2019 IEEE International Conference on Big Data (Big Data)

Natural Gas Price Prediction with Big Data

Yuanyuan Tang
School of Information
Renmin University of China
Beijing, China
253850183@qq.com

Qingmei Wang
School of Information
Renmin University of China
Beijing, China
wqm@ruc.edu.cn

Wei Xu
School of Information
Renmin University of China
Beijing, China
weixu@ruc.edu.cn

Mingming Wang
School of Information
Renmin University of China
Beijing, China
mingmwang@263.net

Zhaowei Wang
School of Information
Renmin University of China
Beijing, China
807933844@qq.com

***Abstract*—In recent decades, commodities have attracted the attention of a large number of investors and financial professionals, and their prices have become very important for investors and policy makers. Although natural gas is widely used as a clean energy, there is less literature on price research and most of them use historical prices and fundamental factors to analyze natural gas price. Based on historical prices, this paper analyzes whether internet search data and news sentiment can improve forecasting ability, and compare which can achieve better prediction results. The experimental results show that both internet search and news sentiment contain additional information to improve the prediction effect, and internet search data can achieve better prediction results.**

Keywords—natural gas, price prediction, internet search, news sentiment, deep learning

I. INTRODUCTION

With the impact of global pollution, natural gas has been widely used as the cleanest energy source and the second most active energy commodity in fossil fuels. It is not only the main fuel for civilian use, but also an important input for industrial applications and power generation. In order to hedge the risks of the natural gas spot market, a large number of hedging strategies such as natural gas futures and options can be used. In recent years, energy giants have turned to natural gas, and natural gas futures trading has soared, but the natural gas futures market has proven to be one of the most volatile markets [1]. Natural gas prices are of great concern to investors, scholars, and policy makers and the high volatility of the natural gas market makes accurate predictions of natural gas prices and their movements critical.

Most literature has examined some fundamental factors that can fluctuate natural gas prices, such as natural gas storage [1-3], seasonal factors [4]. Other studies use historical prices data to analyze and predict the price of natural gas [5, 6]. However, with the rapid development of big data technology and practice, the use of big data for commodity sales and price forecasts has attracted widespread attention in academia and the business societies. There is less research on natural gas than on existing stock markets and other commodities such as crude oil.

As we all know, the futures market and the stock market have very similar properties. In the past decade, researchers have explored the changing patterns of internet data on stock markets from different perspectives, including search data, news events, and social media data. The economic intuition behind it is that internet activity represents the attention of investors, and it can be considered an important factor in price volatility [7]. So search indices are widely used in financial markets to measure investor concerns. Search behavior is conducted privately, subjectively, voluntarily, and will not be manipulated, so it can reflect the investor's attention more objectively and truly. What's more, search data has already been heavily used in stock forecasts [8], sales forecasts [9], and other commodities price forecasts such as crude oil [10].

In addition, online news is also often used in stock forecasts. Stocks prices are subject to the company's fundamentals, company disclosures [11] and company-related news [12]. Similarly, for energy commodities, natural gas prices are also affected by many other objective factors, such as national policies, shale revolutions, and economic and political events, supply and demand shocks [13], and these information can be obtained from the news. A small amount of literature indicates that news has an impact on natural gas price movements, but there are still conflicts. Grundmann et al. [14] showed that the recognition of text news has a weak influence on the natural gas price trend. But Alfano et al. [15] indicated there is a significant positive effect of news sentiment on the natural gas price. And Borovkova et al. [16] investigated the impact of news sentiment on the price dynamics of natural gas futures. By augmenting all models with news sentiment variables, they found that including news sentiment in volatility models significantly improves volatility forecasts.

Based on historical prices, this paper not only focuses on subjective investor concerns, but also on objective news events. This paper proposes a natural gas price prediction model based on search data and news sentiment to analyze whether they have predictive power on natural gas prices. Besides, we compares and analyzes which can achieve better prediction performance. We collect web search data, natural gas futures price data and online news data from Google Trends, the Energy Information Administration website and Yahoo Financial News respectively. News sentiment is obtained through lexicon-based sentiment analysis method, and a deep neural network is used to predict natural gas price. The experimental results show that based on historical prices, both web search volume and news sentiment can improve the forecasting effect, and the effect of adding network search volume is better than news sentiment.

In the remainder of this article, we will review the relevant literature in the second part. In the third part, we present a natural gas price prediction model based on Google search data and online news sentiment, and elaborate on the news sentiment extraction process and artificial neural network regression method. The fourth part introduces and displays the

978-1-7281-0858-2/19/$31.00 ©2019 IEEE
978-1-7281-0858-2/19/$31.00 ©2019 IEEE

data. The fifth part analyzes the empirical process and results. Finally, it summarizes and discusses the research and proposes the future work plan.

II. Related Work

Previous researchers have done a lot of research on natural gas price forecasts, search volume, and news analysis. In this section, we will briefly review previous work.

A. Natural Gas Price Forecast

In the early literature, most researchers used econometric methods to predict and analyze natural gas prices. Buchananan et al. [17] is one of the earlier literatures that provide a way to predict the movement of the spot price of the natural gas market. Nguyen et al. [5] used wavelet transform and adaptive models to predict natural gas futures price. Hailemariam and Smyth [13] used structural allogeneic autoregressive VAR (SHVAR) models to analyze the main drivers of natural gas price volatility in the United States. With the development of big data technology, more and more big data models are used in forecasting. Ceperi et al. [18] gave a short-term prediction of the Henry Hub spot natural gas price based on the classical time series model and machine learning method. The results show that the machine learning method with feature selection algorithm is much better. Herrera et al. [19] used machine learning to forecast energy commodities price, and the results showed that machine learning models are superior to traditional econometric methods and have the ability to predict inflection points.

B. Search Trend

Internet search data has been widely used as a tool to measure the attention of investors or consumers in different markets. Xu et al. [8] studied the role of the Google search index in stock market volatility forecasting and found that its combination with other macroeconomic variables can enhance its forecasting effect, indicating that Google Trends contains useful information on stock market volatility forecasts. D'Amuri and Marcucci [20] used Google search data to predict the US monthly unemployment rate index and found that the model based on search data is significantly better than other models. In addition, Geva et al. [9] used search trend data and social media data to predict vehicle sales, and the results showed that model's accuracy based on cheap search data can be comparable to the model based on social media data. In the energy commodity market, Li et al. [10] discussed the relationship between the Google search volume index, different trader positions and crude oil price. Tao et al. [21] used the internet search data-driven model to predict crude oil prices. The results showed that the model had better prediction accuracy than the time series model. Similarly, Afkhami et al. [22] confirmed the effectiveness of Google's search activity and found it to be more predictive of energy commodity price volatility. Rao and Srivastava [23] used Google search volume index and Twitter sentiment to predict oil and gold price, and proved that search volume index has a better predictive performance. What's more, Campos et al. [24] found a significant positive correlation between search volume and oil volatility. The search volume is not only a good proxy for traditional macro financial variables, but also carries additional information about historical information.

C. News Analysis

A large body of literature indicates that news sentiment has a good predictive power for financial markets such as stocks. Atkins et al. [25] showed that financial news can more effectively predict stock volatility, so derivatives can be priced by quantifying volatility. Ding et al. [12] extracted events from news texts and proposed a deep learning method based on event-driven stock market forecasting. The results showed that the method has nearly 6% improvement in stock forecasting. For energy commodities, news is often used in the prediction of oil and crude oil. Schmidbauer and Rosch [26] studied the impact of OPEC's statement on oil production on changes in oil prices. Similarly, Alfano et al. [27] examined the effect of news sentiment on crude oil prices for different investor types according to the noise trader approach. Their finding suggested that news sentiment not only has a significant positive effect on the noise residual, but also on the fundamental price. Mensi et al. [28] studied the impact of different OPEC news announcements on crude oil market conditions and volatility through the ARMA GARCH model. The results showed that some announcements about production cuts and maintenance yields have a significant impact on crude oil market reporting and volatility. Li et al. [29] offered a wealth of online news data to help predict oil price movements. Similarly, Li et al. [30] proposed a deep learning model that predicts the crude oil price based on the sentiment of online news, and indicated that the textual features of the news and financial characteristics are complementary.

By reviewing the previous literature, we find that most studies use historical price data or other fundamental factors to conduct predictive analysis. Second, there is less literature on the use of search data to predict natural gas price. Third, as mentioned before, using news to predict natural gas price still has conflicts. Therefore, based on historical price information, this paper uses neural network to analyze the predictive power of search data and news sentiment on natural gas futures prices, and compares which can achieve better prediction results.

III. Methodology

The structure of our research model is shown in figure 1, including three steps: (a) data collection, (b) news sentiment analysis, and (c) neural network regression.

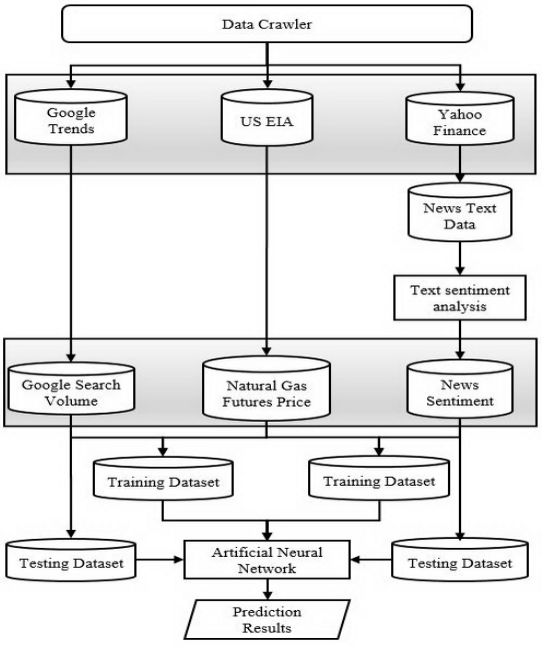

Fig. 1. Structure of the research model

A. Lexicon-based Sentiment Analysis

The sentiment analysis methods of texts are usually divided into two categories: supervised and unsupervised. For supervised methods, each item in the corpus is marked positive or negative, while unsupervised methods do not need labels, so lexicon-based sentiment analysis is widely used. Feuerriegel and Gordon [11] used Loughran-Mcdonald finance-specific dictionary [31] to calculate regulatory disclosures sentiment for short-term stock index forecasting. Besides, SentiWordNet has also been extensively applied to extract positive and negative emotions from texts in large number of researches [32]. Since our dataset is unlabeled, we also use the Loughran-McDonald finance-specific dictionary in the financial market to perform sentiment analysis on news texts. The dictionary is specifically designed to extract information from financial news and has evolved into an application standard for financial related research.

We remove the duplicate news, then we do text preprocessing, including tokenization, stopword removal, lemmatization, and stemming. Then, we match the sentiment words in the news text according to the dictionary. Finally, we get the number of positive and negative words in the daily news, and calculate the sentiment score of the daily news according to formula (1). Where N_{pos} stands for the number of positive words, N_{neg} stands for the number of negative words, and S is the sentiment score of news.

$$S = \frac{N_{pos} - N_{neg}}{N_{pos} + N_{neg}} \quad (1)$$

B. Artificial Neural Network for Regression

Artificial neural network was first developed in 1943 by McCulloch and Pitts [33], which simulates the structure of human brain neurons interconnected. In recent decades, neural networks have become a research hotspot in the field of artificial intelligence. No matter the computing power, data volume and algorithm performance have made great progress, it has been widely used in various fields. Many studies [6, 18] have also used neural networks to predict the prices of energy commodities and have achieved good results.

The neural network starts from the training data and constantly adjusts the parameters to achieve the best results. When the model is training, ANN adopts the back-propagation method to adjust the weights between neurons in the previous layer and neurons in the next layer. The function from the input to the output is:

$$y = f(\sum wx + b) \quad (2)$$

Where f is the activation function, w is the weight, b is bias, x is the input and y is the output. Through the learning process, we can get the trained model and the model is adopted in testing dataset to validate the effectiveness and efficiency of our method.

IV. Data

Natural gas futures price data, internet search data, news data and their basic statistics are described in this section. Table 1 shows a statistical description of each variable.

A. Natural Gas Futures Price Data

We collect daily NYMEX natural gas futures price from the US Energy Information Administration website (https://www.eia.gov/), which was created in 1977 and regularly releases energy information on production, reserves, demand, import and export, and price. It is the main source of information for US energy data, analysis and forecast. Natural gas futures prices are for 1 month, 2 months, 3 months and 4 months futures contracts. Figure 2 depicts natural gas futures prices for four contracts from January 2013 to June 2019, covering 1,638 records.

Fig. 2. Natural gas futures prices

B. Internet Search Data

We use Google Trends to collect internet search data. Based on Google Search, Google Trends displays a specific search term search for each region of the world. Google Trends provides search volume for specific search items on a monthly, weekly, and daily starting in 2004. This paper uses the daily search for natural gas from January 2013 to June 2019, covering 2,372 records.

C. News Data

According to the search for the keyword of natural gas, we collected natural gas related news from January 2013 to June 2019 of Yahoo Finance, covering 972 records. Yahoo Finance is one of the famous financial news and financial research websites in the United States. It provides financial information on stocks, news, financial reports, etc. According to media analysis company comScore, Yahoo Finance has become the most visited commercial and financial news website, and the average number of monthly visits has reached more than 100 million in this year.

TABLE I. Descriptive Statistics of Variables

Variables	Obser vations	Mean	Standard deviation	Max	Min
Natural gas future1 price	1,638	3.171	0.718	6.149	1.639
Natural gas future2 price	1,638	3.202	0.683	5.465	1.767
Natural gas future3 price	1,638	3.231	0.659	4.846	1.867
Natural gas future4 price	1,638	3.249	0.637	4.82	1.944
Search volume	2,372	63.501	17.052	100	23
News sentiment	972	-0.019	0.198	1	-1

V. Empirical Analysis

A. Evaluation Criteria

In our research, by estimating the error between predicted and actual values, we use two common evaluation criteria to verify the validity and performance of the prediction for natural gas futures price with different features: mean absolute error (MAE) and root mean square Error (RMSE). The smaller they are, the more accurate the prediction model is. Where N is the number of data in the test set, and y_i and $y_i^{'}$ are the real and predicted values respectively.

$$MAE = \frac{1}{N}\sum_i^N \left| y_i - y_i^{'} \right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_i^N \left(y_i - y_i^{'} \right)^2} \tag{4}$$

B. Regression Results

We use different features to predict the natural gas futures price, which is historical price, historical price and internet search data, historical price and news sentiment, and the lag of each feature is 5. We use 70% of the data as the training set, 30% of the data as the test set, and average the 100 predictions to get the final evaluation result. Table 2 shows the regression results. First, based on historical prices, both internet search volume and news sentiment can increase the performance of predictions, which indicates that they can provide additional information. Second, the results show that the predictive effect of adding the internet search volume is much better than the forecast effect of adding news sentiments. According to the forecast results, after adding the search volume, the four contracts RMSE increased by 18.50% on average, and the news sentiment increased by 14.40% on average.

TABLE II. Comparison of ANN Prediction Results with Different Features

Contract	Variables in Predictive Model	ANN Regression Results	
		MAE	***RMSE***
Future1	Only historical data	0.1138	0.1576
	Historical data + Internet search	0.0891	0.129
	Historical data + News sentiment	0.0956	0.1368
Future2	Only historical data	0.1115	0.1498
	Historical data + Internet search	0.0951	0.1314
	Historical data + News sentiment	0.1002	0.137
Future3	Only historical data	0.1215	0.1578
	Historical data + Internet search	0.0946	0.1266
	Historical data + News sentiment	0.0987	0.133
Future4	Only historical data	0.1207	0.1608
	Historical data + Internet search	0.0875	0.1225
	Historical data + News sentiment	0.0902	0.1284

VI. Conclusion and Future Work

Compared with the previous research, this paper applies the internet search volume and news sentiment to the natural gas futures price forecast, and uses artificial neural network to predict the futures price. The results show that although historical prices can already predict future prices better, internet search data and news sentiment can also provide additional information. What's more, we find that the model added internet search data performs better than the news sentiment.

However, the current research still has certain limitations. Firstly, this paper uses the existing sentiment dictionary to analyze news sentiment, the method and accuracy of text mining can be further improved. Secondly, this paper only analyzes the text of the news and the information such as pictures and videos appearing in the news can be consider in the future. Thirdly, this paper does not consider the impact of news and search of other factors on natural gas prices. And we think that one of the reasons why search data can achieve better results than news sentiment may be the sparseness of news data.

Based on the above discussion, we also propose put forward an outlook for future work. Firstly, this paper only compares the prediction effects of search data and news sentiment. Subsequent work can add them to the predictive model with historical data to see if higher accuracy can be achieved. Secondly, a large number of research mentioned that natural gas prices can be affected by other commodities such as crude oil. Subsequent research could further consider more Google Trends data and news using more keywords, such as other commodities (e.g., crude oil), macroeconomics (e.g., economic growth, CPI), and climate (e.g., cold wave).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 71771212, U1711262).

References

[1] I. Ergen, and I. Rizvanoghlu, "Asymmetric impacts of fundamentals on the natural gas futures volatility: an augmented GARCH approach," Energy Economics, vol. 56, pp. 64-74, 2016.

[2] X. Mu, "Weather, storage, and natural gas price dynamics: fundamentals and volatility," Energy Economics, vol. 29 no. 1, pp. 46–63, 2007.

[3] S. Z. Chiou-Wei, S. C. Linn, and Z. Zhu, "The response of us. natural gas futures and spot prices to storage change surprises: fundamental information and the effect of escalating physical gas production," Journal of International Money and Finance, vol. 42(C), pp. 156-173, 2014.

[4] Q. Ji, H. Y. Zhang, and J. B. Geng, "What drives natural gas prices in the united states? – a directed acyclic graph approach," Energy Economics, vol. 69, pp. 79-88, 2017.

[5] H. T. Nguyen, and I. T. Nabney, "Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models," Energy, vol. 35 no. 9, pp. 3674-3685, 2010.

[6] G. P. Herrera, M. Constantino, B. M. Tabak, H. Pistori, J. J. Su, and A. Naranpanawa, "Long-term forecast of energy commodities price using machine learning," Energy, pp. 214-221, 2019.

[7] N. Vlastakis, and R.N. Markellos, "Information demand and stock market volatility," J. Bank. Financ, vol. 36, pp. 1808–1821, 2012.

[8] Q. Xu, Z. Bo, C. Jiang, and Y. Liu, "Does google search index really help predicting stock market volatility? evidence from a modified mixed data sampling model on volatility," Knowledge-Based Systems, vol. 166, pp. 170-185, 2018.

[9] T. Geva, G. Oestreicher-Singer, N. Efron, and Y. Shimshoni, "Using forum and search data for sales prediction of high-involvement projects, " MIS Quarterly, vol. 41 no. 1, pp. 65-82, 2017.

[10] X. Li, J. Ma, S. Wang, and X. Zhang, "How does google search affect trader positions and crude oil prices?" Economic Modelling, vol. 49, pp. 162-171, 2015,

[11] S. Feuerriegel, and J. Gordon, "Long-term stock index forecasting based on text mining of regulatory disclosures," Decision Support Systems, vol. 112, pp. 88-97, 2018.

[12] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," International Conference on Artificial Intelligence, AAAI Press, 2015.

[13] A. Hailemariam, and R. Smyth, "What drives volatility in natural gas prices?" Energy Economics, vol. 80, pp. 731-742, 2019.

[14] T. Grundmann, C. Felden, and M. Pospiech, "Forecasting the natural gas price trend - evaluation of a sentiment analysis," 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 160-164. 2016.

[15] S. J. Alfano, M. Rapp, N. Pröllochs, S. Feuerriegel, and D. Neumann, "Driven by news tone? understanding information processing when covariates are unknown: the case of natural gas price movements," Social Science Electronic Publishing, 2015.

[16] S. Borovkova, and D. Mahakena, "News, volatility and jumps: the case of natural gas futures," Social Science Electronic Publishing, vol. 38 no. 1, pp. 1-26, 2013.

[17] W. K. Buchanan, P. Hodges, and J. Theis, "Which way the natural gas price: an attempt to predict the direction of natural gas spot price movements using trader positions," Energy Economics, vol. 23 no. 3, pp. 279-293, 2001.

[18] E. Ceperi, S. Zikovic, and V. Ceperic, "Short-term forecasting of natural gas prices using machine learning and feature selection algorithms," Energy, vol. 140, pp. 893-900, 2017.

[19] G. P. Herrera, M. Constantino, B. M. Tabak, H. Pistori, J. J. Su, and A. Naranpanawa, "Long-term forecast of energy commodities price using machine learning," Energy, vol. 179, pp. 214-221, 2019.

[20] F. D'Amuri, and J. Marcucci, "The predictive power of google searches in forecasting unemployment," Temi Di Discussione, 2013.

[21] R. Tao, X. Zhang, and L. Zhan, "Forecasting crude oil prices based on an internet search driven model," 2018 IEEE International Conference on Big Data (Big Data), 2018.

[22] M. Afkhami, L. Cormack, and H. Ghoddusi, "Google search keywords that best predict energy price volatility," Energy Economics, vol. 67, pp. 17-27, 2017.

[23] T. Rao, and S. Srivastava, "Modeling movements in oil, gold, forex and market indices using search volume index and twitter sentiments," Computer Science, 2012.

[24] I. Campos, G. Cortazar, and T. Reyes, "Modeling and predicting oil vix: internet search volume versus traditional macro-finance variables," Energy Economics, vol. 66, pp. 194-204, 2017.

[25] A. Atkins, M. Niranjan, and E. Gerding, "Financial news predicts stock market volatility better than close price," The Journal of Finance and Data Science, vol. 4, pp. 120-137, 2018.

[26] H. Schmidbauer, and A. Rosch, "Opec news announcements: effects on oil price expectation and volatility," Energy Economics, vol. 34 no. 5, pp. 1656-1663, 2012.

[27] S. J. Alfano, S. Feuerriegel, and D. Neumann, "Is news sentiment more than just noise?" SSRN Electronic Journal, 2014.

[28] W. Mensi, S. Hammoudeh, and S. Yoon, "How do OPEC news and structural breaks impact returns and volatility in crude oil markets? further evidence from a long memory process," Energy Economics, vol. 42, pp. 343-354, 2013.

[29] J. Li, Z. Xu, L. Yu, and L. Tang, "Forecasting oil price trends with sentiment of online news articles," Procedia Computer Science, vol. 91, pp. 1081-1087, 2016.

[30] X. Li, W. Shang, and S. Wang, "Text-based crude oil price forecasting: a deep learning approach," International Journal of Forecasting, vol. 35 no. 4, pp. 1548-1560, 2018.

[31] T. Loughran, and B. Mcdonald, "Textual analysis in accounting and finance: a survey," Journal of Accounting Research, vol. 54 no. 4, pp. 1187-1230, 2016.

[32] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta. DBLP.

[33] W.S. McCulloch, and W. Pitts, "A logical calculus of ideas immanent in nervous activity," Bull. Math. Biophys, vol. 5, pp. 115–133, 1943.