approach as CEEMDAN-LSTM_att-ADD. At the same time, other three ensemble models were further compared with the approach to verify the validity of it.

For LSTM, we set 16 as the number of hide layer nodes after experimental comparison. In addition, we chose 4 and 100 as the value of batch size and the value of epochs, respectively. For SVR, we used Radial Basis Function (RBF) as the kernel function of the model. For AdaBoost and Random Forest, 50 sub-models and 20 sub-models were chosen.

In addition, we conducted the operation of data normalization, which is helpful for time series forecasting. Data normalization can speed up the training of objective function and unify the dimension of sample. In this paper, we used the Min-Max normalization for all models, which maps the raw values in the range of [0, 1]. The normalization formula is as equation (17):

$$Y_{\text{norm}} = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}}, \tag{17}$$

where $Y_{\max}$ and $Y_{\min}$ are the maximum and minimum values of the original dataset, respectively. Y and Y_{norm} are the raw values and the normalized values, respectively. After the prediction sequence is obtained, we can carry out the inverse normalization operation.

In order to explore the ability of model prediction in advance, we chose horizons from 1 to 6 and conducted multi-step-ahead prediction. The forecast horizon means the number of days ahead of the predicting day. For example, the horizon at 6 means that we used Y_t, (t = 1, 2, 3,..., T) to predict Y_{t+6}. For predicting the value of crude oil price using a price time series Y_t, (t = 1, 2, 3,..., T) and a news sentiment series X_t, (t = 1, 2, 3,..., T), it can be formulated as equation (18):

$$y_{t+h} = f\left(Y_t, Y_{t-1}, Y_{t-2}, \cdots, Y_{t-(l-1)}, X_t, X_{t-1}, X_{t-2}, \cdots, X_{t-(l-1)}\right), \tag{18}$$

where Y_t, y_t represents the predicted value of Y_t at time t and l represents the lag order. As suggested by Zhou *et al.* [38] and shown by our comparative experiments, we set 6 as the lag order.

All the experiments were conducted by Python3.7 on a 64-bit Windows 10 professional edition with an i7-8565U CPU @1.8 GHz and 8 GB RAM. The experiments have been performed with the following Python libraries. Packages pandas and numpy were used for data processing while packages BeautifulSoup, requests and jieba were used for crawling and word segmentation. As for the models, we utilized packages sklearn and Keras for machine learning and deep learning models, respectively.

4.4 Results and analysis

4.4.1 Single models

The MAPE and RMSE values for single models on WTI crude oil prices are shown in Table 2 and Table 3, respectively. For MAPE values, LSTM with attention achieved the lowest values in all horizons. For RMSE values, LSTM with attention got the best performance in 5 out of 6 cases. And for all models, MAPE values and RMSE values increased when the horizon increased.

4.4.2 Ensemble models

The results are shown in Tables 4 and 5. As for the ensemble models, CEEMDAN-LSTM_att-ADD achieved the lowest (the best) values among all models in all horizons for MAPE values and RMSE values. Compared with the above single models, the performance of ensemble models improved significantly. As for LSTM-attention model, the MAPE values and RMSE values at least reduced 50%. But for AdaBoost model and Random Forest model, the results did not improve obviously, even worse sometimes. This shows that CEEMDAN has little effect on some ensemble models.

Table 2. The MAPE values by different single models on WTI crude oil daily prices.

Horizon	LSTM_att	SVR	AdaBoost	RF
One	**0.0229**	0.0361	0.0250	0.0245
Two	**0.0325**	0.0405	0.0358	0.0325
Three	**0.0381**	0.0456	0.0410	0.0393
Four	**0.0447**	0.0502	0.0464	0.0454
Five	**0.0497**	0.0545	0.0501	0.0530
Six	**0.0537**	0.0586	0.0550	0.0575

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

Table 3. The RMSE values by different single models on WTI crude oil daily prices.

Horizon	LSTM_att	SVR	AdaBoost	RF
One	**1.4455**	2.1408	1.5949	1.5690
Two	2.0848	2.4467	2.2113	**2.0590**
Three	**2.4754**	2.7494	2.5719	2.5356
Four	**2.7841**	3.0050	2.8739	2.8457
Five	**3.0893**	3.2504	3.0883	3.2762
Six	**3.3587**	3.4988	3.3713	3.5635

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

Table 4. The MAPE values by different ensemble models on WTI crude oil daily prices.

	CEEMDAN			
Horizon	LSTM_att	SVR	AdaBoost	RF
One	**0.0095**	0.0295	0.0281	0.0201
Two	**0.0142**	0.0310	0.0322	0.0248
Three	**0.0145**	0.0316	0.0354	0.0273
Four	**0.0147**	0.0328	0.0403	0.0323
Five	**0.0226**	0.0347	0.0434	0.0351
Six	**0.0239**	0.0368	0.0467	0.0396

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

4.4.3 Ensemble models with news sentiment index

After the news sentiment index taken into consideration, the MAPE values and RMSE values of all models on WTI crude oil prices almost improved. The results could be seen from Tables 6 and 7. For CEEMDAN-LSTM_att-ADD, the MAPE values and RMSE values

Table 5. The RMSE values by different ensemble models on WTI crude oil daily prices.

CEEMDAN				
Horizon	LSTM_att	SVR	AdaBoost	RF
One	**0.5879**	1.8764	1.8408	1.4335
Two	**0.9270**	1.9819	2.1427	1.6910
Three	**0.9405**	2.0447	2.3145	1.8269
Four	**0.9387**	2.1205	2.6146	2.0612
Five	**1.3900**	2.2489	2.8004	2.2277
Six	**1.5075**	2.3766	3.0138	2.5068

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

Table 6. The MAPE values by ensemble models with news sentiment index on WTI crude oil daily prices.

With news sentiment index CEEMDAN				
Horizon	LSTM_att	SVR	AdaBoost	RF
One	**0.0065**	0.0291	0.0276	0.0173
Two	**0.0116**	0.0306	0.0312	0.0221
Three	**0.0129**	0.0311	0.0359	0.0262
Four	**0.0143**	0.0321	0.0411	0.0322
Five	**0.0208**	0.0337	0.0427	0.0359
Six	**0.0231**	0.0347	0.0461	0.0398

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

both improved significantly, achieving 46.15% reduction and 46.46% reduction, respectively in horizon one. In other horizons, the values also improved obviously, which proved the function of news sentiment index. And for CEEMDAN-SVR, the MAPE values and RMSE values improved slightly and the effect is not as good as the effect of LSTM with attention, showing that the attention mechanism could better combine the price series and sentiment index. For CEEMDAN-AdaBoost-ADD and CEEMDAN-RF-ADD, the results changed unsteadily with the horizon increasing, which further proved that LSTM with attention could utilize the sentiment index effectively.

As for the statistical test, the results of the DM test for the ensemble methods on WTI are shown in Table 8, statistically confirming the above results. Firstly, the superiority of the proposed CEEMDAN-LSTM_att-ADD was validated from the perspective of statistics. Specifically, the p-values of the CEEMDAN-LSTM_att-ADD were far below 0.01 in all cases. This demonstrates that the proposed CEEMDAN-LSTM_att-ADD performed statistically better than other benchmark models at a confidence level of 99.9% in most cases when it was treated as the model of testing the target.

4.5 Discussion

In addition, in order to explain the impact of the weight of sentiment words in news texts, the impact of lag order and the impact of number of news texts dataset samples, we conducted three groups of comparative experiments. The results can be seen from Table 9, Figure 5 and Figure 6, respectively.

Table 7. The RMSE values by ensemble models with news sentiment index on WTI crude oil daily prices.

With news sentiment index CEEMDAN				
Horizon	LSTM_att	SVR	AdaBoost	RF
One	**0.4014**	1.7183	1.9010	1.1280
Two	**0.7587**	1.8312	2.1330	1.4607
Three	**0.8485**	1.8882	2.3769	1.7392
Four	**0.9055**	1.9808	2.6660	2.1096
Five	**1.3654**	2.1267	2.8289	2.3699
Six	**1.4562**	2.2093	3.0841	2.6136

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

4.5.1 The impact of the weight of sentiment words

For simplicity, the first comparative experiment set the single model LSTM with attention to baseline. "+ News" represents the news sentiment index was added and "+ News + TFIDF" represents the news sentiment index was added and TFIDF value of sentiment words were used as their weights. Some interesting results can be found from Table 9. When we add the sentiment index to LSTM_att model, the MAPE values got smaller only in horizons four and six while all RMSE values did not get improved. The result showed that untreated sentiment index could not make the performance better. And we continued to add TFIDF value of sentiment words, the situation changed. It achieved the best values in 4/6 cases for MAPE and 3/6 cases for RMSE. The results illustrated that it might cause noise if the sentiment index was not processed.

4.5.2 The impact of lag order

To further analyze the impact of the lag order on LSTM_att + News + TFIDF, we chose one-step-ahead prediction with the lag order in the range of 1–12. The results were shown in Figure 5. It can be seen that the MAPE values and RMSE values slightly improved with the lag order increasing from 1 to 6. After the lag order 6, the values did not change obviously or became worse. The reason might be that more noises were introduced to the model. By contrastive analysis, we can choose a suitable lag order 6 for forecasting crude oil prices accurately.

4.5.3 The impact of number of news texts dataset samples

In order to further illustrate the impact of number of news texts dataset samples on the prediction results, we randomly eliminate part of news samples in the step of

Table 8. The DM test results for ensemble models on WTI crude oil daily prices.

CEEMDAN		Benchmark Model						
Horizon	Test Model	SVR(s)	AdaBoost(s)	RF(s)	LSTM_att	SVR	AdaBoost	RF
One	LSTMatt(s)	−15.33 (0.0000)	−12.76 (0.0000)	−10.54 (0.0000)	−9.21 (0.0000)	−20.34 (0.0000)	−18.87 (0.0000)	−17.63 (0.0000)
Two	LSTMatt(s)	−13.64 (0.0000)	−10.45 (0.0000)	−9.23 (0.0000)	−8.98 (0.0014)	−17.76 (0.0000)	−16.22 (0.0000)	−16.85 (0.0000)
Three	LSTMatt(s)	−12.78 (0.0000)	−11.98 (0.0000)	−8.46 (0.0003)	−9.34 (0.0000)	−15.34 (0.0000)	−17.26 (0.0006)	−15.37 (0.0000)
Four	LSTMatt(s)	−11.65 (0.0000)	−10.12 (0.0000)	−7.39 (0.0000)	−7.54 (0.0025)	−13.77 (0.0000)	−15.39 (0.0000)	−13.94 (0.0000)
Five	LSTMatt(s)	−10.98 (0.0000)	−10.76 (0.0000)	−8.11 (0.0031)	−6.33 (0.0000)	−14.34 (0.0000)	−14.36 (0.0000)	−13.55 (0.0000)
Six	LSTMatt(s)	−9.35 (0.0000)	−7.58 (0.0000)	−6.87 (0.0045)	−7.85 (0.0000)	−12.65 (0.0000)	−13.93 (0.0000)	−11.58 (0.0000)

Table 9. The impact of the news index for LSTM_att.

	MAPE			RMSE		
	LSTM_att	+ News	+ News + TFIDF	LSTM_att	+ News	+ News + TFIDF
One	0.0229	0.0267	**0.0228**	**1.4455**	1.7150	1.4597
Two	0.0325	0.0345	**0.0310**	**2.0848**	2.3392	2.0890
Three	0.0381	0.0400	**0.0374**	**2.4754**	2.7808	2.5367
Four	0.0447	**0.0435**	0.0437	2.7841	2.9340	**2.7411**
Five	0.0497	0.0497	**0.0483**	3.0893	3.3694	**3.0208**
Six	0.0537	**0.0511**	0.0538	3.3587	3.3902	**3.3317**

Note: The values in bold signify the minimum MAPE or RMSE values of different models for each horizon.

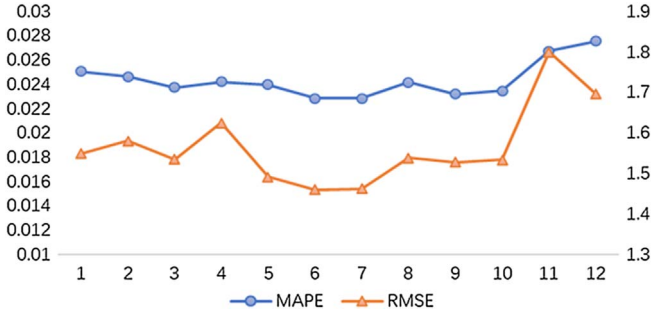

Fig. 5. The impact of lag on WTI crude oil prices with one-step-ahead prediction (Note: left and right scales on the vertical axis stand for, respectively, MAPE and RMSE).

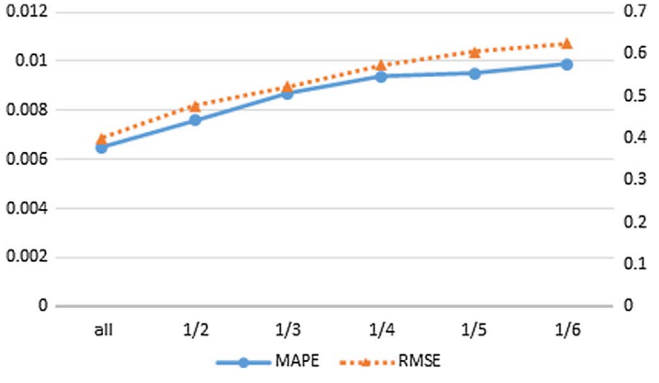

Fig. 6. The impact of number of news texts dataset samples on WTI crude oil prices with one-step-ahead prediction (Note: left and right scales on the vertical axis stand for, respectively, MAPE and RMSE).

constructing news sentiment index. The results were shown in Figure 6. We selected 1/2, 1/3, 1/4, 1/5, 1/6, respectively as the proportions of remaining samples to original samples. It can be seen that when the number of remaining samples decreased, the MAPE values and RMSE values correspondingly increased, which means enough news texts samples is necessary for constructing good news sentiment index.

5 Conclusion

The nonlinearity and nonstationarity of crude oil prices make it a challenging task for forecasting time series accurately. Traditional methods including statistical methods

and AI-based models usually cannot achieve satisfactory results when the forecasting is performed on raw crude oil prices. To handle the issue, this paper proposed a novel forecasting approach for crude oil prices that combines CEEMDAN, LSTM with attention mechanism and addition (namely CEEMDAN-LSTM_att-ADD), under the well-known "decomposition and ensemble" framework. Inspired by the correlation between the news sentiment and crude oil prices, a news sentiment index based on Chinese crude oil news texts was constructed and added to the prediction of crude oil prices.

After the results and analyses above mentioned, the conclusions are as follows: 1) Compared with other models, LSTM with attention mechanism model was very powerful for forecasting crude oil prices in both single models and ensemble models; 2) The ensemble models significantly improved the forecasting accuracy when compared with single models, owing to the strategy of decomposing and ensemble; 3) The attention mechanism could utilize the sentiment index effectively and the results improved significantly when combining the price series and sentiment index for CEEMDAN-LSTM_att-ADD.

In the future, we can improve our work from the following three aspects. First, a more detailed syntax analysis can be carried out for crude oil news text to remove the text noise and make the calculation of emotional indicators more accurate. In addition, for different components after decomposition, different models are adaptively selected according to their properties to improve the prediction results. Finally, exploring the impact of a sentiment index in level term but also in variance term using structural VAR models a valuable research direction.

Acknowledgments. This work was supported by the graduate innovation fund of Shanghai University of Finance and Economics (under Project No. CXJJ-2020-428).

References

1 Galyfianakis G., Garefalakis A., Mantalis G. (2017) The effects of commodities and financial markets on crude oil, *Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles* **72**, 1, 3.
2 Wang Y., Wei Y., Wu C. (2011) Detrended fluctuation analysis on spot and futures markets of West Texas Intermediate crude oil, *Phys. A, Stat. Mech. Appl.* **390**, 5, 864–875.
3 Abledu G.K., Agbodah K. (2012) Stochastic forecasting and modelling of volatility of oil prices in Ghana using ARIMA time series model, *Eur. J. Bus. Manag.* **4**, 16, 122–131.
4 Baumeister C., Kilian L. (2012) Real-time forecasts of the real price of oil, *J. Bus. Econ. Stat.* **30**, 2, 326–336.
5 Shin H., Hou T., Park K., Park C.K., Choi S. (2013) Prediction of movement direction in crude oil prices based on semi-supervised learning, *Decis. Support Syst.* **55**, 1, 348–358.
6 Yu L., Xu H., Tang L. (2017) LSSVR ensemble learning with uncertain parameters for crude oil price forecasting, *Appl. Soft Comput.* **56**, 692–701.
7 Tang M., Zhang J. (2012) A multiple adaptive wavelet recurrent neural network model to analyze crude oil prices, *J. Bus. Econ.* **64**, 4, 275–286.
8 Zhao Y., Li J., Yu L. (2017) A deep learning ensemble approach for crude oil price forecasting, *Energy Econ.* **66**, 9–16.
9 Kristjanpoller W., Minutolo M.C. (2016) Forecasting volatility of oil price using an artificial neural network GARCH model, *Expert Syst. Appl.* **65**, 233–241.
10 Safari A., Davallou M. (2018) Oil price forecasting using a hybrid model, *Energy* **148**, 49–58.
11 Yu L., Zhao Y., Tang L. (2014) A compressed sensing based AI learning paradigm for crude oil price forecasting, *Energy Econ.* **46**, 236–245.
12 Zhang X., Lai K.K., Wang S. (2008) A new approach for crude oil price analysis based on empirical mode decomposition, *Energy Econ.* **30**, 905–918.
13 Xing F.Z., Cambria E., Welsch R.E. (2018) Natural language based financial forecasting: a survey, *Artif. Intell. Rev.* **50**, 1, 49–73.
14 Xing F.Z., Cambria E., Welsch R.E. (2018) Intelligent asset allocation via market sentiment views, *IEEE Comput. Intell.* **13**, 4, 1–20.
15 Oladosu G. (2009) Identifying the oil price-macroeconomy relationship: An empirical mode decomposition analysis of US data, *Energy Policy* **37**, 12, 5417–5426.
16 King K., Deng A., Metz D. (2012) *An econometric analysis of oil price movements: The role of political events and economic news, financial trading, and market fundamentals*, Bates White Economic Consulting.
17 Bohn T.A. (2017) Improving long term stock market prediction with text analysis, *Electronic Thesis and Dissertation Repository*, The University of Western Ontario. Available at https://ir.lib.uwo.ca/etd/4497.
18 Li X., Wang C., Dong J., Wang F., Deng X., Zhu S. (2011) Improving stock market prediction by integrating both market news and stock prices, in *Database and Expert Systems Applications, 22nd International Conference, DEXA 2011, Bilbao, Spain, August 29 – September 2, 2011, Proceedings, Part II*, A. Hameurlain, S.W. Liddle, K.-D. Schewe, X. Zhou (eds), Springer-Verlag, Berlin Heidelberg, pp. 279–293.
19 Li T., Hu Z., Jia Y., Wu J., Zhou Y. (2018) Forecasting crude oil prices using ensemble empirical mode decomposition and sparse bayesian learning, *Energies* **11**, 7, 1882.
20 Yu L., Dai W., Tang L. (2016) A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting, *Eng. Appl. Artif. Intel.* **47**, 110–121.
21 Li J., Wang J. (2020) Stochastic recurrent wavelet neural network with EEMD method on energy price prediction, *Soft Comput.* **24**, 17133–17151.
22 Abdollahi H. (2020) A novel hybrid model for forecasting crude oil price based on time series decomposition, *Appl. Energy* **267**.
23 Wu Y., Wu Q., Zhu J. (2019) Improved EEMD-based crude oil price forecasting using LSTM networks, *Physica A Stat. Mech. Appl.* **516**, 114–124.
24 Devitt A., Ahmad K. (2007) Sentiment polarity identification in financial news: A cohesion-based approach. Association for Computational Linguistics, in: A. Zaenen, A. Van Den Bosch (eds), *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 984–991.
25 Das S.R., Chen M.Y. (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web, *Manage Sci.* **53**, 9, 1375–1388.
26 Bollen J., Mao H., Zeng X. (2011) Twitter mood predicts the stock market, *J. Comput. Sci.* **2**, 1, 1–8.

27 Lechthaler F., Leinert L. (2012) Moody oil-What is driving the crude oil price? *CER-ETH Economics working paper series 12/168*, CER-ETH-Center of Economic Research (CER-ETH) at ETH Zurich.
28 Alfano S.J., Feuerriegel S., Neumann D. (2015) Is news sentiment more than just noise?
29 Dong Z., Dong Q. (2003) HowNet – a hybrid language and knowledge resource, in C. Zong (ed), *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, IEEE, pp. 820–824.
30 Yan J., Bracewell D.B., Ren F., Kuroiwa S. (2008) The creation of a Chinese emotion ontology based on HowNet, *Eng. Lett.* **16**, 1, 166–171.
31 Liu J., Xu J., Zhang Y. (2013) An approach of hybrid hierarchical structure for word similarity computing by HowNet, in: R. Mitkov, J.C. Park (eds), *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing, pp. 927–931.
32 Huang N.E., Shen Z., Long S.R. (1999) A new view of nonlinear water waves: The Hilbert Spectrum, *Ann. Rev. Fluid Mech.* **31**, 1, 417–457.
33 Wu Z., Huang N.E. (2009) Ensemble empirical mode decomposition: A noise-assisted data analysis method, *Adv. Adapt. Data Anal.* **1**, 1, 1–41.
34 Torres M.E., Colominas M.A., Schlotthauer G., Flandrin P. (2011) A complete ensemble empirical mode decomposition with adaptive noise, in: *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, pp. 4144–4147.
35 Hochreiter S., Schmidhuber J. (1997) Long short-term memory, *Neural Comput.* **9**, 8, 1735–1780.
36 Mnih V., Heess N., Graves A., Kavukcuoglu K. (2014) Recurrent models of visual attention, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (eds), *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. **2**, MIT Press, Cambridge, MA, United States, pp. 2204–2212.
37 Bahdanau D., Cho K., Bengio Y. (2015) Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (eds), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, International Conference on Learning Representations,* ICLR.
38 Zhou Y., Li T., Shi J., Qian Z. (2019) A CEEMDAN and xgboost-based approach to forecast crude oil prices, *Complexity* **2019**, 1–15.