LITERATURE SURVEY ON CRUDE OIL PRICE PREDICTION:

MEMBERS:

MUTHU KUMARAN S U(922119106060) SSMIET,DINDIGUL.(TL)

MUNIYAPPAN P(922119106058) SSMIET,DINDIGUL.

MUTHU KUMAR S(922119106059) SSMIET,DINDIGUL.

MUTHUSURYA R(922119106061) SSMIET,DINDIGUL.

NIZAMDEEN A(922119106068) SSMIET,DINDIGUL.

ABSTRACT:

The literature on forecasting the ‘‘black gold’’ price is vast. This paper provides a literature review on the various techniques that have been used to forecast crude oil price. We mainly focused on the researches that have utilized artificial neural network models in their forecasting study. Therefore, a detail description of this model is presented in this paper.

INTRODUCTION:

Crude oil plays a significant role in the global economy, for nearly one-third of the world’s energy consumption comes from it. Also, oscillations in oil prices significantly affect the economy of oil-exporting and oil-importing nations . Accurate oil price forecasting would help policymakers adopt proper policy and make appropriate decisions regarding energy resources. However, crude oil price prediction has been a challenging problem in forecasting research because oil prices are affected by many factors. Except for the fundamental market factors, such as supply, demand, and inventory, oil price fluctuation is strongly influenced by economic development, conflicts, wars, and breaking news . For example, oil producers were paying buyers to take the commodity off their hands over fears that storage capacity could run out in May 2020, and WTI oil price even turned negative for the first time in history on 20 April 2020. Another recent example is that crude oil price movements have exhibited a stronger correlation with the severe degree of the COVID-19 pandemic . The challenge is characterizing and modeling such nonlinear and nonquantitative factors because most of such information is contained in raw texts.

Overall, existing research on crude oil price forecasting can be categorized into three main classes: econometric models (including time-series models), machine learning or deep learning

methods, and hybrid approaches. Among time series approaches, Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional (GARCH), Empirical Mode Decomposition (EMD), and Complete Ensemble Empirical Mode Decomposition (CEEMD) are primarily used . For example, uses CEEMD to decompose the original oil prices into five nonlinear and three volatile components (IMFs). Then the author uses MS-GARCH to model and forecast volatile components and SVM-PSO (Support Vector Machine - Particle Swarm Optimization) to model and predict nonlinear components, respectively. Lastly, the linear addition of these forecasts gives a more reliable estimation of the oil price. Econometric models, especially structural models, focus on how specific economic factors and the behaviors of economic agents affect the future values of crude oil prices . For example, proposes a Self-Exciting Threshold Auto-regressive-SETAR model dealing with structural breaks in oil price longitudinal data, treatment, and forecasting.

Support vector machines (SVMs) and neural networks (NNs) are the most typical machine learning methods due to their extraordinary ability in modeling nonlinearity and volatility . However, shallow architectures are insufficient to model complex patterns with numerous factors . Recently, deep learning (DL) has become a mainstream approach in various fields. DL approaches explore complicated structures and patterns in large data sets using the backpropagation algorithm, which indicates how a machine should change its internal parameters . Deep learning can represent highly nonlinear and highly varying functions; thus, DL-based approaches have also been widely used in oil price forecasting .

Hybrid methods integrate the methods mentioned above and thus utilize their advantages synthetically. Reference assemble time series decomposing models and AI models to enhance the forecasting performance. The reason why hybrid methods usually

achieve better results than single models lies in two aspects. On the one hand, time-series approaches or econometric models specialize in capturing the linearity and volatility in price time series. On the other hand, AI models specialize in nonlinear and non-stationary characteristics.

This section reviews existing typical literature on crude oil price forecasting and summarizes the literature in Table 1. More details about the leading technologies used in the proposed framework AGESL are followed.

Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional (GARCH) are two mainstream models in time series predicting. ARIMA was introduced by Box and Jenkins (1976), which is a linear combination of past values and past residuals. Bollerslev (1986) presented GARCH, which is a generalized form of the Autoregressive Conditional Heteroscedastic (ARCH) model initiated by Engle (1982) [8]. The main difference is that ARIMA forecasts future values from past values and past residuals

Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional (GARCH) are two mainstream models in time series predicting. ARIMA was introduced by Box and Jenkins (1976), which is a linear combination of past values and past residuals. Bollerslev (1986) presented GARCH, which is a generalized form of the Autoregressive Conditional Heteroscedastic (ARCH) model initiated by Engle (1982) [8]. The main difference is that ARIMA forecasts future values from past values and past residuals. Whereas GARCH focuses on the time-varying variance of residuals, also being called time-varying volatility. They can be separately used to model the time series or integrated to enhance the prediction performance.

Literature	Features	Models
[24] (2006)	Price	SVM
[13] (2008)	Price	EMD and Neural Network
[8] (2014)	Price	ARIMA, GARCH, SVM
[2] (2016)	Price	Grid-GA-based Least squares support vector regression
[19] (2017)	Price and news text sentiment	SVM, Decision Tree, and Back Propagation Neural Networks (BPNN)
[25] (2017)	Price	Stacked denoising autoencoders and bagging
[26] (2018)	Price, sentiment,and text statitics	ARIMAX
[9] (2019)	Price	CEEMD, ARIMA, and sparse Bayesian learning
[10] (2019)	Price	EEMD and Long Short Term Memory (LSTM)
[20] (2019)	Price and web text sentiment	Ridge regression, LASSO, support vector regression (SVR), BPNN, and random forest (RF)
[27] (2019)	Price and news topics	A two-layer NMF (non-negative matrix factorization) model and GIHS (giant information history simulation)
[28] (2019)	Price, sentiment, topics and market data	Convolutional Neural Network (CNN), Latent Dirichlet Allocation (LDA), Random Forest (RF), SVR and Linear Regression
[1] (2020)	Price	CEEMD, SVM, PSO, and MS-GARCH
[18] (2020)	Price	WPD (wavelet packet de-noise)), EMD, and GARCH-M models
[29] (2020)	Price and news sentiment	Vector auto regression and Kalman (VAR) filtering framework
[21] (2021)	Price, news topic, and sentiment	LDA, SeaNMF, and AdaBoost RT
[3] (2021)	Price, news text, and Google trends	CNN, BPNN, MLR (multiple linear regression), SVM, and GRU (gated recurrent unit)
[22] (2021)	Price, oil market data, and news headlines	CNN, BPNN, SVM, RNN, and LSTM

$$ct=c+\sum i=1p\alpha ict-i+\varepsilon t+\sum j=1q\theta j\varepsilon t-j \quad \ldots\ldots\ldots\ldots (1)$$

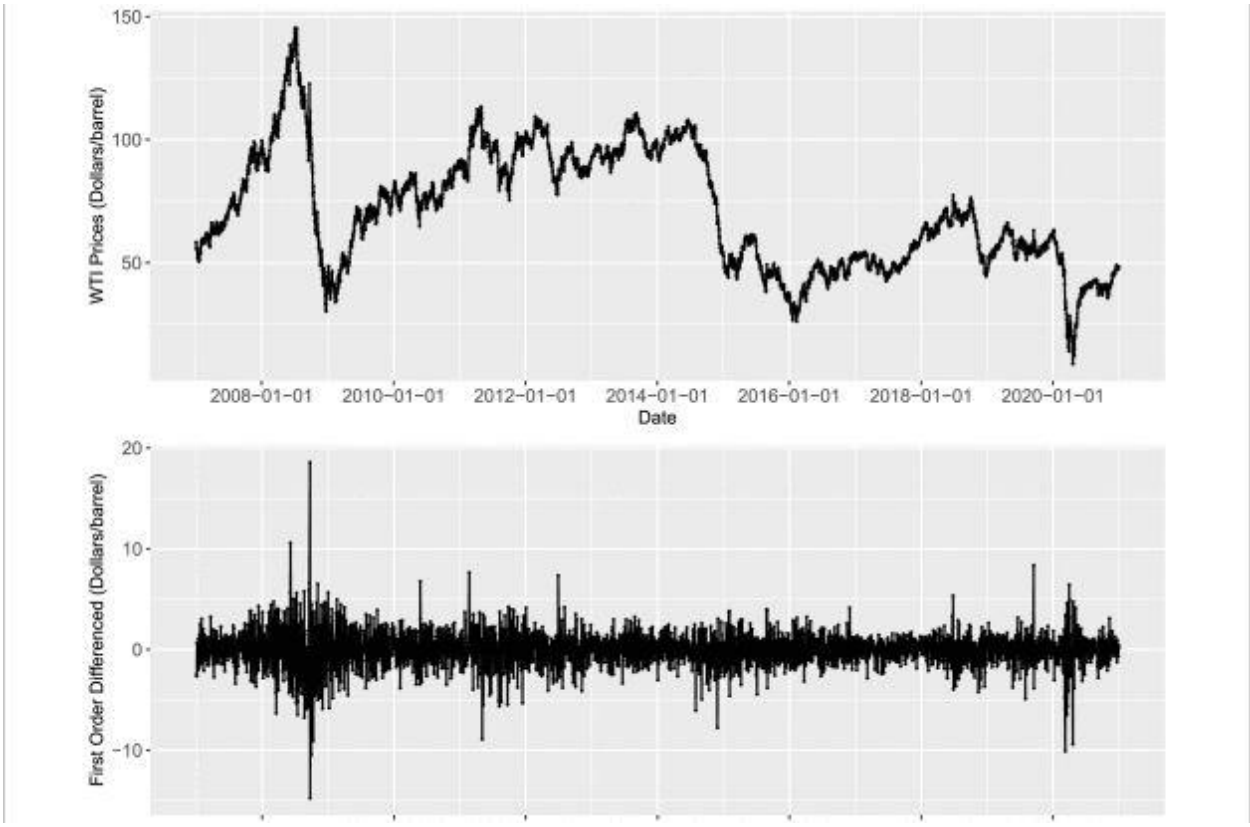
$$at=\sigma t2=ut\sim\sigma tut, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$\alpha 0+\sum i=1m\alpha ia2t-i+\sum j=1s\beta j\sigma 2t-j, \quad \ldots\ldots\ldots\ldots (3)$$

$$IID(mean=0,variance=1), \quad \ldots\ldots\ldots\ldots\ldots (4)$$

West Texas Intermediate (WTI) is one of the most critical global crude oil price indexes and can reflect global crude oil price movements in time. In this study, WTI daily prices are

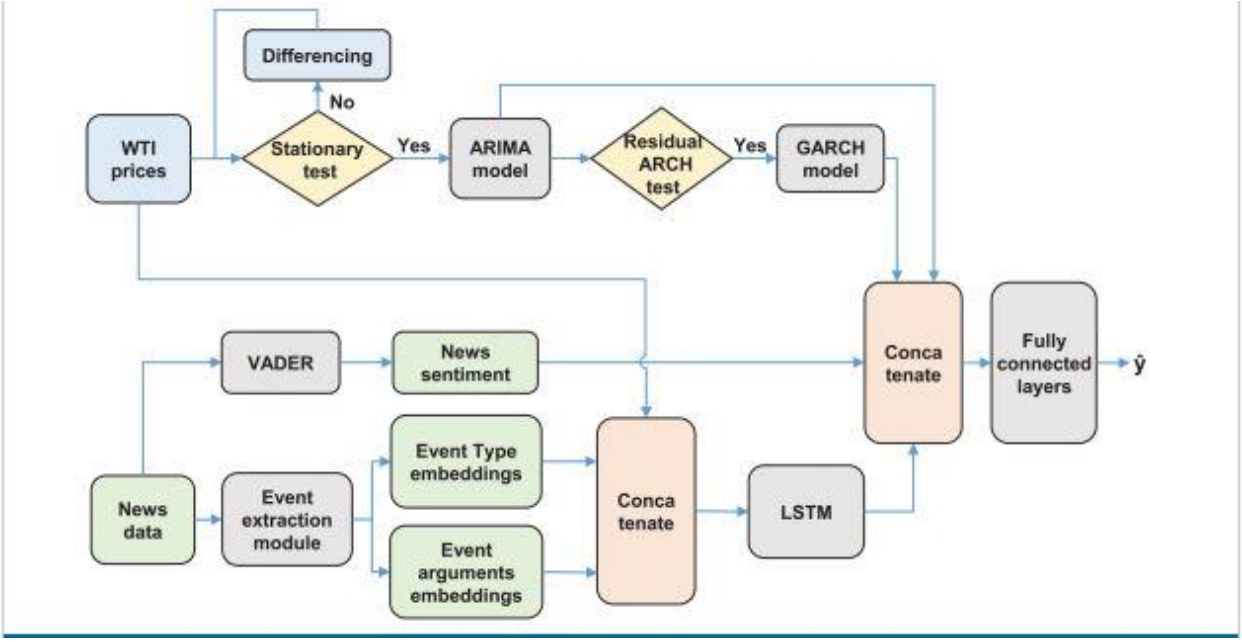
investigated. There are a total of 3522 observations ranging from January 2007 to December 2020. The original WTI series and its first difference series are illustrated



Our Approach

This section introduces our crude oil price prediction framework AGESL, as illustrated in Fig. 4. Overall, AGESL contains five main modules: Mean price prediction (ARIMA model), Volatility prediction (GARCH model), Event extraction module, Sentiment analysis module, and LSTM module. Four kinds of features are used as input of AGESL, involving two data sources: historical prices and news text. Three features, historical prices excluded: sentiment scores, news types, and news arguments are extracted from news text. Mean price prediction and Volatility prediction modules are responsible for predicting future mean prices and fluctuation, mainly depending on historical price data or its varieties. The event extraction module contains a neural latent

variable network and Bayesian inference model, with the responsibility to extract latent event type embeddings and event arguments. We concatenate event type embeddings, event arguments embeddings, and historical prices as input of the LSTM module. Then we concatenate the output of the ARIMA model, GARCH model, LSTM model, along with the formatted news sentiment sequence together as the input of the last fully connected layers. The last fully connected layers integrate and weigh the individual predictive model outputs, affording a synthetical prediction. The modules of AGESL are described step by step as follows.



Mean Price Prediction:

ARIMA module is responsible for predicting WTI future mean prices. Five steps are needed to fulfill this task. First, the Augmented Dicky-Fuller (ADF) test should be carried out to check the stationarity of the WTI daily price series. Second, differencing operation is used to generate a stationary time series. Third, the Autocorrelation function (ACF), Partial Autocorrelation function (PACF), and Information Criterion (AICC, AIC, and BIC) are incorporated to determine the AR and MA orders. Fourth, the

Box-Pierce test is performed to check whether the residuals are independent white-noise, and if not, model parameters are readjusted. Lastly, the fitted model is used to forecast the future price rollingly.

Conclusion

Crude oil prices may be affected by some factors that are hard to quantify, such as political events, regional conflicts, and policies of oil-exporting countries, which are frequently reported in online news. In this paper, we propose a new hybrid framework, AGESL, for predicting crude oil prices whose focus is on capturing these features to enhance the prediction accuracy. For this purpose, we utilize an open-domain Event Extraction algorithm and other NLP technologies to extract news events (e.g., latent event types and news arguments) and news sentiment from news text. The proposed AGESL integrates the strength of linear and nonlinear modules. The proposed AGESL outperforms the other benchmark models in the empirical study, including the single-channel input models (ARIMA, SVM, LSTM) and multi-channel input models (LSTM-Sent, ARIMA-GARCH-Sent, LSTM-Event).

Despite the promising capability of the proposed AGESL framework, it may be further developed from the following perspectives. First, from the export-import view, the world's major economies consume the vast majority of crude oil, so more economic or financial exogenous variables, such as economic growth, industrial added value, or financial indices for these economies need to be considered into the framework. Second, from the view of practical applications, market risk should be taken into account to minimize a market trader's potential loss. For example, the VaR method could help the proposed AGESL report a forecast value and the corresponding risk at a certain confidence level.

REFERENCE:

https://www.researchgate.net/publication/278032369_Forecasting_Crude_Oil_Price_Using_Artificial_Neural_Networks_A_Literature_Survey

https://ieeexplore.ieee.org/document/9599721/references