Crude oil price and the trend Prediction

Jongbae Yoon

December 13th, 2022

1 Abstract

Crude oil is closely related to our lives. Not only individuals but also lots of businesses and industries heavily rely on the crude oil price, so predicting crude oil price has always been an important issue in economics. The aim of this study is to build a computational model that predict the oil price more accurately. For that, this project focuses on two aspects: fundamental analysis and technical analysis. For technical analysis, neural network models are compared, and the LSTM (Long Short-Term Memory) model outperforms CNN (Convolutional Neural Network) model. For fundamental analysis, this project finds the most impactful factors that determine the oil: 'World Steel Production', 'Kilian Index', 'ISM index', 'CRB Index', 'Dollar Index', 'Number of terrors in Middle East and Northern Africa' are turned out to be the most impactful factors in crude oil price. Then compare the relationship between the prediction of crude oil price and these features by analyzing the movement of the values. The result shows that the prediction values of these major factors have similar trend with the crude oil price prediction. Thus, LSTM models for the features can be used as indicators to enhance the accuracy of crude oil prediction.

2 Introduction

Crude oil is a naturally occurring unrefined petroleum product composed of hydrocarbon deposits and other organic materials, and it can be refined to produce usable products such as gasoline, diesel, and different forms of petrochemicals [1] (Rayan et al., 2021). It is one of the most valuable commodities in the world and its pricing attributes to the world economy in different levels. Oil price fluctuations have an indisputable effect on the world economy at different levels. Khalid and Sarwar studied the impacts of crude oil price on micro to macro levels in Pakistan and reported significant correlations between oil price and several macro- and micro- economic indexes [2] (M.Khalid et al., 2018). Many studies have attempted to model oil price fluctuations, however developing a robust model to predict oil price volatilities is not an easy task due to the intricacies and highly dynamic nature of the oil industry. Traditionally, statistical time series models such as ARIMA (Auto Regressive Integrated Moving Average) are used

to predict the oil price. However, due to the availability of large datasets along with increased computing power, neural network methods are widely used not only in domains such as image and language processing, but also in time-series analysis domains such as stock price prediction and crude oil price prediction. It is reported that some of the neural network models such as LSTM, CNN outperform traditional statistical models even in time series analysis.

For time series data analysis, there are two aspects to be considered: fundamental analysis and technical analysis. For crude oil price prediction, the fundamental analysis is implemented through analyzing the factors that have impacts on the crude oil price. Analyzing micro and macro economies, Supply Demand, and geopolitical issues are the examples of fundamental analysis. On the other hand, technical analysis is implemented through analyzing time series data using computational and mathematical model analysis. Analyzing past oil prices and trading volumes are the examples of technical analysis. Most of the economics studies only focus on the fundamental analysis, while majority of the machine learning / neural network studies only work on technical analysis. This project focuses on both perspectives of the analyses. The aim of this study is summarized as follows:

1. Introduce factors that are known to have impacts on Crude oil price.
2. Find the most impactful factors among others.
3. Find the best performing model on the crude oil price using neural network models.
4. Apply the model on the most influential factors identified.
5. Find the relationships between crude oil price and the features.

The rest of this paper is organized as follows: Section 3 introduces related works and literature reviews. Section 4 explains key methods used in this paper such as dataset and the techniques / algorithms applied. Section 5 narrates the experimental design, explains the overall architecture of the study. Section 6 shows the results and explains insights / limitations of the study. Section 7 summarizes the overall paper and explains potential ethical implications of this project.

3 Related Work

3.1. Details of crude oil price

Like most commodities, the fundamental driver of crude oil price is supply and demand in the market, and market sentiment toward the physical product. In addition, oil futures contracts, which are traded heavily by speculators, play a dominant role in price determination. Cyclical trends in the commodities market may also play a role. As crude oil differs in quality and availability depending on where it comes from, producers and traders need a reliable benchmark against

which to judge the correct price. Two of the most widely used benchmarks are Brent crude and WTI (West Texas Intermediate) crude oil. Brent Crude oil originates from the North Sea between the Shetland Islands and Norway, and 2/3 of all crude oil is priced using Brent Crude as the benchmark. West Texas Intermediate (WTI) origins primarily from Oklahoma, Texas, Louisiana, and North Dakota, and it accounts for the rest 1/3 of all oil pricing benchmark. For this project, WTI crude oil price is used. [3] U.S EIA (Energy Information Administration) explains that Crude oil prices are heavily determined by global supply and demand. Economic growth is one of the most critical factors affecting petroleum product. Growing economies increase demand for energy in general and especially for transporting goods and materials from producers to consumers. [4] (Hong Miao et al., 2018) introduces various features known to determine the crude oil price. They apply various machine learning models and find out that 'LASSO' model returns the most accurate results among others. Then, it identifies 6 most impactful features; 'World Steel Production', 'Kilian Index', 'ISM index', 'CRB Index', 'Dollar Index', 'Number of terrors in Middle East and Northern Africa'.

3.2. CNN (Convolutional Neural Network) in time-series analysis

CNN is one of artificial neural network model and each node of the model receives input only from a small neighborhood in previous layer and shares parameters. It has proven its performance in the areas like computer vision, such as face recognition, image classification, and it is most widely used in these areas. However, CNN is not limited to such areas. Recently CNN is applied to time-series forecasting problems such as predicting stock market and electricity consumption prediction. [5] (Sheng et al., 2018) tried to predict the Chinese stock market using CNN, especially used Conv1D function to handle 1-Dtime series data. They did the binary classification instead of regression, so the model returns binary values one or zero to show whether the stock price movement would go up or down. They evaluated the CNN model with different stock data and concluded that the CNN model is robust, and the result can be reliable even if the source data is 1-D sequential. [6] (Omer et al., 2018) converts 1-D financial time series into a 2-D image-like data representation to utilize the power of CNN to predict the stock price. They introduce CNN model to determine "Buy" and "Sell" points in stock price using 15 different technical indicators with different time intervals. The results show that CNN performs remarkably well even over long periods. They conclude that the proposed CNN based model outperformed MLP (Multi-Layer Perceptron), and even LSTM based model on short and long out-of-sample periods.

3.3. LSTM (Long Short-Term Memory) in time-series analysis

Recurrent Neural Network (RNN) models have feedback connections to the previous data, but the memory of past in RNN diminishes as time goes further.

LSTM is one of the RNN models and it is designed to overcome the problem of memory loss by having 'Forget Gate' in its architecture. LSTM is the most popular deep learning model used in time series prediction, especially in stock price and crude oil price prediction. [7] (Adil et al., 2020) tried to predict the 'Google' and 'Nike' stock price using LSTM and found that the proposed model based on LSTM can trace the opening prices for both assets. [8] (Kexian et al., 2022) tried to compare the performance of ARIMA, ANN and LSTM models on crude oil price. They built traditional time-series model, ARIMA (AutoRegressive Integrated Moving Average) model, and two of Neural Network based models, ANN (Artificial Neural Network) based model and LSTM based model. They compared the performance of the models and concluded that the LSTM model demonstrated higher forecasting accuracy and better forecasting stability for different timescales than ARIMA and ANN based models. [9] (Anita et al., 2020) used LSTM based model to predict four of Indian based companies' stock price. They focus on the importance of number of layers and hyper-parameters of LSTM model and compare the results. [10] (Qihang Ma, 2020) compared the ARIMA, ANN, and LSTM on stock price. The paper mentions that ARIMA and ANN have been widely used in time-series data forecasting, but these models cannot measure the continuity of the trends. However, due to its characteristics of feedback connection, LSTM makes it easier to find development trends through the back propagation of historical prices and the current prices. They conclude that the LSTM model performs better than ANN and ARIMA models. They assume the reason of the superiority of LSTM is the improvement of the LSTM model on the problem of vanishing gradient.

4 Method

4.1. Dataset

West Texas Intermediate (WTI) crude oil spot price is the response, and 27 potential factors classified into six broad groups. Our data spans the period from January 04, 2002 to September 25, 2015. There are 6 broad groups, and each group has several potential predictors.

4.1.1. Supply factors

Oil price heavily depends on supply and demand. We consider the following supply factors for our potential predictors:

Global crude oil production: This includes both OPEC and non-OPEC crude oil production.
Global crude oil export: This factor measures the potential capacity for production.
OPEC surplus crude oil production capacity: This factor is an indicator of general market supply conditions. Surplus production capacity can help mitigate the oil price fluctuation, especially due to the shortage of the supply, thus sta-

bilize the global market.
Crude oil inventory: The accumulation of crude oil stock gives the market a great flexibility in responding to short-term supply shortages. Both global and the U.S. crude oil closing stock are included.
U.S. refinery utilization rate: Refining rate plays an important role in determining crude oil prices because lower refinery utilization rate will lead to a preference for higher quality crude oil, putting upward pressure on prices.
Baltic exchange dirty and clean tanker index: Baltic exchange dirty tanker index indicates the cost of shipping unrefined petroleum oil, while clean tanker index indicates the cost of shipping refined products without heavy residual components.

4.1.2. Demand factors

The below predictors are considered under the demand category:

GDP (Gross Domestic Product): Global economic growth is closely related with the demand of oil. We consider the GDPs of the U.S., China, and Europe, which together account for more than 60Kilian index: This is an index of global real economic activity in industrial commodity markets.
Steel production: Steel production is a reliable indicator of global economic activity, and we use world steel production, as well as the production in the U.S., China and the Europe.
ISM manufacturing index: This index is a monthly indicator of U.S. economic activity based on a survey of purchasing managers at more than 300 manufacturing firms.
Global crude oil imports: Global crude oil imports reflects the state of the economy. There is a positive relationship between global crude oil imports and the price of the crude oil.

4.1.3. Financial market factors

The following three factors are considered under financial market factors:

U.S. interest rates: Crude oil prices have usually a negative relationship with interest rates. We consider the three-month treasury bill rate and the federal fund rate.
Exchange rate: We use the U.S dollar index, computed as the weighted geometric mean of the dollar's value relative to other major currencies.
Stock market: We use SP 500 index and MSCI world index. Stock market and oil prices are linearly related to each other.

4.1.4. Commodity market factors

Crude oil prices and other industrial commodities prices are correlated with each other. We consider two possible predictors in our forecasting models.

SP GSCI non-energy index: This is a composite index of commodities that mea-

sures the performance of the commodities market.
CRB raw industrial materials index: This measures the aggregated price direction of 22 sensitive basic commodities whose markets are believed to be sensitive to changes in economic conditions.

4.1.5. Speculative factors

Speculative factor is the act of conducting a financial transaction that has substantial risk of losing value but also holds the expectation of a significant gain or other major value. We use the ratio of trading volume of crude oil futures contracts to the oil production as a potential predictor of the speculative factor.

4.1.6. Geopolitical factor

We use the total number of terrorist attacks in the middle east Asia and north Africa where most of the OPEC countries are located in.

4.2. LASSO (Least Absolute Shrinkage and Selection Operator)

To find the most impactful factors among others, Lasso model was applied based on the results from [4] (Hong Miao et al., 2018). Lasso is a linear regression analysis technique that performs both feature selection and regularization to enhance the accuracy of the model. Lasso adds "absolute value of weights" term to the cost function as penalty. It not only overcomes overfitting issues, but it also helps feature selection by removing the features with zero slope which means less impactful features. The below is the formula of Lasso.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Figure 1: Lasso

4.3. CNN(Convolutional Neural Networks)

CNN is one of the artificial neural network models. Each node of the CNN model receives inputs only from a small neighborhood in previous layer and it also shares parameters with neighbor nodes. Due to this characteristics, Convolutional layers dramatically reduce number of model parameters. This overcomes the limitations of the Fully Connected Neural network models such as overfitting, and increased training time caused by excessively large number of parameters. It has shown that it outperforms other neural network models in the areas like computer vision, such as face recognition, image classification. For image processing works, Conv2D function is widely used to handle 2-D matrix input data, and the kernels and poolings are 2-D windows as below figure.

Since this project handles 1-D time series data, Conv1D function is used to

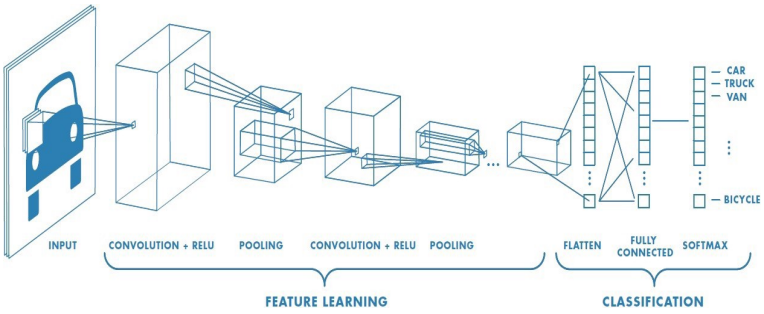

Figure 2: Structure of CNN - Conv2D

handle 1-d array input. Thus, 1-d kernel as well as 1-d pooling window are applied as below figure.

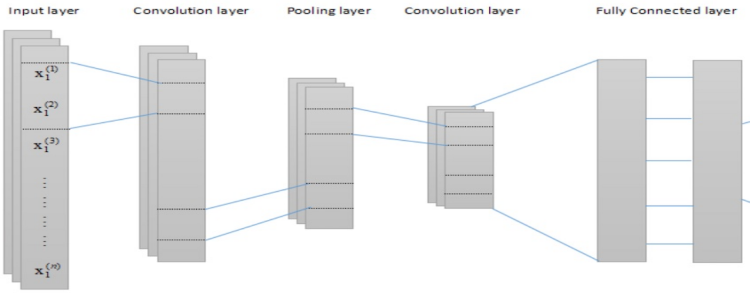

Figure 3: Sturcture of CNN - Conv1D

3.4. LSTM(Long Short Term Memory)

RNN is one of Neural Network where the output from the previous step is fed as an input to the current step, widely used for handling sequential data or time series data. However, the memory of past in RNN diminishes as the number of previous steps are increased. This problem is known as 'Vanishing Gradient Problem'. LSTM is one of the RNN models and it is designed to overcome this vanishing gradient problem by having a 'Forget Gate' in its architecture (Figure3). LSTM is among the most popular deep learning models used in time series prediction, especially in stock price and crude oil price prediction.

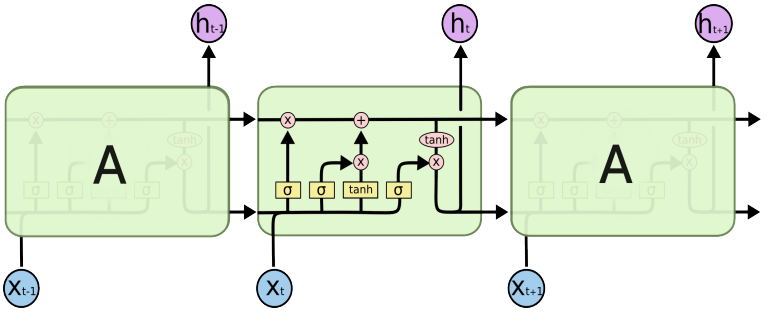

Figure 4: Structure of LSTM

5 Experimental Design

5.1. Data Preprocessing

The dataset has 716 samples with 51 columns. There are different intervals of the feature values. Some of the features have daily interval, while the other features have weekly, monthly, even quarterly interval. Thus, the interval of the dataset was set to weekly basis. For the features having daily values use the value of the last day of a week, and the features having monthly and quarterly were interpolated accordingly. 'Date', and 'STORAGE CAP' columns are dropped since they do not add any values to the model. To handle the time series dataset, 'MinMaxScaler' other than 'StandardScaler' is applied for data normalization. The dataset is split into training and test dataset by 70:30 respectively. To train the models, we need inputs and outputs of data, and since the dataset is converted to weekly dataset, and one calendar year has 52 weeks, the input of the dataset is values of previous 52 weeks, and the output is the value of the last week of a year. Thus, the input dataset has a shape of (449, 52, 1) and the output dataset has a shape of (449,1).

5.2. Find most impactful factors

To find the most impactful factors among others, Lasso method was applied. Lasso adds an additional penalty term to the cost function which keeps the estimated value of the regression coefficients small. LASSO helps find the most influential features by removing the features with zero slope. The following 6 features are found to be the most influential features in crude oil price prediction.

- World Steel Production
- Kilian Index
- ISM Manufacturing Index
- CRB Raw Materials Index

- Dollar Index
- Total amount of terrorist attack in the Middle East and North Africa

All these features are positively correlated to the WTI crude oil price.

5.3. Modeling

Various studies have shown that neural network models outperform traditional statistical time-series models in domains such as stock price prediction or crude oil price prediction. [8] (Kexian et al., 2022) found that the LSTM model demonstrated higher forecasting accuracy and better forecasting stability for different timescales than ARIMA and ANN based models. [10] (Qihang Ma, 2020) also found that the LSTM model outperforms ANN and ARIMA models in stock price prediction. CNN model and LSTM model are considered in this project.

The CNN model for this project has 3 convolutional layers with 52 filters and each CNN layer has kernel Size =3. After each convolutional layer, Maxpooling with pool size = 2 is followed. Since the input data is 1-D time series data, Conv1D function is applied, and kernel and pooling also have 1-D window. After three convolutional layers, 2 Fully Connected layers are followed with 52 nodes for the first layer, and 1 node for the second one. 0.2 Dropout is applied between FC layers. The summary of the CNN model is as follows:

Layer (type)	Output Shape	Param #
conv1d_51 (Conv1D)	(None, 50, 52)	208
max_pooling1d_35 (MaxPooling1D)	(None, 25, 52)	0
conv1d_52 (Conv1D)	(None, 23, 52)	8164
max_pooling1d_36 (MaxPooling1D)	(None, 11, 52)	0
conv1d_53 (Conv1D)	(None, 9, 52)	8164
flatten_8 (Flatten)	(None, 468)	0
dense_77 (Dense)	(None, 52)	24388
dropout_114 (Dropout)	(None, 52)	0
dense_78 (Dense)	(None, 1)	53

Total params: 40,977
Trainable params: 40,977
Non-trainable params: 0

Figure 5: CNN Model Summary

The LSTM model for this project has 4 LSTM layers with 52 units, and the input shape is set to 52 x 1 which is equal to the size of a single row. Between the LSTM layers, 0.2 Dropout is applied. After LSTM layers, 2 Fully connected layers are applied with 52 nodes for the first layer, and 1 node for the second layer. 0.2 Dropout is applied between FC layers. The summary of the LSTM

model is as follows:

Layer (type)	Output Shape	Param #
lstm_76 (LSTM)	(None, 52, 52)	11232
dropout_109 (Dropout)	(None, 52, 52)	0
lstm_77 (LSTM)	(None, 52, 52)	21840
dropout_110 (Dropout)	(None, 52, 52)	0
lstm_78 (LSTM)	(None, 52, 52)	21840
dropout_111 (Dropout)	(None, 52, 52)	0
lstm_79 (LSTM)	(None, 52)	21840
dropout_112 (Dropout)	(None, 52)	0
dense_75 (Dense)	(None, 52)	2756
dropout_113 (Dropout)	(None, 52)	0
dense_76 (Dense)	(None, 1)	53

Total params: 79,561
Trainable params: 79,561
Non-trainable params: 0

Figure 6: LSTM Model Summary

5.4. Comparison of the models

Fit WTI crude oil price training data on both CNN and LSTM models created in the section above and evaluate each model on test dataset. 100 epochs are applied for training and Early stop is also applied with the patience number 3. Then compare the performance of each model. The evaluation metric is validation loss values. Based on the evaluation metric, the better performing model which returns the lower validation loss value is selected. Using this selected model, we predict the WTI crude oil price on the test dataset, and the predicted values are stored for the further analysis.

5.5. Finding the relationship

The selected model is trained on each of the 6 most impactful features and return the predictions on the test dataset. These values are stored as well. As briefly mentioned in the introduction, one of the goals of this project is to find the relationship between WTI crude oil price and the most impactful features. The basic idea of finding the relationship is follows. If the model predicts the WTI crude oil price would go up, and the 6 key features also predict the same direction, then the prediction result would be reliable. However, if the model predicts the WTI crude oil price would go up, but the key features predict the other way, then this could be less reliable than the former case, and this could be a warning that the model might return a wrong prediction. Thus, with this idea, we can enhance the accuracy of the prediction, and lower the risk from the inaccurate predictions. For this work, binary dataset is created, which is that if this week value is higher than the last week value, it's 1, and 0 otherwise. Then find the index of the values that incorrectly predict and find the values of

6 features on the same index. Count the numbers of the case that each feature has different binary values with the binary value of the crude oil price. If this value is significantly larger than the numbers that each feature has the same binary values, then the models for the features can be used as an indicator to enhance the accuracy of the crude oil price prediction.

6 Experimental Results

6.1. Finding the best model

As described in the section 5. Experimental design, two neural network models, CNN and LSTM models are built and the validation loss of each model on the test dataset are compared to find the best model for WTI crude oil price prediction.

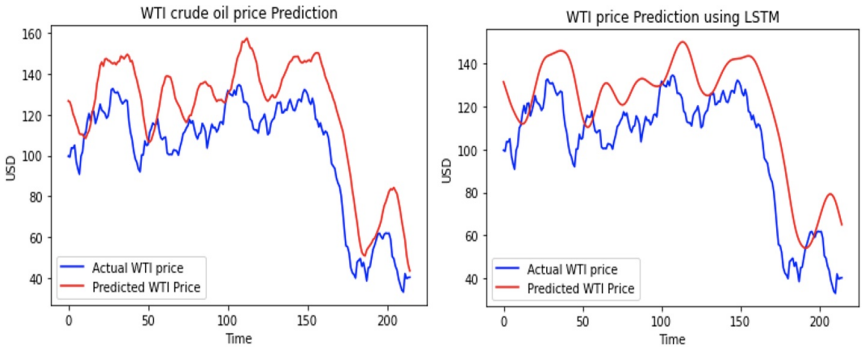

Figure 7: WTI crude oil prediction using CNN and LSTM

LSTM model returned 0.0148 as the validation loss, while CNN model returned 0.0155 for the same. These models trained several times, yet most of the times, LSTM model returns smaller validation loss over CNN model. As per the result, we select LSTM model as the best performing model on the given WTI crude oil price dataset.

6.2. Find the relationship

The same model is applied to WTI crude oil price as well as each of the 6 major factors to compute the prediction. As explained in the section 5. Experimental design, the binary table is created using the predicted values, if the value is higher than the previous week's store 1, and if not, store 0. The LSTM model

made prediction errors 84 times out of 215. The below table shows that the number of times that whether each feature predicts the same direction with the WTI prediction or each feature predicts the opposite direction for those 84 times.

Binary Values	1 (Opposite direction)	0 (Same direction)	Percentages (%)
World Steel	45	39	54%
Kilian Index	47	37	56%
ISM index	55	29	65%
CRB Rind Index	53	31	63%
Dollar Index	51	33	61%
Terrors	54	30	64%

Figure 8: Table of binary values

The table shows that when crude oil price prediction model made wrong predictions, more than half of the times, features predict the opposite direction. This means that when WTI crude oil price model made errors, the features' models warn that the prediction could be incorrect by returning the binary value 1.

6.3. Limitations

The binary numbers returned from the predictions are barely larger than the half (42). This number is smaller than expected and is not sufficient to claim that the features models and their prediction results can be used as indicators to enhance the accuracy of the predictions. One of the potential reasons for this limitation is that the dataset is too small for neural network modeling. The data has only 716 samples, and that is divided into training and test datasets, thus the LSTM model is trained and built on 449 samples. This is not even close to the enough number of samples for neural network modeling. I think with larger dataset with more complicated LSTM model, higher percentages of the indicators are expected. Adding more datasets is one of the top priorities for the next research. For the next research, with the much larger dataset, I'd like to use 2-d image like CNN model as done in [6] (Omer et al., 2018), and compare the performance with LSTM based model.

7 Conclusion

In this project we build a neural network model to predict the crude oil price. For accurate forecast, this project focuses on two aspects: fundamental analysis and technical analysis. For technical analysis, neural network models are compared. The result shows that LSTM model outperforms 1-D CNN model. For fundamental analysis, this project finds the most impactful factors that determine the

oil: 'World Steel Production', 'Kilian Index', 'ISM index', 'CRB Index', 'Dollar Index', 'Number of terrors in Middle East and Northern Africa' are turned to be the 6 most impactful factors in crude oil price. After the model is selected, the predicted values are converted to binary values to find the relationship between the prediction of crude oil price and these features by analyzing the movement of the values. The result shows that the prediction values of these major factors have similar trend with the crude oil price prediction. Thus, LSTM models for the features can be used as indicators to enhance the accuracy of crude oil prediction.

As briefly mentioned in the section 6. Experimental Results, the binary numbers returned from the predictions are barely over the half. This number is not sufficient to claim that the features models and their prediction results can be used as indicators to enhance the accuracy of the predictions. Thus, there remains a question that whether these feature prediction values can be used to predict the crude oil price. In terms of accountability, this model has apparent limitations. However, these indicators are just additional supplements. Regardless of the performance of the models of these factors, the crude oil price prediction model stays the same and returns the same predicted values. In addition, even the crude oil price model is just an indicator to see the future trend and can't be directly used for serious investment decisions, yet. Hence, there is no foreseeable ethical issues with this study.

8 Bibliography

[1] Rayan H. Assaad, Sara Fayek, "2. Predicting the Price of Crude Oil and its Fluctuations Using Computational Econometrics: Deep Learning, LSTM, and Convolutional Neural Networks", DOI: 10.2478/erfin-2021-0006

[2] M. Khalid, S. Sarwar, R. Waheed, and M. Amir, "Role of Energy on Economy The Case of Micro to Macro Level Analysis," Econ. Bull., vol. 38, no. 4, pp. 1905–1926, 2018. .

[3] EIA(U.S. Energy Information Administration), Oil and Petroleum products explained – Oil price and outlook

[4] H. Miao, S. Ramchander, T. Wang, and D. Yang, "Influential factors in crude oil price forecasting," Energy Econ., vol. 68, pp. 77-88, 2017, doi: 10.1016/j.eneco.2017.09.010.

[5] Sheng Chen, and Hongxiang He, "Stock Prediction Using Convolutional Neural Network", doi:10.1088/1757-899X/435/1/012026

[6] Omer Berat Sezer, Ahmet Murat Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion ap-

proach", https://doi.org/10.1016/j.asoc.2018.04.024

[7] Adil MOGHAR, Mhamed HAMICHE, "Stock Market Prediction using LSTM Recurrent Neural Network", DOI:1016/j.procs.2020.03.049

[8] Kexian Zhang and Min Hong, "Forecasting crude oil price using LSTM neural networks", DOI: 10.3934/DSFE.2022008

[9] Anita Yadav, C K Jha, Aditi Sharan, "Optimizing LSTM for time series prediction in India stock market", DOI: 10.1016/j.procs.2020.03.257

[10] Qihang Ma, "Comparison of ARIMA, ANN and LSTM for Stock Price Prediction", https://doi.org/10.1051/e3sconf/202021801026