Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 5903057, 18 pages
https://doi.org/10.1155/2020/5903057

Research Article

How Does Internet Information Affect Oil Price Fluctuations? Evidence from the Hot Degree of Market

Lu-Tao Zhao[1,2] **Shi-Qiu Guo**[1] **Jing Miao**[1] **and Ling-Yun He**[2,3,4]

[1]*School of Mathematics and Physics, University of Science and Technology Beijing, Beijing, China*
[2]*Center for Energy and Environmental Policy Research, Beijing Institute of Technology, Beijing, China*
[3]*School of Economics, Jinan University, Guangzhou 510632, China*
[4]*School of Economics and Management, Nanjing University of Information Science and Technology, Nanjing 210044, China*

Correspondence should be addressed to Ling-Yun He; lyhe@amss.ac.cn

Received 28 March 2020; Revised 11 November 2020; Accepted 18 November 2020; Published 1 December 2020

Academic Editor: Alfred Peris

Not only the fundamentals of supply and demand but also international oil prices are affected by nonfundamental indicators such as emergencies. With the development of big data technology, many unstructured and semistructured factors can be reflected through Internet information. Based on this, this paper proposes a HD-based oil price forecasting model to explore the impact of Internet information on international oil prices. Firstly, we use LDA and other methods to extract topics from massive online news. Secondly, based on conditional probability and correlation, the positive hot degree (PHD) and negative hot degree (NHD) of the oil market are constructed to realize the quantitative representation of Internet information. Finally, the SVAR method is established to explore the interactive relationship between HD and oil prices. The empirical results indicate that PHD and NHD have a better ability to predict international oil prices compared with Google Trends which is widely used in the other research. In addition, PHD has a significant positive impact on oil prices and NHD has a negative impact. In the long term, PHD accounts for 51.00% of oil price fluctuations, ranking the first among relevant influencing factors. The findings of this paper can provide support to investors and policy-makers.

1. Introduction

As a strategic energy source, oil has both commodity, financial, and political attributes [1, 2]. The fluctuation of oil prices will have an important impact on economic growth, stock exchange rates, bond markets, and national security, so the forecasting of crude oil prices has received much attention [3, 4]. However, crude oil price prediction is a typical Nondeterministic Polynomial Complete (NP-C) problem, and the indicators affecting its price fluctuations are complex [5–7], not only being related to the supply and demand of fundamentals, but also to the USD exchange rate, emergencies, market speculation, and big country games [8, 9]. The fluctuations of nonfundamental factors mostly lead to psychological changes in investors, triggering market speculation [10, 11], which further causes changes in fundamental supply and demand. Faced with such a complex system in the oil market, how to assess the price trend is key [12, 13], however, many indicators affecting the crude oil market are difficult to quantify directly, and investors' psychological changes are hard to capture in time. Therefore, it is necessary to discover new indicators to characterise the volatility of the oil market quickly and accurately. With the advent of big data era, investor behaviors are increasingly influenced by Internet information orientation. Some studies propose the use of Internet information to quantify investors' speculative behaviors [14–16].

2. Literature Review

In fact, several studies have proved that Internet information can promote the prediction of commodity price trends. The current research is mainly divided into two parts: on the one hand, more traditional research investigates planned and easily identifiable news cases, macroeconomic reports, income reports, and so forth to characterise the impact of Internet information on asset prices [17,18]. For example, Yuan [19] uses Dow Jones record-breaking events and front-page news to characterise the stock market attention and concludes that investors would generally sell stocks when widely concerned, which has a negative impact on prices; Schmidbauer and Rösch [20] put the announcements issued after the OPEC meeting as dummy variables into a Generalized Autoregressive Conditionally Heteroskedastic (GARCH) model to assess the impact of the announcement on oil price fluctuations, and the results show a significant effect. On the other hand, most of the more advanced research uses Google Trends to conduct related research [21,22]. Among them, Yao et al. [23] used the principal component analysis (PCA) method to combine the Google Trends to characterise oil market investor attention, and based on the Structural Vector Autoregression (SVAR) model, the results show investor attention has a significant negative impact on crude oil prices. Wang et al. [24] constructed an Internet concern index by analysing the correlation between Google Trends and oil prices and predicted oil prices by combining Extreme Learning Machine (ELM) methods, which improves the accuracy of forecasting; Gao et al. [17] explored the impact of Internet attention on China's stock market through Qihoo 360's search index and found that Internet attention contributes to the spread of information to stock prices and weakens information asymmetry.

As a free and open tool, Google Trends has the advantages of easy accessibility, timeliness, and objectivity. It is favored by researchers and has achieved good research results in the field of price forecasting [25]. However, Google Trends, an indicator of investor attention, still has certain limitations. For example, Li et al. [26] had shown that minors and nonprofessional investors tend to search industry news through Google and experienced investors in the oil market will choose a more professional platform to obtain the first news. In addition, the motivation of users to actively search is often derived from the continuous reporting and fermentation of news; that is, news reports will lead the change of Google Trends and have the first timeliness. Moreover, Google Trends is repetitive, and multiple searches of the same user will be recorded, which will cause data bias.

Given the abovementioned limitations of Google Trends, news reports issued by professional media can play a complementary role [27]. Studies have shown that news reports often have a large influence, disseminating major events in the energy market to various groups, and have an important role in promoting price predictions [28,29]. However, owing to the fact that news is a text variable, there are technical barriers to data preprocessing and quantification. Based on the above analysis, a summary of the characterisation methods of the oil market's Internet information is shown in Table 1.

As Internet information is uneven and difficult to use directly, natural language processing (NLP) technology and text mining technology are used: firstly, these methods can quickly grab a large amount of information on the Internet; then, denoise the extracted Internet information (news, social network feeds, etc.) to enhance data availability; finally, adopt the characteristics of text information in relevant, quantitative ways to explore the relationship between Internet information and financial price series data [30,31]. In terms of data acquisition, web crawlers such as Scrapy, Puppeteer, and Selenium are currently often used [32]. Among them, Selenium is a packaged tool for data acquisition in the programming software Python [33]. The principle is to use a virtual browser to open the specified web page and locate the data according to the CSS function. This will involve some basic web page knowledge and any of us can use it for free. In addition, we can learn more about the operating principle of Selenium through the website "https://selenium-python.readthedocs.io/." In terms of data application, Wang et al. [34] used the Term Frequency-–Inverse Document Frequency (TF–IDF) method to represent the Internet text as a feature vector and put it into the Autoregressive Integrated Moving Average (ARIMA) model to predict stock prices and obtain a better prediction; Ho et al. [35] extracted emotional information from online news and put it into the Fractionally Integrated Generalized Autoregressive Conditionally Heteroskedastic (FIGARCH) and Regime-Switching GARCH models to analyse the dynamic relationship between emotion and stock return rate, and the conclusion shows that news emotion can better reduce yield volatility; Füss et al. [36] proposed the use of information density to measure the response of prices to online news and then to characterise uncertainties in the market and analyse its relationship with market price "jump." Lee [37] used word2vec (a word embedding method) to represent news headlines as vectors and introduced a recurrent convolutional neural networks (RCN) model for deep mining of stock market information. The final results show that the information is more analytical and more conducive to price forecasting.

On the whole, Internet information is hard to capture and quantify in time. Although in the financial market, research on the deep mining of Internet information to assist the forecasting of price fluctuations has made some breakthroughs, research into oil markets mostly uses Google Trends. In addition, the content expressed in the form of Internet information is diversified, and different topics will have different effects on prices. The aforementioned studies take into account all the information on the Internet, do not implement any filtering of fraudulent or irrelevant information, and do not focus on analysing different types of Internet information, which will lead to subjectivity and bias in the results. Therefore, the application of Internet information extraction in oil price changes remains to be further studied.

Based on the above problems, we use the probabilistic topic model of NLP technology to extract the topic of news reports about the oil market and classify it automatically. Not only does this method filter out invalid information of Internet news but also it realises text clustering and topic factor mining. More importantly, this method can mine the topic hot degree based on conditional probability to achieve

Table 1: Representation and characteristics of Internet information.

Source	Advantages	Disadvantages	Literatures
Special issues	More detailed analysis of certain types of major events	Not sustainable	[18–20]
Google trends	Easy access to data	Locality and repeatability	[21–23,26]
Internet news	Strong transmission, with first-time effectiveness	Difficult to quantify	[27–29]

the quantification of text data. Compared with traditional search volume and news volume processing, this method is more rational and interpretable. Considering that the use of linear regression does not readily capture the dynamic relationship between sequences [38], the current relatively hot machine learning models can only provide a result of prediction accuracy, whose economic significance and interpretability of the model are poor. However, the SVAR model can better avoid the above problems. The model is often used to predict the interconnected time series system and analyse the dynamic impact of random disturbances on the variable system, so as to explain the impact of various economic shocks on the formation of economic variables. It has been widely used. Applied to the literature of energy economics and policy modeling; for our research, SVAR can analyse the impact of crude oil prices and influencing factors at the same time and can provide rich quantitative results based on the impact of these factors on crude oil prices. Thus, we use an SVAR model to explore the dynamic effects between hot degree (HD) and oil prices [39,40]. Generally, the SVAR model is developed based on the hypothesis that all variables are stationary [41]; however, most variables cannot meet the constraint conditions. Research indicates that the results of the unit root test are sensitive to the size of the dataset [42,43]. Fortunately, the method proposed by Toda and Yamamoto [44] to estimate the SVAR model is less restrictive, whose advantage is that it does not need to consider the stability of variables, or the single and cointegration relationships. The main resolved problems of this paper are as follows:

(1) How does the hot degree extracted based on the probability topic model affect the oil price trend? How long will this impact last?

(2) Compared with the traditional influencing factors, how much does the contribution of the hot degree to oil price changes?

(3) Compared with Google Trends that is widely used in other research, can the hot degree extracted here better explain oil price fluctuations?

3. Methods

In order to solve the quantification of Internet news and analyse how the Internet information affects oil price fluctuations, this paper builds a HD-based oil price forecasting model based on LDA, SVAR, and other methods. The model framework is shown in Figure 1.

First, the web news related to the international oil market is obtained; then, cleaned news is modeled in the oil market topic generation method, and the topic hot degree is obtained by means of the probability matrix. Next, through analysing the correlation between the crude oil price and topic hot degree, identify the tendency, and the positive and negative hot degree is selected. Finally, the supply and demand factors and hot degree (HD) are put into the SVAR model to explore the impact of the HD on international oil prices and make a comprehensive comparison with the effect of Google Trends. Next, the model in this paper will be explained in detail.

3.1. Oil Market Text Analysis. The theoretical basis for text analysis of the oil market is a probabilistic topic model, and its development is a process of continuous improvement. The main idea of this model is to regard text as the polynomial distribution of several topics, and the topic is the polynomial distribution of words. The earliest theory is the Latent Semantic Analysis (LSA) model, proposed by Scott et al. [45] in 1990; subsequently, Hofmann [46] proposed the Probability Latent Semantic Analysis (PLSA) model in 1999 to better describe polysemy in texts. Blei et al. [47] raised the Latent Dirichlet Allocation (LDA) model on the basis of PLSA in 2003; the main contribution is to add the Dirichlet prior distribution, which effectively solves the overfitting problem caused by too many parameters in the PLSA model. Currently, a large number of papers have proved that the LDA model shows excellent performance in text topic extraction [48–50], which can not only obtain a qualitative output of the topic keywords but also obtain a quantitative output of the topic probability values. In view of this, we selected the LDA model as the topic extraction model used on crude oil market text to extract the topic heat of web information. The process of generating simulated text based on the LDA model is as follows:

Let $\theta \in \Theta = \{\theta_1,\theta_2,\ldots,\theta_N\}$ be the topic distribution of text $d \in D = \{d_1,d_2,\ldots,d_N\}$ generated from probability $P(\theta,\alpha)$ sampling in the Dirichlet distribution α; $z_{i,j}$ the topic of the j-th word position of news item d_i generated with probability $P(z_{i,j}|\theta)$ as sampled; $\varphi_{z_{i,j}}$ the word distribution of topic $z_{i,j}$ generated from probability $P(w|zt;n\beta)$ samples in the Dirichlet $\varphi_{z_{i,j}}$ distribution β; and $w \in W = \{w_1, w_2,\ldots,w_M\}$ a word generated by the sampling from. In the oil market text generation process, the joint probability is defined as follows.

Definition 1. The joint probability of generating oil market text based on the Dirichlet distribution and the polynomial distribution is given by the following equation:

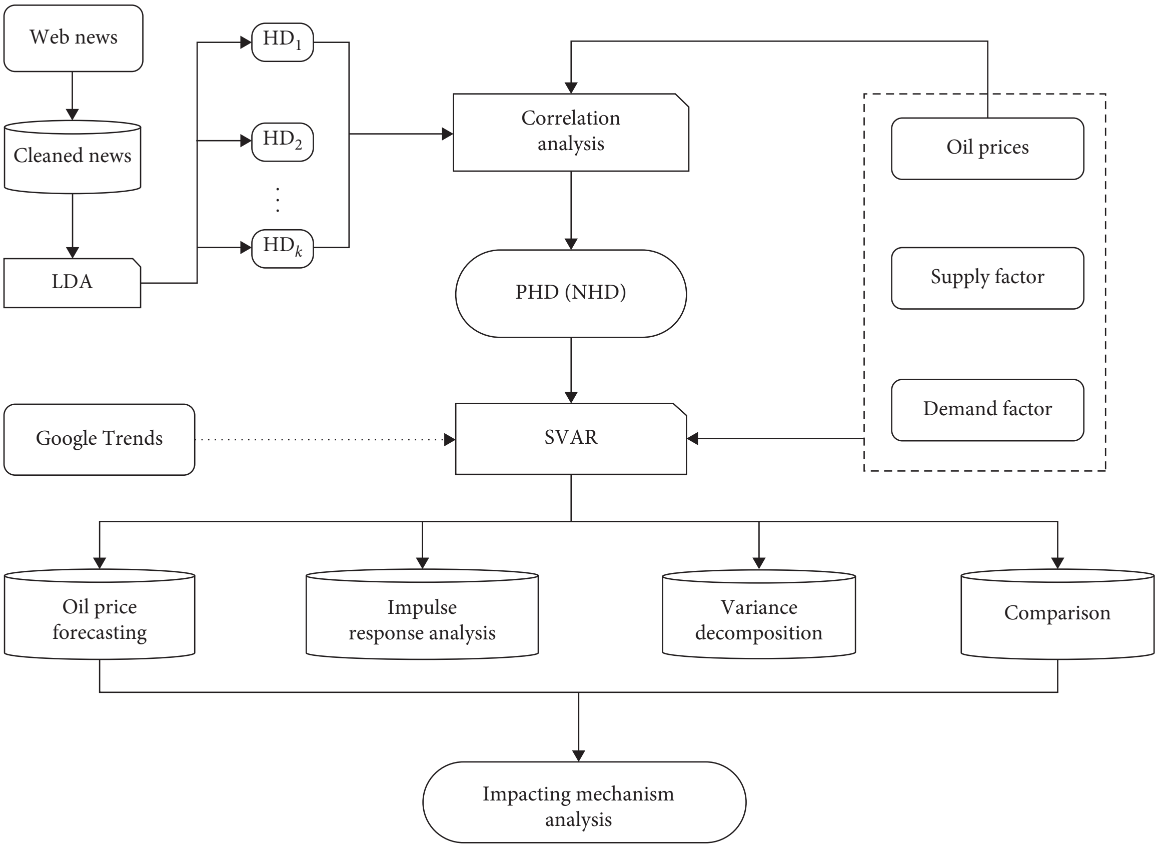


Figure 1: Framework of a HD-based oil price forecasting model.

$$P(w) = \int_{\theta\in\Theta}\left(\prod_{w\in W}\sum_{z\in Z}P(w|zt;n\beta)\right)P(z|\theta))P(\theta;\alpha)\mathrm{d}\theta, \tag{1}$$

where $P(\theta;\alpha)$ is Dirichlet, $P(z|\theta)$ is a polynomial parameterised by θ, $P(w|zt;n\beta)$ is a polynomial over the words.

Since the log-likelihood function of the LDA model contains latent variables and cannot be estimated by simple maximum likelihood estimation method, the Expectation–Maximization (EM) algorithm proposed by Dempster et al. [51] in 1977 is adopted. At this point, the unstructured and semistructured data can be structured, and the optimal value of $p(w|z)$ and $p(z|\theta)$ is obtained. Then, the topic of each text and the keywords under each topic can be determined according to the probability value.

3.2. Oil Market Hot Degree Extraction Method. Based on the value of $p(w|z)$ and $p(z|\theta)$, the probability that each text corresponds to each topic and each topic corresponds to each word is obtained. Next, we propose the definition of topic hot degree based on the changes of the topic over time.

Definition 2. At time t, the probability of each text corresponding to each topic is summed and divided by the total number of texts to get the hot degree of each topic at this moment, as defined in the following equation:

$$\begin{gathered}\mathrm{HD}_{tj} = \frac{\sum_{i=1}^{n} p_{ij}^{t}}{n},\\ p_{ij}^{t} = \begin{cases} p(z_j|d_i^t), & p(z_j|d_i^t) > w,\\ 0, & \text{else},\end{cases}\end{gathered} \tag{2}$$

where HD_{tj} is the hot degree of topic j at time t, n is the number of texts, and $p(z_j|d_i^t)$ is the probability of text d_i^t corresponding to the topic z_j at time t; the smaller the value is, the less the content of the text is related to the topic. In addition, the threshold w is set here to avoid invalid information, when the probability which is less than the value of w does not participate in the calculation. On the whole, HD_{tj} forms a time series HD_t in continuous time, which in turn quantifies the Internet information. The higher the HD_t, the more texts related to the topic, that is, the Internet attention of the topic is higher. On this basis, we give the definition of positive and negative hot degree.

Definition 3. Calculate the linear correlation coefficient between crude oil prices and HD_j $(j = 1, 2, \ldots, k)$ obtained from Definition 2. The topic hot degree is divided into positive topic hot degree set Φ_{phd} and negative topic hot degree set Φ_{nhd}, as defined in the following equation:

$$\begin{cases} \Phi_{\text{phd}} = \left\{ \text{HD}_j, \text{corr}(\text{HD}_j, \text{price}) > = \lambda \right\}, \\ \Phi_{\text{nhd}} = \left\{ \text{HD}_j, \text{corr}(\text{HD}_j, \text{price}) < = -\lambda \right\}, \end{cases} \quad (3)$$
$$j = 1, 2, \ldots, k,$$

where corr $(\text{HD}_j, \text{price})$ is the correlation coefficient between HD_j $(j = 1, 2, \ldots, k)$ and crude oil price; λ is the threshold value to ensure the relevance of the data; and k is the number of topics.

Next, perform a principal component analysis (PCA) on the topical hot degree in the set Φ_{phd}, and obtain the first principal component. Then, the positive hot degree (PHD) is obtained within the range of 0–100 by the maximum and minimum normalisation; this method is very similar to the definition of Google Trends. The definition of negative hot degree (NHD) is the same as PHD, as shown in equation (4). At this point, two indicators, PHD and NHD, are used to characterise the tendency of the news to the trend of oil prices, and the two are collectively referred to as HD:

$$\begin{cases} \text{PHD} = \text{MinMaxScale}(\text{PCA}(\Phi_{\text{phd}}, 1)) * 100, \\ \text{NHD} = \text{MinMaxScale}(\text{PCA}(\Phi_{\text{nhd}}, 1)) * 100, \end{cases} \quad (4)$$

where PCA stands for principal component analysis and the number "1" stands for obtaining the first principal component.

On the basis of the definition of joint probability, topic hot degree, and HD tendency, the PHD and NHD extraction algorithm for the massive Internet information about the oil market is as follows.

3.3. Crude Oil Price Forecasting Based on Hot Degree. To explore the relationship between Internet information and oil price fluctuations, we define a vector $Y_t = (\text{Supply}_t, \text{Demand}_t, \text{Web\_Index}_t, \text{Price}_t)$, where $\text{Supply}_t, \text{Demand}_t$ represents the supply and demand factors of the oil market, Web_Index_t is the hot degree factor proposed in this paper, and Price_t is the international crude oil prices. Compared with the traditional VAR model, the SVAR model can capture the contemporaneous correlation between variables and can reflect the response of the model system to the independent perturbation shock, thus better explaining the fluctuation of oil prices. Therefore, we define the SVAR model related to the oil market as follows:

$$B_0 Y_t = \beta + \sum_{i=1}^{p} B_i Y_{t-i} + u_t, \quad (5)$$

where B_0, β, B_i are the vector parameters to be estimated, p is a lag order, and u_t is a structural innovation. Assuming that B_0 is invertible, the SVAR model is simplified to the following equation:

$$Y_t = B_0^{-1}\beta + \sum_{i=1}^{p} B_0^{-1} B_i Y_{t-i} + \varepsilon_t, \quad (6)$$

where ε_t represents the residual vector of the simplified SVAR model and $\varepsilon_t = B_0^{-1} u_t$ (see equation (7)). According to Kilian and Lee [52], the restrictions on B_0^{-1} mean that it is in lower triangular form:

$$\varepsilon_t = \begin{bmatrix} \varepsilon_t^{\text{Supply}} \\ \varepsilon_t^{\text{Demand}} \\ \varepsilon_t^{\text{Web\_index}} \\ \varepsilon_t^{\text{Price}} \end{bmatrix} = B_0^{-1} u_t = \begin{bmatrix} b_{11} & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 \\ b_{31} & b_{32} & b_{33} & 0 \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} \begin{bmatrix} u_t^{\text{Supply}} \\ u_t^{\text{Demand}} \\ u_t^{\text{Web\_index}} \\ u_t^{\text{Price}} \end{bmatrix}, \quad (7)$$

where b_{ij} represents the response coefficient of the ith variable's response to the jth variable's structural shock; the larger the coefficient, the greater the impact on the whole system; and 0 means that the current position has no response to a specific impact.

Next, we will explore the impact of hot degree (HD) on oil prices through the impulse response function (IRF) and variance decomposition (VD) of the SVAR model. The IRF is used to calculate the response of the whole system when the error term of the Internet information changes. The VD is used to analyse the contribution (measured by variance) of each structural shock to oil price changes and further evaluate the importance of different shocks.

4. Empirical Analysis

4.1. Data Sources. The sources of Internet news are uneven, including social networking sites and new media, but scientific research needs to ensure the security, normalisation, and universality of information sources. Therefore, this paper uses "oil," "oil price," "oil market," "crude oil," "OPEC," "WTI," and "Brent" as keywords to crawl the news published by UPI, Reuters, Oil price, and World oil which are authoritative online media as the source of oil market Internet information, and the key technology is Python's selenium framework. The total number of news items is 220,362; after text preprocessing and information filtering, we finally obtained 88,763 oil market-related news items from January 2012 to June 2019.

Meanwhile, the SVAR model adopted here considers four variables: in addition to the HD extracted from the Internet in this paper, it also includes global oil supply, global oil demand, and Brent crude oil prices, which are all monthly data. Among them, oil supply is represented by global oil production; on the demand side, the Purchasing Manager Index (PMI) has become an important evaluation indicator of world economic operations and a barometer of world economic changes [53]. Accordingly, oil demand is represented by PMI. The specific sources are shown in Table 2. In addition, the samples are split into two subsets when forecasting, with the data from January 2012 to December 2018 being regarded as the training set and the data from January 2019 to July 2019 forming the test set.

4.2. Topic Generating of Oil Market. Firstly, text preprocessing is undertaken, including four steps of removing invalid text, abnormal vocabulary, stop words, and word form conversion. Specifically, due to network connection faults and other reasons, some news is empty that needs to be

Table 2: Variable description.

Name	Meaning	Source	Unit	Time interval
Supply	Global oil production	Wind database	Thousand barrels per day	2012.01–2019.06
Demand	Purchasing managers index	Wind database		
Price	Brent oil prices	https://www.eia.gov/dnav/pet/hist/RBRTED.htm	USD per barrel	

deleted. Internet news obtained at a first pass is prone to containing garbled, or abnormal, characters. To avoid interference with information quality, it is also necessary to remove such erroneous data. Different word forms increase the time complexity of the model and require conversion. Stop words are a relatively complex part of the data: stop words in the oil market text data include not only general stop words (a, an, the, etc.) but also a large number of words that are less relevant to the oil market (year, time, week, etc.), thereby interfering with the results. Through the analysis of the preliminary results, 34 words without specific information other than the common stop words are added to the stop words lists to form a dedicated stop words dictionary of the oil market.

In addition, the number of topics is the key factor in determining the topic extraction effect of LDA model. Based on the literature [54–56], the number of topics k is valued as 5, 8, 10, 12, and 15, through the repeated execution of Algorithm 1, combined with the perplexity of the LDA model; the results show that the optimal number of topics is 8. We filter the 50 keywords most similar to the topic to illustrate the meaning of each topic and select 10 words that have practical significance. The final topic generation effect is shown in Table 3.

Taking Topic 8 as an example, it includes keywords such as Iran, attack, and sanction. It can be speculated that the topic is related to the Middle East situation. Overall, the news topics include market economy, exploration and development, government intervention, and military war. These topics are closely related to the composition of the oil market, and this is basically consistent with the factors influencing the oil market proposed by Miao et al. [6] and Huang et al. [9] which corroborates the effectiveness of the news topics extracted in this paper. But the meaning of each topic is pluralistic, and it is difficult to summarise it with simple words.

4.3. Hot Degree Extraction of Oil Market. For the news-topic probability output matrix of LDA model, based on Definition 2, the probability threshold w is set to be 0.1, 0.2, and 0.3, respectively. After executing Step 5 of Algorithm 1, the HD of each topic is obtained in monthly units of time. The correlation coefficient between $\text{HD}_j\,(j = 1, 2, \ldots, k)$ and Brent oil prices is calculated. It can be seen from Figure 2, and the correlation is the strongest when w is set to 0.1. Looking at Figure 2, the deeper the circle colour and the larger its area, the stronger the correlation. It can be found that the correlation between the hot degree of each topic is small, and the numerical value remained between 0.03 and 0.65, indicating that the information contained in each topic is basically independent of other items; it also verified that the topic clustering of the LDA model is better.

Looking at the first column of Figure 2 to explore the correlation between crude oil prices and topic hot degree, we can find that $\text{HD}_1, \text{HD}_4, \text{HD}_6$ have a negative correlation with crude oil prices, while $\text{HD}_2, \text{HD}_3, \text{HD}_5, \text{HD}_7, \text{HD}_8$ have a positive correlation with oil prices. According to Definition 3, set $\lambda = 0.4$; then, $\Phi_{\text{phd}} = \{\text{HD}_2, \text{HD}_3, \text{HD}_7, \text{HD}_8\}, \Phi_{\text{nhd}} = \{\text{HD}_1, \text{HD}_4, \text{HD}_6\}$. In order to avoid the problem of too many parameters failing in the estimation of the SVAR model, based on equation (4), the PHD and NHD are obtained through PCA and maximum and minimum normalisation to represent the hot degree of the oil market. The comparison between the two and the trend of oil prices is shown in Figure 3.

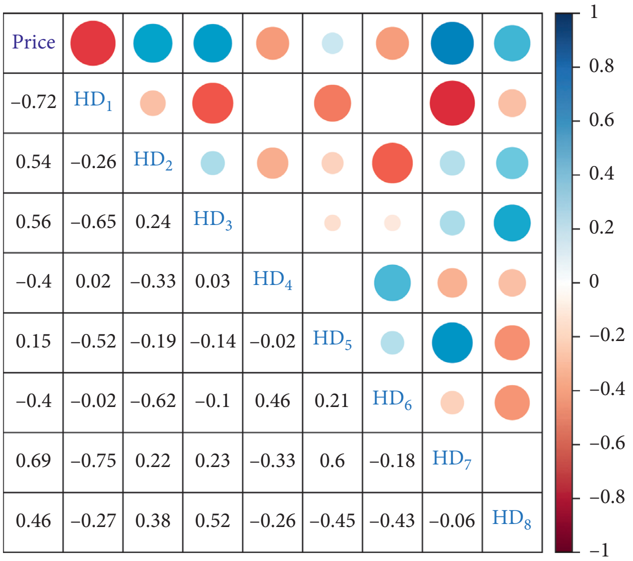
It can be found from Figure 3 that there is a clear codirectional relationship between crude oil prices and PHD. In 2012, oil prices were at a high level, and PHD was also at high levels. In 2014, oil prices plummeted and PHD also fell to low levels. The NHD shows exactly the opposite effect. Next, the dynamic relationship between crude oil prices and HD will be captured in detail through the SVAR model.

4.4. Analysis of the Interactive Relationship between Oil Price and Web Information Index. The input vector of the SVAR model is $Y_t = (\text{Supply}_t, \text{Demand}_t, \text{Web\_Index}_t, \text{Price}_t)$, and Web_Index_t is represented by PHD and NHD, respectively. Before developing model estimation, data preprocess is performed first, including deflation processing and seasonal adjustment of Brent crude oil prices, seasonal adjustment of supply and demand factors, and finally taking the logarithm uniformly. Next, the unit root test is performed on each variable by using the Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests. The results are shown in Table 4. It can be found that all variables are the first-order stationary. Consequently, we use the method proposed by Toda and Yamamoto [44] to estimate the SVAR model. As long as certain conditions are met ($n \geq d_{\max}$, n is the optimal lag for the VAR model, $d_{\max}$ is the maximum single integer order of variables), the model of variables in levels can be established. Moreover, using variables in levels facilitates the capture of long-term information and enhances the explanatory ability of the model [57].

Then, Web_Index_t is represented by PHD first, and the optimal lag order of VAR is determined by considering multiple criteria. The results are shown in Table 5. Among them, Final Prediction Error (FPE), Akaike Information Criterion (AIC), Schwarz Criterion (SC), and Hannan–Quinn (HQ) criteria show that the optimal lag order is 1, and the

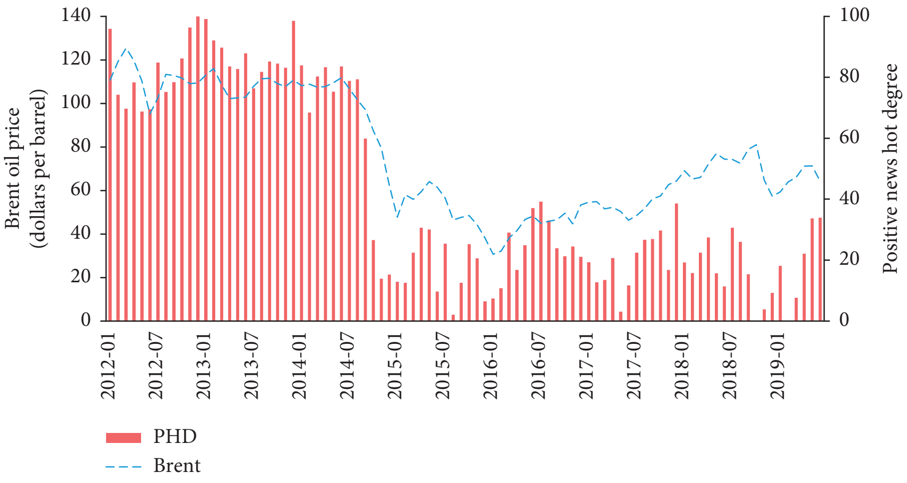
Step 1: Use crawler technology to obtain massive Internet information related to the oil market and then preprocess the Internet information, including details such as removing invalid text, filtering abnormal vocabulary, removing stop words, and converting word form.
Step 2: Vectorise the cleaned oil market text. First, all the words appearing in the texts constitute a dictionary. If the frequency of a word appearing in text d is i, the position of this word is recorded as i; otherwise, it is recorded as 0. Based on this, text d becomes a vector, and all the texts form a word frequency matrix.
Step 3: Select the appropriate number k of topics, use the EM algorithm to estimate the joint probability distribution of text-word, get the probability of $p(z|\theta)$ and $p(w|z)$, and determine 50 words most relevant to each topic according to the probability value to define the realistic meaning of the topic.
Step 4: Investigate whether, or not, the text topic is reasonable and effective. If there is more redundancy in the information or the meaning of topic is ambiguous, repeat Steps 1 to 3 until the topic is reasonable and effective, and the model confusion is small. After meeting the above conditions, output the current topic and get influence factors affecting the oil market.
Step 5: Calculate the heat $\mathrm{HD}_j (j = 1, 2, \ldots, k)$ corresponding to oil market text topics based on equation (2), and realize the quantification of Internet information.
Step 6: Calculate the topic hot degree set Φ_{phd} and Φ_{nhd} based on equation (3), and obtain the value of PHD and NHD based on equation (4). So far, the indicators of news reports on crude oil prices have been extracted.

Algorithm 1: Hot degree extraction algorithm.



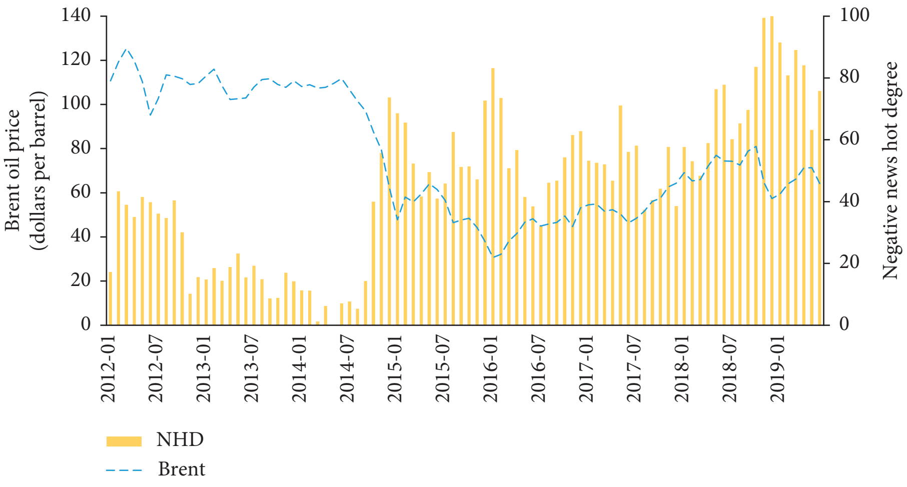
Price								
−0.72	HD_1							
0.54	−0.26	HD_2						
0.56	−0.65	0.24	HD_3					
−0.4	0.02	−0.33	0.03	HD_4				
0.15	−0.52	−0.19	−0.14	−0.02	HD_5			
−0.4	−0.02	−0.62	−0.1	0.46	0.21	HD_6		
0.69	−0.75	0.22	0.23	−0.33	0.6	−0.18	HD_7	
0.46	−0.27	0.38	0.52	−0.26	−0.45	−0.43	−0.06	HD_8

Figure 2: Correlation coefficient between oil prices and HD. Note: “price” represents Brent oil prices and “HD_j” represents the hot degree of topic j.



(a)

Figure 3: Continued.



(b)

Figure 3: Comparison of Brent oil prices and HD. (a) The trend of Brent oil prices and PHD. (b) The trend of Brent oil prices and NHD.

Table 3: The keywords of topics related to oil market news.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Saudi	pipeline	trump	fuel	rig	game	offshore	Iran
futures	Canada	house	power	permian	win	Russia	attack
Dollar	import	federal	change	technology	season	exploration	sanction
Bank	fuel	tariff	enlarge	shale	play	block	Saudi
Arabia	refinery	administration	coal	operations	team	shell	war
Yield	gasoline	rule	click	offshore	score	sea	force
investors	capacity	public	climate	engineer	pass	Mexico	military
European	ship	tax	plant	system	goal	investment	Iraq
Rate	refineries	job	emissions	basin	run	assets	Venezuela
bond	coast	drug	India	deliver	star	stake	security

Table 4: Results of unit root test.

Unit root test		Supply	Demand	PHD	NHD	Price
ADF	Level	−1.591 (0.789)	−2.893 (0.169)	−2.041 (0.270)	−2.076 (0.255)	−1.347 (0.605)
	First differenced	−12.315 (≤0.001)	−9.743 (≤0.001)	−10.422 (≤0.001)	−10.093 (≤0.001)	−7.742 (≤0.001)
PP	Level	−2.092 (0.543)	−2.968 (0.147)	−2.010 (0.282)	−2.089 (0.250)	−1.347 (0.605)
	First differenced	−12.315 (≤0.001)	−9.753 (≤0.001)	−10.616 (≤0.001)	−10.404 (≤0.001)	−7.699 (≤0.001)

The numbers in the table are the values of the t-statistic, and the values in brackets are p values.

maximum single order of the variables in the SVAR model is 1. According to the principle proposed by Toda and Yamamoto, the VAR lag order selected here is 2.

The SVAR model is estimated on the basis of equations (5) to (7), and the impact matrix is shown in equation (8). Observing the impact of various errors in fluctuations of oil prices, the value of b_{41} is smaller than zero, indicating that the fluctuation of global oil supply has a certain negative impact on oil prices. Meanwhile, b_{42}, b_{43} are positive numbers, meaning the changes in oil demand and PHD have positive effects on oil prices. The increase in production causes oil prices to fall, and the increase in demand leads to an increase in oil prices, which is consistent with our usual perception, and the positive effect of PHD on oil prices is consistent with the results in Figure 3, indicating that the larger the PHD, the more bullish news the media reports, and the higher the prices of oil:

$$\begin{bmatrix} b_{11} & 0 & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 \\ b_{31} & b_{32} & b_{33} & 0 \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} = \begin{bmatrix} 0.0046 & 0 & 0 & 0 \\ 0.7174 & 0.0125 & 0 & 0 \\ -762.7574 & 86.3205 & 9.8726 & 0 \\ -3.8176 & 1.1940 & 0.001 & 0.069 \end{bmatrix}. \tag{8}$$

4.4.1. The Forecasting Effect of HD on Oil Prices. This section forecasts crude oil prices based on the VAR model, and the influence of HD factor on the prediction result will be analysed. The input of the model VAR_0 is $Y_t = (\mathrm{Suppy}_t, \mathrm{Demand}_t, \mathrm{Price}_t)$, the input of the model VAR_p is $Y_t^P = (\mathrm{Suppy}_t, \mathrm{Demand}_t, \mathrm{PHD}_t, \mathrm{Price}_t)$, and the input of the model VAR_N is $Y_t^N = (\mathrm{Suppy}_t, \mathrm{Demand}_t, \mathrm{NHD}_t, \mathrm{Price}_t)$. In addition, as mentioned before, related research uses Google Trends to indicate investor attention of the oil market [17,23,24], analyse its impact on oil price fluctuations, and achieve good results. Google Trends and the HD extracted in this paper are also the products of the Internet era. Based on this, the effect of Google Trends and HD to predict crude oil prices will be compared. We deal with Google Trends in the same way as Yao et al. [23] and obtain Google search volume index (GSVI). Therefore, the input of the model VAR_G is $Y_t^G = (\mathrm{Suppy}_t, \mathrm{Demand}_t, \mathrm{GSVI}_t, \mathrm{Price}_t)$.

The prediction is performed to the next step, and the obtained prediction results are shown in Table 6. It can be found that, regardless of the results of Mean Absolute Error (MAE) or Root Mean Square Error (RMSE), the model's prediction effect with HD is significantly better than the original model, and the model VAR_p reduces the MAE by $ 3.7075 compared to the model VAR_0. This is a major improvement, and the prediction effect of VAR_P is better than VAR_N. It works best in the four models. Unfortunately, the model VAR_G with the addition of GSVI has the worst prediction effect and does not play a role in assisting the prediction. It can be seen that the HD factor extracted in this paper can significantly improve the forecasting effect of oil prices and is significantly better than the auxiliary forecasting ability of Google Trends. Next, we will specifically analyse how the HD affects oil prices through the impulse response function (IMF) and variance decomposition (VD) method.

4.4.2. The Influence Timeliness and Explanation Ratio of PHD Shocks on Oil Price Fluctuations. Based on the estimation results of the VAR_p model, the impulse response of oil prices to other variables' shock is shown in Figure 4. From Figure 4(c), it can be found that, within the sample interval, given the shock of one standard deviation of PHD, the oil price showed a significant positive response in the first period, with a value of 1.45%, and the oil price reached the maximum response (5.92%) in the seventh period. The impact persists for a long time, with the most significant impact during the fifth to ninth periods. In addition, the positive impact of the demand factors represented by PMI also gives a positive response to oil prices in the short term, but the oil prices respond more strongly and last longer to PHD shocks, clearly leading the demand factor. This figure indicates that PHD has better timeliness and a higher impact on oil prices than traditional demand indicators. Moreover, the positive shock of oil supply factors has a certain negative impact on oil price fluctuations, and this effect has been present for a long time.

To further analyse the explanatory ratio of PHD on oil price fluctuations, the VD method was used to investigate the variance of oil prices prediction error. The results are shown in Table 7. It can be seen that the error from oil prices itself is over 80% in the first period; as the forecast period increases, the proportion of errors in the supply, demand, and PHD increases. In the short-term, the variance of oil prices prediction error explained by PHD is 3.72% in the first phase, which gradually increased later, and, in the fifth period, the proportion of oil price fluctuations explained by PHD's shocks reaches 28.63%, ranking first among all variables. In the long term, PHD's ability to explain oil price fluctuations still ranks the first in the factors listed in this paper. Therefore, it can be considered that PHD has a strong ability to explain oil price fluctuations, while the demand factor interpretation ability represented by PMI is at a relatively low level, and the supply factor interpretation ability has shown a steady increase.

In summary, the results of IMF and VD both show that the impact of PHD's shocks on oil price has exceeded traditional supply and demand factors. In reality, according to the response cycle and explanatory ratio of crude oil prices to the shock of PHD, it can help policy-makers and investors to reasonably arrange corresponding countermeasures, know the turning point of the event in advance, and make decisions in a timely manner.

4.4.3. The Influence Timeliness and Explanation Ratio of NHD Shocks on Oil Price Fluctuations. Based on the estimation results of the VAR_N model, the impulse response of oil prices to other variables' shock is shown in Figure 5. From Figure 5, given the impact of one standard deviation of NHD, the oil price showed a significant negative response in the first period, and this once again validates our hypothesis. Meanwhile, this impact will probably last for 7 periods, with the most significant impact in periods 1–3, peaking at 3.70% in the second period. Compared with PHD, the impact of NHD on oil prices is more rapid and disappears faster. This also shows that news events that have a suppressing effect tend to have shorter timeliness. In addition, the response of oil prices to the impact of supply factors and demand factors is similar to that in Figure 4 and is consistent with reality and has good stability.

Under the current model, the variance decomposition results of crude oil price prediction errors are shown in Table 8. Similar to the impulse response results, in the short term, NHD shows good performance with a peak of 12.19%. In the long term, NHD's explanation ratio to oil price fluctuations is declining, but it is still superior to demand. In addition, the supply factor's ability to explain price fluctuations is growing steadily, maintaining first in the factors listed in the VAR_N model.

Through the above research, it can be found that the PHD and NHD indicators extracted from web news show better performance in the SVAR model, which is significantly better than the demand factor, and the PHD factor also surpasses the supply factor, which has the largest impact on oil price fluctuations. It answers the second question raised at the beginning of this paper. A third question will be considered next. Can PHD better explain fluctuations in oil prices than Google Trends?

Table 5: Model VAR lag order selection.

Lag	LR	FPE	AIC	SC	HQ
1	NA	0.0000*	8.7636*	8.3039*	8.5787*
2	13.9418	0.0000	−8.5683	−7.6487	−8.1984
3	29.6396*	0.0000	−8.5978	−7.2185	−8.0430
4	15.0406	0.0000	−8.4393	−6.6002	−7.6996
5	13.5677	0.0000	−8.2716	−5.9726	−7.3469

*Lag order selected by the criterion.

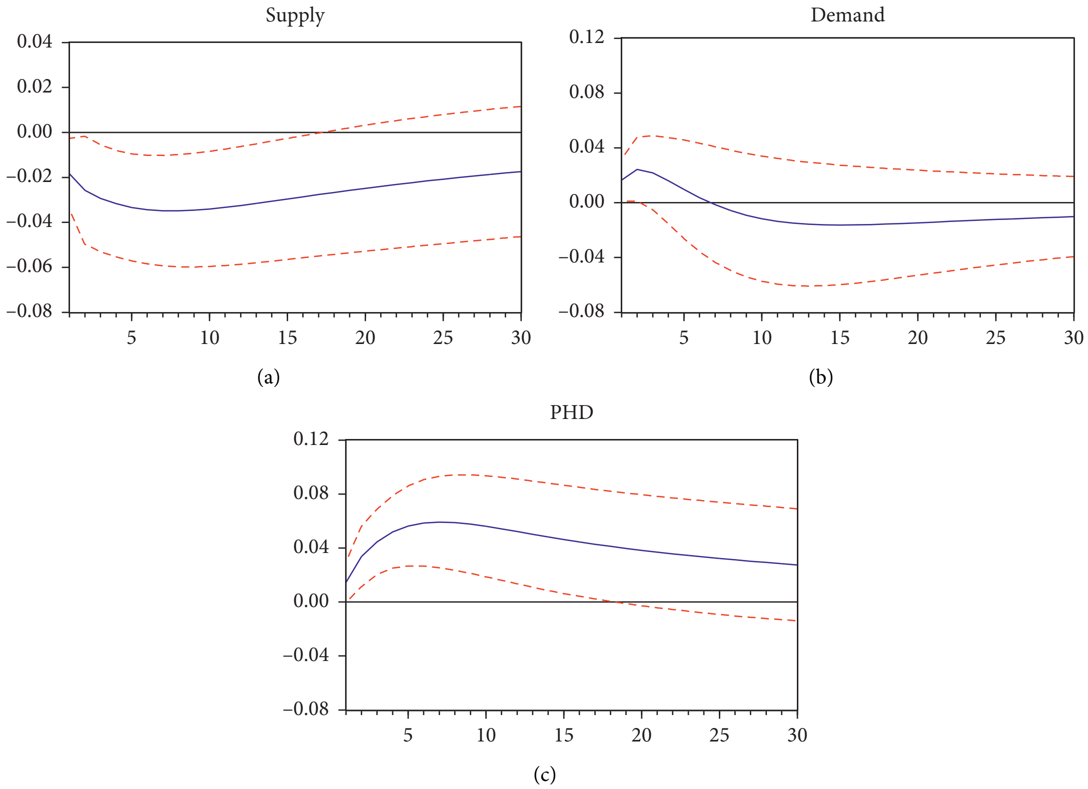


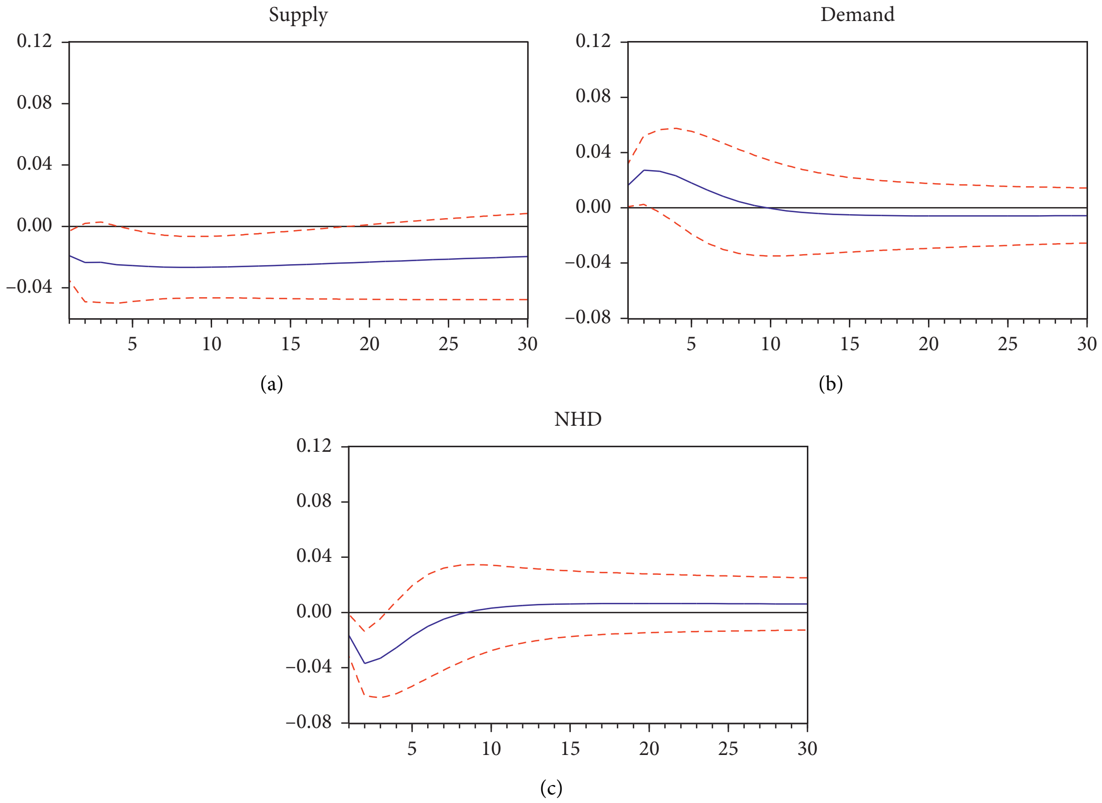
Figure 4: Responses of Brent oil prices to different shocks based on VAR_P. The blue solid lines represent the mean impulse response of Brent oil prices and the red dotted lines denote two standard deviations from the mean. The response period is 30 months. (a) Response of Brent oil prices to a supply shock. (b) Response of Brent oil prices to a demand shock. (c) Response of Brent oil prices to a PHD shock.

Table 6: Model forecasting results.

Model	VAR_0	VAR_p	VAR_N	VAR_G
MAE	5.8798	2.1723	3.0048	7.2573
RMSE	6.5546	2.7979	3.2938	7.4034

4.4.4. The Impact of Google Trends on Oil Price Fluctuations and Comparison. The construction of GSVI based on Google Trends was explained in section (1), and an oil prices forecasting model was made. This section will compare the impact of Google Trends and PHD on oil prices in various aspects. First, the trend comparison effect of Brent oil prices, PHD, and GSVI is shown in Figure 6. It can be seen that the GSVI has a large fluctuation range and has two peaks. The overall trend is negatively correlated with crude oil price. In order to better understand the causal relationship between variables, Granger causality tests were performed on GSVI, PHD, and Brent oil prices. The results are shown in Table 9. Obtaining the Brent oil prices is the Granger cause of the GSVI change, and PHD is the Granger cause of the Brent crude oil price change, which indicates that PHD has a better interpretation of oil prices than GSVI. From the practical point of view, people often have search behaviors due to changes in oil prices, which in turn causes fluctuations in GSVI. On the other hand, changes in oil prices often come from news reports.

Based on the VAR_G model's estimation results, the impulse response and variance decomposition results corresponding to GSVI are obtained. The comparison of IMF and VD results with PHD is shown in Figure 7. According to Figure 7(a), it can be found that, for the shocks of one standard deviation of GSVI, oil prices responded in the first period, similar to the trend of NHD shocks, and the response reduction caused by GSVI was slower. But the positive response brought by PHD is more obvious, both the



(a)

(b)

(c)

Figure 5: Responses of Brent oil prices to different shocks based on VAR_N. The blue solid lines represent the mean impulse response of Brent oil prices and the red dotted lines denote two standard deviations from the mean. The response period is 30 months. (a) Response of Brent oil prices to a supply shock. (b) Response of Brent oil prices to a demand shock. (c) Response of Brent oil prices to a NHD shock.

Table 7: Results of variance decomposition of Brent oil prices based on VAR_P (%).

Period	Supply	Demand	PHD	Price
1	6.02	4.80	3.72	85.46
5	12.41	5.26	28.63	53.70
10	16.69	3.28	43.68	36.35
15	18.45	3.88	47.82	29.86
20	19.23	4.52	49.11	27.14
25	19.62	4.88	49.72	25.79
30	19.83	5.07	50.10	25.00

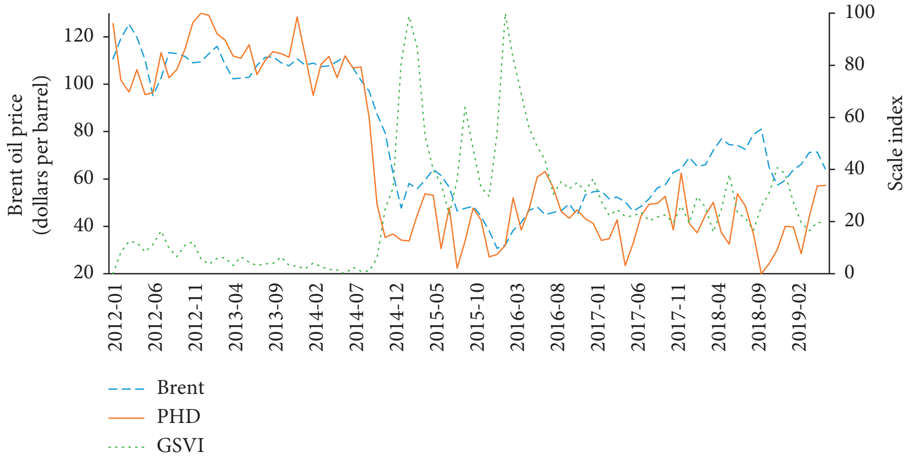


Figure 6: Comparison of oil prices, PHD, and Google trends.

Table 8: Results of variance decomposition of Brent oil prices based on VAR_N (%).

Period	Supply	Demand	NHD	Price
1	6.21	4.59	4.67	84.52
2	6.44	7.10	11.52	74.94
3	6.52	7.61	12.19	73.67
4	7.13	7.68	11.57	73.62
5	7.92	7.43	10.63	74.01
10	12.53	5.67	7.65	74.15
15	16.19	4.92	6.71	72.18
20	18.83	4.65	6.31	70.21
25	20.79	4.53	6.12	68.56
30	22.29	4.49	6.01	67.22

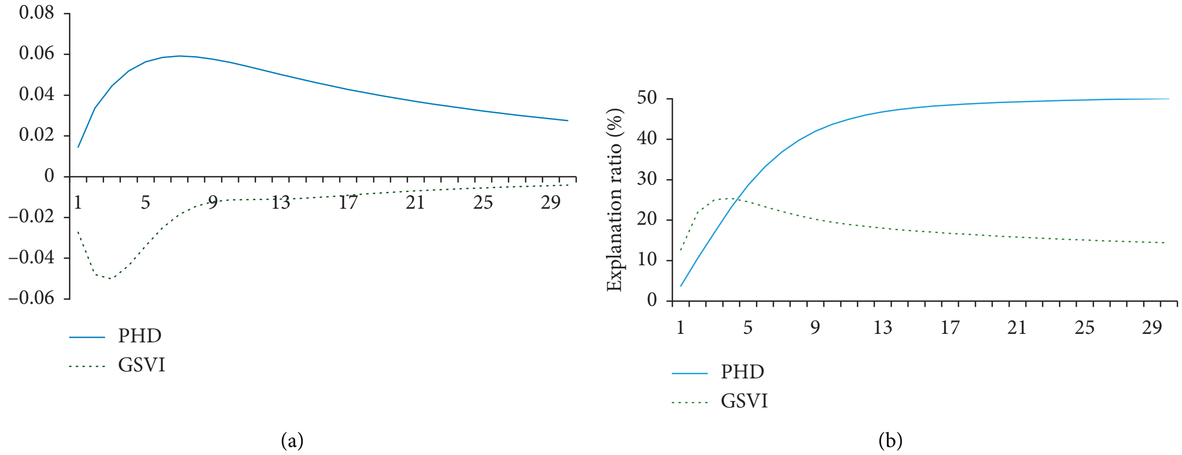


Figure 7: Comparison of the impact of PHD and GSVI on oil prices based on IMF and VD. (a) The response of oil prices based on IMF. (b) Comparison of explanation ratio based on VD.

maximum impulse response and the cumulative impulse response are better than GSVI. According to Figure 7(b), from the perspective of VD, the contribution ratio of GSVI in the early stage to the crude oil price was relatively high, reaching a peak at 25.39%, and gradually stabilized in the later stage. The contribution rate of PHD shocks to oil prices has shown a steady increase, much higher than GSVI. Based on this, PHD shows better performance than GSVI, whether it is predictive power, causality, or explanatory ratio, which further validates the effectiveness of the PHD index proposed in this paper. It also provides new ideas for the current research referring to Google Trends.

4.5. Robustness Analysis. In order to ensure the robustness of the experimental results in this paper, we tested the model based on three aspects: transforming the sample interval, estimating the DCC-GARCH model of dynamic relationship testing, and using a different benchmark of crude oil prices.

4.5.1. Analysis of the Influences of HD on Oil Prices during Different Sample Intervals. Considering the sample range of 2012–2019 is a relatively long period. Thus, the possibility of structural changes cannot be fully ruled out. This section will discuss the possibly different influences of HD on oil prices during different sample intervals.

Firstly, we define the concept of time window: L is called the length of the time window, the sample is divided into a finite number of subintervals using L, s is called the moving step of the window, so each subinterval will move s steps and form a new time window. The schematic diagram is shown in Figure 8.

According to the above definition, the sample range from 2012 to 2019 is divided into 7 time windows, with 48 months as a window for rolling. The prediction is performed to the next step, that is, the window length $L = 48$, and the window moving step length is 6 months; in the other words, $s = 6$. The range of the first window is from January 2012 to December 2015, and the test set is 6 months after the training set. The last window is from January 2015 to December 2018, and the test set is from January 2019 to June 2019. Taking PHD as an example, the obtained prediction results are shown in Table 10.

It can be found that, as the time window gets closer, the predicted error becomes smaller. The possible reason is that with the globalization of the Internet, the number of news is gradually enriched and the timeliness of news reports has increased, which is in line with practical significance. Next,

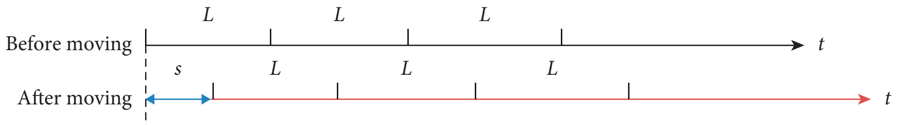


Figure 8: Schematic diagram of the time window concept.

Table 9: Result of Granger causality test.

Null hypothesis	F-statistic	Prob.
Brent does not Granger cause GSVI	6.4332	0.0025**
GSVI does not Granger cause Brent	1.7641	0.1777
Brent does not Granger cause PHD	0.6493	0.5251
PHD does not Granger cause Brent	6.4421	0.0025**

**Significance at 1%.

we will analyse specifically how HD affects crude oil price trends through impulse response analysis and variance decomposition in different windows.

Based on the estimated results of the model, the impulse response of crude oil prices to PHD shocks is shown in Figure 9. In different time windows, given a shock of one standard deviation of the PHD, the oil price has shown a clear positive response in the current period and has persisted for a long time. It is worth noting that between January 2012 and December 2015 in the sample window, oil prices had a negative response to the impact of PHD. This may be due to the US shale oil and gas revolution in 2014, and the fundamental factors of supply are severely larger than demand, leading to global oil prices plummet; even if news that is beneficial to oil prices appears, it will not cause oil prices to rise.

In order to further analyse PHD's ability to explain crude oil price fluctuations, the variance decomposition method is used to analyse the variance composition of crude oil price forecasting errors as shown in Figure 10. Under different time windows, with the extension of the period, the proportion of the influence of PHD gradually increases, and after the fifth period, it slowly decreases and tends to be flat. The same difference is that from January 2012 to December 2015, due to the influence of political events, PHD's ability to explain oil price fluctuations gradually increased after the 9th period.

In summary, the results of impulse response and variance decomposition show that, in the absence of major events, the impact of PHD in different time windows has similar effects on oil prices and forecast errors. It also illustrates the stability of the model in this paper and the validity of the proposed HD index. In reality, based on the response cycle and interpretation ratio of crude oil prices to HD shocks, it is helpful for policymakers and investors to reasonably arrange corresponding countermeasures, know the turning point of the event in advance, and make timely decisions.

4.5.2. The Dynamic Correlations between Oil Prices and HD Variables Based on the DCC-GARCH Model. In addition to the SVAR analysis, there are also other models that can capture the possible dynamic correlations between oil prices and HD variables [58]. To verify the robustness of the model results, we estimate the DCC-GARCH model and the results of the dynamic correlation between oil prices and relevant factors (HD, supply, and demand) are shown in Figure 11.

Firstly, it can be seen from the size of the dynamic correlation coefficient that the correlation coefficients of PHD and Brent are distributed between (0.1, 0.4), which is positively correlated with oil prices and can reach up to 0.4. The correlation coefficient of NHD and Brent is (−0.75, −0.15) and the highest negative correlation can reach 0.75, even exceeding the negative influence of the famous supply factor on Brent.

Secondly, it can be seen from the sign of the correlation coefficient that the correlation between PHD and Brent during the entire sample period is positive. When it shows that the larger the PHD, the more news media reports on good news, which is conducive to the upward movement of oil prices; while the correlation between NHD and Brent is negative, it means that NHD has a significant negative spillover effect on Brent, which is in line with expectations.

Thirdly, it can be seen from the trend of the path diagram of the correlation coefficient that the correlation coefficient between HD and Brent fluctuates more frequently, which means that Brent is more sensitive to HD fluctuations, and small fluctuations in HD will quickly cause Brent to change, but Brent's response to supply and demand was relatively flat. From the above three conclusions, we can conclude that the DCC-GARCH model and the VAR model used in this paper have reached a consistent conclusion.

4.5.3. Analysis of the HD's Ability to Explain the WTI Crude Oil Market. Compared with the Brent crude oil market, WTI crude oil market also occupies an important position in the international trading market. The price range and fluctuation trend of the two markets are different due to differences in their own attributes. To explore the explanatory power of HD for the WTI market, the Price_t series of $Y_t = (\text{Supply}_t, \text{Demand}_t, \text{Web\_Index}_t, \text{Price}_t)$ is replaced by the WTI crude oil prices, and the SVAR model is reestimated. The impulse response results of WTI oil prices to PHD and NHD shocks are shown in Figure 12. According to Figure 12(a), the shocks of PHD also have a positive effect on WTI oil price. The impact has gradually weakened from the second period to the 25th period, and the impact from the third period to the sixth period is the most significant. Figure 12(b) shows that NHD has a suppressive effect on WTI oil price, which is completely consistent with the results of Brent oil prices, further clarifying that HD still has the same cycle and effect on the WTI crude oil market.

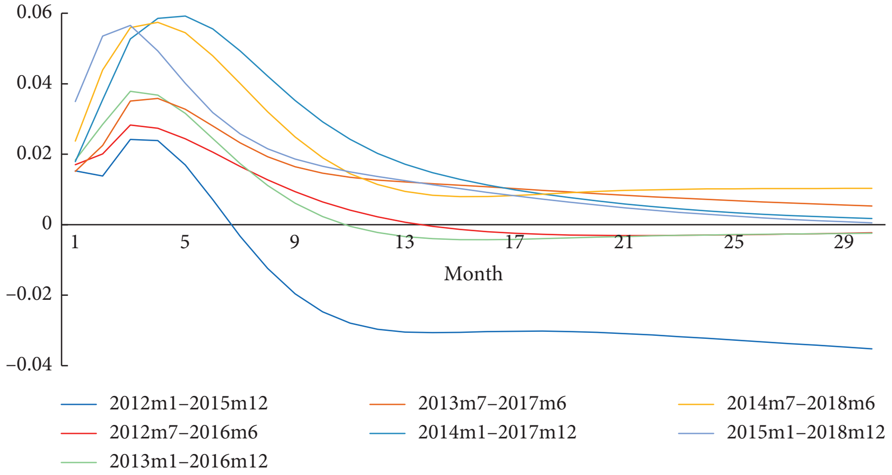


Figure 9: Response of Brent oil prices to a PHD shock.

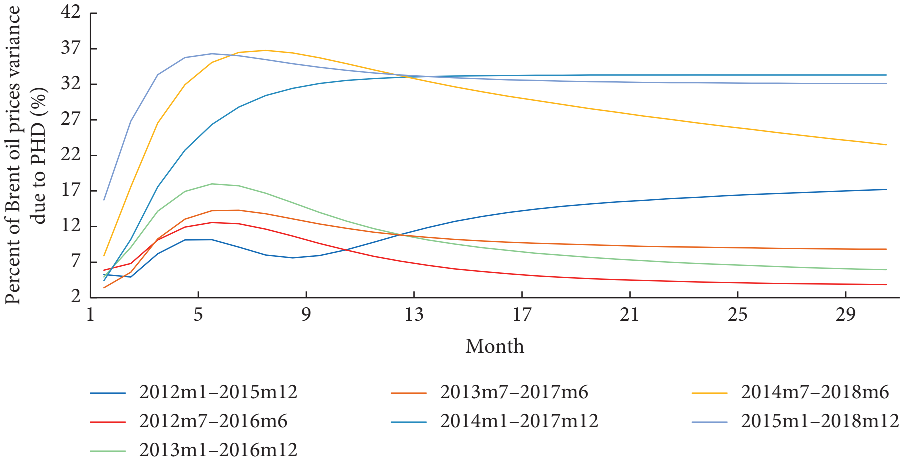


Figure 10: Percent of Brent oil prices variance due to PHD.

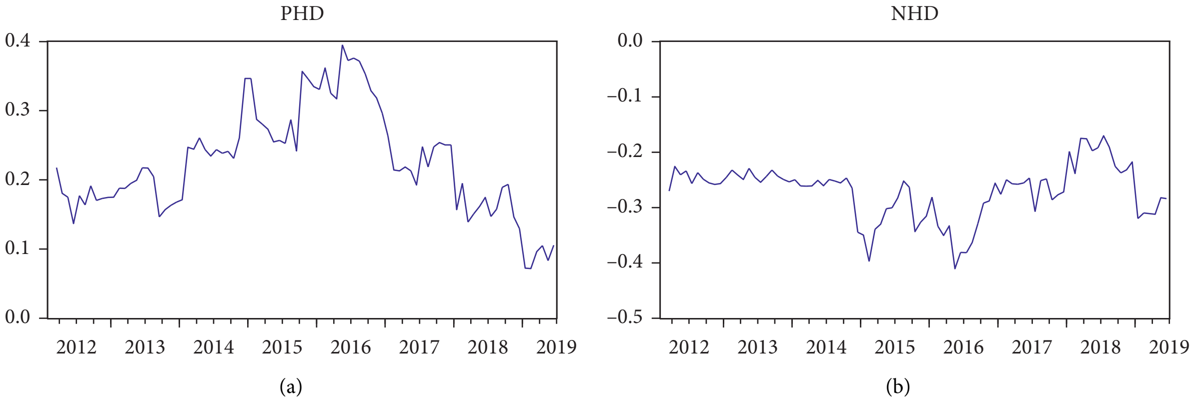
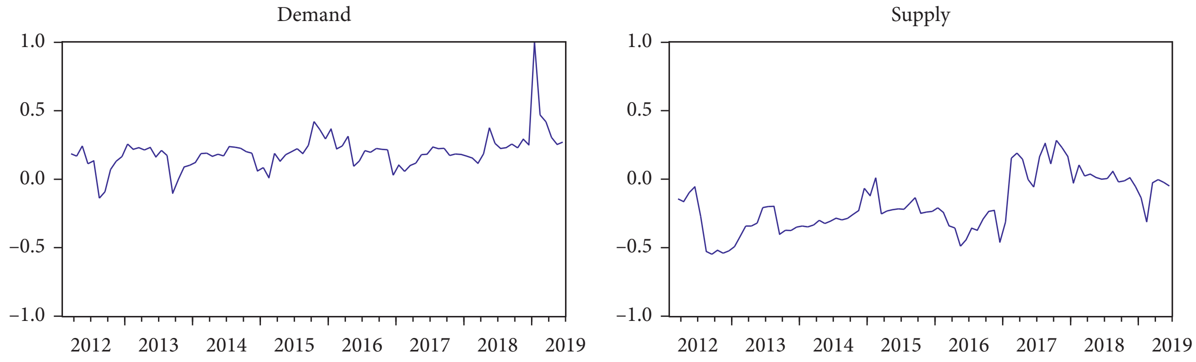


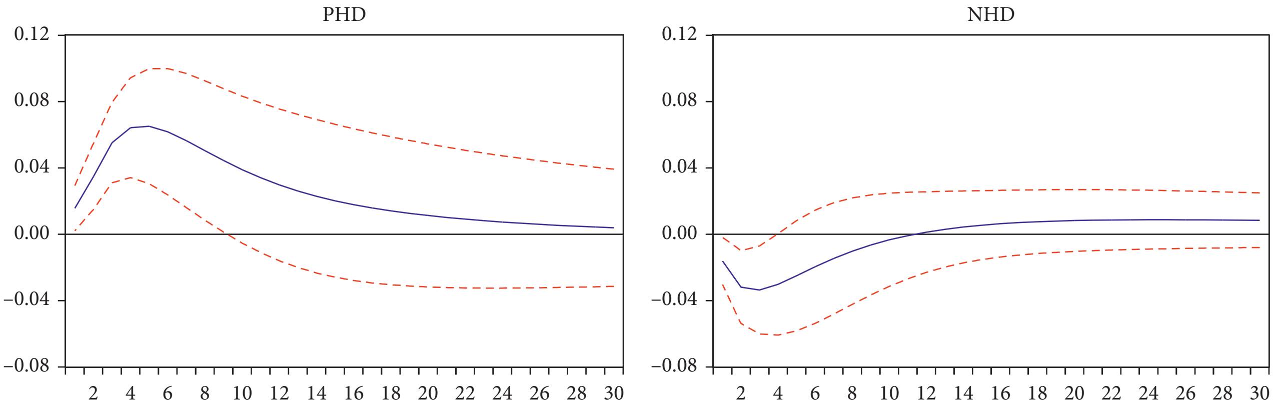
Figure 11: Continued.



(c)

(d)

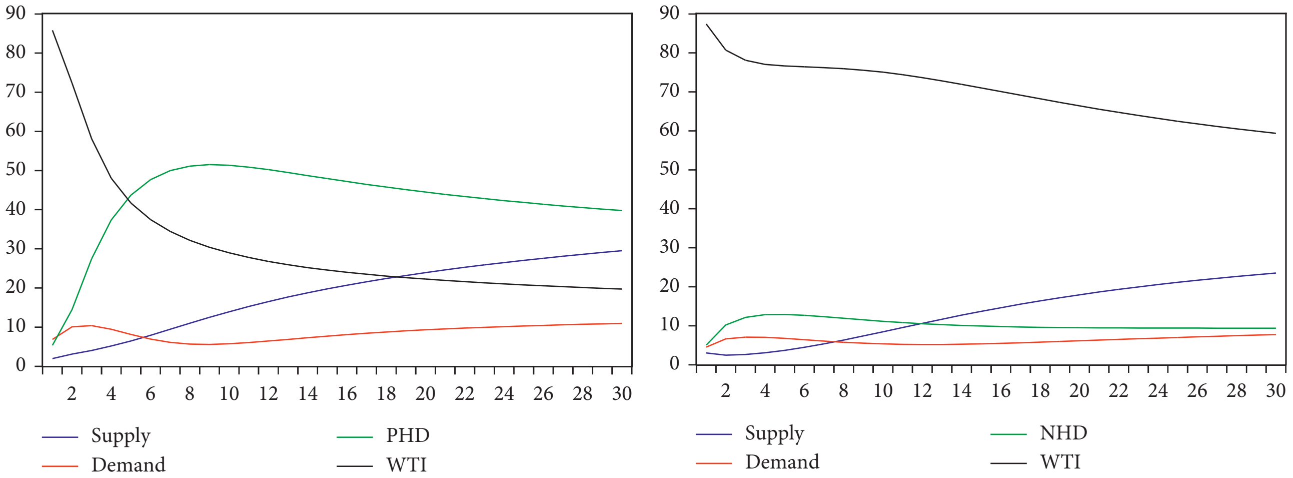
Figure 11: Dynamic correlation series from DCC-GARCH model. (a) Dynamic correlation coefficient of Brent and PHD. (b) Dynamic correlation coefficient of Brent and NHD. (c) Dynamic correlation coefficient of Brent and Demand. (d) Dynamic correlation coefficient of Brent and Supply.



(a)

(b)

Figure 12: Responses of WTI oil prices to different shocks. (a) Response of WTI oil prices to a PHD shock. (b) Response of WTI oil prices to a NHD shock.



(a)

(b)

Figure 13: Results of variance decomposition of WTI oil prices. (a) Percent WTI oil prices variance due to PHD. (b) Percent WTI oil prices variance due to NHD.

Table 10: The forecasting results of VAR_p.

Sample period	MAE	RMSE
2012m01–2015m12	3.47	4.19
2012m07–2016m06	4.50	5.96
2013m01–2016m12	4.65	7.29
2013m07–2017m06	4.27	4.73
2014m01–2017m12	3.07	4.22
2014m07–2018m06	3.67	5.40
2015m01–2018m12	2.02	2.43
Average	3.67	4.89

The results of variance decomposition are shown in Figure 13, the proportion of PHD for WTI oil prices fluctuation reached 51.50% in the long term, slightly higher than the interpretation of Brent oil prices, which also ranks the first among the factors. In addition, the highest proportion of NHD for WTI oil prices fluctuation reached 12.92%. Overall, the explanatory of HD is similar to the Brent price in terms of WTI oil prices, which proves the validity and robustness of HD.

5. Conclusions

Based on the massive Web news of the oil market, this paper uses the LDA model to perform multiple topic extraction and result evaluation on 88763 items of oil news. After filtering duplicate and invalid information, eight topics related to the oil market are obtained. Based on the probability matrix, the definition of the topic hot degree was proposed to characterise how much each topic is being talked about on the web. At the same time, according to the method of correlation coefficient and principal component analysis, the positive hot degree (PHD) and negative hot degree (NHD) were obtained, the quantitative expression of web news was realized, and the bridge between web news reports and crude oil prices was established. Finally, this paper uses the SVAR model to explore the impact of HD on oil prices. The main conclusions obtained are as follows:

(1) The results of the SVAR model show that PHD and NHD can play a role in assisting oil price forecasting. Among them, the shocks of PHD have a significant positive impact on crude oil prices, and this effect persists for a long period, and the most significant impact appears between the second and fifth periods. The shocks of NHD have a significant negative impact on oil prices, and the impact gradually disappeared after 5 periods. In addition, supply shocks have a significant negative impact on oil prices, and demand shocks have a significant positive impact.

(2) In the long term, PHD accounts for 51.00% of all oil price fluctuations, ranking the first among the influencing factors considered. Among them, the contribution of the oil supply factor is 19.83%, and the contribution of the demand factor is only 5.07%. And in the robustness analysis, HD has the same impact on the fluctuation of WTI oil prices, further proving that the HD extracted based on news reports can better explain the fluctuations of the global oil market.

(3) Through comparison, it is found that PHD has better performance than Google Trends in oil prices forecasting, maximum impact value, and maximum explanation ratio, verifying the effectiveness of PHD indicators derived from news reports, and it also provides new ideas for the current research referring to Google Trends.

(4) The robustness check by transforming different sample interval, reestimating DCC-GARCH model, and using different oil price benchmark confirms that our empirical results are robust.

In summary, accurately grasping various influencing factors is the key to improving the accuracy of oil price forecasting. However, the factors that affect oil prices are intricate and complex. Many nonfundamental factors hidden in the Internet text are difficult to characterise through quantitative indicators. Using NLP technology to structure this unstructured Internet information is the main work of this paper, and the importance of this data for oil price forecasting is verified through econometric models. The sudden and large fluctuations in oil prices are of great significance to future inflation and economic growth. At the same time, fluctuations in oil prices will have a significant impact on various key macroeconomic indicators (e.g., fixed investment, consumption, employment, and unemployment). Therefore, the European Central Bank and the Federal Reserve have used future oil prices as an important reference in the decision-making process. For policy-makers, through the method of this paper, an important indicator "HD" has been added as an indicator of nonfundamental factors to help predict future prices, improve the accuracy of forecasts, further adjust market trends in a timely manner, and stabilize market operations. In addition, for investors, based on the HD changes proposed in this paper, they can gain timely insight into the impact of online news volatility on oil prices, which cannot be quantitatively estimated in the past, so as to help specify better investment strategies.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China under Grant nos. 71871020 and 71021002.

References

[1] J.-Y. Wan and C.-W. Kao, “Interactions between oil and financial markets - do conditions of financial stress matter?” *Energy Economics*, vol. 52, pp. 160–175, 2015.

[2] M. K. Tule, U. B. Ndako, and S. F. Onipede, “Oil price shocks and volatility spillovers in the Nigerian sovereign bond market,” *Review of Financial Economics*, vol. 35, pp. 57–65, 2017.

[3] X. Zou, “VECM model analysis of carbon emissions, GDP, and international crude oil prices,” *Discrete Dynamics in Nature and Society*, vol. 2018, Article ID 5350308, 11 pages, 2018.

[4] F. Picciolo, A. Papandreou, K. Hubacek, and F. Ruzzenenti, “How crude oil prices shape the global division of labor,” *Applied Energy*, vol. 189, pp. 753–761, 2017.

[5] L.-T. Zhao, Y. Wang, S.-Q. Guo, and G.-R. Zeng, “A novel method based on numerical fitting for oil price trend forecasting,” *Applied Energy*, vol. 220, pp. 154–163, 2018.

[6] H. Miao, S. Ramchander, T. Wang, and D. Yang, “Influential factors in crude oil price forecasting,” *Energy Economics*, vol. 68, pp. 77–88, 2017.

[7] X. Wang, K. Chen, and X. Tan, “Forecasting the direction of short-term crude oil price changes with Genetic-Fuzzy information distribution,” *Mathematical Problems in Engineering*, vol. 2018, Article ID 3868923, 12 pages, 2018.

[8] M. S. Kim, “Impacts of supply and demand factors on declining oil prices,” *Energy*, vol. 155, pp. 1059–1065, 2018.

[9] S. Huang, H. An, S. Wen, and F. An, “Revisiting driving factors of oil price shocks across time scales,” *Energy*, vol. 139, pp. 617–629, 2017.

[10] K. J. Singleton, “Investor flows and the 2008 boom/bust in oil prices,” *Management Science*, vol. 60, pp. 300–318, 2013.

[11] Y.-J. Zhang and T. Yao, “Interpreting the movement of oil prices: driven by fundamentals or bubbles?” *Economic Modelling*, vol. 55, pp. 226–240, 2016.

[12] C. K. Baumeister, L. Kilian, and X. Zhou, “Are product spreads useful for forecasting? an empirical evaluation of the Verleger hypothesis,” *Macroeconomic Dynamics*, vol. 22, pp. 562–580, 2017.

[13] Y. Zhang, F. Ma, B. Shi, and D. Huang, “Forecasting the prices of crude oil: an iterated combination approach,” *Energy Economics*, vol. 70, pp. 472–483, 2018.

[14] J. Choi, D. Laibson, and A. Metrick, “How does the Internet affect trading? evidence from investor behavior in 401 (k) plans,” *Journal of Financial Economics*, vol. 64, no. 3, pp. 397–421, 2002.

[15] J. Shen, J. Yu, and S. Zhao, “Investor sentiment and economic forces,” *Journal of Monetary Economics*, vol. 86, pp. 1–21, 2017.

[16] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[17] Y. Gao, Y. Wang, C. Wang, and C. Liu, “Internet attention and information asymmetry: evidence from Qihoo 360 search data on the Chinese stock market,” *Physica A: Statistical Mechanics and its Applications*, vol. 510, pp. 802–811, 2018.

[18] J.-F. Guo and Q. Ji, “How does market concern derived from the Internet affect oil prices?” *Applied Energy*, vol. 112, pp. 1536–1543, 2013.

[19] Y. Yuan, “Market-wide attention, trading, and stock returns,” *Journal of Financial Economics*, vol. 116, no. 3, pp. 548–564, 2015.

[20] H. Schmidbauer and A. Rösch, “OPEC news announcements: effects on oil price expectation and volatility,” *Energy Economics*, vol. 34, no. 5, pp. 1656–1663, 2012.

[21] H. Hu, L. Tang, S. Zhang, and H. Wang, “Predicting the direction of stock markets using optimized neural networks with Google Trends,” *Neurocomputing*, vol. 285, pp. 188–195, 2018.

[22] L. Yu, Y. Zhao, L. Tang, and Z. Yang, “Online big data-driven oil consumption forecasting with Google trends,” *International Journal of Forecasting*, vol. 35, no. 1, pp. 213–223, 2019.

[23] T. Yao, Y.-J. Zhang, and C.-Q. Ma, “How does investor attention affect international crude oil prices?” *Applied Energy*, vol. 205, pp. 336–344, 2017.

[24] J. Wang, G. Athanasopoulos, R. J. Hyndman, and S. Wang, “Crude oil price forecasting based on internet concern using an extreme learning machine,” *International Journal of Forecasting*, vol. 34, no. 4, pp. 665–677, 2018.

[25] M. Afkhami, L. Cormack, and H. Ghoddusi, “Google search keywords that best predict energy price volatility,” *Energy Economics*, vol. 67, pp. 17–27, 2017.

[26] 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.

[27] Y. Chai, H. Luo, Q. Zhang, Q. Cheng, C. S. M. Lui, and P. S. F. Yip, “Developing an early warning system of suicide using Google Trends and media reporting,” *Journal of Affective Disorders*, vol. 255, pp. 41–49, 2019.

[28] A. Manela and A. Moreira, “News implied volatility and disaster concerns,” *Journal of Financial Economics*, vol. 123, no. 1, pp. 137–162, 2017.

[29] P. K. Narayan, “Can stale oil price news predict stock returns?” *Energy Economics*, vol. 83, pp. 430–444, 2019.

[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, 2019.

[31] T. H. Nguyen, K. Shirai, and J. Velcin, “Sentiment analysis on social media for stock movement prediction,” *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015.

[32] M. Abukausar, S. V. Dhaka, and S. K. Singh, “Web crawler: a review,” *International Journal of Computer Applications*, vol. 63, no. 2, pp. 31–36, 2013.

[33] S. Gojare, D. R. Joshi, and D. Gaigaware, “Analysis and design of selenium WebDriver automation testing framework,” *Procedia Computer Science*, vol. 50, pp. 341–346, 2015.

[34] B. Wang, H. Huang, and X. Wang, “A novel text mining approach to financial time series forecasting,” *Neurocomputing*, vol. 83, pp. 136–145, 2012.

[35] K.-Y. Ho, Y. Shi, and Z. Zhang, “How does news sentiment impact asset volatility? evidence from long memory and regime-switching approaches,” *The North American Journal of Economics and Finance*, vol. 26, pp. 436–456, 2013.

[36] R. Füss, M. Grabellus, F. Mager, and M. Stein, “Something in the air: information density, news surprises, and price jumps,” *Journal of International Financial Markets, Institutions and Money*, vol. 53, pp. 50–75, 2018.

[37] C.-Y. Lee and V.-W. Soo, “Predict stock price with financial news based on recurrent convolutional neural networks,” in *Proceedings of the 2017 Conference on Technologies and*

Applications of Artificial Intelligence (TAAI), pp. 160–165, Taipei, Taiwan, December 2017.

[38] L. Kilian and D. P. Murphy, "The role of inventories and speculative trading in the global market for crude oil," *Journal of Applied Econometrics*, vol. 29, no. 3, pp. 454–478, 2014.

[39] Y. Wang, C. Wu, and L. Yang, "Oil price shocks and agricultural commodity prices," *Energy Economics*, vol. 44, pp. 22–35, 2014.

[40] J. Tenhofen, G. B. Wolff, and K. H. Heppke-Falk, "The macroeconomic effects of exogenous fiscal policy shocks in Germany: a disaggregated SVAR analysis," *Jahrbücher für Nationalökonomie und Statistik*, vol. 230, pp. 328–355, 2010.

[41] C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980.

[42] D. I. Harvey and D. van Dijk, "Sample size, lag order and critical values of seasonal unit root tests," *Computational Statistics & Data Analysis*, vol. 50, no. 10, pp. 2734–2751, 2006.

[43] J. A. Clarke and S. Mirza, "A comparison of some common methods for detecting Granger noncausality," *Journal of Statistical Computation and Simulation*, vol. 76, no. 3, pp. 207–231, 2006.

[44] H. Y. Toda and T. Yamamoto, "Statistical inference in vector autoregressions with possibly integrated processes," *Journal of Econometrics*, vol. 66, no. 1-2, pp. 225–250, 1995.

[45] S. Deerwester, S. T Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[46] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 289–296, Stockholm, Sweden, July 1999.

[47] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[48] M. Hajjem and C. Latiri, "Combining IR and LDA topic modeling for filtering microblogs," *Procedia Computer Science*, vol. 112, pp. 761–770, 2017.

[49] H. Nabli, R. Ben Djemaa, and I. A. Ben Amor, "Efficient cloud service discovery approach based on LDA topic modeling," *Journal of Systems and Software*, vol. 146, pp. 233–248, 2018.

[50] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and LDA topic models," *Expert Systems with Applications*, vol. 80, pp. 83–93, 2017.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via theEMAlgorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[52] L. Kilian and T. K. Lee, "Quantifying the speculative component in the real price of oil: the role of global oil inventories," *Journal of International Money and Finance*, vol. 42, pp. 71–87, 2014.

[53] T. Afshar, G. Arabian, and R. Zomorrodian, "Stock return, consumer confidence, purchasing managers index and economic fluctuations," *Journal of Business & Economics Research*, vol. 5, 2007.

[54] J. M. Fernandez-Diaz and B. Morley, "Interdependence among agricultural commodity markets, macroeconomic factors, crude oil and commodity index," *Research in International Business and Finance*, vol. 47, pp. 174–194, 2019.

[55] P. C. Beccue, H. G. Huntington, P. N. Leiby, and K. R. Vincent, "An updated assessment of oil market disruption risks," *Energy Policy*, vol. 115, pp. 456–469, 2018.

[56] M. I. Khan, "Falling oil prices: causes, consequences and policy implications," *Journal of Petroleum Science and Engineering*, vol. 149, pp. 409–427, 2017.

[57] U. Soytas, R. Sari, and B. T. Ewing, "Energy consumption, income, and carbon emissions in the United States," *Ecological Economics*, vol. 62, no. 3-4, pp. 482–489, 2007.

[58] K.-Y. Ho, Y. Shi, and Z. Zhang, "It takes two to tango: a regime-switching analysis of the correlation dynamics between the mainland Chinese and Hong Kong stock markets," *Scottish Journal of Political Economy*, vol. 63, no. 1, pp. 41–65, 2016.