International Journal of English Literature and Social Sciences
Vol-6, Issue-4; Jul-Aug, 2021

Journal Home Page Available: https://ijels.com/
Journal DOI: 10.22161/ijels

Peer-Reviewed Journal

Study Oil Price with Market Sentiments:A Literature Review

Yan Han

School of International Trade and Economics, Central University of Finance and Economics, Beijing, China

Received: 12 Jun 2021; Received in revised form: 10 Jul 2021; Accepted: 18 Jul 2021; Available online: 27 Jul 2021

Abstract— *Oil price shows its strong volatility starting from new millennium. However, traditional oil price researches mainly focus on fundamental factors, while omitting the role market sentiments play in shifting oil price. In this paper, we point out the importance of including sentiments in oil price analysis. Most important, we introduce advanced machine learning methods to quantify market sentiment and lead to new direction in oil price research.*

Keywords— ***Market sentiments;Oil price; Machine learning.***

I. INTRODUCTION

Oil is the crown jewel of commodities that is used in a multitude of ways in our lives. World transportation systems need oil to provide energy for vehicle to move, chemical plants require crude oil as raw material to produce base chemicals for industrial use, and even the important ingredients of cosmetics used by women come from crude oil. Particularly in China,as a barren natural resource country, Chinese oil demand keeps increasing with the rapid development of Chinese economy.

However, oil price shows strong volatility, though it's essential to the economy. The volatility becomes more clear after 2000. Three episodes draws our attention, as it shows in Brent price movement in Fig.1. The first period is from 2002 to 2008, when the world economy boomed and the oil price increased from 25 dollar per barrel to 140 dollar per barrel peak price in 2008. There is widespread agreement that this price surge was not caused by oil supply disruptions, but by a series of individually small increases in the demand for crude oil over the course of several years. Kilian (2008), Hamilton (2009), and Kilian and Hicks (2013), among others, have made the case that these demand shifts were associated with an unexpected expansion of the global economy and driven by strong additional demand for oil from emerging Asia in particular. Following a long period of relative price stability, between June 2014 and January 2015 the Brent price of oil fell from 112 dollar to 47 dollar per barrel, providing yet another example of a sharp decline in the price of oil. Baumeister and Kilian (2015) provide the quantitative analysis of the 49 dollar per barrel drop in the Brent price between June and December 2014. They conclude that about $11 of this decline was associated with a decline in global real economic activity that was predictable as of June 2014 and reflected in other industrial commodity prices as well. Finally, the oil price presents violent fluctuation starting from early 2020, when COVID-19 was spreading around

the world. Brent price fell to historical low level in the beginning of 2020 to 10 dollar per barrel, but it steadily rose to 70 dollar per barrel in the second half of 2020, when the world economy begun to recover.Oil price volatility has a negative and significant effect on economy, depressing investment, consumption of durable commodity and aggregate output(John 2010). Look at the violent fluctuation of Brent price after new millennium. Hardly can we imagine such volatility is purely driven by oil market fundamentals. An conspicuous example is that US WTI price fell to negative value in early 2020, and it's obvious that such price volatility is driven by abnormal market sentiment but not the market fundamental at that time. Xiong and Yan(2009),Singleton(2014) and Qadan and Nama(2018) point out that there's a lack of behavior factor in traditional oil price research, while the sentiment of oil traders increasingly plays an important role in oil price movement, because global oil market exhibits several distinct features in contrast to the past.

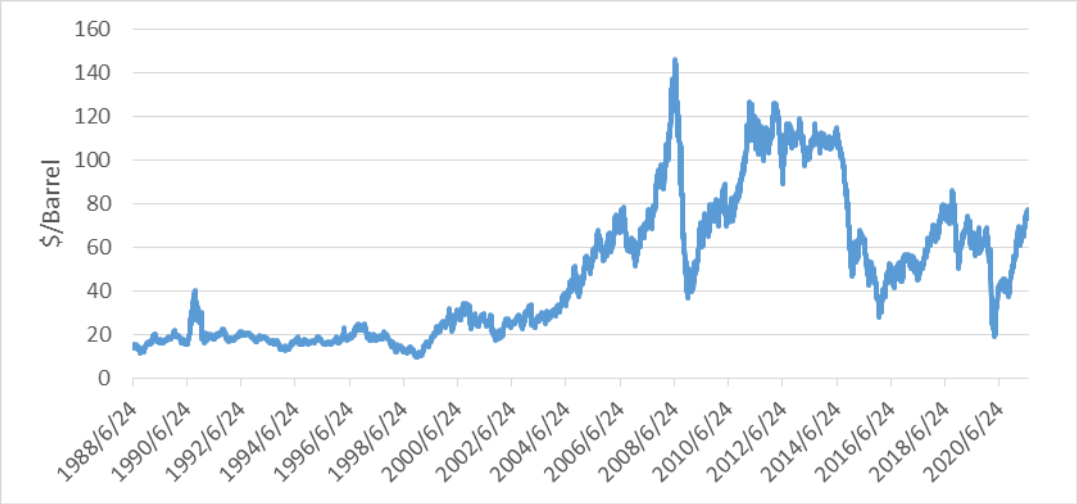

Fig.1: International Brent Price

First,information availability and transition improved greatly in the past 20 years. Internet's swift development caused the humanity to enter into the information age, hence a oil trader can easily get the latest news even if it happens thousands of miles away. Some market intermediaries further enhance the oil market players' access to market dynamics, as they try to gather information and present it in the information platform for traders' reference. For example, advisory companies such as Platts,Bloomberg and Thomas Reuters have analysts around the world, provide real-time update of global market fundamentals news,making traders closely follow up with oil market dynamics. Besides, companies like Wood Mackenzie and IHS make deep insights and professionally forecasting price trend through data analysis.On the other hand,convenient communication tools greatly favor information dispersion.Traders can easily share information via email or telephone calls,and the market become more transparent in consequence. Advanced technology also enables private information available. The application of satellite to supervise oil storage is on example. In this case, satellite can photo the storage tank and measure the liquid level by means of thermal imaging technology. Suppose the the information spread to the market, oil price will be impacted.

Second,numerous participants in the oil market jointly push oil price to go up and down.In the past time as we know,basically in the marketplace are end-users and oil producers,with oil producers supply crude oil to end-users for their own use.Typically the suppliers are countries from OPEC.However,things change dramatically since 2000.In terms of market fundamental,shale revolution make the previous largest oil import country-US,to become a net oil exporter.As a result,the monopoly power of OPEC decreases with diversification of oil supplier.What's more,with the development of global financial market,oil becomes an important asset for investors to diversify

risk.Large investment banks like Goldman Sachs and JP Morgan have their commodity trading department focusing on speculation and hedging business.And we can see significant cash flow from financial firms has enrolled into oil futures market since 2008, indicating that investment banks gradually resort to commodity assets to avoid risk.

Both integration and diversification render sentiments more importance in oil pricing.First of all, information spreads fast so that even a slight disturbance could lead to large price swings as people overact.This phenomenon is analogous to the term once proposed by Maynard Keynes,"animal spirit".The phenomenon of price drift is also likely to occur due to sentiments effect,as surging bullish mood may continuously lead price leaning to the same direction, presenting a tendency price curve.However,most of the literatures to date focus on the demand and supply fundamentals by which the researchers typically set up a SVAR model,overlooking the role that psychology could play in oil trades.

In this article, we review traditional oil market literature and present how sentiments can act in oil price as shown in recent research. Then we list the cutting edge method that can be used to quantify oil market sentiment, which direct the way for future oil price research.

The rest of the paper is structured as follows. Section 2 reviews traditional oil price research methods. Section 3 stresses the drawback if omitting sentiment factor in oil price study. Section 4 introduces important and novel methods that are useful in quantifying oil market sentiment. Section 5 presents conclusion.

II. TRADITIONAL OIL PRICE RESEARCH

Kilian (2009) is the pioneer article that decompose oil shocks to quantify the structural factors that impact oil trading price. Specifically,Kilian (2009) proposes three components that lead to oil price fluctuation:oil supply shocks;shocks of global business cycle;and specific demand shocks from oil market,denoting precautionary demand from crude oil buyers.Though global business cycle is hard to measure,Kilian (2009) take advantage of the Baltic Shipping Index to simulate economy boom and bust.It's also the first ever research that concludes global business cycle is the main factor causing oil price to change,whereas supply shocks only account for a small percentage in oil price fluctuations.The fact that oil price is largelychanged by global aggregate demand explains why oil price surge between 2003 to 2008 did not end with a recession in global economy.

Since Kilian(2009),a large amount of researches have used SVAR methodology to analyze oil price. Kilian et al.(2014) devise a structural oil price forecast model including inventory as proxy for speculative demand.He decomposes oil shocks to flow demand,flow supply and speculative demand.The conclusion differs slightly from Kilian(2009),in which flow supply play a larger role in explaining oil price movement,occupying the capacity of speculative demand shocks.However,global business cycle still explains most of fluctuations in the history of oil price fluctuations.Based on canonical research framework,Aastveit et al.(2015) further disentangle global demand into demand from advanced and developing economy,by which he can make comparison about disparate contributions to oil price by two economies.In order to quantify economic power,they use Industrial Production as the indicator of business cycle.He concludes in large part emerging market lead the oil price increase from 2003 to 2008.Macroecomics variables like interest rates and exchange rates can also impact oil price.When US dollar depreciates,dollar-based crude oil price falls which means oil is cheaper.Kian and Zhou(2019) integrate interest rate and exchange rate to the analysis model by sign restriction methodology.The research for the first time show US exchange rate has significant effect on oil price movement,whereas interest rates only occasionally play a role in oil pricing.

Researchers relies on distinct methodology in extending the classical Kian(2009) structural analysis.In terms of global business cycle,Hamilton(2019) propose to use industrial production as proxy for global business cycle.Through close scrutinization of properties shipping index suggested by Kilian(2009),Hamilton concludes IP is a more credible measurement of global business cycle.In the research of oil elasticities,Caldara et al.(2019) use metal prices to substitute for global aggregating demand,as Bernanke(2016) asserts commodity prices can be the indicator of economic development. Caldara et al.(2019) compares the effect of Kilian's shipping index,industrial

production and metal prices on oil price movement,drawing the conclusion metal prices is outstanding of the three in predicting oil price.

From the early 2000s,oil price has showed its momentum in price swings.Lots of literatures try to analyze the long-term price movement assuming speculative demand could take effect.Kilian and Murphy builds up a SVAR model with inventory to proxy precautionary demand.However,their discovery is basically alike to Kilian(2009),global aggregate demand still accounts most in explaining oil price movement while speculative demand only plays a slight role.In comparison,Juvenal and Petrella(2014) estimate a DSGE model but with the conclusion speculative demand greatly affect the price movement between 2003 to 2008.Hamilton(2009) sets up a oil hedging model,concluding oil speculation to some extent support oil price movement.Smith(2009) lends no support to evidence that speculation is the driving force of oil price movement,because inventory did not change from 2003 to 2008.

Structural VAR methodology evolves in the process of oil price analysis.Kilian(2009) first applies exclusion restriction to illustrate the impact of oil shocks.Since then,Baumeister and Peerman(2012) and Peerman and Van Robays(2009) rely on sign restriction method to quantify demand and supply shocks.Kilian and Murphy(2012) identify oil shocks through a augmented sigh restriction approach.That is,they implement additional empirical bounds into the conventional sign restriction model.Caldara et al.(2019) propose that traditional application of oil elasticity is not acceptable because demand elasticity and supply elasticity are jointly determined. To truly identify the different oil shocks,they minimize the Euclidean distance of estimated elasticity value to empirical results.Baumeister and Hamilton(2019) make further progress in SVAR method.They relax strong parametric assumption as proposed by previous research by introducing uncertainty in the model,concluding supply shock turns out to be the most important factor in driving oil movement.

III. IMPORTANCE IN QUANTIFYING MARKET SENTIMENT

Owing to more integrity of global oil market,beliefs of market participants can quickly reflect in oil trading price fluctuations.However,previous researches mainly attribute oil shocks to fundamental factors,like demand and supply variations,while paying little attention to the role of sentiments in determining oil price.Even though market structure did not change significantly since 2000,oil price demonstrates larger volatility,as Brent price climbed to more than 140 dollar/bbl in 2008 but slump to slightly above 20 dollar/bbl in early 2016.What's the matter?Obviously market sentiments should be responsible for capricious price movement.Singleton(2014) points out,absenting from characterizing market player's sentiments,the result of traditional SVAR analysis could be misleading.For one thing,information friction make market participants hold different market views,so that they may based on their judgement to do speculation business,as Xiong and Yan(2009) demonstratesIn addition,"animal spirits" cast light upon barbarous movement of oil price.According to Banerjee (2009),price drift phenomenon is likely to appear due to market sentiments fluctuations.Angeletos(2013 2018) proposes that we should emphasize on the sentiments impact on business cycle. Further, Qadan and Nama(2018) provide support for sentiment drivers for oil price volatility. They use BW sentiment index of Baker and Wurler(2006) ,EPU index of Baker et al.(2016) and other 7 indicators representing market sentiment, challenging the traditional view that investor sentiment is irrelevant with oil price movement. They find market sentiment impacts both oil return and volatility. Through wavelet approach, Yang(2019) investigates causality and connectedness between economic policy uncertainty and oil price shocks across time scales. He concludes that crude oil price behaves as receivers of information from economic policy uncertainty, and the connectedness intensifies when time scales increase. Thus, it may cause omit variable problem when we fail to consider sentiments in oil price research.

IV. SENTIMENT QUANTIFYING METHODOLOGY

The reason why market sentiments are excluded from traditional oil price analysis framework is understandable.Sentiments is hard to quantify as we cannot observe it.Things change currently thanks to the fast

development of computer technology.Machine learning skills like penalized model and LDA method contributing to numerous textual analysis,meanwhile hardwarecreation such as GPU make high-dimensional calculation a reality.Nontheless little advancement with regard to machine learning has come in oil market analysis.Utilizing massive text from market information providers,I can move textual analysis prevalent in IT field to oil price analysis.The commonly used machine learning methods are listed as below.

1.Dictionary-based method

Dictionary-based method doesn't relate to statistical inference, which mainly constructs $y_i = f(x_i)$,where y_i is the outcome we're interested in and x_i is the text independent variable. The earliest practitioner that use dictionary-based method in economic research is Tetlock(2007).This paper use Harvard-IV vocabulary to calculate the sentiments by Wall Street Journal,then make a principal component analysis to accumulate the sentiments words in each article to form a emotion score. However,the flaw of Tetlock(2007) is each term included in Harvard-IV dictionary is equally weighted.He concludes that bullish sentiment give support to price,while pessimism depresses market price movement.Loughran and Mcdonald(2011) cast doubt on the effectiveness of Harvard-IV dictionary.Because this sort of dictionary is suitable to categorize psychology,thus this may be biased if used in financial analysis.By manually examining the words in 10-K files,the authors create a sentiment dictionary suiting to financial market.In addition,they modify the weighting scheme of Tetlock(2007) based on TF-IDF method.

The most influential economic research to date relates to machine learning should be Baker, Bloom and Davis(2016), which is a typical example of dictionary-based method application. Economic policy uncertainty has the potential to increase risk in economy, depressing investment and other economic activity. The authors use text from news outlets to provide a high-frequency measure of EPU and then estimate its economic effects, the process to create EPU index is as follows. Baker, Bloom, and Davis (2016) define the unit of observation i to be a country– month. The outcome y_i of interest is the true level of economic policy uncertainty. The authors apply a dictionary method to produce estimates y_i based on digital archives of ten leading newspapers in the United States. An element of the input data is a count of the number of articles in newspaper containing at least one keyword from each of three categories defined by hand: one related to the economy, a second related to policy, and a third related to uncertainty. The raw counts are scaled by the total number of articles in the corresponding newspaper–month and normalized to have standard deviation one. The predicted value y_i is then defined to be a simple average of these scaled counts across newspapers.

Hassan et al. (2020) measure political risk at the firm level by analyzing quarterly earnings call transcripts. Their measure captures the frequency with which policy-oriented language and "risk" synonyms co-occur in a transcript. Firms with high levels of political risk actively hedge these risks by lobbying more intensively and donating more to politicians. When a firm's political risk rises, it tends to retrench hiring and investment, consistent with the findings of Baker, Bloom, and Davis (2016) at the aggregate level. Their findings indicate that political shocks are an important source of idiosyncratic firm-level risk.

2.Generative language models

Generative model reverse the data generating process of traditional econometric models $\mathrm{p}(y_i|x_i)$, as it attributes the occurrence of text words to the outcome we're interested in, or $\mathrm{p}(x_i|y_i)$. This makes sense. For example, the oil market sentiment is not induced by text words in oil market reports; rather, it's the sentiment of analysts lead to occurrence of market text. Generative models can be separated to supervised and unsupervised models based on availability of outcome y_i.

2.1Unsupervised generative models

In terms of unsupervised generative models, as we cannot observe attributes y_i, it's necessary for us to construct a structure for relationship between y_i and independent variables x_i. Topic model is a popular structure form, in which y_i is regarded as the latent variable.

A typical generation model implies that each observation x_i is conditionally independently extracted from a possible token vocabulary based on a document-specific token probability vector, such as $q_i = [q_{i1}, \dots q_{ip}]'$. According to the length of the document,

$m_i = \sum_i x_{ij}$, which means the multinomial distribution of the count

$$x_i \sim mn(q_i, m_i) \quad (1)$$

This multinomial model is a basic form for application of generative model. Under the basic form of generative model, the function $q_i = q(y_i)$ builds the structure of distribution of text counts. Blei, Ng, and Jordan (2003) introduce topic model, which now is widely used in the generative setting, where

$$q_i = \theta_1 v_{i1} + \ldots \theta_k v_{ik} \quad (2)$$

Topic modeling has become very popular since the introduction of text analysis. (See a high-level overview of BLEI 2012.) This model is particularly useful in political science (e.g. Grimmer 2010), where researchers have successfully linked political issues and beliefs to estimated latent themes.Bandiera et al.(2019) use a LDA model to examine CEO behavior and firm performance.The authors records activities of many company CEOs and try to acquire the total impact of CEO behavior on firm performance.LDA method gives aid to deal with high-dimensional CEO behaviors (meetings,parties,business trips.etc.) and collapses all characteristics in two categories:leaders and managers.He concludes leader CEOs contributes more to firm performance,and totally 17% of CEOs in the sample are mismatched.Ke et al.(2019) use a supervised topic model to quantify stock market sentiments in order to predict stock price.First,the authors derive the character words use bag-of-word algorithm.Then they use a topic model to derive the sentiment score of each article,with a lasso penalized term added.Finally,they regress sentiments score on stock performance to generate the sentiment-price relationship.

2.2 Supervised generative models

Though attributes y_i is not available in the unsupervised model, we can observe it in the supervised model setting and variable y_i provide support for model estimation. Among all the supervised generative models, naïve Bayes classifier is the commonly used one. This model is based on posteriori probability theory in mathematics, and we illustrate it as below.

Since attribute y_i is available in naïve Bayes setting, we can illustrate the model structure as $\mathrm{p}(x_i|y_i) = \prod_j p_j(x_{ij}|y_i)$. Note that there's conditional independence between tokens j as conforming to posterior probability algorithm. Naive Bayes is so called because it assumes that each input variable is independent. This is a hard assumption to make and is far from satisfactory in real life, but the technique is still very effective for most complex problems.

In order to train the Naive Bayes model, we need to first give the training data and the corresponding classification of these data. So these two probabilities up here, category probabilities and conditional probabilities. They can all be calculated from the training data given. Once calculated, the probabilistic model can use Bayesian principles to predict new data. The calculation process is shown as

$$\mathrm{p}(\mathrm{Y}|x_i) = \frac{p(x_i|Y)\pi_y}{\sum_a p(x_i|a)\pi_a} \quad (3)$$

where π_a is prior probability for a.

Naïve Bayes classifier has been used in economic research. For example,Li(2010) use Naïve Bayesian machine learning method to analyze the impact of financial statement tone on firms' future performance.Based on the 10-K files,Li(2010) classifies performance of firms in four categories:positive,negative,neutral and uncertain.The conclusion is that positive tone of financial statements links with better future performance. However, naïve baysian hypothesize that words are independent with each other,which may not conform to reality.

3.Text regression

Similar to traditional regression methodology, text regression aims to predict y_i by resgressing on x_i, whereas in this case the independent variables are text data. The complication and high dimensionality make traditional econometric methods such as OLS infeasible. Here we introduce some methods that contribute to analyze oil market sentiment.

3.1 Linear text regression

Typically text regression consists of a penalized term to reduce high dimensionality.In this method,the cost

function penalize the deviations of parameters form zero.In consequence,weak parameter at last are deleted to achieve the goal of dimensional reduction. Among all the penalized text regressions, L_1 penalization is the most popular one. It produces sparse solutions, and these solutions have many features to our satisfactory (e.g., Bickel, Ritov, and Tsybakov 2009; Wainwright 2009; Belloni,Chernozhukov, and Hansen 2013; Buhlmann and van de Geer 2011), and the number of nonzero estimated coefficients is an unbiased estimator of the regression degrees of freedom (which is useful in model selection;see Zou, Hastie, and Tibshirani 2007).

Focusing on L_1 penalization, its form is as follows:

$$min\{l(\alpha,\beta) + n\tau\sum_{i=1}^{p}\omega_i|\beta_i|\}(4)$$

Different choices of τ impact the parameters estimation of the model. Large τ leads to simple model estimatesin the sense that most coefficients will be set at or close to zero, while as $\tau \to 0$ we approach maximum likelihood estimation(MLE). Since there is no way to define an

optimal τ a priori, standard practice is to compute estimates for a large set of possible τ and then use some criterion to select the one that yields the best fit. To find a appropriate τ , researchers most often use K -fold cross-validation (CV).

Typically, Kozak (2019) use a elastic penalty regression model to analyze the impact of stochastic discount factor on stock price. Normally economists only use several factors to forecast stock returns, like three factors model of Fama and French(1993),four factors of Hou et al.(2015) and so forth. Kozak et al.(2019) make progress by use penalty model to include a large set of factors into regression.

Besides, two classic dimension reduction techniques—*principal components regression* (PCR) and *partial least squares* (PLS) are popular in linear text regression. PCR consists of a two-step procedure. In the first step, principal components analysis (PCA) combines regressors into a small set of K linear combinations that best preserve the covariance structure among the predictors. In the second step, standard regression is conducted based on the K components. Foster, Liberman, and Stine (2013) use PCR to build a hedonic real estate pricing model that takes textual content of property listings as an input.

PCR fails to consider the ultimate output variable when reducing dimension, whereas PLS overcomes this drawback. PLS performs dimension reduction by directly exploiting covariation of predictors with the forecast target. Suppose we are interested in forecasting variable y_i. PLS regression proceeds as follows. For each element j of the independent variablex_i, estimate the univariate covariance between y_i on x_{ij}. This covariancereflects the attribute's "partial" sensitivity. Next, form a single predictor by averaging all attributes into a single aggregate predictor $\hat{y_i} = \sum_j \varphi_j x_{ij} / \sum_j \varphi_j$, where φ_j denote the covariance between dependent and independent variables. This forecast places the highest weight on the strongest univariate predictors, and the least weight on the weakest. In this way, PLS performs its dimension reduction with the ultimate forecasting objective in mind.

3.2 Nonlinear text regression

Some scholars argue that linear relationships are too restrictive for the complex text data, and some nonlinear methods are put into practice. Here we introduce three types of commonly used nonlinear text regressions.

A popular way to examine nonlinear relationship is support vector machine, or SVM(Vapnik 1995). This is used for text classification problems, the prototypical example being email spam filtering. SVM also begins to show its existence in economic study, as Manela and Moreira(2017) adopt Support vector machine method to seek the relationship between uncertainty and stock price.Through dimensionality reduction they regress uncertainty factors on stock price,concluding wars and government policy coincide with the largest price volatility.

What SVM wants is to find the farthest distance from each kind of sample point to the hyperplane, that is, to find the maximum interval hyperplane. The computation process for support vector machine is shown below. A hyperplane can be described by $W^T x + b = y$,and extending to n-dimensional space, the distance between the point $x(x_1, x_2, \ldots x_n)$ and the hyperplane is $(W^T x + b)/\|\omega\|$,where $\|\omega\| = \sqrt{\omega_1^2 + \ldots \omega_n^2}$. To maximize the distance from support vector to the hyperplane, we have the

optimization problem $\min \frac{1}{2}\|\omega\|^2 \; s.t. \; y_i(W^T x_i + b) \geq 1$. Then SMO(Sequential Minimal Optimization) method can be applied to reach the solution.

More advanced approaches like regression trees and deep learning are also used in text analysis. Regression trees have become a popular nonlinear approach for incorporating multi-way predictor interactions into regression and classification problems. The logic of trees differs markedly from traditional regressions. A tree "grows" by sequentially sorting data observations into bins based on values of the predictor variables. This partitions the data set into rectangular regions, and forms predictions as the average value of the outcome variable within each partition (Breiman et al. 1984). This structure is an effective way to accommodate rich interactions and nonlinear dependencies. Two extensions of the simple regression tree have been highly successful thanks to clever regularization approaches that minimize the need for tuning and avoid overfitting. Random forests (Breiman 2001) average predictions from many trees that have been randomly perturbed in a bootstrap step. Boosted trees (e.g., Friedman 2002) recursively combine predictions from many oversimplified trees.

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain, and its design is effective in deal with complicated data structure, such as text data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy. A main attraction of neural networks is their status as universal approximators, a theoretical result describing their ability to mimic general, smooth nonlinear associations.

V. CONCLUDING REMARKS

There're vast of literatures analyzing oil price. However, in classical energy economic theory, investor sentiment does not play a role in oil price volatility. This paper reviews traditional oil price literatures and challenges this view. Further, we list advanced machine learning skills that are useful to quantify oil market sentiments, including dictionary-based methods, generative methods and text regression. This paper offers a new direction for oil price analysis.

REFERENCES

[1] Aastveit, K. A., H. C. Bjrnland, and L. A. Thorsrud (2015). What Drives Oil Prices? Emerging versus Developed Economies. Journal of Applied Econometrics 30 (7), 1013{1028.

[2] alerie A. Ramey,Matthew D. Shapiro. Displaced Capital: A Study of Aerospace Plant Closings[J]. Valerie A. Ramey;Matthew D. Shapiro,（2001）,109(5).

[3] Baumeister, C., and J.D. Hamilton (2015), "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information," *Econometrica*, 83, 1963-1999.36 https://doi.org/10.3982/ecta12356

[4] Baumeister, C., and J.D. Hamilton (2019), "Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Oil Demand Shocks," *American Economic Review*, 109, 1873-1910. https://doi.org/10.1257/aer.20151569

[5] Baumeister, Christiane and Gert Peersman (2013). "The Role of Time-Varying Price Elasticities in Accounting for Volatility Changes in the Crude Oil Market." *Journal of Applied Econometrics*

[6] Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique, （2016）, Risk and risk management in the credit card industry, Journal of Banking & Finance 72, 218–239.

[7] Caldara, D., Cavallo, M, and M. Iacoviello (2019), "Oil Price Elasticities and Oil Price Fluctuations," *Journal of Monetary Economics*, 103, 1-20.

[8] Gentzkow, M., J. M. Shapiro, and M. Taddy (2016). Measuring polarization in high-dimensional data: method and application to congressional speech. NBER Working Paper No. 22423.

[9] Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, （2019）, Text as data, Journal of Economic Literature 57, 535–74.

[10] Goldberg, Yoav, and Jon Orwant. （2013）. "A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books." In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, edited by Mona Diab, Tim Baldwin, and

Marco Baroni, 241–47. Stroudsburg: Association for Computational Linguistics.

[11] Hamilton, James D., and J. Cynthia Wu. （2014）. Risk premia in crude oil futures prices. *Journal of International Money and Finance* 42: 9-37.

[12] Harvey, Campbell R, and Yan Liu, （2016）, Lucky factors, Technical report, Duke University

[13] Harvey, Campbell R, Yan Liu, and Heqing Zhu, （2016）, ... and the cross-section of expected returns, Review of Financial Studies 29, 5–68.

[14] Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. （2017）. "Firm-Level Political Risk: Measurement and Effects." NBER WorkingPaper 24029.

[15] Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. （2015）. Statistical Learning with Sparsity: The Lasso and Generalizations. New York: Taylor and Francis.

[16] https://doi.org/10.1016/j.jmoneco.2018.08.004

[17] Jegadeesh, N. and D. Wu (2013). Word power: a new approach for content analysis. Journal of Financial Economics 110(3), 712–729.

[18] Juvenal, L. and I. Petrella (2015). Speculation in the Oil Market. Journal of Applied Econometrics 30 (4),621{649.

[19] Kelly, Bryan, Seth Pruitt, and Yinan Su, （2019）, Some characteristics are risk exposures, and the rest are irrelevant, Journal of Financial Economics, forthcoming .

[20] Kenneth J. Singleton(2014).Investor Flows and the 2008 Boom/Bust in Oil Prices.Management Science

[21] Khandani, Amir E, Adlar J Kim, and Andrew W Lo, （2010）, Consumer credit-risk models via machinelearning algorithms, Journal of Banking & Finance 34, 2767–2787.

[22] Kilian, L. and D. P. Murphy (2012). Why Agnostic Sign Restrictions Are Not Enough: Understanding the Dynamics of Oil Market VAR Models. Journal of the European Economic Association 10 (5),1166{1188.

[23] Kilian, L. and D. P. Murphy (2014). The Role of Inventories and Speculative Trading in the Global Market for Crude Oil. Journal of Applied Econometrics 29 (3), 454{478.

[24] Kilian, Lutz (2009). "Not all Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market." *American Economic Review*, 99, 1053–1069.

[25] Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, （2019）, Shrinking the cross section, Journal of Financial Economics, forthcoming .

[26] Li, Feng, （2010）, The Information Content of Forward-Looking Statements in Corporate Filings-A Na¨ıve Bayesian Machine Learning Approach, Journal of Accounting Research 48, 1049–1102.

[27] Lippi, F. and A. Nobili (2012). Oil and The Macroeconomy: A Quantitative Structural Analysis.Journal of the European Economic Association

[28] Loughran, Tim, and Bill McDonald, （2011）, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, The Journal of Finance 66, 35–65.

[29] Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke, （2016）, Deep learning for mortgage risk, Available at SSRN 2799443 .

[30] Smith, James L. （2009）. World oil: Market or mayhem? *Journal of Economic Perspectives* 23:145-164.

[31] Taddy, Matt. （2015）. "Document Classification by Inversion of Distributed Language Representations." In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 2, edited by Chengqing Zong and Michael Strube, 45–49. Stroudsburg: Association for Computational Linguistics. Taddy, Matt. 2017a. "Comment: A Regularization Scheme on Word Occurrence Rates That Improves Estimation and Interpretation of Topical Content." https://github.com/TaddyLab/reuters/raw/master/ comment/comment-AiroldiBischof.pdf.

[32] Tang, K., and W. Xiong, （2011）, "Index Investing and the Financialization of Commodities," working paper, Princeton University.

[33] Tetlock, Paul C, （2007）, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, The Journal of Finance 62, 1139–1168.

[34] Wang, Zhang and Bao, "Uncertainty of economic policy and dynamic adjustment of enterprise capital structure and stabilizing leverage" [J]. China Industrial Economics,2018(12):134-151.

[35] wavelet coherence to geophysical time series. Nonlinear Process. Geophys.（2004）, 11 (5/6),561–566.

[36] Wensheng Kang,Fernando Perez de Gracia,Ronald A. Ratti. Oil price shocks, policy uncertainty, and stock returns of oil and gas corporations[J]. Journal of International Money and Finance,（2017）,70.

[37] Wensheng Kang,Ronald A. Ratti,Joaquin L. Vespignani. Oil price shocks and policy uncertainty: New evidence on the

effects of US and non-US oil production[J]. Energy Economics,（2017）,66.

[38] Wensheng Kang,Ronald A. Ratti. Structural oil price shocks and policy uncertainty[J]. Economic Modelling,（2013）,35.

[39] Xiaoqing Zhou. Refining the workhorse oil market model[J]. Journal of Applied Econometrics,（2020）,35(1).

[40] Xiong, W., and H. Yan, （2010）, "Heterogeneous Expectations and Bond Markets," Review of Financial Studies, 23, 1433–1466.

[41] Xu,Jiang and Yang, "The correlation between the fluctuation of international gold price and crude oil price and Shanghai stock index based on the wavelet analysis method" [J]. Financial Forum,（2019）,24(06):54-61.

[42] Zalla, R., "Economic Policy Uncertainty in Ireland," [R].（2016）,working paper, 20 September.

[43] Zheng Tracy Ke, Bryan T. Kelly, Dacheng Xiu(2019) "Predicting Returns With Text Data" NBER working paper